

MCGRAW-HILL ENCYCLOPEDIA OF SCIENCE & TECHNOLOGY



www.MHEST.com



Plasma (physics) — Pythagorean theorem

Plasma (physics)

A gas of charged and neutral particles which exhibits collective behavior. All gases become ionized at sufficiently high temperatures, creating what has been called a fourth state of matter, together with solids, liquids, and gases. It has been estimated that more than 99% of the universe is in the plasma state. On the Earth, plasmas are much less common. Lightning is a familiar natural manifestation, and fluorescent lights are a practical application. Plasma applications and studies make use of an enormous range of plasma temperatures, densities, and neutral pressures (Fig. 1). They extend from plasma processing applications at relatively low temperatures (such as plasma etching of semiconductor chips at low pressure, or plasma cutting torches at atmospheric pressure) to studies of controlled fusion at very high temperatures. See FLUORESCENT LAMP.

Plasma physics is a many-body problem that can be described by a combination of Newton's laws and Maxwell's equations. The charged particles in plasmas are usually ions, both positive and negative, and electrons. Plasmas are normally quasineutral; that is, the net positive ion charge density approximately equals the net negative charge density everywhere in the bulk of the plasma. Quasi-neutrality refers to charge density and does not imply equal densities of electrons and ions since ions can be multiply charged and can also have negative charge. In space and fusion plasmas, plasmas are normally magnetized, while in application plasmas on Earth, such as plasma processing, both magnetized and unmagnetized plasmas are employed. See MAXWELL'S EQUA-TIONS; NEWTON'S LAWS OF MOTION.

Plasma parameters. It is convenient to keep track of plasma properties in terms of characteristic lengths, frequencies, and velocities. Among these are the Debye length, the electron and ion plasma frequencies, the electron and ion gyrofrequencies and gyroradii, the electron and ion thermal velocities, the

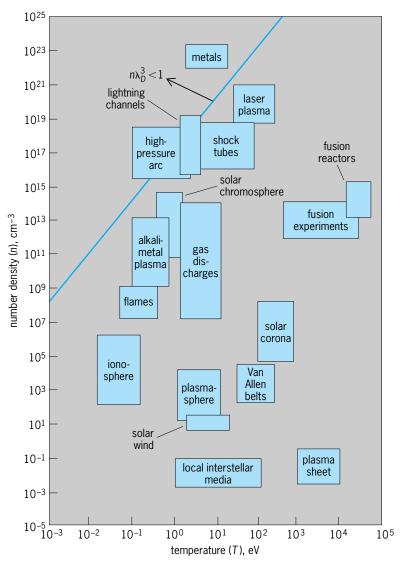


Fig. 1. Approximate number densities (*n*) and temperatures (*T*) of some typical plasmas. The temperature is measured in electronvolts (eV), with 1 eV = 11,605 K. The region above and to the left of the diagonal line represents strongly coupled plasmas ($n\lambda_D^3 < 1$, where λ_D is the Debye length). The figure uses logarithmic scales.

ion sound velocity, the Alfvén velocity, and various collision lengths. The definition of a plasma depends on several of these characteristic parameters, and the magnitude of ratios of these parameters to system size or applied frequencies determines most plasma behavior.

The simplest plasma is a collisionless, unmagnetized collection of ions and electrons with no significant currents. Such plasmas have quasineutral regions and nonneutral regions. The nonneutral regions are highly localized. They are usually located near boundaries (where they are known as sheaths), but are sometimes located within the plasma (where they are known as double layers).

The plasma potential, ϕ , is described by Poisson's equation. In a one-dimensional system, this equation reduces to Eq. (1), where ρ is the net charge den-

$$\frac{d^2\phi}{dx^2} = -\frac{\rho}{\varepsilon_0} \tag{1}$$

sity and ε_0 is the permittivity of free space. From Poisson's equation, it can be shown that the spatial dependence of the plasma potential in nonneutral regions scales with the Debye length, given by Eq. (2), regardless of the geometry of a particular

$$\lambda_D = \sqrt{\frac{\varepsilon_0 T_e}{ne^2}} \tag{2}$$

situation (see **table**). Here *n* is the value of the electron density in the quasineutral region, *e* is the electron charge, and T_e is the electron temperature, measured in electronvolts (1 eV corresponds to 11,605 K, and Boltzmann's constant $k_B = 1$). The dimensions of nonneutral regions are the order of the Debye length when $e\delta\phi$ is of the order of the electron temperature (where $\delta\phi$ is a characteristic variation of the potential), and depend on the Debye length when $e\delta\phi$ is much greater than the electron temperature. In quasineutral regions where the charge density is approximately 0, the potential is normally constant or, at most, slowly varying with a constant electric field over distances comparable to the characteristic dimension, *L*, of a system. In terms of the Debye

Convenient expressions for evaluating plasma parameters Parameter Symbol Value[®] $7.4 \times 10^3 \sqrt{\frac{T_e}{p}}$ Debye length λD $f_{pe} \equiv \frac{\omega_{pe}}{2\pi}$ Electron plasma frequency 9.0*√n* $f_{ce} \equiv \frac{\Omega_e}{2\pi}$ $2.8 imes10^{10}\,B$ Electron cyclotron frequency $4.2 imes 10^5 \sqrt{T_e}$ Electron thermal velocity Ve $10^4 \sqrt{\frac{T_e}{U}}$ C_{s} Ion sound velocity $2.2 \times 10^{16} \frac{B}{\sqrt{n}}$ VA Alfvén velocitv

^{*}Lengths are expressed in meters (m); frequencies are in hertz (Hz); velocities are in m/s; electron temperatures (T_e) are in electronvolts (eV); particle densities (n) are in m⁻³; masses (μ) are in atomic mass units (μ); and magnetic fields (B) are in tesla. length, quasineutral means that electric fields must be much weaker than $T_e/(e\lambda D)$. See POTENTIALS.

Plasma phenomena begin to be apparent only when the system's characteristic dimension is greater than the Debye length. Further, for a gas to be a plasma, there should be at least one electron within each Debye-length volume, λ_D^3 .

The electron thermal velocity is given by Eq. (3),

$$v_e = \sqrt{\frac{T_e}{m_e}} \tag{3}$$

where m_e is the electron mass. Dividing this velocity by the Debye length yields the (electron) plasma frequency, given by Eq. (4), which is the characteris-

$$\omega_{pe} = \frac{v_e}{\lambda_D} = \sqrt{\frac{ne^2}{\varepsilon_0 m_e}} \tag{4}$$

tic frequency of electrons in unmagnetized plasmas. Similarly, the characteristic frequency of ions, called the ion plasma frequency, is given by Eq. (5), where

$$\omega_{pi} = \sqrt{\frac{ne^2}{\varepsilon_0 m_i}} \tag{5}$$

 m_i is the ion mass. Multiplying the ion plasma frequency by the Debye length gives a characteristic ion velocity, called the ion sound velocity and denoted c_s , which depends on the electron temperature rather than the ion temperature.

In the presence of collisions, it is still meaningful to consider the system to be a plasma so long as the collision frequency, v, defined as the reciprocal of the mean collision time, is smaller than the plasma frequency. This idea can also be expressed by the statement that the mean free path for collisions is longer than the Debye length. If the neutral pressure is too high, electron and ion neutral collisions can dominate the properties of the charged gas.

In magnetized plasmas (with magnetic field *B*), electrons and ions gyrate with characteristic frequencies (cyclotron frequencies) given by Eqs. (6).

$$\Omega_e \equiv \frac{eB}{m_e} \qquad \Omega_i \equiv \frac{q_i B}{m_i} \tag{6}$$

Multiplying the ratio Ω_t/ω_{pi} by the speed of light yields another important characteristic velocity, the Alfvén velocity, denoted v_A .

The three criteria for the existence of a plasma are thus given by inequalities (7). While the charged par-

$$L > \lambda_D$$
 (7*a*)

$$n\lambda D^3 > 1 \tag{7b}$$

$$v < \omega_{pe}$$
 (7c)

ticles in a typical room ($n = 10^6/\text{m}^3$, $T_e = 0.025$ eV) satisfy inequalities (7*a*) and (7*b*), they are not a plasma because they do not satisfy inequality (7*c*).

Collective behavior refers to the plasma properties not present in single-particle motion. Collective behavior is a distinguishing characteristic of a plasma. It consists of flows, waves, instabilities, and so forth. Common examples are fluctuations in the aurora, generation of microwaves in devices such as magnetrons and klystrons, and reflection of electromagnetic waves from the ionosphere. *See* AU-RORA; KLYSTRON; MAGNETRON; RADIO-WAVE PROPA-GATION.

Charged-particle gases. Curiously, very high density collections of equal numbers of ions and electrons are not plasmas. According to Eq. (2), the product $n\lambda_D^3$ is proportional to $1/\sqrt{n}$, so at high plasma density this product is less than 1, and inequality (7*b*) is no longer satisfied. Such systems are referred to as strongly coupled plasmas (even though, strictly speaking, they are not plasmas at all).

A collection of either electrons or ions can exhibit properties similar to those of an electrically neutral plasma if the charged-particle density is sufficiently large. For such so-called plasmas, the Debye length and the characteristic frequency of electrons or ions, ω_{pe} or ω_{pi} , can still be defined, and collective behavior is still exhibited when the Debye length is less than the system's characteristic dimension. So-called pure electron plasmas or pure ion plasmas are unconfined in an unmagnetized system. However, particle traps consisting of a combination of electric and magnetic fields can be used to confine the charges. *See* PARTICLE TRAP.

Plasmas based on carbon- or silicon-containing gases can give rise to polymer formation within the plasmas. As a consequence, macroscopic insulating particles can grow within the plasma, reaching sizes the order of micrometers. The insulating particles charge negative, with thousands of electron charges or more, so near the particles the plasma can consist of many positive ions and very few electrons. Unlike other plasmas, dusty plasmas consist of negative ions with a large variety of masses and chargeto-mass ratios.

Visual characteristics. The visual appearance of a plasma depends on the kind of ion present, the electron temperature, and the plasma density. Some plasmas are invisible. Curiously, if a plasma is present and not glowing, it is either very hot or very cold. For example, an H⁺ plasma, or any other relatively hot plasma with fully stripped ions, contains atomic nuclei with no electrons, so there is no atomic physics and no optical emission or absorption. If plasma electrons and ions are very cold, there is insufficient energy to excite optical transitions. The glow often associated with plasmas indicates only where visible energy transitions are excited by energetic electrons or perhaps absorption of ultraviolet radiation, and may have little to do with the presence of bulk plasma. In fusion plasmas, the edges are often copious sources of emission associated with the dissociation and ionization of hydrogen and edge-generated impurities, while much of the hotter core plasma is fully ionized and invisible.

Direct-current glow-discharge plasmas originate from electrons created by secondary electron emission due to ion bombardment of a negatively biased cathode. The secondary electrons are accelerated through the cathode sheath potential (called the cathode fall) to energies the order of 1 keV, and partially ionize the neutral gas, releasing additional energetic electrons in a multiplicative process. The energetic electrons also undergo inelastic collisions with neutrals which result in optical emission that contributes to the so-called glow. *See* GLOW DIS-CHARGE; SECONDARY EMISSION.

Single-particle motion. The understanding of plasma physics begins with an understanding of the motion of single charged particles in a combination of electric and magnetic fields (*E* and *B*), produced by a combination of external fields and the motion of the charged particles themselves. The motion of a single particle, with mass *m*, charge *q*, and velocity \mathbf{v} , is governed by the Lorentz force, as given in Eq. (8).

$$m\frac{d\mathbf{v}}{dt} = q\left(\mathbf{E} + \mathbf{v} \times \mathbf{B}\right) \tag{8}$$

From the perpendicular component of Eq. (8), it can be shown that the charged particles gyrate about magnetic field lines with a characteristic frequency (the cyclotron frequency) given by Eqs. (6). Ions rotate about the magnetic field in the clockwise direction, while electrons rotate counterclockwise with the magnetic field pointing outward. *See* ELECTRIC FIELD; PARTICLE ACCELERATOR.

In addition to the motion parallel to the magnetic field and the gyromotion about the magnetic field, there are drifts perpendicular to the magnetic field. For a general force, **F**, in the presence of a magnetic field, the perpendicular drift velocity is given by Eq. (9).

$$\mathbf{v}_D = \frac{\mathbf{F} \times \mathbf{B}}{qB^2} \tag{9}$$

Two specific examples of perpendicular drifts are the **E** × **B** and curvature drifts. Substituting $\mathbf{F} = q\mathbf{E}$ into Eq. (9) gives Eq. (10). This **E** × **B** drift is orthog-

$$\mathbf{v}_D = \frac{\mathbf{E} \times \mathbf{B}}{B^2} \tag{10}$$

onal to both the magnetic field and the electric field. Curiously, it is independent of the particle's mass and charge. Given a perpendicular electric field, particles can walk across a magnetic field. Forces associated with magnetic-field curvature give rise to a curvature drift in the direction orthogonal to the magnetic field, and to the radius of curvature of the magnetic field lines.

Magnetic mirroring. For gyro motion in a slowly changing magnetic field, which is approximately periodic, it can be shown that the ratio of the perpendicular energy to the magnetic field is approximately constant. This means that a charged particle moving parallel to a magnetic field and gyrating about the field will gyrate faster as the magnetic field increases. If the magnetic field changes in space and is constant in time, the total energy is conserved. For a sufficiently large magnetic field, a point is reached where the total energy equals the perpendicular energy, so that the parallel energy goes to zero and the

particle reflects. This is known as magnetic mirroring.

Magnetic mirroring is the chief natural mechanism of charged-particle confinement. For example, this process confines charged particles in the ionosphere and magnetosphere. The magnetic field lines that connect the north and south magnetic poles of the Earth provide a mirror magnetic field which increases as either pole is approached. In the absence of collisions, a particle moving along and gyrating about such a magnetic field is magnetically confined, if it has a sufficiently large velocity perpendicular to the magnetic field. The Van Allen belts are composed of such mirror-trapped charged particles. The source of these particles is the solar wind, a stream of charged particles continuously emitted by the Sun.

Distribution functions. The number of particles per unit volume at position **x** and time *t* between velocities **v** and **v** + d**v** can be specified by the distribution function f(**v**,**x**,*t*) d**v**. This distribution function satisfies the Boltzmann transport equation, which expresses the rate of change of *f* as the sum of a term due to external forces, a term due to diffusion, and a term due to collisions of the particles. If the collision term is set equal to zero, the Vlasov equation results. *See* BOLTZMANN TRANSPORT EQUATION.

Under many situations the distribution functions of plasma electrons and ions are approximately maxwellian, which means that each distribution function that varies with **v** is proportion to the exponential of -E/T, where *E* is the kinetic energy of the particle in question. For maxwellian distribution functions, the average energy is equal to $3/_2T$. *See* BOLTZMANN STATISTICS; KINETIC THEORY OF MATTER.

Fluid equations. It is often convenient to describe the plasma in terms of fluid variables such as the plasma species densities $n_j(\mathbf{x},t)$, the fluid velocity $\mathbf{u}(\mathbf{x},t)$, and the pressure tensor $P(\mathbf{x},t)$. This procedure, together with Maxwell's equations, provides the simplest description of a plasma. The simplest fluid equations are the continuity equation and the equation of motion. The fluid equation of motion resembles the single-particle equation of motion with the addition of terms representing a pressuregradient force on a single particle and a frictional force. *See* FLUID-FLOW PRINCIPLES.

For given electric and magnetic fields \mathbf{E} and \mathbf{B} , determined by Maxwell's equations, the equations of continuity and motion are two equations with three unknowns, n, \mathbf{u} , and \mathbf{P} . Solutions to the fluid equations can be obtained by giving another equation for the extra unknown, an equation of state (11), where

$$p = C(nm)^{\gamma} \tag{11}$$

p is the pressure and *C* is a constant. For isothermal compression $\gamma = 1$, while for adiabatic compression γ is the ratio of specific heats, equal to (2 + N)/N, where *N* is the number of degrees of freedom. Even in this simple description, the equations are nonlinear. *See* POLYTROPIC PROCESS.

Nonlinear physics. Many of the equations describing plasma physics are nonlinear. The usual approach is to linearize the equations. This procedure simply ignores all nonlinear terms. However, solutions exist that are inherently nonlinear and not superpositions of linear solutions.

Sheaths are a simple example of a nonlinear structure. Ion acoustic solitons are another example of a nonlinear structure. Ion acoustic solitons are pulselike modes which result from a balance of nonlinearity, which tends to steepen pulses (since largeamplitude components travel faster), and dispersion, which tends to broaden pulses. The velocity of solitons, measured in the frame moving at the ion acoustic velocity, is proportional to the soliton amplitude, b, while the soliton width, w, is proportional to $1/\sqrt{b}$, so the product bw^2 is a constant of motion. These solitons can survive collisions with other solitons with no change in shape. In this respect, they behave like linear pulses even though they are inherently nonlinear. See NONLINEAR PHY-SICS: SOLITON.

MHD equations. For fully ionized plasmas, it is convenient to describe the plasma as a single fluid together with Maxwell's equations. This gives the magnetohydrodynamic (MHD) equations, which are used to describe plasma equilibria and plasma waves and instabilities. Their relative simplicity has made them ideal for solutions of fusion problems in complicated geometries, and they have been widely used to describe astrophysical plasmas and magnetohydrodynamic energy conversion. *See* MAGNETOHYDRODYNAMIC POWER GENERATOR; MAGNETOHYDRODYNAMICS.

Waves in plasmas. Plasmas can support an impressive variety of electrostatic and electromagnetic waves not present in the absence of plasma. The waves are distinguished by their frequency, the presence or absence of dc magnetic fields, and the plasma temperature and density. Wave characteristics also depend on whether the plasma is bounded, and if it is, how it is bounded. Relatively simple expressions exist only for unbounded (that is, infinite), unmagnetized plasmas or cold plasmas with zero temperature. Wave characteristics depend on the ratio of the wave frequency to the characteristic frequencies of the plasma, ω_{pi} , ω_{pe} , Ω_i , and Ω_e .

Unmagnetized plasmas. In unbounded, unmagnetized plasmas with finite electron temperature and zero ion temperature, only two modes can exist: ion acoustic waves with frequencies less than the ion-plasma frequency, ω_{pi} , and Langmuir waves with frequencies greater than the electron plasma frequency, ω_{pe} . Both modes are electrostatic with the fluctuating electric field parallel to the wave propagation direction. No waves can exist between the ion and electron plasma frequencies.

Alfvén waves. In the presence of a magnetic field, and at frequencies less than the ion cyclotron frequency, Ω_i , the solutions to the wave equation are known as Alfvén waves. They come in three varieties: magnetosonic, shear Alfvén, and slow Alfvén waves (ion acoustic waves). For low frequencies (much less than Ω_i) the magnetosonic and shear Alfvén waves have similar values of frequency versus wave number. As the frequency is increased, the solutions to the wave equation split into two separate branches. Near Ω_i , one branch is known as the ion cyclotron waves, while the other continues to be known as the magnetosonic waves, or fast waves. *See* ALFVÉN WAVES.

Whistlers. At frequencies between $\sqrt{\Omega_i \Omega_e}$ and the electron cyclotron frequency, Ω_e , the electromagnetic wave solutions are known as whistlers (**Fig. 2**). Whistler waves are observed to propagate along the

magnetic field lines connecting the Earth's magnetic poles. They are at audio frequencies and make a whistlinglike sound when the signals they generate are played over a loudspeaker. This phenomenon occurs because the phase and group velocities differ, with the higher-frequency components arriving first.

Beinstein waves. For frequencies greater than Ω_i , electrostatic waves propagating across the magnetic field are found to be associated with the harmonics of Ω_i and are known as ion Bernstein waves. These waves can exist over a range of frequencies between

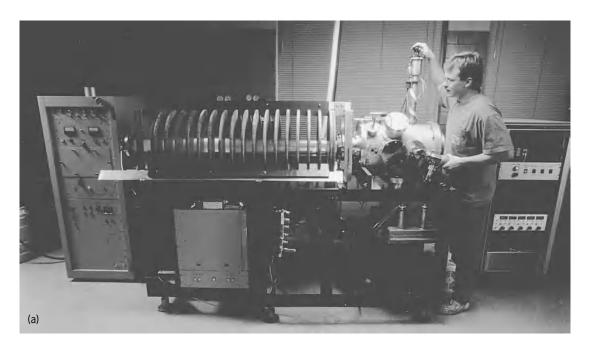




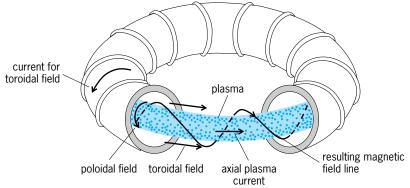
Fig. 2. Helicon plasma, which is generated by whistler waves, which then propagate in the plasma. (a) Plasma source. (b) Plasma in a Pyrex chamber mounted inside the magnetic field coils. (*Bruce Fritz, University of Wisconsin Engineering Center for Plasma Aided Physics*)

successive harmonics of Ω_i with new modes appearing at each harmonic. Electrostatic waves known as electron Bernstein waves are also found to be associated with the harmonics of Ω_e .

Collisionless Landau damping. It might be guessed that waves in a collisionless plasma would not damp, since there is no energy dissipation. This does not turn out to be true. Starting with the Vlasov equation, L. D. Landau showed that langmuir waves damp when the derivative of the distribution function evaluated at the phase velocity of the wave is negative. This condition is satisfied for the common case of a maxwellian electron distribution.

Plasma sources. Much of the progress in the understanding of plasma physics has been the result of the development of new plasma sources. Discharge tubes, Penning discharges, Q machines, multidipole confined plasmas, magnetron discharges, stellarators, and tokamaks are examples. The key elements shared by all plasma sources are ionization, confinement, equilibrium, stability, and heating.

lonization. Ionization is the key to plasma production and can be accomplished in many different ways. The most common approach is to employ energetic electrons with energies greater than the ionization potential of the gas being ionized. In dc glow dis-



(a)

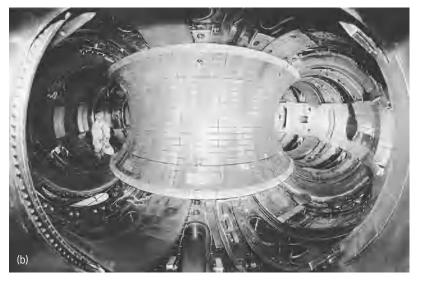


Fig. 3. Tokamaks. (a) Diagram of tokamak plasma and magnetic field. (b) Interior of the reaction chamber of the Tokamak Fusion Test Reactor (TFTR).

charges, electrons produced by ion secondary electron emission are accelerated by the cathode sheath potential, as are electrons created by thermionic emission in hot-cathode plasmas. Electrons can also pick up energy by reflecting from oscillating radiofrequency sheath electric fields, or by cyclotron resonance in magnetic fields, or from collisions with other energetic electrons. *See* ELECTRICAL CONDUC-TION IN GASES; GAS DISCHARGE; IONIZATION POTEN-TIAL; THERMIONIC EMISSION.

Several other approaches involving collisions, which do not require energetic electrons, also exist. These techniques include photoionization, ion-neutral charge exchange, surface ionization, and Penning ionization. Ions can also be produced in the dissociation of molecules. Yet another mechanism, called critical ionization velocity, is instability driven, and occurs when the kinetic energy of the neutral gas atoms streaming perpendicular to a magnetic field exceeds their ionization potential. *See* ION SOURCES; IONIZATION; PHOTOIONIZATION.

Confinement. A vacuum chamber provides the simplest approach to confinement. In an unmagnetized plasma, electrons are lost more rapidly than ions, and the plasma acquires a net positive charge. The excess positive charge appears in a sheath at the plasma boundary with the bulk plasma potential more positive than the boundary potential. The decrease in potential at the boundary provides plasma electron confinement, reducing their loss rate to balance the ion loss rate. The ion loss current is found to lead to the particle confinement time given by Eq. (12),

$$\tau = \frac{V}{A} \sqrt{\frac{m_i}{T_e + T_i}} \tag{12}$$

where A is the loss area and V is the plasma volume. For tabletop-size devices used in plasma processing, τ is of the order of 0.1 millisecond.

Addition of a uniform magnetic field reduces the loss rate of ions and electrons transverse to the magnetic field, but has no effect on losses parallel to the magnetic field because the Lorentz force has no components along this field. Effective confinement by magnetic fields requires that the ion and electron gyroradii be small compared to device dimensions. Plasma transport across the magnetic field can still occur as a result of collisions or of perpendicular drifts.

Open magnetic-field confinement devices, in which the magnetic field intersects a boundary, can take advantage of nonuniform magnetic fields to enhance axial confinement by providing magnetic mirrors. However, charged particles which do not have sufficiently large perpendicular velocities fall within a loss cone, in velocity space, and are not confined. Scattering in velocity space removes particles from the confined region.

In closed magnetic confinement devices, magnetic fields do not intersect a boundary, but charged particles can still flow across magnetic fields. A toroidal magnetic field can be produced by a set of toroidal field coils (**Fig. 3**). Perpendicular

plasma drifts cause a vertical separation of electrons and ions and a corresponding vertical electric field. Combined with the toroidal magnetic field, this results in an outward $\mathbf{E} \times \mathbf{B}$ drift. Toroidal plasma confinement devices avoid this problem by twisting the magnetic field helically by adding a poloidal magnetic field. The $\mathbf{E} \times \mathbf{B}$ drift (which is always outward with respect to the major radius of the torus) changes from outward to inward (with respect to the minor radius of the torus) as the magnetic line of force goes from the low magnetic field, outside the torus, to the high field inside the torus and averages to zero. In stellarators, the twist in the magnetic field is accomplished by using external conductors. In tokamaks (Fig. 3), it is produced by current in the plasma itself.

Equilibrium. In the absence of magnetic fields (both inside and outside the plasma), an equilibrium can be achieved by establishing a pressure balance between plasma and edge walls or edge gas.

When a plasma column is confined by a magnetic field, it can be shown that Eq. (13) is satisfied, where

$$p_{\perp} + \frac{B^2}{2\mu_0} = \frac{B_0^2}{2\mu_0} \tag{13}$$

 p_{\perp} is the plasma pressure perpendicular to the magnetic field, B_0 is the magnetic field just outside the plasma, and μ_0 is the permeability of free space. Diamagnetic currents generated at the plasma edge reduce the magnetic field in the plasma. The quantity $B^2/(2\mu_0)$ can be regarded as the magnetic-field pressure. Its ratio to the plasma perpendicular pressure is called the plasma beta, given by Eq. (14). Equilibrium

$$\beta \equiv p_{\perp} \left(\frac{B_0^2}{2\mu_0} \right) \tag{14}$$

exists only when $\beta \leq 1$, and $\beta = 1$ corresponds to zero magnetic field inside the plasma. *See* DIAMAGNETISM.

Stability. The existence of an equilibrium does not guarantee that a particular configuration is stable. There are two classes of instabilities, macroscopic and microscopic. Macroscopic or magnetohydrodynamic instabilities are of relatively low frequency and involve large-scale motions of the plasma, such as ripples in the plasma edge. The energy for the instability comes from the plasma expanding in a magnetic field. Microinstabilities are of relatively high frequency and get their free energy from changes in the distribution function $f(\mathbf{v}, \mathbf{x}, t)$ toward a uniform maxwellian distribution function, the distribution function with the lowest free energy.

An important example of a magnetohydrodynamic instability is the Rayleigh-Taylor instability. The equilibrium between a uniform plasma column and a uniform magnetic field is not stable in the presence of a transverse force applied away from the plasma. The growth rate of the instability is proportional to the square root of the effective gravity associated with the transverse force. The gravitational force can be the result of curved magnetic field lines. The gradient of the magnetic field points away from the center of curvature of curved magnetic field lines, so there can be good or bad curvature. Plasmas are stable when this gradient points toward the plasma, and unstable when it points away from the plasma.

A similar instability results from an attempt to support a heavy fluid by a light fluid. When a glass of water is turned upside down, the upward air pressure on the water exceeds the downward force of the water so that an equilibrium exists, but the water falls out of the glass never the less. It is apparent that the equilibrium of a heavy fluid (water) supported by a light fluid (air) is not stable.

Plasma heating. Combined with collisions, dc electric fields can heat electrons. The electrons are accelerated by the electric field, and the energy is randomized in direction by collisions with other electrons, ions, or neutrals. Resonant heating can also be employed. In the presence of a circularly polarized electric field rotating with electrons at the electron cyclotron frequency, electrons can be spun up to high energies without the need for collisions. A similar approach can be used to heat ions. A third approach to heating plasmas is to use charge exchange on energetic neutral beams which converts energetic neutrals to energetic ions and cold ions to cold neutrals.

Ohmic heating, the combination of an applied electric field and Coulomb collisions, is a very successful heating technique in fusion plasmas. The resistivity averaged over a maxwellian electron distribution function is known as the Spitzer resistivity. For an electron temperature of 1 keV, this resistivity is 2.5×10^{-8} ohm-m, comparable to the resistivity of copper (2×10^{-8} ohm-m).

Ohmic heating has provided tokamaks with an easy way to heat plasmas and an easy path to relatively high density and high temperature, which may progress all the way to fusion reactors. However, the resistivity is proportional to the $-\frac{3}{2}$ power of the electron temperature, and thus becomes much less effective at high temperatures.

Plasma processing. Plasma processing can be defined as the collection of techniques which make use of plasmas to create new materials or to modify properties of existing materials. It is used in a large variety of applications including semiconductor etching, preparing plastic surfaces to accept ink, depositing polymers, depositing diamond films, and hardening artificial hip joints. The technique has its foundations in plasma physics, chemistry, electrical and chemical engineering, and materials science.

Plasma processing employs a wide range of plasma densities and neutral pressures, but a narrow range of electron temperatures, from 1 to 10 eV. At the lowest neutral pressures, less than 10^{-1} pascal (10^{-3} torr) , plasma effects are most important and plasmas can generally be treated as collisionless. The most interesting application in this neutral pressure regime is ion implantation into objects immersed in plasmas. As the pressure is increased, through the range 10^{-2} to 10 Pa $(10^{-4} \text{ to } 10^{-1} \text{ torr})$, many-body effects begin to be important and the density of free

radicals (active chemical species) increases. Plasma etching of semiconductors is an important application in this pressure regime. At higher pressures, in the range from 10 Pa to 10 kilopascals (10^{-1} to 10^2 torr), a variety of plasma deposition techniques are important. Applications at higher pressure, from 10 kPa (10^2 torr) to several atmospheres (1 atm = 10^2 kPa), include plasma torches, plasma sprays, high-pressure deposition, and high-pressure arc lamps. *See* FREE RADICAL.

Plasma-source ion implantation. Low-energy ion beams produce secondary electron emission, radiation damage, and sputtering at material surfaces. Raising the ion energy results in ion implantation. Ions are embedded far into material surfaces, changing the surface properties. This process differs from coatings in that a layer 50–100 nanometers or more thick is transformed. Conventional ion implantation employs an accelerator to produce the energetic ions. Plasma-source ion implantation (PSII) is one of many names used to describe the process of ion implantation into objects immersed in a plasma. It provides ion acceleration in a sheath formed at the material surface.

A series of short, pulsed negative-bias voltages, δV , with $e\delta V$ much greater than the plasma electron temperature, are applied to an object immersed in the plasma. This results in a thick expanding sheath, with thickness much greater than the Debye length, during the pulses. Plasma-source ion implantation has the advantage that sheaths are conformal to objects, so ion implantation is as easy to use with objects with relatively complicated shapes as it is for planar objects. *See* ION IMPLANTATION.

Plasma etching. Chemical etching of semiconductor device structures undercuts masks by approximately 0.5 μ m, so plasma etching has been required to achieve structures less than 1.0 μ m in size. The directed ion acceleration at the plasma-wafer sheath boundary gives rise to the anisotropy in the etch. In many situations, a passivating polymer layer is deposited on the side walls of structures, preventing chemical etching. At the bottom of a trench, the passivating layer is sputtered away by the directed ion flux.

The most widely used etch tools have employed parallel-plate capacitive discharges operating at 13.56 MHz, a frequency that is designated for general use, at neutral pressures the order of 10 Pa (10^{-1} torr) . This system is chemically etched, enhanced by the presence of ion beams. Several different approaches can be used to obtain highdensity, low-pressure etching plasmas needed for finer linewidths. Among them are inductive discharges (Fig. 4) and helicon (bounded whistler wave) discharges (Fig. 2) operating at 13.56 MHz, and electron cyclotron resonant (ECR) discharges operating at 2.45 GHz, a frequency for which magnetron sources are widely available and relatively inexpensive because it is used in microwave ovens. See INTE-GRATED CIRCUITS; MICROWAVE.

Plasma deposition. The walls of devices that employ plasmas are often found to be coated. The deposited

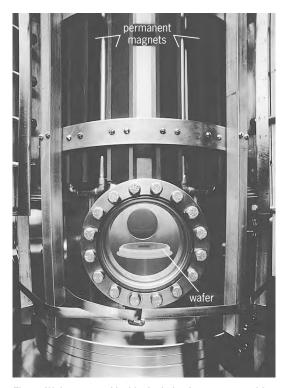


Fig. 4. Wafer mounted inside the inductive source used for plasma etching. The surface magnetic field produced by the permanent magnets helps to confine the plasma. (Bruce Fritz, University of Wisconsin Engineering Center for Plasma Aided Physics)

species can originate from sputtering by plasma ions, accelerated by the plasma-object sheath at the edge of objects immersed in the plasma, or from deposition of molecular species created in the plasma. For example, gas-phase carbon difluoride (CF₂) can lead to poly(tetrafluoroethylene)-like polymer formation on silicon dioxide (SiO₂). Graphite can originate either by sputtering of carbon into a mostly noncarbon plasma or from a plasma based on a carboncontaining gas such as methane (CH₄). It is also possible to deposit diamond and diamondlike films. At atmospheric pressure, plasma sprays provide a way to deposit films as thick as several millimeters. Plasma sprays work by injecting powder into arc jets. The powders become molten and are ejected supersonically onto substrates, where they stick. See VAPOR DEPOSITION.

Thermal plasmas. At pressures greater than 10^2 Pa (1 torr), electron, ion, and neutral gas temperatures are almost in equilibrium and plasma properties are dominated by heat transport. Important applications involve arcs and plasma torches.

When the current in a discharge is increased by reducing the current-limiting resistance in external circuits, glow-discharge plasmas can make a transition to being an arc. Compared to glow discharges, which require voltages of several hundred volts, arcs are self-sustaining at relatively low voltages, the order of 10 V. This occurs because the cathode emission mechanism differs from those in glow discharges. Field emission of electrons combined with thermionic emission result in cathode-spot current densities as high as 10^{10} A/m², compared to 10-100 A/m² for glow discharges. *See* ARC DISCHARGE; FIELD EMISSION.

Arcs are used in plasma torches, plasma sprays, lighting, gas heaters, and many other applications. Arcs often have a negative resistance characteristic, with current decreasing with increasing voltage. Arcs have high column luminosity at pressures above 1 kPa (10 torr). At pressures greater than or equal to atmospheric pressure, this luminosity is used in lighting applications. *See* ARC LAMP.

Arcs tend to be unstable, and the arc may wander about or move along an electrode. Several stabilization schemes are used to confine arcs to a restricted region. In arc lamps, the glass or quartz tube walls provides stabilization. In other applications, the glass tube is replaced by a stack of insulated, water-cooled copper disks or by an intense vortex of swirling gas or liquid. Magnetically stabilized arcs employ external magnetic fields to stabilize, deflect, and drive them.

Plasma torches have replaced chemical flames as heat sources in many applications and have many advantages. Temperatures determined by resistive heating in the arc plasma can be much higher than those found in chemical flames. Operation with inert gases avoids chemical reactions. In addition, torches can have very high thermal efficiencies, reaching as high as 92%. *See* ARC HEATING.

Controlled fusion. Controlled fusion aims at taking advantage of nuclear fusion reactions to generate net power. Cross sections for fusion reactions become significant only for energies greater than 10 keV and for light nuclei. The reaction that is likely to power the first fusion reactors, because it has the highest cross section for energies less than 30 keV, is reaction (15), employing deuterium and tritium plasmas.

$$D + T \to n (14.1 \text{ MeV}) + \alpha (3.5 \text{ MeV})$$
 (15)

Plasma heating can be accomplished with ohmic heating, neutral beams, radio-frequency fields, and so forth. As the input power is increased, plasma is heated to higher temperatures, and the relative power from the fusion reaction also increases. In this state, the plasma is referred to as burning. For deuterium-tritium fusion, collisions with magnetically confined alpha particles convert power from the fusion reaction to heating power. At high enough temperatures, the plasma ignites; that is, the power provided by the alpha particles to the plasma is sufficient to balance power losses from the plasma. After ignition, the burning plasma no longer requires external heating.

At energies of the order of 10 keV or more, electrons and ions are essentially Coulomb collisionless, even though collisions leading to thermonuclear reactions are the basis of controlled fusion. Curiously, plasmas studied in controlled fusion are close relatives of low-density space plasmas. In spite of the great differences in parameter values, the ratios of characteristic lengths and time scales are comparable and many plasma phenomena are seen only in these two systems.

Advances in fusion studies have been tied to the techniques developed for plasma confinement and heating. Fusion experiments employ either magnetic confinement or inertial confinement, in which fusion reactions take place before the plasma has a chance to expand to chamber boundaries. Magnetic mirrors are an example of open systems, while tokamaks, stellarators, and reversed-field pinches are examples of closed toroidal systems. Most magnetic confinement research experiments are done on tokamaks (Fig. 3). *See* NUCLEAR FUSION. Noah Hershkowitz

Naturally occurring plasmas. Naturally occurring plasmas exist throughout the solar system and beyond. Above the atmosphere, most matter is ionized. The lower-density ionized materials are considered to be plasmas, and they behave in manners very different from the behavior of nonplasmas. Some dense materials, such as stellar matter or electrolytic solutions, are often not considered to be plasmas even though they are ionized; they behave, for the most part, as do ordinary fluids.

As discussed above, two important parameters of a plasma are its number density, n, and its temperature, T_{e} (Fig. 1). The combination of the two parameters determines whether the motions of the electrically charged particles that compose the plasma will be dominated by the large-scale electric and magnetic fields generated by the plasma or whether their motions will be dominated by interparticle collisions. Hot, low-density plasmas are termed collisionless, and cool, high-density plasmas are termed collisional. The behaviors of collisionless and collisional plasmas are quite different. In particular, for collisionless plasmas, a third parameter that is important is the strength of the magnetic field in the plasma. If the magnetic field is strong, the field acts to confine the plasma; if the magnetic field is weak, the movement of the plasma distorts the field and plasma motions can carry embedded magnetic-field structures.

The behavior of naturally occurring plasmas is complicated and often not well understood, and major gaps exist in knowledge about them. Several processes often operate simultaneously, some acting on large spatial scales and long time scales while others act on short spatial scales and fast time scales. Self-generated magnetic and electric fields which are difficult to predict are often important. Consequently, the fundamental dynamics of naturally occurring plasmas is an active area of research.

lonosphere. The upper atmosphere of the Earth, known as the ionosphere, is in a plasma state because of the photoionization by sunlight of the atmospheric gas, and the impact ionization of the gas by particles associated with the aurora. Because plasmas are capable of reflecting electromagnetic waves, the ionosphere is studied for its role in enabling radio communication around the Earth and for its role in disrupting radio communication between the Earth and space satellites. *See* IONOSPHERE.

Solar atmosphere. The solar atmosphere (photosphere, chromosphere, and corona) is in a plasma

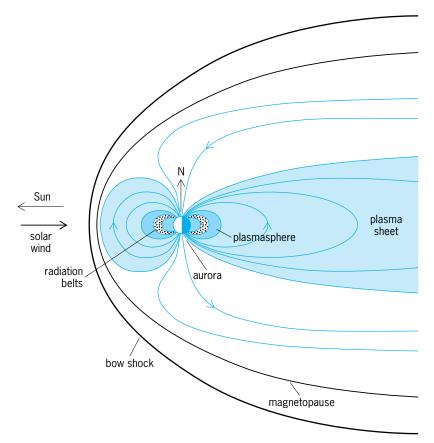


Fig. 5. Earth's magnetosphere. The magnetotail extends for a long distance off the figure to the right, away from the Sun.

state because of the heating of solar material by electromechanical processes driven perhaps by fluid motions at the solar surface. The study of the solar atmosphere is important to gain understanding of the transport of energy from the Sun, the production of solar wind (which can drive geomagnetic storms that disrupt communication), the production of energetic particles and energetic photons from solar flares in the atmosphere (which have environmental consequences), and the production of radio bursts (which interfere with communication). The insight gained from studying the solar atmosphere is used to increase and test the understanding of the atmospheres of other stars. *See* SUN.

Solar wind. The solar wind is a supersonic flow of plasma emanating from the Sun and extending far beyond the solar system, keeping the interstellar plasma pushed away from the Sun. The solar wind is turbulent, with embedded magnetic-field structures convected along with the flow. The self-generated electric field of the solar wind is the driver for geomagnetic disturbances at the Earth, and the embedded magnetic field of the solar wind modulates the ability of cosmic rays to enter the solar system. The solar wind is also the source of some intense radio bursts that interfere with communications.

Besides being important for understanding effects on the Earth, investigations of the solar wind serve to extend the understanding of stellar winds, stellar mass loss, and astrophysical-plasma turbulence. *See* COSMIC RAYS; SOLAR WIND. **Magnetospheric plasmas.** The Earth's magnetosphere is the region of space where the magnetic field of the Earth controls the motion of the plasmas; the magnetosphere forms a barrier to the solar-wind flow. The interaction of the solar wind draws the magnetosphere out into a long tail (**Fig. 5**). The size, shape, and content of the magnetosphere varies with time; solar-wind-driven geomagnetic disturbances (substorms and storms) can reconfigure the magnetosphere on time scales of tens of minutes. Plasmas in the magnetosphere have two sources: solar-wind plasma that is captured by the magnetosphere, and ionospheric plasma that evaporates into the magnetosphere.

Two of the major plasmas of the magnetosphere are the plasmasphere and the plasma sheet (Figs. 1 and 5). The plasmasphere is a region of denser, cooler plasma that originates from the ionosphere. If the plasmaspheric plasma is removed by a geomagnetic disturbance, it refills from the ionosphere in a matter of days. Within the plasmasphere lie the Van Allen radiation belts, which are regions where very energetic electrons and ions are magnetically trapped for years. *See* VAN ALLEN RADIATION.

The plasma sheet is a region of lower-density, hotter plasma that has its source in the distant magnetotail where solar-wind plasma and evaporated ionospheric plasma are magnetically captured. The plasma-sheet plasma flows up the nightside magnetotail toward the Earth, then flows around the dawn and dusk sides of the Earth, through the dayside magnetosphere, then out into the solar wind, with the entire trip through the magnetosphere taking only hours.

Some of the major topics of study for magnetospheric plasmas are the energy flow from the solar wind to the Earth, the mechanics of geomagnetic disturbances, the sources and movements of energetic particles, the causes of the aurora and auroral radio emissions, and the effects of magnetospheric plasma on satellite operations. The plasmas of other planetary magnetospheres (Mercury, Jupiter, Saturn, Uranus, and Neptune) are studied to understand the sources of magnetospheric radio emissions, and the production of energetic charged particles by magnetospheres. *See* JUPITER; MAGNETOSPHERE; NEPTUNE; SATURN; URANUS.

Lightning. The hot, luminous channels of lightning strokes are dense plasmas. The plasma is produced by impact ionization of air molecules by the free electrons that carry the lightning currents. Typically, tens of thousands of amperes flow in channels that are only millimeters in radius. Lightning-channel plasmas are studied to determine the dynamics of the channels, to obtain current-versus-voltage information for lightning, to discern the effects of lightning on atmospheric chemistry, and to ascertain the sources of the intense radio emissions from lightning. *See* LIGHTNING.

Metals and semiconductors. The electrons that provide the electrical conduction of metals are in a plasma state, and the mobile electrons and electron holes in semiconductors exhibit plasma properties. The reflection of light by a metal is an example of a plasma process (like the reflection of radio waves by the ionosphere). These solid-state plasmas differ from other plasmas because quantum-mechanical effects often are significant. The study of solid-state plasmas is important for the development of amplifiers and oscillators and for the comprehension of conductivity and superconductivity. *See* ELECTRICAL CONDUCTIVITY OF METALS; FREE-ELECTRON THEORY OF METALS; SEMICONDUCTOR; SUPERCONDUCTIVITY.

Interstellar medium. The hydrogen plasma of the interstellar medium (denoted in astronomy as H II regions) is produced by photoionization of interstellar gas by starlight and by the collisional ionization of interstellar gas by supernova-explosion plasmas, with an additional contribution from the ionized winds from stars. The interstellar plasma plays roles in the formation of stars from interstellar gas, in the dynamics and evolution of galaxies, and in the transport of cosmic rays. *See* INTERSTELLAR MATTER.

Other astrophysical plasmas. In addition to the interstellar medium, the study of plasma astrophysics includes (1) the high-velocity plasmas of supernova explosions, which are of interest for their interactions with interstellar gases and for their possible production of cosmic rays; (2) the electron-position plasmas believed to exist in pulsar magnetospheres, which are thought to give rise to the optical and radio emissions of pulsars; and (3) the radio-emitting extragalactic-jet plasmas, which emanate from the cores of large galaxies. Because of inaccessibility, the critical parameters of most astrophysical plasmas (temperatures, densities, and magnetic-field strengths) are not known. *See* ASTROPHYSICS, HIGH-ENERGY; GALAXY, EXTERNAL; PULSAR; SUPERNOVA.

Some of the major plasma-physics issues that are under study with naturally occurring plasmas are the energization of charged particles, the reconnection of magnetic fields (temporal changes in magneticfield topology), the production of magnetic fields by dynamos, the production of electromagnetic waves, the interaction between waves and particles, and the transport of mass, momentum, and energy across magnetic fields.

Naturally occurring plasmas are in general difficult to measure. The solar-wind, ionospheric, and magnetospheric plasmas are diagnosed by singlepoint measurements by rockets and satellites; the solar atmosphere and all astrophysical plasmas are unreachable and must be diagnosed by the light and radio waves that they emit; and lightning is unpredictable and inhospitable to instruments and must be diagnosed primarily by the light that it emits. As a consequence of limited diagnostics, theoretical analysis and laboratory-plasma experiments play supporting roles in the investigations of naturally occurring plasmas. Joseph E. Borovsky

Bibliography. F. F. Chen, Introduction to Plasma Physics and Controlled Fusion, 2d ed., 1984; H. Friedman, Sun and Earth, 1986; M. C. Kelley, The Earth's Ionosphere, 1989; D. M. Manos and D. L. Flamm (eds.), Plasma Etching, 1989; G. K. Parks, Physics of Space Plasmas, 1991; L. Spitzer, Physical *Processes in the Interstellar Medium*, 1978; H. Stix, *Waves in Plasmas*, 1992; M. A. Uman, *The Lightning Discharge*, 1987; J. Wesson, *Tokamaks*, 2d ed., 1997.

Plasma diagnostics

Techniques to measure the properties and parameters of ionized gases. Plasmas are gases in which a sufficient proportion of the atoms are ionized that the resulting mixture of free electrons and positively charged ions exhibits collective behavior resulting from electromagnetic interactions. *See* PLASMA (PHYSICS).

Plasma properties. The simplest properties of a plasma that one might wish to measure are the density (*n*) of particles per unit volume, the fluid velocity ($\langle \mathbf{v} \rangle$), and the temperature (*T*). These gaseous parameters may be different for the different particle species in the plasma. There are normally at least two such species, electrons (subscript *e*) and ions (subscript *i*); but often multiple ion species and neutral particles are present. (The brackets around **v** indicate that it is the average value of the particle velocities of all the particles, or of the particles of a particular species.) Because plasma temperatures are generally high, from a few thousand degrees Celsius upward, they are usually expressed in energy units such as electronvolts (1 eV = 1.6×10^{19} J $\simeq 11,600^{\circ}$ C).

Plasmas often possess nonthermal particle distributions. That is, the velocity distribution function of electrons or ions, f_s (s = e, i), which measures the numbers of electrons or ions that have various velocities, has not simply the Maxwell-Boltzmann form corresponding to thermal equilibrium at some temperature but has some more complicated form $f_s(\mathbf{v}_s)$. In that case an ideal that may sometimes be approached is to measure the entire velocity distribution $f_s(\mathbf{v}_s)$ and not just the parameters n, $\langle \mathbf{v} \rangle$, and T, which are moments of this distribution. *See* BOLTZMANN STATISTICS; KINETIC THEORY OF MATTER.

The electromagnetic fields **E** and **B** are also essential parameters that affect the plasma. They are affected in turn by the charge and current densities, ρ and **j**, which are simple sums of the densities and velocities of the plasma species weighted by their charge per particle, $q: \rho = \sum_{s} q_s n_s; \mathbf{j} = \sum_{s} q_s n_s \langle \mathbf{v}_s \rangle$.

Diagnostic techniques. Techniques to measure the plasma parameters, and their space- and time-variation, are based on a wide variety of physical principles. Direct invasion of the plasma using solid probes is possible only for relatively cold and tenuous plasmas; otherwise the probe will be damaged by excessive heat flux. However, powerful noncontact diagnostics are based on measurements external to the plasma of electromagnetic fields, waves propagated through or scattered by the plasma, photons emitted by plasma electrons or atoms, and particle emission, transmission, or scattering. The applicability of different techniques is determined by the plasma parameters, the accessibility of the plasma, and the acceptability of any perturbation of the plasma. For

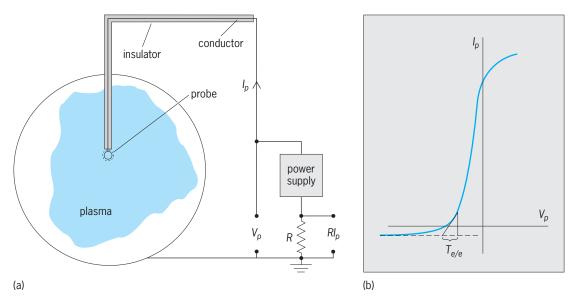


Fig. 1. Langmuir probe. (a) Probe setup. I_p = probe current; V_p = probe potential. (b) Current-voltage (I_p - V_p) characteristic. The base of the small right triangle, T_e/e , is the electron temperature in electronvolts.

example, direct plasma probing is feasible for cold laboratory plasmas and, by spacecraft, for solarsystem plasmas, but not for the core of hot fusion research plasmas (because probes will not survive), for stellar or interstellar plasmas (because they are inaccessible), or for some industrial processing plasmas (because they might be contaminated by the probe). *See* NUCLEAR FUSION; SOLAR WIND.

Electric probes. Direct probing of plasma species is occasionally based on collection of specific atoms from the plasma as surface deposits that are subsequently analyzed by surface analysis techniques. By far the majority of probe measurements, however, are based on measuring the electric current carried to the probe by the plasma particles. The simplest and most direct such measurement, called the Langmuir probe, consists of an electrode of some well-defined geometry (plane, cylinder, and sphere are all used). The potential of this probe is varied and the resulting probe current is measured (**Fig. 1**).

The Debye length λ_D is an important parameter that determines the length scale over which significant electric charge density can exist in a plasma. In the advantageous situation where the Debye length is much smaller than the probe, the probe surface is surrounded by a thin "sheath" of thickness about $5\lambda_D$. Ions are accelerated toward the probe in the region beyond the sheath and enter the sheath at the sound speed, $c_s = [(ZT_e + T_i)/m_i]^{1/2}$, where m_i is the ion mass and Z its charge. Since the sheath is thin, the ion electric current density j_i to the probe is $Zen_{is}c_s$. Here, *e* is the electron charge and n_{is} is the ion density at the sheath edge, which is related to the electron density at the sheath edge n_{es} by $Zn_{is} = n_{es}$. This perturbed electron density is approximately half the unperturbed density n_e far from the probe, because of the dynamics of the acceleration up to the sheath. Therefore, the ion current density is $J_i = \frac{1}{2} e n_e c_s$. Assuming the probe potential V_p to be negative, the electron current to the probe is approximately the random one-way current of a maxwellian electron distribution, $-en_e(T_e/2\pi m_e)^{1/2}$, reduced by a Boltzmann factor, $\exp(eV_p/T_e)$. The total current in then given by Eq. (1). The electron temperature

$$j = \frac{1}{2} e n_e c_s - e n_e (T_e/2\pi m_e)^{1/2} \exp(eV_b/T_e) \quad (1)$$

in electronvolts, T_e/e , can thus be deduced from the logarithmic slope of the probe current with respect to V_p . A simple graphical construction is shown in Fig. 1*b*. The density can then be deduced from the ion saturation current, which is the value of *j* at large negative values of V_p , where the second term in the right-hand side of Eq. (1) is small. Some appropriate value for T_i/T_e must be assumed.

The Langmuir probe cannot provide an independent measurement of the ion temperature because the ions are usually the attracted species and their current is almost independent of probe potential. More elaborate plasma flux measurements can be performed with devices called gridded energy analyzers, which can obtain energy distribution functions of both electrons and ions at the expense of far greater complexity, greater probe size and hence plasma perturbation, and often loss of absolute density calibration.

Electric and magnetic fields. Langmuir probes can provide measurement of the electrostatic potential in the plasma, and hence the electric field, although not quite directly, since the "floating potential" adopted by a probe at zero current is actually a few times T_e below the unperturbed space potential. The magnetic field can be measured by sensors based either on simple induction in a coil for pulsed fields, or typically on the Hall effect in a semiconductor for static fields. The former approach dominates fusion plasma research because of its linearity, noise insensitivity, and robustness, while the latter dominates space measurements in which sensitivity and dc response are essential. Magnetic probes suffer the same

limitations as Langmuir probes associated with overheating for internal plasma measurements. However, in plasmas that are sufficiently tenuous that they do not perturb the field, the internal magnetic field can often be deduced from external measurements by solving the governing magnetostatic equation, $\nabla \times \mathbf{B} = \mu_0 \mathbf{j}$, with $\mathbf{j} = 0$. If the plasma carries current, this equation can be used in the other way, to deduce the current density j from measurements of the magnetic field **B**. Any plasma whose pressure, p(equal to $n_e T_e + n_i T_i$), is comparable to the internal magnetic field pressure, $B^2/2\mu_0$, carries appreciable current because it must satisfy the equilibrium equation $\mathbf{j} \times \mathbf{B} = \nabla p$. The plasma pressure thus tends to reduce the magnetic field inside it. A measurement of this "diamagnetic effect" provides a direct plasma pressure measurement. A more indirect estimate of plasma temperature can be obtained if both the current density and electric field are available, giving the plasma electrical conductivity $\sigma = i/E$ (parallel to B), because the conductivity is a function of electron temperature, equal to $\sigma \simeq 1300 T_e^{3/2} \Omega^{-1} \mathrm{m}^{-1}$ (for singly charged ions and negligible neutral collisions), where the electron temperature T_e is expressed in electronvolts. See ELECTRIC FIELD; HALL EFFECT; MAGNETIC INSTRUMENTS; MAGNETISM.

Refractive index measurements. Propagating an electromagnetic wave through the plasma and measuring the plasma's effect on it is a powerful nonperturbing diagnostic technique. The plasma's presence modifies the refractive index, kc/ω , where ω and **k** are the wave's frequency and wave vector and *c* the velocity of light. For the so-called ordinary wave, which propagates in a magnetic-field-free plasma, or propagates perpendicular to the ambient magnetic field **B**₀ with its electric field **E** polarized along **B**₀ (that is $\mathbf{k} \cdot \mathbf{B}_0 = 0$ and $\mathbf{B}_0 || \mathbf{E}$), the refractive index at high frequency is given by Eq. (2). Here, ω_p , called

$$kc/\omega = \left(1 - \frac{\omega_p^2}{\omega^2}\right)^{1/2}$$
(2)

$$\omega_p = \sqrt{\frac{n_e e^2}{\epsilon_0 m_e}} \tag{3}$$

the plasma frequency, is given by Eq. (3), where ϵ_0 is the permittivity of vacuum. The measurement of the ordinary-wave refractive index thus provides a measure of the plasma electron density. The measurement requires a wave frequency larger than the plasma frequency for the plasma of interest; otherwise the index becomes imaginary and the wave is not transmitted. In fact, a wave frequency up to 20 times higher than ω_p is best for preventing excessive refraction but retaining adequate sensitivity. Coherent sources from microwave to x-ray are used, depending on the plasma to be measured, with farinfrared lasers dominating the magnetic fusion applications. *See* LASER; REFRACTION OF WAVES.

The index measurement is performed by measuring the phase difference caused by the plasma along an optical path, using some form of interferome-

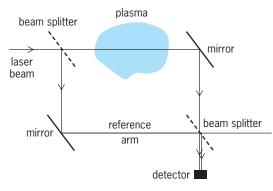


Fig. 2. Plasma interferometer with simple Mach-Zehnder configuration.

ter (**Fig. 2**); the Michelson or Mach-Zehnder configurations are most popular. The phase shift is approximately proportional to the average plasma density along the optical path. A variety of welldeveloped heterodyne techniques exist for obtaining the phase shift unambiguously in real time at high speed. Multiple optical paths are often used to obtain spatial profile information; and in the limit, a two-dimensional interferometric image of the plasma can be obtained, representing the plasma line density projected along the probing beam. *See* INTERFE-ROMETRY.

If the plasma density somewhere along the optical path rises to a value such that $\omega_p = \omega$, reflection occurs at that point. By detecting the reflected wave and measuring its phase or preferably group delay (that is, the round-trip propagation time), the location of the reflection point can be determined. This "reflectometer" measurement can also give density profile information if performed over an appropriate range of wave frequencies, but is generally less versatile and more sensitive to plasma fluctuations than interferometry.

At propagation angles other than perpendicular to **B**, the magnetic field enters into the refractive index, which offers the diagnostic opportunity for measuring **B**. The predominant effect is Faraday rotation: the progressive angular displacement of the angle of wave polarization as it propagates. The rotation is proportional to the path-integral of $n_e(\mathbf{B} \cdot \mathbf{k})$; so if the density is determined independently, the component of **B** along the propagation direction is obtained. This technique has been used in the laboratory, and also for interstellar radiation, providing one of the few measurements of the interstellar magnetic field.

Thomson scattering. In addition to causing the phase shifts measured by interferometry, plasma particles scatter electromagnetic waves by Thomson scattering. Collection of the scattered radiation provides an important diagnostic technique. The scattering can take place either incoherently (the powers of light scattered from individual particles add) or coherently (scattered electric fields add), depending on whether the quantity $2\lambda_D k \sin \theta$ is much greater than 1 (incoherent) or not (coherent). Here θ is the scattering angle.

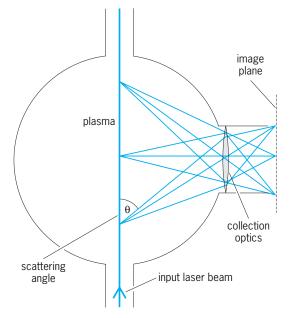


Fig. 3. Thomson scattering apparatus that uses imaging of the scattered light to give a spatial profile of the electron temperature T_e and the electron density n_e .

Incoherent scattering is purely from the electrons. Each scattered photon is Doppler-shifted by an amount corresponding to the velocity of the electron from which it scatters. Consequently there is a one-to-one correspondence between the scattered spectrum and the electron velocity distribution. Unfortunately the Thomson scattering cross section is so small (6.652 \times 10⁻²⁹ m²) that typically only perhaps one photon in 10^{13} might be scattered into the collection optics. Therefore high-energy pulsed laser sources are essential, great care to eliminate spurious signals must be taken, and even so the measurement is difficult and the total detected spectrum is often only a few hundred photons. It is therefore very rare for details of the electron distribution other than temperature and density to be obtainable. However, for those parameters that are measured, the spatial localization defined by the input and collection crossed optical paths permits highly detailed and unambiguous profiles to be obtained (Fig. 3).

Coherent scattering samples the correlations between electrons, notably their density fluctuations. Therefore coherent scattering is used to measure plasma turbulence or to sense high-amplitude heating waves inside the plasma. Also, since electrons form correlated shielding clouds around the ions of the plasma, coherent scattering is able to sense the velocity of these clouds and hence measure ion parameters such as temperature. *See* SCATTERING OF ELECTROMAGNETIC RADIATION.

Plasma radiation. Electromagnetic wave emission from the free electrons in the plasma occurs predominantly from acceleration either by the ambient magnetic field (cyclotron or synchrotron emission) or by collisions with other plasma particles (bremsstrahlung).

Cyclotron emission in magnetically confined

plasmas provides a spatially resolved measurement of electron temperature because the resonant frequency (eB/m_e) identifies the location of emission when *B* is a known function of position; and in the usual situation that the plasma is optically thick, the intensity is equal to the blackbody level, $\omega^2 T_e/8\pi^3 c^2$ (the Rayleigh-Jeans approximation), that is, it is proportional to the electron temperature. In astrophysical plasmas, cyclotron radiation from relativistic electrons is often intense enough for terrestrial detection, especially when the distribution function is such that collective amplification of the waves occurs. The radiation provides information about both the electrons and the magnetic field. See CYCLOTRON RESONANCE EXPERIMENTS; HEAT RADIA-TION; PARTICLE ACCELERATOR.

Bremsstrahlung from optically thin (hot and tenuous) plasmas has a photon energy spectrum that resembles the electron energy distribution; so, for example, its logarithmic slope gives the electron temperature. The spectrum for kiloelectronvolttemperature plasmas is measured by x-ray techniques such as pulse-height analysis or approximately using foil filters. Large, dense, or cold plasmas are often optically thick to bremsstrahlung, in which case their spectrum is blackbody and provides electron temperature. Alternatively, by backlighting the plasma (for example with x-rays), their bremsstrahlung (collisional) absorption provides a measure of line-averaged density. This last technique is particularly important for laser-produced plasmas and inertial fusion research. See BREMSSTRAHLUNG.

Plasma spectroscopy. Line radiation arises from electronic transitions between bound levels of atomic species in the plasma. The characteristic wavelengths serve to identify the species, and the intensities are proportional to the excited-state densities. Therefore line radiation is the primary method for determining the atomic and molecular composition of plasmas. Excitation of the atom from the ground state usually occurs by electron impact, requiring an electron energy greater than the excitation energy. Therefore the ratio of the intensities of lines from different excited states whose excitation energies differ by an amount comparable to the electron temperature, is a measure of electron temperature. The rate of excitation is proportional to electron density; however, in order to eliminate from the measured intensity its dependence on the atomic density, a line ratio is generally chosen for measurement. If the population of one excited level of the line pair chosen is substantially affected by collisional deexcitation, because it is a metastable level for example, or is populated by recombination, then the ratio is sensitive to electron density. See ATOMIC STRUCTURE AND SPECTRA; SCATTERING EXPERIMENTS (ATOMS AND MOLECULES).

Lines are broadened by the Doppler effect arising from thermal atom velocities. So the spectral line shape reflects the atom velocity distribution and, especially, its width gives the atom temperature. If the atoms are charged (because of being a partially ionized species), they are collisionally well coupled to the other ions in the plasma and often, therefore, represent the overall ion temperature. Moreover, if the plasma is flowing, its ion velocity can be obtained from the spectral line shift. At higher densities, broadening can also arise from collisions. This Stark broadening gives a width approximately proportional to $n^{2/3}$ and provides a density measurement. Spectral lines are split by magnetic fields (Zeeman effect) and electric fields (Stark effect), and the different components of the resulting multiplets have specific polarization with respect to the field. This effect thus provides an opportunity to measure both the field's amplitude and direction. *See* DOPPLER EFFECT; SPECTROSCOPY; STARK EFFECT; ZEEMAN EFFECT.

Heavy-particle diagnostics. Neutral atoms are continually formed from plasma ions by charge exchange with other neutrals or recombination with electrons. The neutrals formed may escape promptly from the plasma, providing a means for detection of the ion distribution using a neutral-particle spectrometer. In fusion research plasmas, fusion reactions give rise to megaelectronvolt-energy neutrons and charged products. These too provide information on the ion temperature and density via the reaction rate and the emitted spectrum.

Energetic neutral or ion beams, generated specifically for the purpose, can be used for active plasma probing. Their attenuation is generally proportional to plasma density; but more important specialpurpose applications include the internal measurement of potential (by local beam ionization in the plasma) and the provision of localized atomic species for other spectroscopic diagnostics. I. H. Hutchinson

Bibliography. O. Auciello and D. Flamm, *Plasma Diagnostics*, Academic Press, Boston, 1989; R. H. Huddlestone and S. L. Leonard, *Plasma Diagnostic Techniques*, Academic Press, New York, 1965; I. H. Hutchinson, *Principles of Plasma Diagnostics*, Cambridge University Press, New York, 1987; W. Lochte-Holtgreven, *Plasma Diagnostics*, North-Holland, Amsterdam, 1968.

Plasma propulsion

The imparting of thrust to a spacecraft through the acceleration of a plasma (ionized gas). A plasma can be accelerated by electrical means to exhaust ve-

locities considerably higher than those attained by chemical rockets. The higher exhaust velocities (specific impulses) of plasma thrusters usually imply that, for a particular mission, the spacecraft would use less propellant than the amount required by conventional chemical rockets. This means that for the same amount of propellant a spacecraft propelled by a plasma rocket can increase in velocity over a set distance by an increment larger than that possible with a chemical propulsion system. An increasing number of near-Earth satellites have benefited from the propellant savings of plasma thrusters and rely on them for attitude control and station-keeping maneuvers. Generally, the larger the required velocity increment, the more beneficial is plasma propulsion compared to propulsion with rockets having lower exhaust velocities. Consequently, many ambitious missions-such as deep-space exploration including Moon tours and orbiters to outer planets, sample return missions, and heavy cargo and piloted missions-may be enabled by plasma propulsion. See PLASMA (PHYSICS); SPECIFIC IMPULSE.

Plasma propulsion is one of three major classes of electric propulsion, the others being electrothermal propulsion and ion (or electrostatic) propulsion. In electrothermal devices the working fluid is a hot nonionized or very weakly ionized gas. All three electric propulsion systems offer high specific impulses; the plasma devices are further distinguished by their high thrust density (thrust per unit exhaust area) as indicated in the **table**. *See* ELECTROTHERMAL PROPUL-SION; ION PROPULSION; SPACECRAFT PROPULSION.

Types of plasma thrusters. One common type of electric thruster in which a plasma is generated is the ion thruster. However, in the ion thruster only the plasma ions are accelerated, and therefore this variant of plasma propulsion is not discussed here. The three most mature types of plasma propulsion devices are the Hall thruster, the magnetoplasma dynamic thruster (MPDT), and the pulsed plasma thruster (PPT).

Although these three types of plasma thrusters share the distinction of having been in use, or in the case of the MPDT having been tested, in space, there are many other promising plasma propulsion concepts that have been advanced to various degrees of technological readiness over the past few decades. These concepts differ in [1] their mode of operation (pulsed or steady-state) [2] methods of generating

Propulsion system	Working fluid	Acceleration	Exhaust velocity, km/s (mi/s)	Thrust density, N/m ² (lbf/ft ²)	Efficiency, %
Plasma	Plasma	Electromagnetic, or electromagnetic with thermal expansion	10–60 (6–40)	10–1500 (0.2–31)	10–60
Electrothermal	Hot gas, weakly ionized	Thermal expansion	2–10 (1.2–6)	10–100 (0.2–2)	40
lon or electrostatic	lons	Electrostatic	30-100 (18.7-60)	0.1-4 (0.002-0.1)	60–90

the plasma (electrode-based discharge, coil-based induction, or antenna-based radiation) [3] plasma acceleration mechanisms (Lorentz forces; entrainment by plasma sheet; or traveling waves, whereby an alternating magnetic field is synchronized with the electric-field pulses to drive the plasma), [4] expansion through magnetic nozzles (whereby a heated plasma is exhausted through an expanding magnetic field), [5] implementation of the magnetic field (applied or self-induced), [6] propellantfeeding schemes (gas-fed or solid-fed, primarily polytetrafluoroethylene), [7] geometry (rectangular, coaxial, pinch, parallel-plate, constant, or variablearea geometry), and other details of implementation.

Hall thruster. The most successful plasma thruster to date is the Hall thruster. It has been gaining acceptance in the West since the early 1990s, after three decades of improvement in the former Soviet Union. More than 200 Hall thrusters have been flown so far on about 30 spacecraft, mostly by the former Soviet Union and Russia. In a Hall thruster, (**Fig. 1**) a mostly radial magnetic field is applied inside the acceleration chamber. A hollow cathode, placed outside the chamber, is kept at a few hundred volts below the potential of an annular anode located upstream in the device, thus setting a mostly axial electric field inside the chamber. As the cathode is heated, it emits electrons, which drift in the opposite direction of the electric field and enter the device, where they are subjected to the radial magnetic field and the axial electric field. This leads to an azimuthal (that is, about the device's axis of symmetry) drift for the electrons, which constitutes an electric current called the Hall current. The drifting electrons collide with propellant gas molecules (typically xenon) injected in the device through the anode. These impacts produce a plasma that finds itself acted on by an electromagnetic force, called the Lorentz force, which results from the interaction of the radial magnetic field and the azimuthal Hall current that are perpendicular to each other. The plasma is thus accelerated and exhausted out of the thruster at high exhaust velocities (from 10 km/s or 6 mi/s to above 50 km/s or 30 mi/s depending on the available power). The acceleration mechanism can also be described, equivalently, to be the result of the action of the applied electric field on the newly born ions (whose mass is too large to allow them to be magnetized) with the electrons acting as a neutralizing cloud whose momentum is coupled to the ions via interparticle electrostatic fields. Since the entire plasma is accelerated, and not only the positive ions, there is no space-charge buildup as in the ion thruster, and the device can have a thrust density many times that of the ion engine. See HALL EFFECT; MAGNETISM; MAGNETOHYDRODYNAMICS.

Hall thrusters can attain thrust efficiencies (ratio of propulsive power of the exhaust jet to input electric power) in excess of 50% and are thus attractive for many space applications. They have been developed

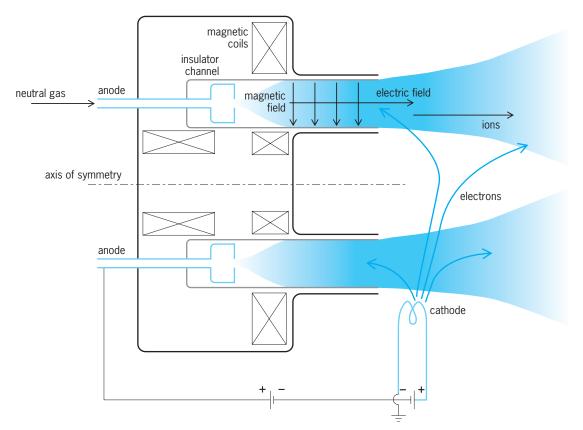


Fig. 1. Hall thruster schematic, showing the electric circuit, the applied electric and magnetic fields, and typical trajectories of ions and electrons. The electrons entering the cylindrical device execute azimuthal drifts (about the shown axis of symmetry) and collide with the injected neutral gas to produce the plasma.

for power levels of the order of 1 kW. Although they had considerable development in the former Soviet Union, a number of research problems remain, including scaling them to high power (in excess of 10 kW) for applications requiring high thrust, high specific impulse, or low power (below 50 W) for micropropulsion; lowering the divergence of the exhaust beam of plasma (which can adversely impact the performance of the device and the integrity of exposed spacecraft surfaces); and understanding the basic processes controlling the various oscillations, plasma waves, and instabilities that are excited in the plasma.

Magnetoplasmadynamic thruster. One of the most promising and thoroughly studied electromagnetic plasma accelerators is the magnetoplasmadynamic thruster. In this device the plasma is both created and accelerated by a high-current discharge between two coaxial electrodes. The discharge is due to the breakdown of the gas as it is injected in the interelectrode region. The acceleration process can be described as the result of the action of a body force acting on the plasma (Fig. 2). This body force is the Lorentz force created by the interaction between the mostly radial current conducted through the plasma and the magnetic field. The latter can be either externally applied by a magnet or self-induced by the discharge, if the current is sufficiently high. The reliance on the radial component of the current (which is aligned with the electric field) instead of the Hall component (perpendicular to both electric and magnetic fields) allows operation at much higher plasma densities (and thrust densities) than with the Hall thruster.

Microscopically, the acceleration process can be described as the momentum transfer from the electrons, which carry the current, to the heavy particles through collisions. Such collisions are responsible for the creation of the plasma (ionization) and its acceleration and heating (Joule heating). Even though charged particles (ions and electrons) in an electromagnetic field drift at the same velocity irrespective of their mass, the electrons recover faster from a collision because of their small mass and thus are the prime beneficiaries of the energy in the electromagnetic field. In other words, the electrons carry the current responsible for the acceleration of the plasma in the magnetoplasmadynamic thruster. Indeed, the thrust T of a purely electromagnetic coaxial and cylindrical self-field plasma accelerator is directly and strongly related to that current by Eq. (1), where b is a constant dependent on the

$$T = bJ^2 \tag{1}$$

ratio of the electrode radii and J is the total current conducted through the plasma. The thrust of a purely electromagnetic device is therefore independent of the mass-flow rate and type of propellant. The electromagnetic thrust of experimental prototypes of the magnetoplasmadynamic thruster obeys this law when operated at full ionization, which corresponds to operation at current levels at and above

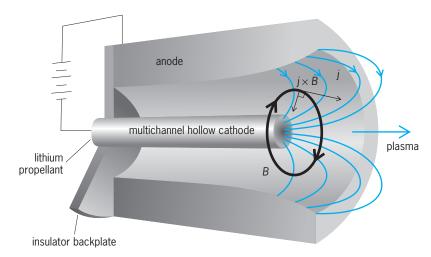


Fig. 2. Self-field, coaxial, magnetoplasmadynamic thruster. Here the current density *j* (color curves) and the induced azimuthal magnetic field *B* (black circle) interact to produce the mostly axial Lorentz force density. Electrodes are typically made of tungsten.

 $J_{\rm fi}$, the current required for effective full ionization of the plasma. This is the regime of $\xi \ge 1$, where ξ is the nondimensional current defined by Eq. (2) and $J_{\rm fi}$ in amperes is given by Eq. (3). Here \dot{m} is

$$\xi \equiv \frac{J}{J_{\rm fi}} \tag{2}$$

$$J_{\rm fi} = \left(\frac{\dot{m}}{b}\right)^{1/2} \left(\frac{2\epsilon}{M}\right)^{1/4} \tag{3}$$

the mass-flow rate of propellant into the device in kilograms per second, and
$$M$$
 and ϵ are the atomic mass and first ionization potential of the propellant in kilograms and joules, respectively. *See* IONIZATION

POTENTIAL. For typical magnetoplasmadynamic thruster geometry, b is about 0.17 N/kA2 (0.038 lbf/kA2), and for operation with 3 g/s of argon the nominal condition $\xi = 1$ is attained at J = 12 kA [from Eqs. (2) and (3)]. Furthermore, Eq. (1) implies an electromagnetic thrust of 25 newtons. At these conditions, the voltages associated with the back electromotive force, plasma heating, and electrode sheaths add up to about 80 V. Power levels exceeding 1 MW are thus required for Eq. (1) to hold at these operating conditions. It is, however, the universal scaling parameter ξ that determines the applicability of the electromagnetic thrust law [Eq. (1)] and not the power level. In other words, it is possible to attain the regime of electromagnetic acceleration at considerably more modest power levels as long as $\xi \ge 1$ is satisfied. This can be achieved by reducing the propellant mass-flow rate, and some devices are operated at power levels as low as 15 kW. However, the power losses associated with the electrodes and ionization dominate at lower power, and thrust efficiencies exceeding 20% are still attainable only with the megawatt-class devices or with the use of lithium as propellant. Below about 250 kW, an applied magnetic field is generally needed to improve the performance.

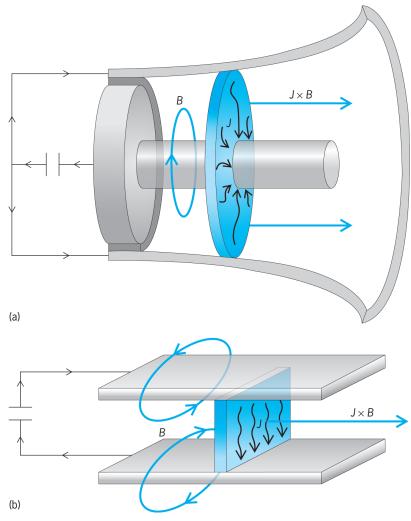


Fig. 3. Schematic of two configurations of the pulsed plasma thruster, showing the current sheet, the induced magnetic field, and the resulting electromagnetic force. (a) Coaxial configuration. (b) Parallel-plate configuration.

Research has demonsrated the advantage of using lithium vapor as propellant, injected, along with small amounts of barium vapor, from a multichannel hollow cathode (Fig. 2). Lithium (which has a low ionization potential) offers the advantage of low ionization losses, and barium, which has a low work function, ensures a low cathode operation temperature, which lowers the cathode evaporation rate, leading to longer lifetime. Such a magnetoplasmadynamic thruster variant is sometimes called the lithium Lorentz force accelerator (Li-LFA).

The scaling of the various power sinks with the input power is an area of active research, and of major concern for improving efficiency. Other research areas include the understanding of various mechanisms controlling the efficiency as well as an operation-limiting phenomenon called onset that occurs when ξ exceeds a critical value. Operation above onset can lead to severe damage to the anode and to strong voltage fluctuations.

Pulsed plasma thrusters. This type of plasma thruster, called the PPT, relies on pulsing a high-current gas discharge on a time scale of a few microseconds. This is done by quickly discharging the

energy stored in a capacitor into a gas that is either injected (gas-fed PPT) or vaporized from a solid propellant (ablative PPT), typically polytetrafluoroethylene. The quick rise in current through the gas forces the formation of a plasma sheet through which the current flows as shown in Fig 3, where two possible configurations are illustrated. The current sheet is subjected to the Lorentz force produced by the interaction of the current it carries and the induced magnetic field. This causes the current sheet to accelerate downstream, ionizing and entraining ambient gas as in a snowplow. The exhausted plasma and gas give the propulsive impulse. The pulsed operation of such electric propulsion devices allows powerlimited spacecraft to take advantage of high-current plasma acceleration at the low average power available on current spacecraft (hundreds of watts).

Depending on the gas pressure in the device, the gas injection mode, the rate of current rise, and other implementation details, acceleration by thermal effects can dominate over electromagnetic acceleration, or the acceleration can be due to a so-called deflagration effect where the current pattern spreads over the discharge chamber and exerts a body force that extends over a large volume of the plasma. This latter mode can lead to higher exhaust velocities and efficiencies.

Another type of plasma thruster that is an inherently pulsed device, the pulsed inductive thruster (PIT), relies on using coils placed outside the discharge to induce a current sheet in a gas injected from a downstream nozzle. The PIT has the advantage of foregoing the use of electrodes, which often are the life-limiting factor for plasma thrusters. Another pulsed inductive thruster variant, the Faraday accelerator with radio-frequency assisted discharge (FARAD), relies on injecting a plasma, instead of a gas, from an upstream pre-ionization radio-frequency plasma source and on an applied magnetic field to control the plasma injection and acceleration processes.

Propellants for plasma thrusters. Plasma acceleration allows a wider choice of propellants than any other type of propulsion. Plasma-propelled spacecraft thus have the potential of using propellants that are abundant in their environment rather than from Earth alone. There are, however, some constraints. For MPDTs, although Eq. (1) implies that the magnitude of the electromagnetic thrust is independent of the propellant, the attainability of the nominal condition [$\xi = 1$, Eqs. (2) and (3)], compatibility with thruster components, requirements of rate kinetics (for the case of a hybrid thruster), storability, and toxicity restrict the choice of appropriate propellants. Laboratory MPDTs and PPTs have been tested with argon, krypton, xenon, helium, hydrogen, air, deuterium, ammonia, hydrazine, lithium, and polytetrafluoroethylene, among other propellants and mixtures of propellants. For Hall thrusters, the plasma dynamics constrain the choice of propellants to gases of high atomic mass and reatively low ionization potential. Xenon and bismuth have been the most favored propellant.

Plasma thruster use in space. More than 200 plasma thrusters have been flown to date, mostly by the former Soviet Union and Russia, and mostly of the Hall thruster type. This excludes the dozens of ion thrusters presently used on geosynchronous spacecraft. A few PPTs of the solid-fed [polytetrafluoroethylene-ablative] type were used on Russian and U.S. spacecraft in the 1970s and 1980s. The first NASA spacecraft to use an ablative PPT was the EO-1 satellite launched in 2001. These ablative PPTs have a modest thrust and efficiency. The thrust level, however, is very reproducible and can be precisely controlled, thus making these systems ideal for fine-tuning spacecraft attitude control. A gas-fed quasi-steady pulsed magnetoplasmadynamic thruster was tested in space in 1998 onboard the Japanese spacecraft SFU; but MPDTs, which are inherently high-power devices, are still awaiting the availability of high electric power in space before their adoption in actual missions. With technical advancements, successful flight tests, and the advent of a new era of deep-space exploration, plasma propulsion is presently entering its era of wide application. See SPACE POWER SYSTEMS. Edgar Y. Choueiri

Bibliography. R. L. Burton and P. J. Turchi, Pulsed plasma thruster, J. Propul. Power, 14(5):716-735, 1998; E. Y. Choueiri, A critical history of electric propulsion: The first 50 years, J. Propul. Power, 20(2):193-203, 2004; E. Y. Choueiri, Fundamental difference between the two variants of Hall thrusters, Phys. Plasmas, 8(11):5025-5033, 2001; R. G. Jahn, Physics of Electric Propulsion, Dover Publications, 2006; H. R. Kaufman, Technology of closed-drift thrusters, AIAA J., 23(1):78-80, 1985; A. Kodys and E. Y. Choueiri, A Critical Review of the State-of-the-Art in the Performance of Applied-Field Magnetoplasmadynamic Thrusters, AIAA Pap. 05-2005-4247, 2005; M. Martinez-Sanchez and J. E. Pollard, Spacecraft electric propulsion-An overview, J. Propul. Power, 14:(5), 688-699, 1998.

Plasmid

A circular extrachromosomal genetic element that is ubiquitous in prokaryotes and has also been identified in a number of eukaryotes. In general, bacterial plasmids can be classified into two groups on the basis of the number of genes and functions they carry. The larger plasmids are deoxyribonucleic acid (DNA) molecules of around 100 kilobase (kb) pairs, which is sufficient to code for approximately 100 genes. There is usually a small number of copies of these plasmids per host chromosome, so that their replication must be precisely coordinated with the cell division cycle. Their replication genes are clustered within a small segment of the genome. Larger plasmids usually can mediate their own transfer from cell to cell by bacterial conjugation. The genes mediating transmissibility are clustered in a segment of approximately 25 kb. The other regions (around 70 kb) of these larger plasmids can accommodate as many as 70 additional genes which determine a variety of properties (see **table**). The plasmids in the second group are smaller in size, about 6–10 kb. These plasmids may harbor 6–10 genes and are usually present in multiple copies (10–20 per chromosome). Although the smaller plasmids do not carry transfer genes, many of them can be mobilized to recipient cells if a larger transmissible plasmid is present in the same host cell.

Distribution. Plasmid has been identified in a large number of bacterial genera. Some bacterial species harbor plasmids wih no known functions (cryptic plasmids) which have been identified as small circular molecules present in the bacterial DNA. The host range of a particular plasmid is usually limited to closely related genera. Some plasmids, however, are much more promiscuous and have a much broader host range.

Functions. The functions specified by different bacterial plasmids are usually quite specialized in nature. Moreover, they are not essential for cell growth since the host bacteria are viable without a plasmid when the cells are cultured under conditions that do not select for plasmid-specified gene products. Plasmids thus introduce specialized functions to host cells which provide versatility and adaptability for growth and survival. Plasmids which confer antibiotic resistance (R plasmids) have been extensively characterized because of their medical importance.

Plasmid incompatibility. A fundamental characteristic of plasmids by which they have been classified into specific groups is the ability of different plasmids to coexist stably in a host cell. Some pairs of plasmids, isolated independently from nature, can coexist in the same host cell: this is referred to as compatibility. Other pairs of plasmids are called incompatible since they are not inherited stably when present in the same host cell. Based on these criteria, over 25 plasmid incompatibility groups have been identified in gram-negative bacteria, and more than a dozen incompatibility groups have been reported in gram-positive bacteria. R plasmids are represented in many different incompatibility groups, whereas other plasmid-specified functions are not as widely distributed among incompatibility groups. Experiments have shown that plasmid incompatibility is a result of the control of plasmid replication. Plasmids can be changed into a new incompatibility group as a result of a single base change in replication gene which controls the plasmid copy number.

R plasmids. The large increase in multiply drugresistant bacterial pathogens has resulted from increasing use of antibiotics in chemotherapy and animal husbandry. Gram-negative and gram-positive bacterial species have acquired resistance to a variety of antibiotics such as penicillins and cephalosporins, chloramphenicol, aminoglycosides (amikacin, gentamycin, kanamycin, streptomycin, spectinomycin, tobramycin), tetracyclines, sulfonamides, trimethoprim, and MLS (macrolide, lincomycin, and streptogramin B-type) and to toxic heavy metals such as mercuric, cadmium, and silver ions. R plasmids do not confer resistance to all these antibiotics and

	Conferred by plasmids of		
Property	Gram-negative bacteria	Gram-positive bacteria	
Autonomous replication	+	+	
Genetic transfer	+	+	
Bacteriocin production and resistance	+	+	
Antibiotic resistance	+	+	
Resistance to heavy metals	+	+	
Resistance to agents that damage DNA	+		
Metabolism of hydrocarbons and aromatic compounds	+		
	(Pseudomonas)		
Sugar fermentation	+	+	
Hydrogen sulfide production	+		
Protease and urease production		+	
Toxin production	+	+	
Virulence and colonization factors	+		
Hemolysin production	+	+	
Restriction/modification	+	+	
Antibiotic production	Streptomyces coelicor		
Tumorigenicity in plants	Agrobacterium tumefaciens		

heavy metals simultaneously, but transmissible R plasmids conferring resistance to four to six antibiotics are not uncommon. Smaller R plasmids usually confer resistance to one or two antibiotics or heavy metals. *See* ANTIBIOTIC.

The biochemical mechanisms of plasmiddetermined antibiotic resistance have been elucidated in the majority of instances. At least six different types of resistance mechanisms have been described: (1) cleavage of a specific bond in an antibiotic, as in the inactivation of penicillins or cephalosporins by β -lactamases; (2) inactivation by chemical modification, as in the acetylation of chloramphenicol; (3) synthesis of antibiotic-insensitive substitute enzymes, as in the case of sulfonamide or trimethoprim resistance; (4) alteration of a macromolecular target site, as in methylation of 23S ribonucleic acid (RNA) to confer erythromycin or lincomycin resistance; (5) lower cell permeability, as in the case of tetracycline resistance; and (6) sequestering of the antibiotic so that it cannot bind to its target site as in the case of the binding of fusidic acid by chloramphenicol acetyltransferase (the enzyme which acetylates chloramphenicol). In this case the same gene confers resistance to two structurally unrelated antibiotics by two entirely different mechanisms. Many of the aminoglycoside antibiotics are inactivated by several different

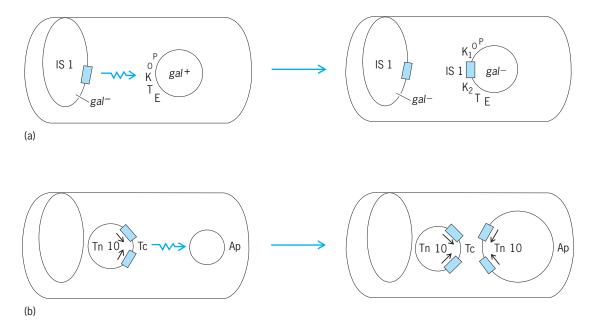


Fig. 1. Translocation of insertion sequences (IS) and transposable elements (Tn). (a) At left an IS 1 element is initially present in the host chromosome which carries a mutation in the *gal* genes which mediate galactose metabolism. The cell can metabolize galactose owing only to the presence of *gal* genes on a plasmid. After insertion of IS 1 into the *gal* genes on the plasmid, these genes are inactivated; the cell at right cannot metabolize galactose. (b) Tn 10, which mediates tetracycline resistance (Tc), translocates from one plasmid to another plasmid which also confers ampicillin resistance (Ap). In both insertion sequence and transposable element translocation, a copy of the element is retained in the donor genome.

types of chemical modification by specific enzymes. *See* DRUG RESISTANCE.

Insertion sequences and transposable elements. Many plasmids have been found to contain genetic elements called insertion sequences and transposable elements. Insertion sequences are specific DNA sequences that exist on bacterial genomes and have the remarkable property of being able to translocate (jump) spontaneously to new sites on the same or other genomes in a cell (chromosome, plasmid, or virus) by a mechanism which is independent of the normal recombination system (Fig. 1). The location of the insertion sequence elements on several bacterial plasmids suggests that they may play a seminal role in plasmid evolution. These insertion sequence elements appear to serve as recombination sites for the fusion of separate genomes, each of which carries a group of specific genes, to form more complex genomes specifying all of the fun of each of the components. Insertion sequence elements may thus serve as linkers in the assembly of different modules of genetic information in the evolution of plasmids and other genomes.

The alarming spread of multiple-drug resistance among bacterial pathogens also appears to be a manifestation of the genetic plasticity made possible by DNA insertions. Most drug resistance genes on bacterial plasmids are flanked by repetitious DNA sequences, usually in inverted orientation, which are analogous to insertion sequences or may even be insertion sequences. This entire unit, referred to as a transposable element, can translocate to new sites in a manner similar to insertion sequence elements with the important difference that the intervening drug resistance genes are transposed in addition to the flanking repetitious sequences. Transposable elements can transpose at low frequency from one genome to another and thus lead to a broad dissemination of the same drug resistance genes among different plasmids. See TRANSPOSONS.

Rearrangements and evolution. Insertion sequence and transposable elements also mediate various types of plasmid deletions and rearrangements, including the amplification of specific plasmid segments to increase the number of copies of specific drug resistance genes (Fig. 2). Analysis of the molecular properties of certain transmissible, multipledrug-resistance plasmids have indicated that these plasmids have evolved by incorporation of genes mediating genetic transfer and various drug resistance genes into a small plasmid containing the replication functions. There are now many examples of the acquisition of specific functions by plasmids by transposition. This ensures the continued evolution of extrachromosomal genetic systems which provides extraordinary versatility and adaptibility to host cells.

Genetic engineering. Plasmids have played a seminal role in the spectacular advances in the area of genetic engineering. Individual genes can be inserted into specific sites on plasmids in cell cultures and the recombinant plasmid thus formed introduced into a living cell by the process of bacterial transformation.

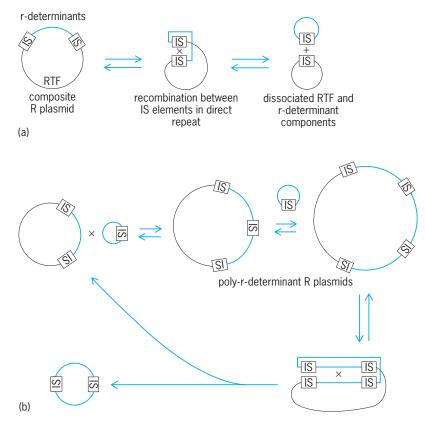


Fig. 2. Plasmid deletions and rearrangements mediated by insertion sequence (IS) elements. (a) The resistance (r-determinants) component of the plasmid is flanked by direct repeats of an IS element. By genetic recombination between the homologous IS elements, the composite plasmid can dissociate into the r-determinants and the resistance transfer factor (RTF) components. (b) Extra copies of the r-determinants can be incorporated into composite R plasmids by using any region of homology to form the multiple, tandem repeats. These poly-r-determinant R plasmids confer higher levels of antibiotic resistance. These molecules can also delete the extra copies of r-determinants to reform the basic composite structure.

Usually both the plasmid vector and the gene to be cloned have been cleaved by a restriction endonuclease which recognizes a specific DNA base sequence of four or six bases. The breaks introduced by most restriction endonucleases in the individual strands of duplex DNA are staggered with the result that the restriction fragments which are formed contain four-base protruding ends of complementary base sequence. These cohesive ends provide the specificity by which the fragment to be cloned is inserted into a single, specific site on the plasmid vector. The singlestrand interruptions can then be sealed by the action of the enzyme ligase. Some restriction endonucleases cleave polynucleotide bonds of bases directly opposite to each other in duplex DNA; such "blunt" ends on restriction fragments can also be joined by ligase to form recombinant DNA molecules, although the ligation step is not as efficient. Plasmid and viral vectors have been developed in both prokaryotes and eukaryotes which make possible a high level of expression of cloned genes. In some cases the vector itself contains the regulatory elements controlling the expression of genes which are not normally expressed in the host cell so that a high level of expression of heterologous genes can be obtained. See GENETIC ENGINEERING. Robert H. Rownd Bibliography. J. Grinsted and P. M. Bennett (eds.), *Plasmid Technology*, 2d ed., 1990; K. G. Hardy, *Bacterial Plasmids*, 1984; K. G. Hardy, *Plasmids*, 1994; R. W. Old and S. B. Primrose, *Principles of Gene Manipulation*, 5th ed., 1994; R. H. Round (ed.), *Advances in Plasmid Molecular Biology*, 1988.

Plasmin

A proteolytic enzyme that can digest many proteins through the process of hydrolysis. Plasmin is found in plasma in the form of plasminogen, which is an inert precursor or zymogen; one site of synthesis is the liver. Plasma also contains inhibitors that limit either the transformation of plasminogen to plasmin (plasminogen activator inhibitors) or the action of plasmin (principally α_2 -antiplasmin and α_2 -macroglobulin).

Activation of plasminogen. In the laboratory, plasminogen can be converted to plasmin by extracts of bacteria, notably streptokinase, a protein that is formed by certain beta-hemolytic streptococci. Plasminogen can also be activated by factors that are present in the body. Tissue plasminogen activators, which are proteolytic enzymes derived from a variety of normal and neoplastic cells, can convert plasminogen to plasmin, as can urokinase, an enzyme found in urine and plasma. Exposure to organic solvents, particularly chloroform, will also lead to the formation of plasmin in human plasma, which is rich in plasminogen.

Plasma itself has the potentiality of activating plasminogen, because it contains intrinsic activators. A tissue plasminogen activator is produced in blood vessel walls, from which it is released into the bloodstream following vascular injury. Further, the process of blood coagulation may foster the activation of plasminogen. Activated Hageman factor (factor XII), activated plasma thromboplastin antecedent (factor XI), plasma kallikrein (an enzyme derived from plasma prekallikrein), and activated protein C, all evolving during blood clotting, bring about reactions leading to the formation of plasmin. Moreover, the presence of a fibrin clot enhances activation of plasminogen by many agents.

Physiologic actions of plasmin. Plasmin can act on many protein substrates. It liquefies coagulated blood by digesting fibrin, the insoluble meshwork of the clot; it also digests fibrinogen, the precursor of fibrin, rendering it incoagulable. The digestion products of fibrinogen and fibrin are anticoagulant substances that further interfere with the clotting process. Plasmin also inactivates several other protein procoagulant factors, particularly proaccelerin (factor V) and antihemophilic factor (factor VIII). Other substrates attacked by plasmin include gamma globulin and several protein or polypeptide hormones found in plasma.

The action of plasmin may be related to certain body defenses against injury. Its capacity to digest fibrin is thought to bring about liquefaction of intravascular clots (thrombi) that might otherwise cause damage by reducing blood flow to tissues. It can convert the first component of complement, a group of plasma proteins important in immune reactions, to the proteolytic enzyme C1 esterase (C1). It can also liberate polypeptide kinins (for example, bradykinin) from plasma precursors. The kinins can reproduce such elements of the inflammatory process as pain, dilatation and increased permeability of small blood vessels, and the migration of leukocytes. In addition to these physiologic actions, plasmin hydrolyzes casein, gelatin, denatured hemoglobin, and certain synthetic esters of arginine and lysine. *See* COMPLEMENT; IMMUNOLOGY; INFLAMMATION.

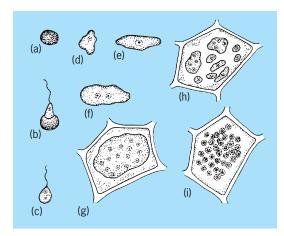
In the test tube, clots formed from plasma or its euglobulin (water-insoluble) fraction dissolve spontaneously, through the action of plasmin. Ordinarily such fibrinolysis is slow, but if the plasma is prepared from the blood of individuals who have been under stress or have severe chronic liver disease, fibrinolysis may be much more rapid. Similarly, the administration of such substances as epinephrine (adrenalin) or des-amino-D-arginine vasopressin (DDAVP) may enhance fibrinolysis. These test-tube phenomena have not been related conclusively to events in the body, but fibrinolysis, although rarely significant, apparently does occur in the body, either as a primary event or secondary to widespread coagulation within the patient's blood vessels (disseminated intravascular coagulation) and may then be responsible for or contribute to hemorrhage. This fibrinolytic purpura has been treated with agents inhibiting either the activation of plasminogen, such as epsilonaminocaproic acid or tranexamic acid, or the action of plasmin, such as aprotinin, a polypeptide that is derived from various tissues. A deficiency of α_2 -plasmin inhibitor has been associated with a hemophilialike disorder, presumably because clots that might stop bleeding readily dissolve. Conversely, a tendency to form intravascular clots (thrombi) has been related to deficiencies or abnormalities of the proteins that foster fibrinolysis.

Fibrinolytic agents, principally tissue plasminogen activator, urokinase, streptokinase, and their derivatives, have been used therapeutically to dissolve thrombi and thereby restore blood flow in the affected vessels. Their chief application is in treating myocardial infarction, dissolving clots that occlude the coronary arteries that supply blood to cardiac muscle and thereby limiting damage by restoring circulation through these vessels. They have also been used to dissolve thrombi in other blood vessels and to dissolve pulmonary emboli, that is, thrombi that have broken off from peripheral veins and traveled to the pulmonary arteries so as to occlude these vessels and thereby damage pulmonary tissue (pulmonary infarction). See BLOOD; ENZYME; FIBRINOGEN; INFARC-TION; RESPIRATORY SYSTEM DISORDERS; UROKINASE. Oscar D. Ratnoff

Bibliography. V. J. Marder and S. Sherry, Thrombolytic therapy: Current status, *N. Engl. J. Med.*, 318:1512–1520, 1585–1600, 1988; O. D. Ratnoff and C. H. Forbes (eds.), *Disorders of Hemostasis*, 3d ed., 1996.

Plasmodiophorida

Protozoa composing an order of Mycetozoia, endoparasitic in plants and primarily causing, for example, club root of cabbage, powdery scab of potatoes, and galls in *Ruppia*. Underground portions of host plants are invaded by young parasites, often a flagellate which sometimes arises from a freshly excysted ameba. Becoming intracellular, the young parasite grows, apparently without phagotrophy, and develops into a plasmodium. At maturity, the plasmodium produces uninucleate cysts (spores) which are released upon degeneration of the damaged cell (see **illus.**). Under favorable conditions,



Typical life cycle of Plasmodiophorida. (a) Uninucleate cyst (spore). (b) Excystment. (c) Flagellate stage. (d) Ameboid stage. (e) Ameboid stage supposedly formed by syngamy. (f) Binucleate ameboid stage. (g) Growing plasmodium inside plant cell. (h) Products of plasmotomy. (i) Uninucleate stages formed from plasmodium (or from products of plasmotomy), prior to encystment. (After R. P. Hall, Protozoology, Prentice-Hall, 1953)

the released spores hatch into uninucleate stages which become the infective forms. In the invaded area, the host's tissue commonly undergoes hypertrophy to form a gall. There are a number of genera, including *Plasmodiophora, Spongospora*, and *Sorosphaera. See* PROTOZOA; RHIZOPODEA; SARCO-DINA; SARCOMASTIGOPHORA. Richard P. Hall

Plasmodiophoromycota

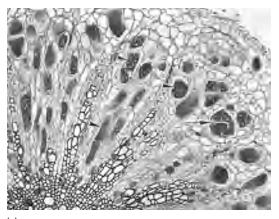
Obligate endoparasites (parasites restricted to living inside a host) of flowering plants, green and yellow-green algae, seaweeds, diatoms, water molds, and ciliates. The Plasmodiophoromycota (plasmodiophorids) are prevalent in both soil and aquatic habitats, where their biflagellate zoospores (independently motile spores) disperse the pathogen and infect hosts. A plasmodium develops within the host cell, often causing cell enlargement and proliferation, resulting in hypertrophy (abnormal enlargement) of host tissue. These organisms include devastating pathogens of food crops, such as *Plasmodiophora brassicae*, which causes clubroot (a plant disease involving malformed and abnormally enlarged roots) of crucifers (for example, mustard, cabbage, radishes, turnips, cauliflower, and broccoli). Plasmodiophoromycota were once thought to be kin to fungi, but recent molecular systematic studies have revealed that they are actually protozoa, members of the phylum Cercozoa and class Phytomyxea. *See* PARASITOLOGY; PROTOZOA.

Life cycle. The life history of plasmodiophorids is complex and includes (1) a flagellated motile stage, the zoospore; (2) an infectious stage, the encysted zoospore (cyst); (3) an unwalled assimilative and feeding stage, the multinucleate plasmodium; (4) a zoospore proliferation stage, the zoosporangium; and (5) a resistant stage, the resting spore. Zoosporangia and resting spores, cleaved from their respective plasmodia, cluster together with a characteristic organization in sori (spore balls).

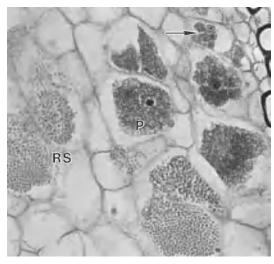
Uninucleate resting spores that survive adverse growth conditions in host tissue debris or directly in the soil germinate with primary zoospores when growth conditions are conducive. Zoospores are uninucleate and microscopic, 3-5 micrometers in diameter. They bear two apically inserted smoothsurfaced flagella. Flagellar basal bodies are especially long, with globules scattered in their core, and are positioned at an angle to each other. Two microtubular roots are associated with each basal body and extend underneath the cell membrane. The anteriorly directed flagellum does most of the work in zoospore motility, with the other, longer, posteriorly directed flagellum essentially serving as a rudder. Plasmodiophorid zoospores have a distinct method for penetration into the host. The zoospore swims to a suitable host, loses its flagella, crawls along the host surface, produces a wall around itself, and adheres to the host. Inside the cyst, a new structure shaped like a bullet (the stachel) forms within a channel (the rohr). A vacuole in the cyst expands and pushes the stachel into the host cell wall, piercing it. The parasite's protoplast is then injected into the host cytoplasm through the pore. See CILIA AND FLAGELLA.

The protoplast enlarges into a multinucleate plasmodium (**Fig.** 1*a*) as it absorbs nutrients from the host. There have been several reports that plasmodia are phagotrophic, ingesting portions of host cytoplasm. Typically the infected host cell enlarges, resulting in swelling of host tissue, as in the distorted club-shaped roots of cabbage infected with *P. brassicae*.

The primary plasmodium then segments (Fig. 1*a*) into thin-walled zoosporangia, which release secondary zoospores. The site of cell and nuclear fusion (sexual reproduction) is controversial, but fusion between secondary zoospores has been reported. Secondary zoospores reinfect host cells with additional zoosporangial plasmodia or with secondary plasmodia (the resting spore producing plasmodia). After secondary plasmodial expansion, the plasmodium cleaves into uninucleate masses with thick chitinous cell walls, the resting spores (Fig. 1*b*). Transmission electron microscopy has revealed synaptonemal complexes (ribbonlike protein structures



(a)



(b)

Fig. 1. Clubroot of cabbage. (a) Cross section of root infected with *Plasmodiophora brassicae*. The thick-walled, hollow cells radiating out like fingers are the water transport system of the host. Primary plasmodia (arrowheads) are scattered in cells of the root cortex. Notice the primary plasmodium (~50 μ m in diameter) in an enlarged host cell beginning to segment (arrow) into masses that will become thin-walled zoosporangia. (b) More detailed cross section of infected root. The secondary plasmodium (P) [~50 μ m in diameter] is beginning to segment into uninucleate masses (arrow) that will become resting spores. Characteristic of *Plasmodiophora*, resting spores (RS) are scattered in the host cell.

that extend the length of synapsing chromosomes and aid in exchange pairing) in resting-spore plasmodia, evidence that reduction division (meiosis) occurs in the secondary plasmodium. The lesions of powdery scab of potatoes and the deformed tissue of clubroot of crucifers are often invaded secondarily by bacteria, rotting host tissue and releasing plasmodiophorid resting spores. *See* PLANT PATHOL-OGY.

Taxonomy. There are approximately 12 currently recognized genera, but exploration of new habitats is leading to discovery of new taxa. Genera are differentiated based on the organization of their resting spores, especially as spore balls. For example, *Plasmodiophora* resting spores lie free in host cells (Fig. 1*b*). *Spongospora* forms spongelike spore balls,

traversed by pockets and channels (**Fig. 2**), and is distinguished from *Sorosphaera*, which forms hollow spore balls. Typically infections result in hypertrophy of host tissue, and hypertrophy in clubroot of crucifers is believed to be caused by the rerouting of the plant hormone precursor, glucobrassicin, into auxin (any organic compound promoting plant growth). However, some infections, as with *Ligniera*, there is little hypertrophy.

Several characters distinguish plasmodiophorids: (1) cruciform nuclear division, where the persistent nucleolus elongates toward the mitotic poles as chromosomes align along the equatorial plate, giving a crosslike configuration; (2) centrioles positioned end to end during interphase; (3) multiple globules in the core of flagellar basal bodies; and (4) relatively simple microtubular root systems in zoospores. *See* PLASMODIOPHORIDA.

Importance. Plasmodiophorids are economically important as pathogens of crop plants. For example, Spongospora subterranea causes powdery scab of potato, as well as crookneck disease of watercress. On potatoes, thick corky lesions become filled with thousands of spore balls containing resting spores. This disease has become more prevalent as a greater range of nonresistant potato varieties is produced. The economic loss to this disease has also increased with consumers' demands for high-quality potatoes. Plasmodiophora brassicae causes clubroot of crucifers and distorts root tissue, resulting in galls (abnormal growths) on twisted roots that cannot transport nutrients and water effectively to the aboveground portion of plants. Diseased crucifers, consequently, become stunted, yellowed, and wilted. See AGRICULTURAL SCIENCE (PLANT); AGRI-CULTURAL SOIL AND CROP PRACTICES.

In addition to being the causal agents of diseases, plasmodiophorids can serve as vectors (agents capable of transferring pathogens from one organism to another) for plant viruses. *Spongospora subterranea* zoospores transmit potato mop top virus. *Polymyxa graminis* is the vector for wheat spindle streak mosaic virus and barley yellow mosaic virus. *Polymyxa betae* is the vector for beet necrotic yellow vein virus. *See* PLANT VIRUSES AND VIROIDS.



Fig. 2. Powdery scab of potato. Longitudinal section through the outer portion of a potato infected with *Spongospora subterranea*. Notice the spongelike spore balls (sori) traversed by irregular channels (arrow). Each resting spore is $\sim 4 \ \mu m$ in diameter. The spore balls are 45-80 μm in diameter.

The future. Plasmodiophorid diseases are difficult to control because of the longevity of their resting spores in the soil. Most chemicals used to control fungal diseases are not effective against plasmodiophorid, protozoan diseases of plants. Use of resistant plant varieties, long-cycle crop rotation, and sanitation are the three most effective controls of plasmodiophorid diseases.

The impact of plasmodiophorid parasites of algae and oomycetes on environmental balance is poorly studied. Modern molecular environmental studies are revealing new plasmodiophorids and their near relatives in newly explored habitats, such as oceanic thermal vents. Clearly, these modern methods will expand our understanding of the diversity of plasmodiophorids and of their role in the environment. Martha J. Powell

Bibliography. C. J. Alexopoulos, C. W. Mims, and
M. Blackwell, *Introductory Mycology*, 4th ed., 1996;
D. J. S. Barr, The phylogenetic and taxonomic implications of flagellar rootlet morphology among zoosporic fungi, *BioSystems*, 14:359–370, 1981; J. P. Braselton, *Plasmodiophoromycota, The Mycota VII, Part A: Systematics and Evolution*, pp. 81–91, 2001;
D. P. Dylewski, Phylum Plasmodiophoromycetes, *Handbook of Protoctista*, pp. 399–416, 1990;
J. S. Karling, *The Plasmodiophorales*, 2d ed., 1968.

Plasmon

The quanta of waves produced by collective effects of large numbers of electrons in matter when the electrons are disturbed from equilibrium. Plasmon excitations are easily detected in metals because they have a high density of electrons that are free to move. The result of plasmon stimulation by energetic electrons is seen in **Fig. 1**. The graph shows the probability of energy losses by fast electrons transmitted through a thin aluminum foil. The number of detected electrons in a beam is plotted against the energy loss during transit through the foil. Each energy-loss peak corresponds to excitation of one or more plasmons. Within experimental error, the peaks occur at integral multiples of a fundamental loss quantum.

The name plasmon is derived from conventional plasma theory. A plasma is a state of matter in which the atoms are ionized and the electrons are free to move. At low particle densities this means an ionized gas, or classical plasma. In a metal the conduction electrons are also free to move, and there are many analogies to the classical low-density plasma. Similar to sound waves in a low-density gas of charged particles, the electron gas in a metal exhibits plasmon phenomena, that is, electron density waves. In both types of physical plasma, the frequency of plasma-wave oscillation is determined by the electronic density. Because of the high electron density in a metal (about 10²³ electrons/cm³), the frequencies are typically in the ultraviolet part of the electromagnetic spectrum. Quantum mechanics dictates that

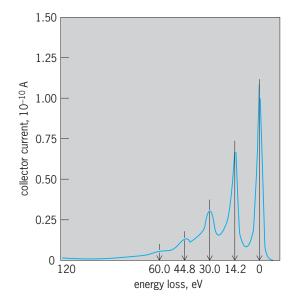


Fig. 1. Number of detected electrons in a beam versus their energy loss during transit through a thin aluminum foil. (Number of electrons is expressed as a current; 10^{-10} A = 6.7 × 10^8 electrons per second.) Peaks at approximate multiples of 14.2 eV correspond to energy donated to plasmons in the aluminum. (After T. L. Ferrell, T. A. Callcott, and R. J. Warmack, Plasmons and surfaces, Amer. Sci., 73:344–353, 1985)

the energy of a plasmon is quantized, and one quantum of energy of a plasmon is given by its frequency multiplied by Planck's constant. *See* FREE-ELECTRON THEORY OF METALS; PLASMA (PHYSICS); QUANTUM ME-CHANICS.

Bulk and surface plasmons. Metallic objects support two kinds of plasmons: bulk and surface plasmons. Bulk plasmons are longitudinal density waves (like sound waves) that propagate through the volume of a metal. Electron density waves can also propagate at the interface between a metal substrate and a dielectric (surface plasmons). This fact can be easily derived from the electromagnetic theory developed by James Clerk Maxwell. Such electronic surface waves are strongly coupled to the electromagnetic field (**Fig. 2**). They are officially termed

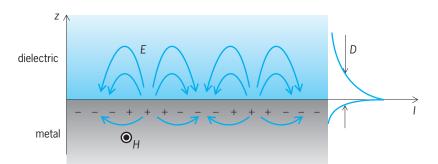


Fig. 2. Surface plasmon-polariton propagating at a metal-dielectric interface. Both the magnetic (*H*) and electric (*E*) field components of this wave are shown. The coupling of the electromagnetic fields to the electrons in the metal is also visualized by the charge density waves near the metal surface that creates a region of slightly negative charge (where electrons bunch up) and positive charge (where electrons have moved away). The electromagnetic field intensity, *I*, of a surface plasmon is highest at the surface and decays exponentially away from the interface. The diameter of the beam (optical mode), *D*, can be significantly smaller than the wavelength of light.

surface plasmon-polaritons, but the term surface plasmon is more commonly used by many scientists. The charge fluctuations are extremely localized in the *z* direction (perpendicular to the interface) within the Thomas-Fermi screening length of about 0.1 nm. The corresponding electromagnetic wave is transverse magnetic (TM) in nature and has a beam (mode) size, *D*, that is not limited by diffraction in the direction normal to the metal surface. The field intensity, *I*, of such a wave has a maximum at the surface and decays exponentially away from the interface. *See* ELECTROMAGNETIC WAVE TRANSMISSION; MAXWELL'S EQUATIONS.

History of plasmon research. Well before scientists set out to study the unique optical properties of metal nanostructures, these optical properties were employed by artists to generate vibrant colors in glass artifacts and to stain church windows. A famous example is the Lycurgus cup dating back to the Byzantine Empire (A.D. 4th century). Scientific studies in which surface plasmons were observed date back to the beginning of the twentieth century. In 1902 Robert W. Wood observed unexplained features in optical reflection measurements on metallic gratings. In 1904 Maxwell Garnett described the bright colors observed in metal-doped glasses using the newly developed Drude theory of metals, and the electromagnetic properties of small spheres as derived by Lord Rayleigh. In an effort to develop further understanding, in 1908 Gustav Mie developed his now widely used theory of light scattering by spherical particles. See SCATTERING OF ELECTROMAG-NETIC RADIATION.

In 1956 David Pines theoretically described the characteristic energy losses experienced by fast electrons traveling through metals, and attributed these losses to collective oscillations of free electrons in the metal. In analogy to earlier work on plasma oscillations in gas discharges, he called these oscillations plasmons. In 1957 Rufus Ritchie studied electron energy losses in thin films and showed that plasmon modes can exist near the surfaces of metals. This study represents the first theoretical description of surface plasmons. In 1968 Ritchie and coworkers described the anomalous behavior of metal gratings in terms of surface plasmon resonances excited on the gratings. In the same year Andreas Otto as well as Erich Kretschmann and Heinz Raether presented methods for the optical excitation of surface plasmons on metal films, making experiments on surface plasmons easily accessible to many researchers.

In the late 1970s and 1980s, surface plasmons at optical frequencies attracted much attention because the high field intensity near the interface could be exploited in nonlinear optical experiments such as surface-enhanced Raman scattering, secondharmonic generation, and surface-enhanced fluorescence. This has been an active area of research for many years, and the physics of surface plasmonpolaritons on two-dimensional (2D) films is now fairly well understood. *See* NONLINEAR OPTICS.

Since the late 1990s, notable developments have

taken place in the field of plasmon optics. These developments have come about primarily because of the ability to fabricate nanoscale metallic structures in a controlled fashion. Thomas Ebbesen showed that optical transmission through a subwavelength aperture in a metal film is strongly enhanced when the incident light resonantly couples to surface plasmons. John Pendry speculated that a thin silver film might act as a "perfect" lens capable of producing images with a spatial resolution beyond the diffraction limit. Furthermore, theory and experiments demonstrated that metal nanowires, ordered arrays of metal nanoparticles, and grooves in metal films can serve as light guides that can manipulate the flow of light at the nanoscale. These findings are expected to have a major impact on the semiconductor industry and on the fields of biology, medicine, energy, and the environment. See NANOSTRUCTURE; NANOTECHNOL-OGY

Optical excitation of surface plasmons. From basic electromagnetic theory, it follows that surface plasmon waves have a shorter wavelength than light waves of the same frequency. In the quantum picture, this means that surface plasmons possess a larger momentum than photons of the same energy. For this reason, free-space light beams in air cannot simply couple to surface plasmon waves and conserve both energy and momentum. Kretschmann realized that photons in a high-refractive-index material, such as glass, possess a larger momentum and can be used to excite surface plasmons at a metal-air interface. He showed that surface plasmons can be excited by directing light through a glass prism at a thin metal film deposited on the prism surface. He found that there was one specific, resonant angle for which significant energy from the light beam could be transferred to surface plasmons. At this angle the intensity of the reflected beam was significantly reduced, consistent with the fundamental law of energy conservation. From the early studies by Wood, it was also found that surface plasmons can be excited by means of a metal grating structure. When gratings are used, the momentum mismatch between photons and surface plasmons is taken up by the grating.

For the excitation of surface plasmons in nanoscale particles, a grating or a prism is not needed. Incident light typically induces an oscillating charge density on the surface of a particle or microstructure, with opposite charges at opposing sides. For particles small compared to the wavelength of the light, this is a purely dipolar charge distribution. The dipolar surface plasmon frequency for a sphere in vacuum is given by the bulk plasmon frequency divided by $\sqrt{3}$ (assuming a simple model for the electronic properties of the metal). The resonance frequencies for most metals are in the ultraviolet, visible, or nearinfrared part of the electromagnetic spectrum. Metal nanoparticles suspended in a liquid or embedded in a glass film strongly absorb light at their surface plasmon frequency and therefore exhibit vibrant colors. The surface plasmon resonance frequency can be easily tuned by changing the shape of the particle

or the refractive index of the medium it is embedded in. Peter Nordlander and Naomi Halas provided a very intuitive model that can be used to predict the optical response of complex metallic nanostructures and systems. Called the plasmon hybridization model, it shows intriguing similarities to the wellestablished theories on molecular orbital formation from atomic orbitals. This model can predict the location of plasmon resonances in complex systems based on knowledge of the resonant behavior of elementary building blocks. It thus provides a powerful tool for optical engineers in the design of functional plasmonic nanostructures. *See* COLLOID; COLLOIDAL CRYSTALS; DIPOLE.

Applications of surface plasmons. One of the most well-established applications of surface plasmons is biochemical sensors that can detect the presence of specific biochemical compounds. Selectivity can be obtained by the use of compounds or molecular impressions that bind only the specific target compound. By coating a thin metal film on a prism with antibodies, and tracking the shift in the surface plasmon excitation angle as antigens from a solution bind to the antibodies, the concentration of antigens can be determined. Similarly, since the surface plasmon excitation angle is sensitive to the index of refraction of the overcoating, it is possible to distinguish monomers from dimers just by exposing the thin film to the solution containing both, the monomer form having a different index from that of the dimer. The indices must be known from other measurements, and the thin film must be protected from deleterious effects of the solution. See ANTIBODY; ANTIGEN; BIOELECTRONICS; TRANSDUCER.

An important contribution of surface plasmons to analytical chemistry and biology is in Raman scattering. A portion of visible light incident on a molecule is frequency-shifted in accordance with the vibrational and rotational spectrum of the molecule. This constitutes the Raman signal that is characteristic of the scattering molecule. When a molecule is sitting on the surface of a metallic nanostructure, the fields can be substantially enhanced due to the resonant excitation of surface plasmons. This in turn leads to a surface-enhanced Raman spectroscopy (SERS) signal. This signal can be many orders of magnitude stronger than the Raman signal obtained without the presence of the metal. This effect has enabled the successful implementation of Raman spectroscopy on trace amounts of chemical compounds and even single molecules. See RAMAN EFFECT.

Promising developments in biology and medicine followed the demonstration by Halas and Jennifer West in 2003 that the resonant properties of gold nanoshells could enable fast whole-blood immunoassays. It has been speculated that nanoshells could play a role in cancer treatments.

There is significant interest in the use of surface plasmon waves to transport information on a chip with nanoscale plasmon waveguides. Currently, two of the most daunting problems preventing significant increases in processor speed are thermal and resistance-capacitance (RC) delay time issues associated with electronic interconnection. Optical interconnects, on the other hand, possess an almost unimaginably large data-carrying capacity and may offer interesting solutions for circumventing these problems. Optical alternatives may be particularly attractive for chips with more distributed architectures in which a multitude of fast electronic computing units (cores) need to be connected by high-speed links. Unfortunately, dielectric photonic devices are limited in size by the fundamental laws of diffraction to about half a wavelength of light, and tend to be at least 1 or 2 orders of magnitude larger than their nanoscale electronic counterparts. This obvious size mismatch between electronic and photonic components presents a major challenge for interfacing these technologies. It thus appears that further progress will require the development of a radically new chip-scale device technology that can facilitate information transport between nanoscale devices at optical frequencies and bridge the gap between the world of nanoscale electronics and microscale photonics. Metallic or "plasmonic" interconnects may provide new solutions. The device technology based on plasmonic interconnects and components has been termed plasmonics. An increased synergy could be attained by integrating plasmonic, electronic, and conventional photonic devices on the same chip and taking advantage of the strengths of each technology. See INTEGRATED OPTICS.

One of the most promising developments in surface plasmon science is the superlens proposed by Pendry. A superlens is an optical element capable of imaging objects of subwavelength dimensions. The simplest version of a superlens consists of a thin slab of metal. Superlenses exhibit enormous potential for commercial applications, including ultrahighresolution imaging, high-density memory storage devices, and nanolithography.

Surface plasmons also play an important role in the extraordinary optical transmission (EOT) phenomenon observed in thin metallic films perforated with an array of subwavelength holes. Ebbesen demonstrated that significantly more light is transmitted through the tiny holes in the metal film than was previously predicted by classical diffraction theory. The reason is that the hole arrays scatter light into surface plasmon waves. First, these surface plasmons waves on the illuminated side of the metal couple to surface plasmon waves on the other side of the metal through the holes. In a second stage, these surface plasmons get scattered again to light waves propagating away from the metal. The exact transmission spectrum can be tuned by adjusting the period and the symmetry of the hole array. This phenomenon has attracted much attention to the field of plasmonics.

The energy losses imparted by materials on a charged particle beam can be analyzed in much the same way as in the original observations of plasmons. This analysis can be used to form high-resolution chemical maps in transmission electron microscopy. This technique has a number of applications in materials analysis of inorganic, organic, and biomaterials. *See* ELECTRON MICROSCOPE.

Mark L. Brongersma; Pieter G. Kik Bibliography. W. L. Barnes, A. Dereux, and T. Ebbesen, Surface plasmon subwavelength optics, Nature, 424:824-830, 2003; T. W. Ebbesen et al., Extraordinary optical transmission through subwavelength hole arrays, Nature, 391:667-669, 1998; J. B. Pendry, Negative refraction makes a perfect lens. Phys. Rev. Lett., 85:3966-3969, 2000; E. Prodan et al., A hybridization model for the plasmon response of complex nanostructures, Science, 302:419-422, 2003; H. Raether, Surface Plasmons, 1988; J. Takahara et al., Guiding of a one-dimensional optical beam with nanometer diameter, Opt. Lett., 22:475-477, 1997; R. W. Wood, On a remarkable case of uneven distribution of light in a diffraction grating spectrum, Phil. Mag., 4:396-402, 1902; R. Zia, J. A. Schuller, and M. L. Brongersma, Plasmonics: The next chip-scale technology, Mater. Today, 9:20-27, 2006.

Plaster

A plastic mixture of solids and water that sets to a hard, coherent solid and that is used to line the interiors of buildings. A similar material of different composition, used to line the exteriors of buildings, is known is stucco. The term plaster is also used in the industry to designate plaster of paris.

Plaster is usually applied in one or more base (rough or scratch) coats up to ${}^{3}\!/_{4}$ in. (2 cm) thick, and also in a smooth, white, finish coat about ${}^{1}\!/_{16}$ in. (0.2 cm) thick. The solids in the base coats are hydrated (or slaked) lime, sand, fiber or hair (for bonding), and portland cement (the last may be omitted in some plasters). The finish coat consists of hydrated lime and gypsum plaster (in addition to the water). *See* LIME (INDUSTRY); MORTAR; PLASTER OF PARIS.

J. F. McMahon

Plaster of paris

The hemihydrate of calcium sulfate, composition $CaSO_4 \cdot 1/_2 H_2O$, made by calcining the mineral gypsum, composition $CaSO_4 \cdot 2H_2O$, at temperatures up to $482^{\circ}F$ (250°C). It is used for making plasters, molds, and models. For a description of the casting method of ceramic forming *see* CERAMICS.

When the powdered hemihydrate is mixed with water to form a paste or slurry, the calcining reaction is reversed and a solid mass of interlocking gypsum crystals with moderate strength is formed. Upon setting there is very little (a slight contraction) dimensional change, making the material suitable for accurate molds and models.

The chemical reaction of hydration requires 18.6 lb of water for each 100 lb of plaster of paris; any water in excess of this makes the mixture more fluid and it eventually evaporates, leaving a porous structure. In general, the greater the porosity, the lower

the strength of the set plaster. Plaster of paris molds for ceramic casting must have a certain minimum porosity in order to absorb water from the slip; thus, for this application, the amount of water added represents a compromise between the conflicting properties of high strength and high porosity.

Diverse types of plaster, varying in the time taken to set, the amount of water needed to make a pourable slip, and the final hardness, are made for different uses. These characteristics are controlled by the calcination conditions (temperature and pressure) and by additions to the plaster. For example, hydrated calcium sulfate ($CaSO_4$ · 2H₂O) greatly accelerates setting time and therefore the use of utensils contaminated with set plaster is to be avoided. *See* GYPSUM; PLASTER. J. F. McMahon

Plastic deformation of metals

The permanent change in shape of a metal object as a result of applied or internal forces. This feature permits metals to be formed into pipe, wire, sheet, and so on, with an optimal combination of strength, ductility, and toughness. The onset of noticeable plastic deformation occurs when the applied tensile stress or applied effective stress reaches the tensile yield strength of the material. Continued plastic deformation usually requires a continued increase in the applied stress. The effects of plastic deformation may be removed by heat treatment. *See* MECHANICAL PROPERTIES OF METAL; PLAS-TICITY.

Mechanisms of plastic deformation in metals are understood in terms of crystal structure and incremental slip between atomic planes. These aspects are useful to understand: (1) regimes of stress and temperature for plastic deformation; (2) types of plastic stress-strain response; (3) reasons why metals exhibit plastic deformation; (4) mechanisms of plastic deformation in metal crystals and polycrystals; (5) strategies to strengthen metals against plastic deformation; (6) effect of strengthening, temperature, and deformation rate on ductility and toughness.

Deformation regimes and plastic stress-strain response. Plastic deformation in metals is observed macroscopically as a departure from linear-elastic stress-strain response. **Figure 1** shows that plastic deformation predominates over other deformation mechanisms when the applied stress is greater than the yield strength. Unlike creep deformation, plastic deformation usually does not increase with time unless the applied stress is also increased. When the applied stress is less than the yield strength, other deformation mechanisms may dominate. *See* CREEP; ELASTICITY; MECHANICAL PROPER-TIES OF METAL; STRESS AND STRAIN.

Tensile-engineering stress-strain (σ - ε) curves exhibit several features of plastic deformation (**Fig. 2**). The response depends on single versus polycrystalline structure, alloy content, temperature, strain rate, loading geometry, as well as sample size, shape,

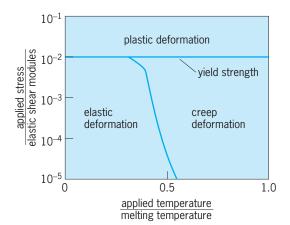


Fig. 1. Deformation mechanism map. (After M. F. Ashby and D. R. H. Jones, Engineering Materials I: An Introduction to Their Properties and Applications, 2d ed., Butterworth-Heimann. 1996)

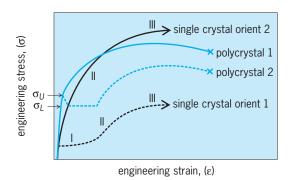


Fig. 2. Stress-strain curves for metals. (P. M. Anderson, The Ohio State University, 2005)

and orientation. For polycrystalline metals, departure from linear-elastic response can be gradual (polycrystal 1) or quite abrupt (polycrystal 2). Similarly, single-crystal samples can have an abrupt departure (single crystal orient 1) or a gradual one (single crystal orient 2). In the latter case, three distinct regimes of plastic deformation are noted. *See* SINGLE CRYS-TAL.

Reasons for plastic deformation. Metals are unique in that they tend to form crystals of densely packed atoms. This occurs in metals since the valence or outer electrons have a continuous succession of possible energy levels. These electrons create a "sea" of negative charge that pulls positively charged nuclei into densely packed crystal structures such as facecentered cubic (fcc), hexagonal close-packed (hcp), and body-centered cubic (bcc). Metallic glasses are an exception, since they are noncrystalline or amorphous. *See* CRYSTAL STRUCTURE; METALLIC GLASSES; METALS.

Under stress, metal crystals are prone to shear between densely packed planes. **Figure** *3a* shows two densely packed layers typical of metal crystals with fcc or hcp crystal structures. The upper (lighter) plane is positioned into depressions created by the lower (darker) plane. Plastic deformation occurs by shearing one plane relative to another, as depicted in Fig. 3*b*. Complete shearing by an amount **b** restores dense packing, since the upper plane moves into an equivalent set of depressions. During the process, atomic bonds are broken and made. Densely packed slip planes are preferred in part because the depressions are shallow and allow for "smoother" slip. In many metal crystals, the unit slip **b** is smoothened by a two-step process of partial slip **p1**, followed by partial slip **p2** (Fig. 3*b*). *See* SHEAR.

Experimental evidence confirms that slip produces plastic deformation in metals. **Figure 4** shows numerous slip traces along the surface of a copper-aluminum crystal deformed in tension. Since the length of unit slip **b** is only 2–3 angstroms (1 angstrom = 10^{-10} m), planes must shear repeatedly to produce macroscopically visible slip traces. Formation of slip traces during plastic deformation turns shiny metal surfaces to a dull state. In contrast, ceramic materials are brittle and lack plastic deformation at room temperature. This is due to the lack of closely packed planes and to the prohibitively

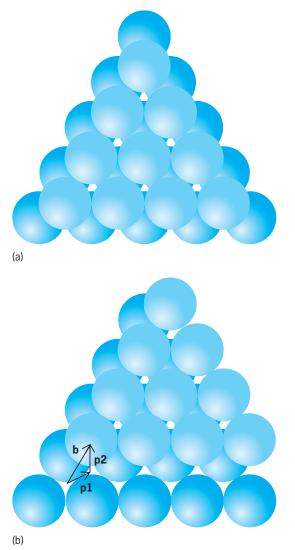


Fig. 3. Pair of closely packed planes in (a) initial and (b) sheared positions. (P. M. Anderson, The Ohio State University, 2005)

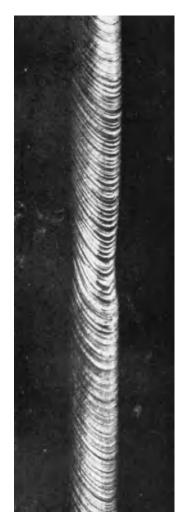


Fig. 4. Slip traces along the surface of a copper-aluminum crystal deformed in tension. (From C. F. Elam, The Distortion of Metal Crystals, Oxford University Press, London, 1935)

large energy needed to break bonds during shearing. *See* BRITTLENESS; CERAMICS.

Incremental slip and theoretical strength. Experimental evidence also demonstrates that atomic planes slip incrementally. The line ABCD in Fig. 5 is known as a dislocation and identifies the boundary between slipped and unslipped portions. Dislocations are lines along which crystal packing is disrupted and atomic bonds are distorted. Figure 6a shows several dislocations as viewed by a transmission electron microscope. Dislocations move under applied load, indicating the advance of slip between two planes. For example, movement of dislocation ABCD to the left (Fig. 5) corresponds to motion of additional atoms from unslipped to slipped positions. In reality, dislocation lines are not as atomically sharp as depicted in Fig. 5. Rather, atoms near dislocations displace to close gaps (for example, along lines BC and CD) and smoothen the transition. These displacements become negligible with increasing distance from the dislocation (Fig. 3). See CRYSTAL DE-FECTS.

The existence of dislocations is also supported

by the small applied tensile stress needed for plastic deformation. The yield strength of many metals varies from 10^{-3} to 10^{-5} times the elastic modulus and agrees well with theoretical estimates of stress to move dislocations along slip planes. In contrast, theoretical estimates of stress to simultaneously shear atomic planes are orders of magnitude larger than measured yield strengths. A convenient analogy is that sliding an entire rubber-backed carpet across a floor can be quite difficult. However, the carpet can be slid easily (and incrementally) by introducing a wrinkle (dislocation) along one edge and pushing the wrinkle to the opposite edge.

Some nanometer-scale crystals have nearly perfect atomic stacking and thus no dislocations. The premise is that dislocation-free crystals will be unable to slip incrementally and therefore approach theoretical strength. Indeed, single-crystal whiskers a few micrometers in diameter exhibit near-theoretical strength, and metal cylinders with diameters <1 μ m are observed to have extraordinary compressive yield strength. *See* CRYSTAL WHISKERS.

Single crystals and polycrystals. Several features of stress-strain curves (Fig. 2) can be understood qualitatively based on dislocation concepts. In single crystals, plastic stress-strain response is anisotropic, or dependent on crystal orientation. If crystal orientation favors slip on just one crystallographic plane and direction, three stages of deformation are often observed (single crystal orient 1). During stage I, a single slip plane and direction, called the primary slip system, operate. Primary slip dislocations move until arrested by other primary slip dislocations or until portions of dislocations (for example, AB in Fig. 5) move onto different crystallographic planes. The latter process, called cross slip, prevents movement of dislocations. Arrest of dislocations increases deformation stress and thus work hardens the metal. The slope $d\sigma/d\varepsilon$, or work-hardening rate, is small in stage I, and a moderately uniform distribution of interacting primary dislocations is produced. A second or conjugate slip system is activated when the

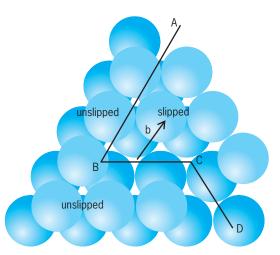


Fig. 5. Incremental slip between closely packed planes. Dislocation line ABCD separates slipped and unslipped regions. (P. M. Anderson, Ohio State University, 2005)

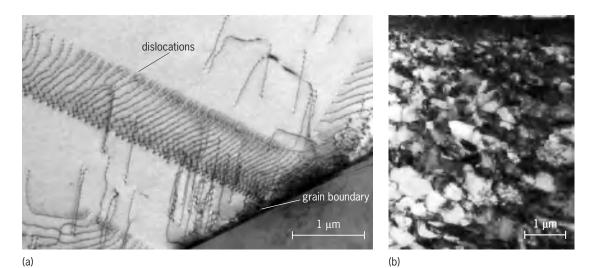


Fig. 6. (a) Arrays of parallel dislocation lines positioned near a grain boundary (lower right) in α -Ti(AI) (courtesy M. Mills and N. Thirumalai, The Ohio State University, 2005). (b) Dislocation cell structure in oxygen-free high conductivity copper deformed by sliding the top surface against M2 tool steel (R. Divakar and D. A. Rigney, The Ohio State University, 2005).

primary slip system becomes sufficiently hardened. Primary and conjugate dislocations interact strongly during stage II so that the work-hardening rate is often large. Cells of entangled dislocations form (Fig. 6*b*), the size of the cells decreasing with continued deformation. During stage III, a minimum dislocation cell size is reached and remains constant with deformation, so that $d\sigma/d\varepsilon$ approaches 0. Single crystals may be oriented to activate two or more primary slip systems initially (single crystal orient 2). In this case, stage I deformation ($\varepsilon > 1$ or 2), additional stages are observed during which $d\sigma/d\varepsilon$ increases slightly or remains constant (stage IV) and then decreases and reaches 0 (stage V).

Maximum plastic strength is inversely proportional to the minimum dislocation cell size achieved during stage III. Minimum size is controlled, in part, by escape of dislocations from cell walls via cross slip. Cross slip is easier for metals with large stackingfault energy, or energy associated with partial slip **p1** of a unit area of atoms (Fig. 3*b*). Consequently, metals with large stacking-fault energy, such as aluminum, exhibit easier cross slip and cannot achieve small stage III cell size or large maximum plastic strength.

Polycrystals tend to have larger initial yield strength (Fig. 2) since they contain internal boundaries between crystals that act as barriers to dislocation motion. Yield is often gradual due to a range of crystal (grain) orientations (polycrystal 1). The first grains to yield have larger values of stress to shear crystallographic planes. In some cases, abrupt yield occurs via breakaway of dislocations from impurity elements such as carbon and nitrogen in steel (polycrystal 2). The upper yield point (σ_U , Fig. 2) is the breakaway stress. Following breakaway, dislocations move at a reduced stress or lower yield point (σ_L , Fig. 2). This phenomenon can be repeated if the sample is unloaded and impurities are allowed to diffuse back to dislocations. *See* GRAIN BOUNDARIES; MECHANICAL PROPERTIES OF METAL.

Stress-strain curves exhibit work hardening due to increasing dislocation density ρ_d during plastic deformation. Density ranges from 10³ to 10⁵ cm of dislocation line length per cubic centimeter of material for carefully produced crystals to greater than 10^{10} cm/cm³ after large plastic deformation. Dislocations generate new dislocations during plastic deformation and interact with one another via the internal stress they create. The increase in deformation stress is proportional to $Gb(\rho_d)^{1/2}$, where *G* is elastic shear modulus and *b* is the magnitude of unit slip vector **b** (Fig. 3*b*). Work hardening of pure metals can increase the deformation stress to more than ten times the initial yield strength.

Tensile engineering stress-strain curves exhibit a negative slope or softening near the end of a test due to a geometric instability called necking (Fig. 2). During testing, elongation in the tensile direction reduces the cross-sectional area, due to the volume-conserving nature of crystallographic slip. A neck, or region of smaller cross-sectional area, arises when softening due to reduction in cross-sectional area cannot be offset by work hardening. The neck becomes more pronounced until plastic failure occurs, usually by linking internal pores along a flat region of the neck center, coupled with angled shearing around the neck perimeter. Engineering stress peaks when true strain equals the work-hardening exponent n. However, true stress properly accounts for area reduction and continues to increase due to work hardening. See MECHANICAL PROPERTIES OF METAL.

Strategies to strengthen metals against plastic deformation. Four elementary strategies are used to make dislocation motion difficult and thereby increase the stress needed for plastic deformation. Grain-size strengthening involves the reduction of grain (crystal) size in polycrystals, so that grain boundaries provide more numerous trapping sites for dislocations. The contribution to deformation stress scales as $(\text{grain size})^{-1/2}$. Solution strengthening involves adding impurity atoms to pure metals (alloying). Impurity atoms interact with dislocations via the local distortion or strain they induce. The contribution to deformation stress scales as (impurity concentration)^{1/2}. Precipitate strengthening is also the result of alloying to the extent that a new phase is produced inside the host phase. Precipitates induce strain and provide interfaces that can trap dislocations. The contribution to deformation stress scales as (precipitate spacing)⁻¹. Cold work increases dislocation density via deformation at lower temperature and larger deformation rate. The contribution to deformation stress scales as (dislocation density) $^{1/2}$ or as (plastic strain)ⁿ, where n is the work hardening exponent. See ALLOY.

Two additional strategies to strengthen metals against plastic deformation are to increase strain rate and decrease temperature. These are effective since dislocation motion involves kinetic processes related to bond breaking and forming during shear. In addition, motion can be assisted or suppressed by the kinetics of impurity or vacancy diffusion to or from dislocations. Deformation stress scales as (strain rate)^{*m*}, where *m* is the strain-rate sensitivity.

Effect of strengthening, temperature, and deformation rate. Cold work and other mechanisms to increase deformation stress often decrease ductility and toughness (Fig. 7). Fracture toughness is also reduced, since restricting dislocation motion often reduces the ability to blunt-out sharp cracks and dissipate energy. The effects of cold work are reversed by annealing, which can reduce dislocation density via dislocation annihilation or formation of new crystals, respectively. *See* HEAT TREATMENT (METAL-LURGY).

Dislocation motion and therefore ductility and toughness are affected by crystal structure. Metals with fcc structure, such as copper, have several orientations of closely packed planes and directions, so that they are easily deformed and require a large energy for failure under impact, regardless of test temperature (**Fig. 8**). In contrast, high-strength metals that use one or more strengthening strategies exhibit small impact energy. Certain metals with hcp structure such as zinc and bcc structure such as iron, chromium, tungsten, and molybdenum display a transition from brittle behavior at lower temperature to ductile behavior at higher temperature. The ductile-to-brittle transition temperature is positioned between the brittle and ductile regimes.

The unique characteristics of hcp and bcc metals stem in part from crystal structure. Hcp metals have only one orientation of closely packed planes and therefore rely on non-densely packed planes for general deformation. Bcc metals have no closely packed planes. Thus, the critical stress for dislocation motion in bcc metals (and some ceramics such as silicon) is typically larger than for fcc materials and is dependent on temperature. An outcome is that ductility, toughness, and deformation stress are also dependent on temperature. However, the transition temperature for each metal is not fixed, but a func-

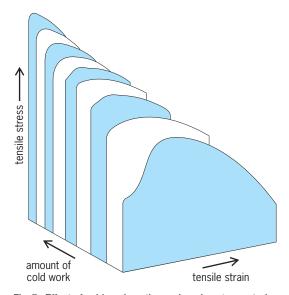


Fig. 7. Effect of cold work on the engineering stress-strain behavior of a low-carbon steel. [ASM Handbook, vol. 1: Properties and Selection: Irons, Steels, and High Performance Alloys (Carbon and low-alloy steels: Cold-finished steel bars: Mechanical Properties), ASM International, 1990]

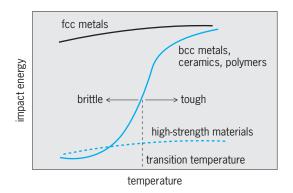


Fig. 8. Variation of impact energy with temperature for different materials. (*After C. R. Barrett, W. D. Nix, and A. S. Tetelman, The Principles of Engineering Materials, Prentice Hall,* 1973)

tion of alloy cleanliness, grain size, prior cold work, rate of deformation, and even sample geometry. Thus, increasing the rate of deformation or including notches in the specimen geometry will shift the curve to the right, thereby creating a larger temperature regime over which the metal is brittle.

Peter M. Anderson

Bibliography. M. F. Ashby and D. R. H. Jones, *Engineering Materials 1: An Introduction to Their Properties and Applications*, 2d ed., Butterworth-Heinemann, Oxford, 1996; ASM International, *Atlas of Stress-Strain Curves*, 2d ed., ASM International, 2002; W. D. Callister, *Materials Science and Engineering: An Introduction*, 6th ed., Wiley, 2003; D. Hull and D. J. Bacon, *Introduction to Dislocations*, 4th ed., Butterworth-Heinemann, Oxford, 2001; J. Martin, *Materials for Engineering*, 2d ed., Maney, Leeds, 2002.

Plasticity

The ability of a solid body to permanently change shape (deform) in response to mechanical loads or forces. Deformation characteristics are dependent on the material from which a body is made, as well as the magnitude and type of the imposed forces. In addition to plastic, other types of deformation are possible for solid materials.

One common test for measuring the plastic deformation characteristics of materials is the tensile test, in which a tensile (stretching) load is applied along the axis of a cylindrical specimen, with deformation corresponding to specimen elongation. The load magnitude and degree of deformation are normally converted into meaningful parameters, which include the size and geometry of the body. The load is converted into stress; its units are megapascals (1 MPa = 10^6 newtons per square meter) or pounds per square inch (psi). Likewise, the amount of deformation is converted into strain, which is unitless. The test results are expressed as a plot of stress versus strain. *See* STRESS AND STRAIN.

Typical tensile stress-strain curves are shown for metal alloys and polymeric materials in **Figs. 1** and **2**, respectively. The inset plots show behavior at small strain values. For both materials, the initial regions of the curves are linear and relatively steep. Deformation that occurs within these regions is nonpermanent (nonplastic) or elastic. This means that the body springs back to its original dimensions once the stress is released, or that all of the deformation is recovered. In addition, stress is proportional to strain (Hooke's law), and the slope of this linear segment corresponds to the elastic (Young's) modulus. *See* ELASTICITY; HOOKE'S LAW; YOUNG'S MODULUS.

Plastic (permanent) deformation begins at the point where linearity ceases such that, upon removal of the load, not all deformation is recovered (the body does not assume its original or stress-free dimensions). The onset of plastic deformation is called yielding, and the corresponding stress value is called the yield strength. After yielding, all deformation is plastic and, until fracture, the curves are nonlinear. This behavior is characteristic of many metal alloys and polymeric materials, as shown in Figs. 1 and 2. *See* PLASTIC DEFORMATION OF METAL.

Qualitatively, the plasticity of a material depends on both the yield strength and fracture strain. That is, plasticity increases with a decrease in yield strength (as plastic deformation becomes easier) and an increase in fracture strain (as the degree of plastic deformation increases). Typical metal alloy and polymer yield strengths are indicated on the stress axis of Fig. 1, whereas the strain axis of Fig. 2 compares their fracture strains. Accordingly, the polymer has a higher plasticity since it has a lower yield strength and a greater failure strain. Many polymers are called plastics because of their relatively high degrees of plasticity.

Another aspect of plasticity deals with deformation mechanisms, the atomic or molecular processes that accompany deformation. Plastic deformation involves the net movement of large numbers of atoms or molecules.

The concept of plasticity does not normally relate to ceramic materials such as glasses and metal oxides (for example, aluminum oxide). At room

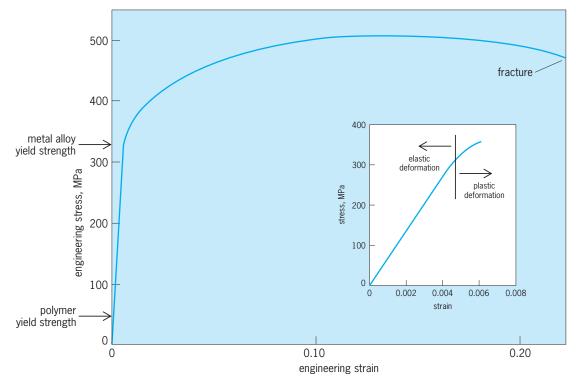


Fig. 1. Plot of stress versus strain to fracture for a typical metal alloy that is deformed using a tensile load.

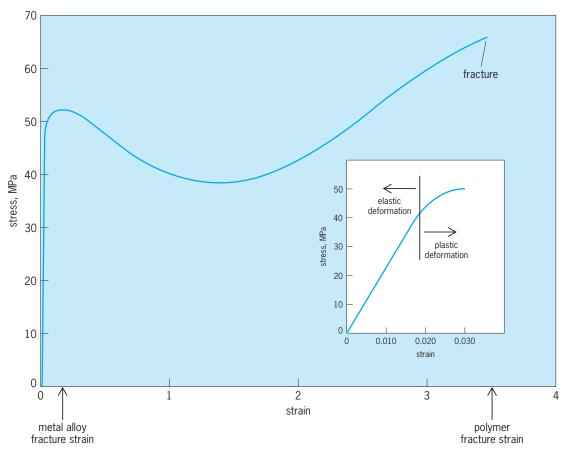


Fig. 2. Plot of stress versus strain to fracture for a typical plastic polymeric material that is deformed using a tensile load.

temperature, virtually all of these materials are brittle; that is, they fracture before deforming plastically. *See* BRITTLENESS. William D. Callister

Bibliography. T. H. Courtney, *Mechanical Behavior* of *Materials*, 2d ed., McGraw-Hill Higher Education, Burr Ridge, IL, 2000; N. E. Dowling, *Mechanical Behavior of Materials*, 2d ed., Prentice Hall, Upper Saddle River, NJ, 1998.

Plastics processing

The methods and techniques used to convert plastics materials in the form of pellets, granules, powders, sheets, fluids, or preforms into formed shapes or parts. Although the term plastics has been used loosely as a synonym for polymers and resins, plastics generally represent polymeric compounds that are formulated with plasticizers, stabilizers, fillers, and other additives for purposes of processability and performance. After forming, the part may be subjected to a variety of ancillary operations such as welding, adhesive bonding, machining, or surface decorating (painting, metallizing).

As with other construction materials, processing of plastics is only one step in the normal design-tofinished-part sequence. The choice of process is influenced by economic considerations, product specifications, number and size of finished parts, and the complexity of postfinishing operations, as well as the adaptability of the plastics to the process. A variety of processes have been used to produce the final plastic parts (**Fig. 1**). Thermoforming (not shown in Fig. 1) uses about 30% of the extruded sheets and films for primarily packaging applications. Among all plastics processing methods, injection molding ranks first in terms of the number of machines, employment, and total product value.

Injection molding. This process consists of heating and homogenizing plastics granules in an injection molding barrel until they are sufficiently fluid to allow for pressure injection into a relatively cold mold, where they solidify and take the shape of the mold cavity. For thermoplastics, no chemical changes occur within the plastic. Injection molding of thermosetting resins (thermosets) differs primarily in that the cylinder heating is designed to homogenize and preheat the reactive materials, and the mold is heated to complete the chemical crosslinking reaction to form an intractable solid. Solid particles, in the form of pellets or granules, constitute the main feed for injection moldable thermoplastics. The major advantages of the injection molding process are production speed, excellent dimensional tolerance, minimal requirements for postmolding operations, simultaneous multipart molding, and design freedom given by the complex geometries produced. In addition to thermoplastics,

thermosets, and elastomers, injection molding can be extended to materials such as fibers, ceramics, and powder metals, processed with polymeric binders. Among the polymer processing methods, injection molding accounts for 32% by weight of all plastics materials processed.

The early injection molding methods were quite similar to the old metal die casting machines. The pellets were fed through a hopper into one end of the machine, heated and fluidized in the cylinder, and injected by a ram into a mold, where cooling occurred. After solidification, the mold was opened, the part or parts were ejected, and the process was repeated.

Molding technology has since undergone an extensive evolution. Technology is now dominated by the reciprocating screw injection machine (Fig. 2), in which the screw acts as a material plasticizer as well as an injection ram. As the screw rotates, it is forced backward by a buildup of viscous plastic at the nozzle end of the cylinder. When an appropriate charge is accumulated, rotation stops and the screw moves forward, acting as an injection ram and forcing the melt into the mold. It remains forward to pack additional material into the mold cavity to compensate for the volumetric shrinkage of the material until it gains enough strength through solidification (thermoplastics) or chemical reaction (thermosets and elastomers) so that the molded part can be subsequently ejected without permanent deformation. When the material at the gate (a small opening through which the material enters the mold cavity from the melt delivery system, such as a sprue and runners in the mold) is frozen, the reciprocating screw starts to rotate and plasticate (soften) the material while it is moving backward, and the next cycle begins. The size of an injection molding machine is rated in one of two ways: by the clamping force (in tons) available to keep the mold closed during injection, or by the maximum amount (in ounces) of polystyrene that can be injected with one complete ram stroke

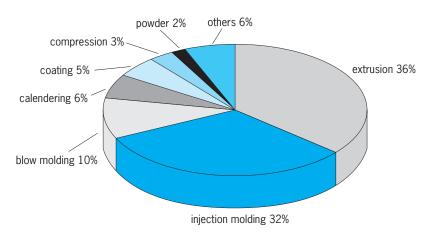


Fig. 1. Plastics consumption by process. (After D. V. Rosato, M.,G. Rosato, and D. V. Rosato, Concise Encyclopedia of Plastics, Kluwer Academic, 2000)

Since the introduction of the reciprocating screw injection molding machine, numerous attempts have been made to develop injection molding processes to produce parts with special design features and properties. Alternative processes derived from conventional injection molding have created additional applications, more design freedom, and special structural features. These efforts have resulted in a number of special injection molding processes, including:

Co-injection (sandwich) molding

- Fusible (lost, soluble) core injection molding Gas-assisted and water-assisted injection
 - molding
- Injection-compression molding
- In-mold decoration and in-mold lamination
- Insert and outsert molding
- Lamellar (microlayer) injection molding
- Liquid silicone injection molding
- Low-pressure injection molding
- Microcellular injection molding
- Multicomponent injection molding
- (overmolding)

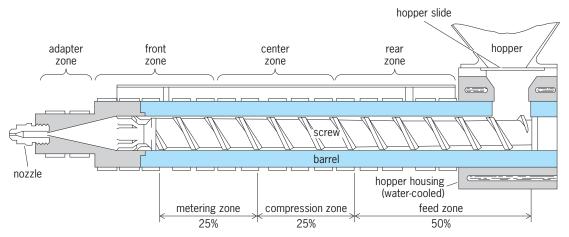


Fig. 2. Typical in-line reciprocating screw showing the three primary barrel zones and their percentages of total length. Note that the screw configuration differs in the three barrel zones. (After W. O. Elliott, Injection molding, Modern Plastics Encyclopedia, vol. 45, no. 14a, pp. 644–661, McGraw-Hill, 1968–1969)

Microinjection molding Multiple live-feed injection molding Powder injection molding Push-pull injection molding Rheomolding Rubber injection molding Structural foam injection molding Thin-wall injection molding

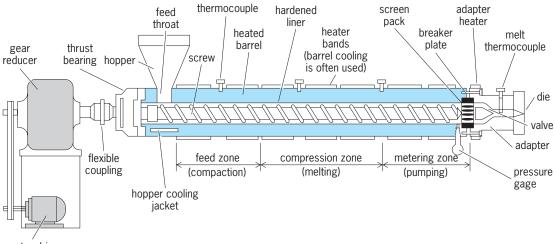
Most processes listed above are for thermoplastics materials. Processes that use thermosetting resins include reactive injection molding, resin transfer molding, and structural reaction injection molding. The development of reaction injection molding allowed for the rapid molding of liquid materials. In these processes (cold or warm), two highly reactive, low-molecular-weight, low-viscosity resin systems are first injected into a mixing head and then into a heated mold, where the reaction to a solid is completed. Polymerization and cross-linking occur in the mold. This process has proven particularly effective for high-speed molding materials such as polyurethanes, epoxies, polyesters, and nylons. See POLYAMIDE RESINS; POLYESTER RESINS; POLYMERIZA-TION; POLYURETHANE RESINS.

Extrusion. In this typically continuous process, plastic pellets or granules are fluidized and homogenized through the rotating action of a screw (or screws) inside a barrel, and the melt is continuously pushed under pressure through a shaping die to form the final product. As material passes through the die, it initially acquires the shape of the die opening. The shape will change once the material exits from the die. The extruded product is called the extrudate, and the typical dimensional expansion exhibited by the extrudate is called extrudate swell (formerly, die swell). Extrudate swell is caused by the viscoelastic behavior of the plastic melt and other factors, such as the inertia effect and the change of velocity distribution. Products made this way include extruded

tubing, pipe, film, sheet, wire and substrate coatings, and profile shapes. The process is used to form very long shapes or many small shapes, which can be cut from the long shapes. The homogenizing capability of extruders is used for plastics blending and compounding. Pellets used for other processing methods, such as injection molding, are made by chopping long filaments of extruded plastic.

The operation of an extruder provides a continuous length resembling that from a toothpaste tube. The great majority of machines are still singlescrew extruders, although twin- or multiscrew extruders have been used as continuous mixing and compounding devices. Extruders consist of essentially seven machine elements (Fig. 3): the feed hopper into which the plastics pellets are charged; the barrel containing the screw and having heating and cooling units on its exterior face and (for some extruders) vent ports; the screw which plasticates, heats, fluidizes, homogenizes, and transports the plastic to the die; the motor and gear reduction system for rotating the screw; the screen pack and breaker plates for filtration of the plastic melt and for backpressure; the die which establishes the extrusion profile; and the ancillary elements and instrumentation for the monitoring of process variables such as pressure, temperature, and screw revolutions

Single-screw extruders are often characterized by the length-to-diameter ratio of the screw, the number of stages (number of consecutive feed, compression, and metering sections in a screw), the compression ratio, and the meter ratio. These parameters provide a means of comparing different extruders. The extrusion process can be combined with a variety of postextruding operations for expanded versatility. Such postforming operations include altering round to oval shapes, blowing to different dimensions, machining, punching, and calendering (a finishing process in which the extruded sheet from



motor drive

Fig. 3. Typical single-screw extruder. Thermoplastic material is fed into the hopper and emerges as a hot melt at the die. (After R. T. Van Ness, G. R. DeHoff, and R. M. Bonner, Extrusion of plastics, Modern Plastics Encyclopedia, vol. 45, no. 14a, pp. 672–700, McGraw-Hill, 1968–1969)

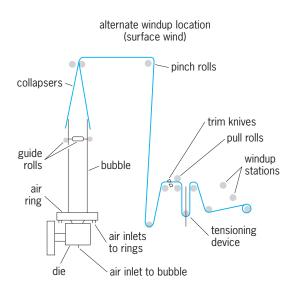


Fig. 4. Extrusion equipment (film blowing machine) for producing blown tubing. The tube is expanded by air (blown) and then is collapsed for windup. (*After R. T. Van Ness, G. R. DeHoff, and R. M. Bonner, Extrusion of plastics, Modern Plastics Encyclopedia, vol. 45, no. 14a, pp. 672–700, McGraw-Hill,* 1968–1969)

an internal batch mixer is passed through pairs of heated rolls to produce a smooth or textured surface). Extrusion can have the highest output rate of any plastics process.

The extrusion process produces pipe and tubing by forcing the melt through a cylindrical die. Wire is coated by running it through the center of an annular opening, so that the melt solidifies directly on the wire. Film can be produced by extruding a thin (less than 0.01 in. or 0.25 mm) sheet over chilled rolls (cast film process), which is then further drawn down to size by tension rolls. Blown film (generally stronger, tougher, and made more rapidly in less bulky equipment than cast film) is made by extruding a tube. In producing blown tubing, the melt flow is turned upward from the extruder and fed through an annular die. As this tube leaves the die, internal air pressure is introduced through the die mandrel, which expands the tube 1.5 to 2.5 times the die diameter and simultaneously draws the film, causing a reduction in thickness (Fig. 4). The resulting sleeve is subsequently slit along one side, making a larger film width than could be conveniently made via the cast film method. Sheet extrusion is similar to cast film, except that thicker sections (greater than 0.01 in. or 0.25 mm) are produced through an elongated die. In extrusion coating, the substrate (paper, foil, fabric, and so forth) is compressed together with the extruded plastic melt by means of pressure rollers so that the plastic impregnates the substrate for maximum adhesion. Profile extrusion is designed for rapid turnover of specially made dies because of the relatively short runs per shape. Control of shrinkage and resultant shape distortion is the most important consideration in profile extrusion. Several rules of thumb should be followed, such as avoiding thick sections, achieving uniform wall thickness, and minimizing the number of hollow sections.

Blow molding. This process consists of melting plastics pellets or granules similar to that of extrusion or injection molding, forming a tube (called a parison or preform) and introducing air or other gas to cause the tube to expand into a free-blown hollow object or, most commonly, against the surface of a spilt mold with a hollow cavity for forming into a hollow object with a definite size and shape. The parison is traditionally made by extrusion (extrusion blow molding), although injection molded preforms (injection blow molding) have gained prominence because they can be mass produced with moldedin features such as cap screws for bottles, do not require postfinishing (for example, trimming), have better dimensional tolerances and wall thicknesses, and can be made unsymmetrical and in higher volumes. In the injection blow molding process, the preforms can be transported from injection mold to blow molding machine directly, or they can undergo a reheating stage in between. In addition to extrusion blow molding and injection blow molding, a retrofitted stretch-blow system for stretch-blow bottle formation can be used in (Fig. 5).

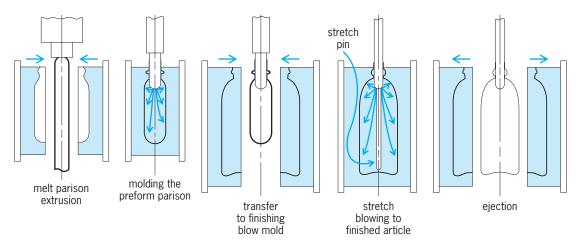


Fig. 5. Retrofitted stretch blow system that uses a preforming mold to condition a blow molded parison before final blow molding of a bottle. (After G. R. Smoluk, Modern Plastics, p. 47, McGraw-Hill, February 1981)



Fig. 6. Common blow-molded products with some preblown, injection molded preforms (tubes) shown in left of center near the foreground.

In blow molding, the mold is cooled with water for solidifying the expanded and thinned parison into the part. After the part is ejected, the pinch-off section of the parison is trimmed. One primary concern during blow molding is controlling the wall thickness distribution within the molded part, as it greatly influences the integrity, performance, and material cost of the final part. Unlike injection molding, which typically requires high-flow materials, resins used in blow molding have a high molecular weight to provide the proper melt viscosity for extrusion and melt strength in the parison for supporting its weight during extrusion as well as for expansion. A variety of blow-molded products are shown in Fig. 6, together with preblown, injection molded preforms (tubes). Bottles still represent a large share (80%) of the blowmolding market, with the remaining 20% considered as industrial blow molding (for example, fuel tanks).

Thermoforming. Plastics sheets into parts are thermoformed through the application of heat and pressure. Heat is used to soften the plastic sheet so that it can be stretched under pressure (vacuum) against one side of a cool mold surface. The pressure can be obtained through use of pneumatics (air) or compression (tooling) or vacuum (**Figs. 7** and **8**). Increasingly, prestretching of the softened sheet and/or a plug is used to improve the wall thickness uniformity of the final part. When the sheet has cooled enough to retain sufficient strength and the shape of the mold, the formed sheet is removed from the mold, and the excess material used to hold and connect the final articles (called trim, web, or skeleton) is cut off. Typically, thermoforming can be further classified into two subprocesses according to the thickness or gauge of the sheet used: thin-gauge thermoforming (sheet is less than 0.06 in. or 1.5 mm) and heavy-gauge thermoforming (sheet thickness greater than 0.12 in. or 3 mm). There is a gray area between thin-gauge and heavy-gauge thermoforming that depends on the way the sheet is presented to the thermoforming press (in rolls or discrete sheets). Tooling for thermoforming is the least expensive of the plastics processes, accounting for the method's popularity. It can also accommodate very large parts as well as small parts, the latter being useful in low-cost prototype fabrication. Thermoformed products are typically categorized as permanent, or industrial products, and disposable products (for example, packaging).

Rotational molding. Rotational molding is used to make hollow objects, such as those produced by blow molding, except that with a closed rotational mold this process is capable of producing articles without openings. In this process, finely ground powders are heated in a rotating mold until melting or fusion occurs. If liquid materials such as vinyl plastisols are used, the process is often called slush molding. With the rotating motion of the mold, the melted or fused resin uniformly coats the inner surface of the mold. When cooled, the mold opens and

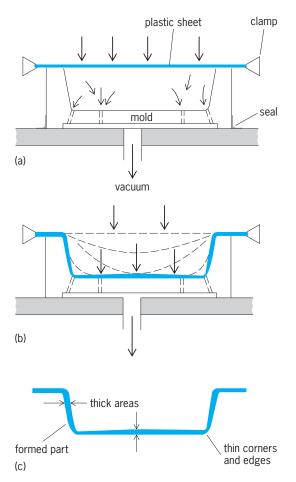


Fig. 7. Straight vacuum forming. (a) The plastic sheet is clamped and heated; the vacuum beneath the sheet causes atmospheric pressure to push the sheet down into the mold. (b) As the plastic contacts the mold, it cools. (c) In formed parts, areas of the sheet reaching the mold last are the thinnest. (*After W. K. McConnell, Jr., Thermoforming, Modern Plastics Encyclopedia, vol. 45, no. 14a, pp. 776–792, McGraw-Hill, 1968–1969*)

a hollow finished part is removed. The processes require relatively inexpensive tooling, are scrap-free, and are adaptable to large, double-walled, hollow parts that are strain-free and uniform in thickness. The processes can be performed by relatively unskilled labor, and the wall thickness is more uniform than for blow molding. However, the finely ground plastics powders are more expensive than pellets or sheets, the thin-walled parts cannot be easily made, and the process is not suited for large production runs of small parts due to long cycle times for heating and cooling of the mold and plastic material.

Compression and transfer molding. Compression molding is one of the oldest molding techniques and consists of charging a plastics powder or preformed plug into a mold cavity, closing a mating mold half, and applying pressure to compress, heat, and cause the plastic to flow and conform to the cavity shape (**Fig. 9**). The process is used primarily for thermosets, bulk molding compounds, and their composite counterparts, sheet molding compounds. Consequently, the mold is heated to accelerate the chemical cross-linking reaction.

Transfer molding is an adaptation of compression molding in that the molding powder or preform is charged to a separate preheating chamber and, when appropriately fluidized, injected into a closed mold (**Fig. 10**). The process predates, yet closely parallels, the early techniques of ram injection molding of thermoplastics. It is mostly used for thermosets and is somewhat faster than compression molding. In addition, parts are more uniform and more dimensionally accurate than those made by compression molding. Many integrated circuit chips are encapsulated by transfer molding of epoxy molding compounds.

Foam processes. Foamed plastics materials have achieved a high degree of importance in the plastics industry. Foams range from soft and flexible to hard and rigid. There are three types of cellular plastics: blown (expanded matrix such as a natural sponge), syntactic (the encapsulation of hollow organic or inorganic microspheres into the matrix), and structural (dense outer skin surrounding a foamed core).

The blowing agents used for producing porous or cellular plastics can be roughly classified into two

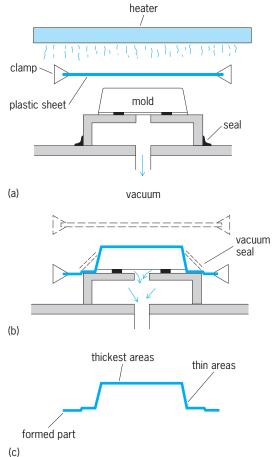


Fig. 8. Drape forming. (a) The plastic sheet is clamped and heated and then drawn over the mold. (b) When the mold has been forced into a sheet and a seal created, vacuum applied beneath the mold forces the sheet over the male mold. Side walls are formed from material draped between the top edges of the mold and the bottom seal area at the base. (c) Final wall thickness distribution. (After W. K. McConnell, Jr., Thermoforming, Modern Plastics Encyclopedia, vol. 45, no. 14a, pp. 776–792, McGraw-Hill, 1968–1969)

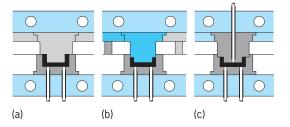


Fig. 9. Three types of compression molds. (a) Flash-type. (b) Positive. (c) Semipositive. (After E. W. Vaill, Thermoset molding, Modern Plastics Encyclopedia, vol. 45, no. 14a, pp. 714–731, McGraw-Hill, 1968–1969)

main categories: physical blowing agents and chemical blowing agents. The amount and type of blowing agent influences the densities of the finished products and their microstructures. Physical blowing agents are volatile liquids or compressed gases that change state during processing to form a cellular structure. Chemical blowing agents are solids that decompose above a certain temperature and release gaseous products. Depending on the blowing agents used and how they are being introduced, cellular plastics or plastics foams can be produced by a number of processes. These include the incorporation of a chemical blowing agent that generates gas through thermal decomposition in the polymer liquid or melt; gas injected into the melt which expands during pressure relief; gas generated as a by-product of a chemical condensation reaction during crosslinking; volatilization of a low-boiling liquid/solvent through the exothermic heat of reaction; dispersion of air by mechanical means (whipped cream); incorporation of nonchemical gas-liberating agents (adsorbed gas on finely divided carbon) into the resin mix which is released by heating; expansion of small beads of thermoplastic resin containing a blowing agent through the external application of heat; and dissolving atmosphere gases, such as nitrogen or carbon dioxide, at the supercritical state into the melt

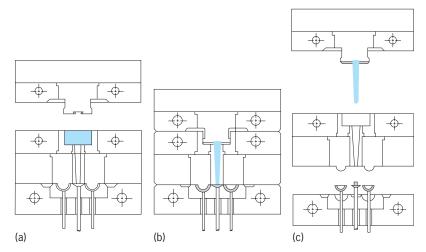


Fig. 10. Transfer molding. (a) In the molding cycle, material is first placed in the transfer pot. (b) Material is then forced through an orifice into the closed mold. (c) When the mold opens, the cull and sprue are removed as a unit, and the part is lifted out of the cavity by ejector pins. (*After E. W. Vaill, Thermoset molding, Modern Plastics Encyclopedia, vol. 45, no. 14a, pp. 714–731, McGraw-Hill, 1968–1969*)

and subsequently introducing cell nucleation and growth during processing.

Structural foams differ from other foams in that they are produced with a hard integral skin on the outer surface and a cellular core in the interior. They are made by injection molding liquefied resins containing chemical blowing agents. The initial high injection pressure causes the skin to solidify against the mold surface without undergoing expansion. The subsequent reduction in pressure allows the remaining material to expand and fill the mold. Co-injection (sandwich) molding permits injection molding of parts containing a thermoplastic core within an integral skin of another thermoplastic material. When the core is foam, an advanced form of structural foam is produced. Microcellular injection molding (also known as the MuCell process) blends gas (usually nitrogen or carbon dioxide) in the supercritical fluid state with polymer melt in the machine barrel to create a single-phase, polymer-gas solution. To facilitate fast and homogeneous mixing of the gas and the polymer and to maintain a single-phase solution prior to injection, the melt pressure inside the screw barrel is typically one order of magnitude higher than the typical backpressure used with the conventional injection molding process. During injection, a sudden pressure drop in the material as it flows through the nozzle triggers the thermodynamic instability of the polymer-gas solution. As a result, the gas emerges from the melt, forming numerous microcells. The size (typically about 10 to 100 micrometers) and density of microcells (10^7 and above) depend strongly on the process conditions and the material system, and are much finer and denser than those of structural foam parts (250 μ m to a few millimeters). This process is capable of producing parts with excellent dimensional stability, using lower injection pressures, shorter cycle times, and less material. It was applied to the extrusion process before seeing wider applications in injection molding. See FOAM.

Reinforced plastics/composites. These plastics have mechanical properties that are significantly improved due to the inclusion of fibrous reinforcements. The wide variety of resins and reinforcements that constitute this group of materials led to the more generalized use of the term composites.

Composites consist of two main components: a fibrous material in various physical forms, and a fluidized resin that will convert to a solid generally through a chemical reaction. Fiber-reinforced thermoplastic materials are typically processed in standard thermoplastic processing equipment.

For components using thermoset resins, the first step in the composite fabrication procedure is the impregnation of the reinforcement with the resin. The simplest method is to pass the reinforcement through a resin bath and use the wet impregnate directly. For easier handling and storage, the impregnated reinforcement can be subjected to heat to remove impregnating solvents or advance the resin cure to a slightly tacky or dry state. The composite in this form is called a prepreg (B stage). This B-stage condition allows the composite to be handled, yet

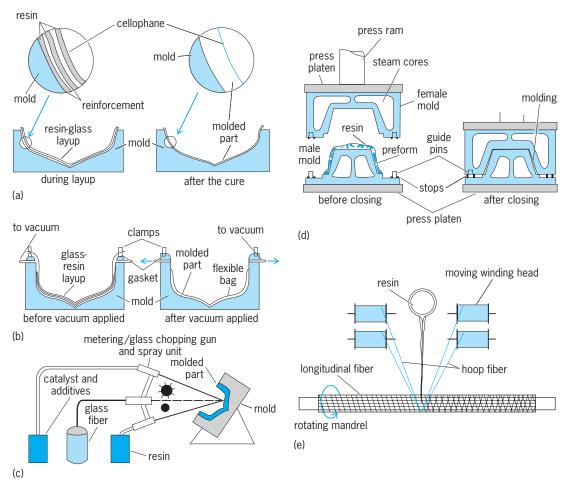


Fig. 11. Techniques for producing reinforced plastics and composites. (a) Hand lay-up technique for reinforced thermosets. (b) Vacuum bag molding method. (c) Spray-up method. (d) Matched metal die molding. (e) Filament winding. (After R. S. Morrison, Reinforced thermosets, Modern Plastics Encyclopedia, vol. 45, no. 14a, pp. 628–637, McGraw-Hill, 1968–1969)

the cross-linking reaction has not proceeded so far as to preclude final flow and conversion to a homogeneous part when further heat or pressure is applied.

Premixes, often called bulk molding compounds, are mixtures of resin, inert fillers, reinforcements, and other formulation additives, which form a puttylike rope, sheet, or preformed shape.

Converting these various forms of composite precursors to final part shapes is achieved in a number of ways. Hand lay-up techniques consist of an open mold onto which the impregnated reinforcement or prepreg is applied layer by layer until the desired thicknesses and contours are achieved (Fig. 11a). The thermoset resin is then allowed to harden (cure). Often the entire configuration will be enclosed in a transparent sealed bag (vacuum bag) so that a vacuum can be applied to remove unwanted volatile ingredients and entrained air to improve the densification of the composite (Fig. 11b). External heat may be applied to accelerate the process. Often, a bagged laminate will be inserted into an autoclave so that the synergistic effects of heat, vacuum, and pressure can be obtained. At times, a specially designed spray apparatus is used which simultaneously mixes and applies a coating of resin and chopped reinforcement onto a mold surface (Fig. 11*c*). This technique is particularly useful for large structures, such as boat hulls and truck cabs, covering complex shapes as readily as simple configurations.

Resin transfer molding and structural reaction injection molding are two additional liquid composite molding processes that are well suited for medium to mass production of medium-to-large, complex, lightweight, and high-performance composite components primarily for the aerospace and automotive industries. In these processes, a reinforced fiber mat (preform) is preplaced in a closed mold to be impregnated by a low-viscosity, reactive liquid resin in a transfer or injection process. Sometimes, vacuum is used to assist the flow of the resin. Resin transfer molding and structural reaction injection molding are similar, with the main differences being the resin used, the mixing and injection set-up, the mold requirements, the cycle time, the fiber volume fraction, and the suitable production volume.

Matched-die compression molding resembles normal compression molding, although the pressures are considerably lower (Fig. 11*d*). Premix molding is essentially the same process, except that premix compounds are used. Pultrusion is a process for the

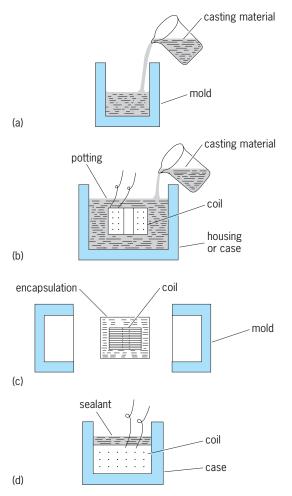


Fig. 12. Low-pressure plastics processes: (a) casting, (b) potting, (c) encapsulation, and (d) sealing. (After H. Lee and K. Neville, Handbook of Epoxy Resins, McGraw-Hill, 1967)

continuous extrusion of reinforced plastics profiles. Strands of reinforcement are drawn (pulled) through an impregnating tank, the forming die, and then a curing area (radio-frequency exposure). Filament winding is a process in which the continuous strands of reinforcement are drawn through an impregnating bath and then wound around a mandrel to form the part (Fig. 11*e*). This technique is most often used to form hollow objects such as chemical storage tanks or chemically resistant pipe. Advanced automated processes, such as ply cutting, tape laying and contouring, and ply lamination, provide improved parts and reduced costs, particularly in the aerospace industry. *See* COMPOSITE MATERIAL; POLYMERIC COM-POSITE.

Casting and encapsulation. Casting is a lowpressure process requiring nothing more than a container in the shape of the desired part. For thermoplastics, liquid monomer is poured into the mold and, with heat, allowed to polymerize in place to a solid mass. For vinyl plastisols, the liquid is fused with heat. Thermosets, usually composed of liquid resins with appropriate curatives and property-modifying additives, are poured into a heated mold, wherein the cross-linking reaction completes the conversion to a solid. Often a vacuum is applied to degas the resultant part for improved homogeneity.

Encapsulation and potting are terms for casting processes in which a unit or assembly is encased or impregnated, respectively, with a liquid plastic, which is subsequently hardened by fusion or chemical reaction (**Fig. 12**). These processes are predominant in the electrical and electronic industries for insulation and protection of components.

Calendering. In the calendering process (**Fig. 13**), a plastic is masticated between two rolls that squeeze it out into a film that passes around one or more additional rolls before being stripped off as a continuous film. Fabric or paper may be fed through the latter rolls so that they become impregnated with the plastic. *See* POLYMER.

Lih-Sheng Turng; Sidney H. Goodman Bibliography. J. Avery, Injection Molding Alternatives: A Guide for Designers and Product Engineers, Carl Hanser Verlag, 1998; A. W. Birley, R. H. Heath, and M. J. Scott, Plastics Materials: Properties and Applications, 1988; J. Brydson, Plastics Materials, 7th ed., 1999; J. H. Dubois and W. I. Pribble, Plastics Mold Design Engineering Handbook, 4th ed., 1987; S. H. Goodman (ed.), Handbook of Thermoset Plastics, 2d ed., 1998; F. Hensen, W. Knappe, and H. Potente (eds.), Plastics Extrusion Technology, Carl Hanser Verlag, 1988; Plastics and Rubber Institute Staff, Plastics and Polymer Processing Automation, 1988; R. G. W. Pye, Injection Mold Design, 4th ed., 1989; C. Rauwendaal, Polymer Extrusion, 4th ed., Carl Hanser Verlag, 2001; L.-S. Turng, Special injection molding processes, in T. A. Osswald, L.-S. Turng, and P. J. Gramann (eds.), Injection Molding Handbook, Carl Hanser Verlag, 2002; A. Whelan, Plastics Molding, 4 vols., 1991; L. J. White, Twin Screw Extrusion, Carl Hanser Verlag, 1991.

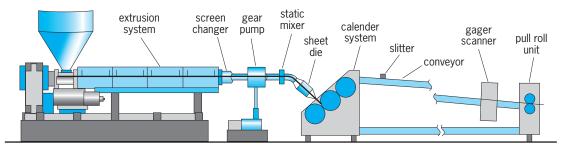


Fig. 13. Calendered sheet extrusion system. (After D. Hanson, Modern Plastics, p. D-29, McGraw-Hill, Mid-November 1998)

Plate girder

A beam assembled from steel plates which are welded or bolted to one another, used to support the horizontal surface of a bridge. Plate girders generally have I-shaped cross sections (Fig. 1). Cross sections with two or more webs that form a U- or box-shape can also be considered plate girders, but are generally classified separately as box girders. Plate girders are commonly used in bridge structures where large transverse loads and spans are encountered. On average, the height (measured from the top of the top flange to the bottom of the bottom flange) of plate girders is one-tenth to one-twelfth the length of the span, which can vary based on the loading and architectural requirements. The dimensions of plate girders are tailored to specific applications by adjusting the width and thickness of the web and flange plates along the length of the beam. By contrast, hot-rolled beams produced at steel mills are formed to specific dimensions according to a standardized list of structural shapes, and have limitations on their size. Plate girders can be produced with much greater depths than hot-rolled beams. The depth is generally considered to be the distance between the centroids of the flanges which defines the effective moment arm in bending. When different-grade steels are used for the flanges and webs, a plate girder is classified as a hybrid plate girder. See BRIDGE; LOADS, TRANSVERSE; STRUCTURAL STEEL.

Splices. Long-span plate girders are often fabricated in segments due to size or weight restrictions at the fabrication facility, during shipping, or at the construction site. These segments are spliced or joined

together by either high-strength bolts or welding. Bolted splices are faster to install and easier to inspect than welded splices, but are readily visible to the observer, which can detract from the esthetics of the structure.

The individual flange and web plates that constitute plate girders commonly consist of multiple plates that are spliced together by welding. Individual plate splices are necessary whenever the plate length required exceeds what is available or when the width, thickness, or material grade of the plate changes. When width or thickness changes occur, the edges of the plates being spliced together are beveled prior to welding to minimize any sharp corners or notches that can cause fatigue cracks (Fig. 1). Since welding affects the material properties of the steel, splices in the flanges and webs should be staggered so two splice welds do not coincide.

Stiffeners. Plate girders commonly have stiffener plates which increase the strength of the web plate. Since plate girder webs are relatively slender, they are susceptible to buckling when subjected to shear and bending forces. The bending strength of webs can be enhanced using horizontal (longitudinal) stiffeners, while the shear strength of webs can be enhanced using vertical (transverse) stiffeners (Fig. 1). Increasing the thickness of the web plate increases the overall web strength and can eliminate the need for stiffeners altogether. The additional material cost of a thicker web plate is often less than the material, fabrication, and installation cost of using stiffeners. In some instances, the use or elimination of stiffeners is based on esthetic reasons alone. Bearing stiffeners are vertical stiffeners that are placed

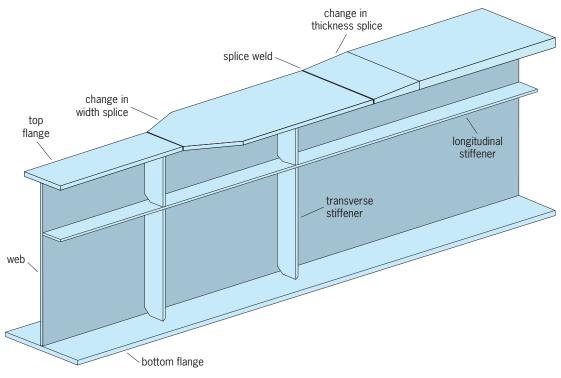


Fig. 1. Plate girder with stiffeners and plate splices.

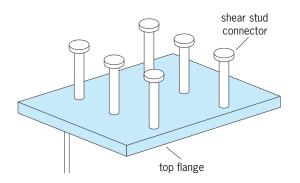


Fig. 2. Shear connectors for composite plate girders.

at locations where concentrated forces are present. Large concentrated forces, such as those at support locations, can cause local buckling of the web. *See* BEAM; SHEAR.

Composite girders. Plate girders that support concrete floors or roadways can be designed as composite or noncomposite girders. For noncomposite plate girders, the concrete is simply laid to rest upon the top flanges. Slip between the concrete and steel is not mechanically prevented, and the structural system behaves as two independent beams. Frictional forces between the concrete and steel are generally ignored as they are very small in comparison to the forces acting to produce slip. Composite plate girders use mechanical shear connectors to prevent slip at the interface (Fig. 2). These shear connectors are attached to the steel and embedded into the concrete. This allows the concrete and steel to act together as a single structural unit, producing a significantly stiffer and more efficient beam. The added expense of composite plate girders is generally offset by increased structural efficiency.

Curved girders. Plate girders can be fabricated with a horizontal curve to accommodate curved bridge structures such as highway interchanges. The narrow flange plates are usually cut and spliced as rectangular plates and then mechanically curved by applying heat. The curved flanges are then welded to the web to form the curved girder. Curved girders add a level of complexity to straight girders because forces due to gravity cause both bending and torsion. *See* HIGHWAY BRIDGES; TORSION. Brian Chen

Bibliography. W. F. Chen, *Handbook of Structural Engineering*, CRC Press, Boca Raton, FL, 1997; W. Schueller, *The Design of Building Structures*, Prentice Hall, Upper Saddle River, NJ, 1996.

Plate tectonics

Plate tectonics theory provides an explanation for the present-day tectonic behavior of the Earth, particularly the global distribution of mountain building, earthquake activity, and volcanism in a series of linear belts. Numerous other geological phenomena such as lateral variations in surface heat flow, the physiography and geology of ocean basins, and various associations of igneous, metamorphic, and sedimentary rocks can also be logically related by plate tectonics theory.

Theory and Evidence

The theory is based on a simple model of the Earth in which a rigid outer shell 30-90 mi (50-150 km) thick, the lithosphere, consisting of both oceanic and continental crust as well as the upper mantle, is considered to lie above a hotter, weaker semiplastic asthenosphere. The asthenosphere, or low-velocity zone, extends from the base of the lithosphere to a depth of about 400 mi (700 km). The brittle lithosphere is broken into a mosaic of internally rigid plates that move horizontally across the Earth's surface relative to one another. Only a small number of major lithospheric plates exist, which grind and scrape against each other as they move independently like rafts of ice on water. Most dynamic activity such as seismicity, deformation, and the generation of magma occurs only along plate boundaries, and it is on the basis of the global distribution of such tectonic phenomena that plates are delineated. See EARTH INTERIOR; LITHOSPHERE.

The plate tectonics model for the Earth is consistent with the occurrence of sea-floor spreading and continental drift. Convincing evidence exists that both these processes have been occurring for at least the last 600 million years (m.y.). This evidence includes the magnetic anomaly patterns of the sea floor, the paucity and youthful age of marine sediment in the ocean basins, the topographic features of the sea floor, and the indications of shifts in the position of continental blocks which can be inferred from paleomagnetic data on paleopole positions, paleontological and paleoclimatological observations, the match-up of continental margin and geological provinces across present-day oceans, and the structural style and rock types found in ancient mountain belts. See CONTINENTAL DRIFT; CONTINENTAL MARGIN.

Plate motion and boundaries. Geological observations, geophysical data, and theoretical considerations support the existence of three fundamentally distinct types of plate boundaries, named and classified on the basis of whether immediately adjacent plates move apart from one another (divergent plate margins), toward one another (convergent plate margins), or slip past one another in a direction parallel to their common boundary (transform plate margins). Figure 1 shows the major plates of the lithosphere, the major plate margins, and the type of motion between plates. Plate margins are easily recognized because they coincide with zones of seismic and volcanic activity; little or no tectonic activity occurs away from plate margins. The boundaries of plates can, but need not, coincide with the contact between continental and oceanic crust. The nature of the crustal material capping a plate at its boundary may control the specific processes occurring there, particularly along convergent plate margins, but in general plate tectonics theory considers the continental crustal blocks as passive passengers riding on the upper surface of fragmenting, diverging, and colliding plates.

The velocity at which plates move varies from plate to plate and within portions of the same plate, ranging between 0.8 and 8 in. (2 and 20 cm) per year. This rate is inferred from estimates for variations in the age of the sea floor as a function of distance from mid-oceanic ridge crests. Ocean-floor ages can be directly measured by using paleontological data or radiometric age-dating methods from borehole material, or can be inferred by identifying and correlating the magnetic anomaly belt with the paleomagnetic time scale.

Divergent plate margins. As the plates move apart from the axis of the mid-oceanic ridge system, the new volcanic material welling up into the void forms a ribbon of new material, which gradually splits down its center as the boundary of plate separation continues to develop. Each of the separating plates thus accretes one-half a ribbon of new lithosphere, and in this way new lithosphere and hence new surface area are added. The process is considered to be continuous, and the boundary at which separation is taking place always maintains itself in the center of the new material. *See* MID-OCEANIC RIDGE.

The accretion at any spreading boundary is usually bilaterally symmetric. The morphology of the ridges is also quite symmetric and systematic. The new material that wells up at the ridge axis is hot and therefore expanded and less dense than the surrounding older material. Consequently, the new material is topographically highest. As new material divides and moves away from the ridge axis, it cools, contracts, becomes denser, and subsides. The densification is caused by the combined effect of pure thermal contraction and thermally driven phase changes. Subsidence is fastest for newly generated oceanic crust and gradually decreases exponentially with time. This observation explains the fact that in cross section the shape of the slope of the ridges is steepest at the ridge axis and gradually decreases down the flanks beneath the abyssal sediments and to the bounding continents. Since all known oceanic lithosphere has been generated by the spreading at a ridge axis, all oceanic lithosphere is part of the midoceanic ridge system. Because of the systematic way in which the morphology of the ridges is formed, most oceanic crust follows the same time-dependent subsidence curve within an error range of about 300 ft (100 m). This means that the same age-versusdepth curve fits nearly all parts of the mid-oceanic ridge system. The ridge axis is found at a depth of 1.71 mi \pm 300 ft (2.75 km \pm 100 m), and oceanic lithosphere that is 30 m.y. old is found at a depth of 2.72 ± 300 ft (4.37 km ± 100 m). There are exceptional areas such as Iceland where the ridge axis is above sea level.

Magnetic lineations and age of oceans. Much of the evidence which leads to the development of the concept of sea-floor spreading, and in turn to an understanding of divergent plate margins and plate tectonics theory, came from analyses of the magnetic properties of the sea floor. Magnetometer surveys

across the sea floor near mid-oceanic ridges reveal a pattern of alternating positive and negative magnetic anomalies (**Fig. 2**). The characteristics of these magnetic anomaly patterns (parallel to and in symmetrically matching widths across the ridge crest) and present-day general knowledge of the Earth's magnetism logically support the conclusion that sea-floor spreading occurs.

The Earth's magnetic field as measured at the surface may be approximately represented by an axial geocentric dipole-in other words, as if a simple two-pole magnet existed coincident with the Earth's rotational axis. The Earth's magnetic field reverses polarity episodically. Over an interval of time, the strength of the dipole field gradually decreases to zero and then gradually increases in the opposite direction. The total transit time from full strength in one direction (polarity) to full strength in the opposite direction may be less than 6000 years. The residence time in any one particular polarity ranges from 10,000 years up to tens of millions of years. Thus, the polarity behavior of the Earth's magnetic field is approximately analogous to a randomly triggered flip-flop circuit.

The top 600 to 1500 ft (200 to 500 m) of the lithosphere is the oceanic basalt, which is accreted at separating boundaries and which contains magnetizable iron minerals. As the newly injected material cools through its Curie point ($\sim 1100^{\circ}$ F or 570 $^{\circ}$ C), it becomes permanently magnetized in the direction of the Earth's magnetic field at that time and place. Because the oceanic crust is continuously being formed and transported in a bilaterally symmetric pattern away from the line of rifting, it is similar to a continuous magnetic tape record of the Earth's magnetic polarity events. At the mid-oceanic ridge axis, the present polarity of the field is recorded. Down the ridge flanks, progressively older oceanic crust is traversed; hence the crust contains in its remanent magnetism a continuously older record of the polarity of the field. See BASALT; CURIE TEMPERATURE.

As a result of sea-floor spreading and polar reversals, the oceanic crust appears magnetically as a set of alternately normally or reversely magnetized strips of basalt arranged in a bilaterally symmetrical pattern about the mid-oceanic ridge axes. Because the rate of separation at most segments of the mid-oceanic ridge axes has varied only slowly with time, the spatial distribution of the stripes is proportional to the temporal history of the polarity of the field. Thus, the polarity of the oceanic crust may be inferred indirectly from the pattern of magnetic anomalies.

Because the occurrence of polarity events is random, they form a unique sequence which is reflected directly in the magnetic anomaly patterns. A unique pattern of magnetic anomalies has been correlated throughout most of the world's oceans. From this pattern a magnetic reversal time scale from the present to 160 m.y. ago has been developed. This geomagnetic time scale has been calibrated by drilling on key anomalies in the oceans and paleontologically determining the age of the sediment-basalt

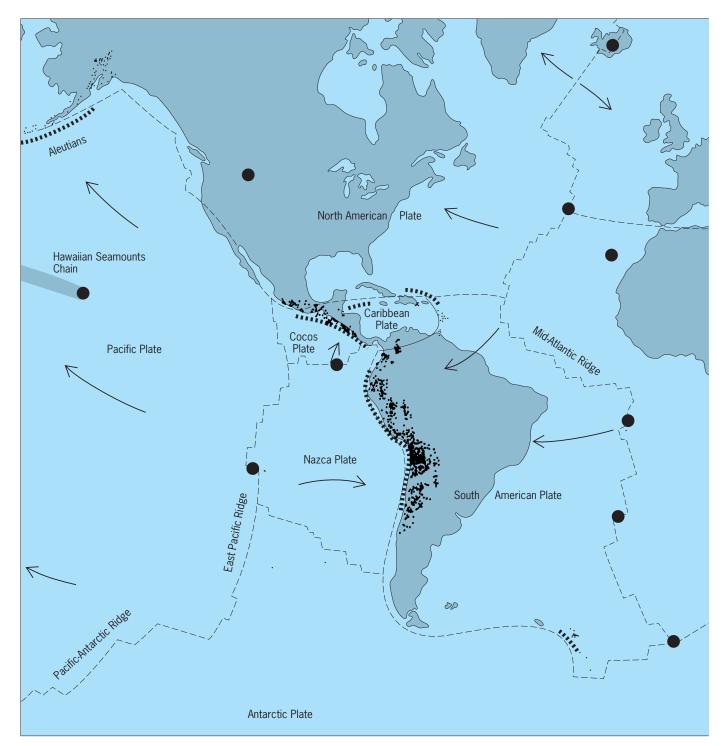


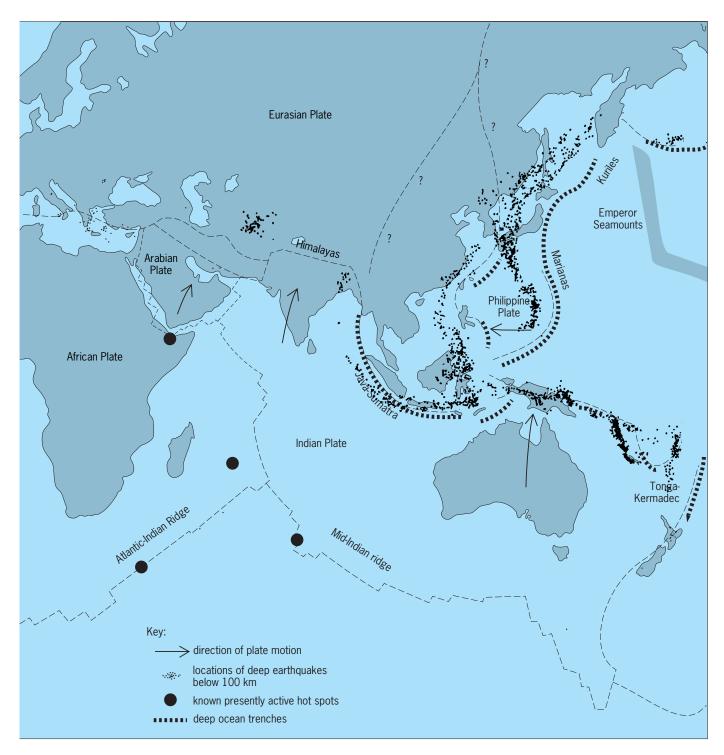
Fig. 1. Tectonic map of the Earth. (After M. Nafi Toksoz, The subduction of the lithosphere, Sci. Amer., 233(5):88–98, November 1975)

interface. Thus a magnetic polarity time scale has been developed, and with it identifiable magnetic anomalies that can be correlated with part of the time scale may be assigned an absolute age. In this way, the spatial distribution of the magnetic anomalies is used to calculate the rates of separation of the plates in the past. *See* GEOMAGNETISM; PALEOMAG-NETISM.

Convergent (destructive) plate margins. Because the Earth is neither expanding nor contracting, the in-

crease in lithosphere created along divergent boundaries must be compensated for by the destruction of lithosphere elsewhere. The rates of global lithosphere construction and destruction must be equal, or the radius of the Earth would change. Compensatory destruction or removal of lithosphere occurs along convergent plate margins (subduction zones) and is accomplished by plate subduction and continental collision. *See* SUBDUCTION ZONES.

Along subduction zones, one plate plunges



beneath another (**Fig. 3**). The downgoing slab is usually oceanic because the relatively buoyant continental lithosphere cannot be subducted beneath the relatively denser oceanic lithosphere. The upper or overriding plate may be continental or island arc lithosphere, and occasionally oceanic plateaus as in the case of Oregon and Alaska.

The dip of the downgoing underthrust slabs varies but averages 45° . Earthquake foci along individual subduction zones indicate the plate underthrust

angle, often referred to as the Benioff plane or Wadati-Benioff zone. This plane dips away from oceanic trenches toward adjacent volcanic arcs and continents, and marks the surface of slippage between the overriding and descending lithospheric plates. Successive belts of shallow (less than 40 mi or 70 km), intermediate (40-200 mi or 70-300 km), and deep-focus (200-400 mi or 300-700 km) earthquakes are associated with subduction zones. Subduction zones are also associated with active

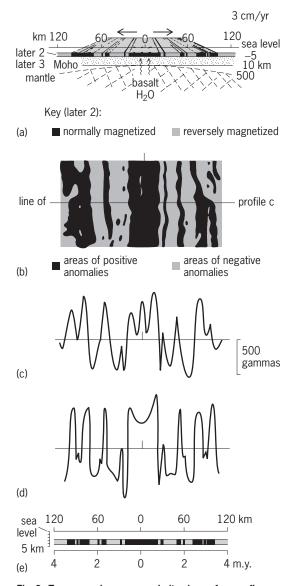


Fig. 2. Tape-recorder conveyor-belt scheme for sea-floor spreading. (a) Schematic of crustal model, applied to Juan de Fuca Ridge, southwest of Vancouver Island. (b) Part of summary map of magnetic anomalies recorded over Juan de Fuca Ridge. (c) Total-field magnetic anomaly profile along the line indicated in b. (d) Computed profile assuming the model and reversal time scale. Intensity and dip of Earth's magnetic field taken as 54,000 gammas (0.54 oersted) and +667°; magnetic bearing of profile 087°; 10 km horizontally is equivalent to 100 gammas vertically. Normal or reverse magnetization is with respect to an axial dipole vector, and the assumed effective susceptibility is 0.01 except for central block and ridge crest (+0.02). (e) Scale applicable to b-d, 1 cm = 0.4 in.; 1 km = 0.6 m; 1 gamma = 1 nanotesla. (After R. A. Phinney, ed., The History of the Earth's Crust, 1968)

volcanism and the development of deep-ocean trenches. These features encircle the Pacific Ocean basin.

A line of andesitic volcanoes usually occurs on the upper plate, forming a chain that is parallel to the trench. The volcanism occurs at that point above the subduction zone where the upper surface of the downgoing lithosphere has reached a vertical depth of approximately 75 mi (120 km). The volcanic materials arise from partial melting of the downgo

ing slab. The volcanoes of the Andes chain of South America and island arcs in the Pacific such as Izu-Bonin and the Mariana have formed in this manner.

Convergent plate margins that develop island arc systems adjacent to trench systems are geologically complex. The region between a volcanic island arc and the trench consists of volcanic rock and sediments. Slumping and turbidity currents transport sediment into the trench axis. The trench sediment may be transferred from the downgoing plate to the continent or island arc. Conversely, a convergent margin without a large sediment supply may be eroded during subduction.

Because of density consideration, subduction requires at least one of the two converging plates to be oceanic. If both converging plates consist of continental lithosphere, continental collision occurs, but not subduction. Continental collision results in compensatory reduction in lithospheric width by folding and compressing the lithosphere into narrower, linear mobile belts. In such collisions, sediments deposited along the continental margins and within the closing ocean basins are compressed into a series of tight folds and thrusts. Fragments of oceanic crust may be thrust up onto adjacent continental rocks (obduction) as ophiolite successions. A classic example of a continental collision belt is the Himalayan belt, produced during the Cenozoic Era by the convergence of the Indian continent with Eurasia.

Like plate divergence, plate convergence produces a distinctive suite of igneous rock types. Subduction zones are marked by the belts of predominantly andesitic volcanoes either in island arcs located landward of the trench system (Japan and the Philippines) or along the rim of overriding continental blocks (the Andes belt). These andesitic volcanic terranes are commonly associated intimately with plutonic igneous rocks, mainly granodiorites. The origin of andesitic magmas, and the predominance of granodiorite plutons within continental blocks to the exclusion of most other igneous rock varieties, was perplexing until the development of plate tectonics theory. Now both seem to be directly related to the generation of parent magmas by the frictional melting of ocean-floor basalt and overlying sediment cover along subduction zones. Partial melting of the lower crust and upper mantle also occurs.

Plate subduction and continental collision can also explain the origin of two other puzzling rock sequences commonly found within mountain belts: mélange and blueschist terranes. Mélange, a heterogeneous assemblage of intensely sheared, poorly sorted, angular blocks set in a fine-grained matrix, is probably generated at shallow depths along subduction zones as the oceanic crust and overlying sediment cover of the descending plate are scraped and crushed against the overriding plate. Blueschist terranes (their dark blue color is imparted by the presence of various low-temperature-high-pressure metamorphic minerals such as glaucophane, lawsonite, and jadeite) occur in belts within mountain chains, parallel with but external to (toward the ocean) the more conventional paired metamorphic

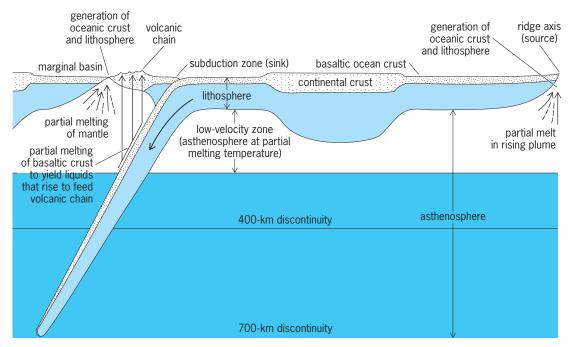


Fig. 3. Cross section of the upper mantle and crust showing a lithospheric plate riding on the asthenosphere. The continent is embedded in the plate and moves with it. Divergent plate margin with generation of oceanic crust and lithosphere is shown on the right; convergent plate margin with subduction is shown on the left. 1 km = 0. 6 mi. (*After J. F. Dewey, Plate tectonics, Sci. Amer.,* 226:56–68, May 1972)

facies of the greenschist terrane. The peculiar physical conditions required by the blueschist facies, great burial depth (in excess of 10 mi or 16 km) but moderate temperature ($390-800^{\circ}$ F or $200-450^{\circ}$ C), should be generated along subduction zones when the descending lithospheric plate is underthrust at a greater rate than the local geothermal gradient can heat it. *See* BLUESCHIST.

Transform plate margins. Transform faults are always strike-slip faults. They occur where the relative motion between the two plates is parallel to the boundary that separates the plates. They may join a ridge to a ridge, a ridge to a trench, or a trench to a trench. Ridge-trench transforms will always change length with time. A trench-trench transform may lengthen, shrink, or remain constant, depending on which of the plates that form the subduction system is the downgoing plate. A transform that joins two ridge axes will not change in length with time.

First, consider ridge-ridge transforms (**Fig.** 4). Earthquake epicenter data show that earthquakes occur only along the ridge axes and the connecting transform. Studies of earthquakes that occur at a ridge-ridge transform fault show that the first motion was strike-slip parallel to the direction of the transform and opposite to the sense of offset of the ridge axis. If the ridge axes are offset leftlaterally, the relative motion across the transform that joins the ridges will be right-lateral as the plates separate. The motion of the fault takes place spasmodically. Deformation occurs on either side of the fault until the elastic limit is reached and the rupture occurs, causing an earthquake. *See* EARTHQUAKE.

Fracture zones are morphologic scars that are fossilized transform faults. They may be used to

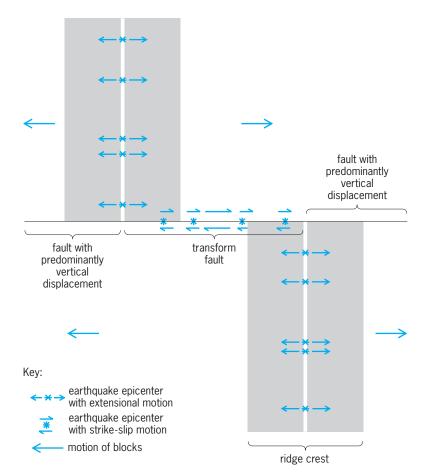


Fig. 4. Idealized map view of a transform fault. The ridge appears to be offset in a left-lateral sense, but the motion on the transform fault between the ridge crests is right-lateral. (*After C. K. Seyfert and L. A. Sirkin, Earth History and Plate Tectonics, Harper Row,* 1973)

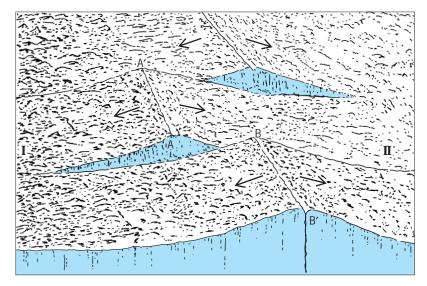


Fig. 5. Ridge-ridge transform fault appears between two segments of ridge that are displaced from each other. (After J. Tuzo Wilson, ed., Continents Adrift and Continents Aground, W. H. Freeman, 1976)

determine the past direction of relative motion of the bounding plates. In **Figs. 5** and **6**, the transform fault A-B offsets two ridge axes, and there is right-lateral strike-slip motion across the transform. All the material to the left of the ridge-transformridge boundary A'ABB' belongs to the torsionally rigid plate I, and all the material to the right of this boundary is part of plate II. Thus, beyond the ends of the transform, to the left of A' and right of B, there is no relative horizontal motion.

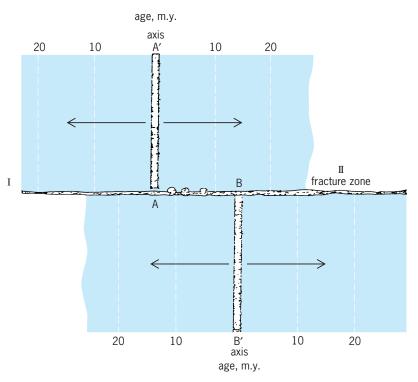


Fig. 6. Molten rock wells up from the deep Earth along a spreading axis, solidifies, and is moved out (shown by arrows). The axis is offset by a transform fault. Between two offset axes, material on each side of the transform fault moves in opposite directions, causing shallow earthquakes. (After J. R. Heirtzler, Sea-floor spreading, Sci. Amer., vol. 219, no. 6, December 1968)

Because of the way in which fracture zones are formed, they also preserve a record of the past direction of relative motion of the plates. The transform fault between plates I and II in Figs. 5 and 6 is in theory a vertical plane between these plates along which horizontal strike-slip motion occurs. The face of plate I that abuts against plate II in effect defined this transform plane. But as plate I moves to the left, this transform plane or face passes by the end of ridge A'A at point A, and new younger material is intruded against it. Together, the transform plane of plate I and the new younger material move off to the left. This process is continuous. The new younger material is more elevated and forms an escarpment whose horizontal trace records the direction of relative movement. Because the rate of subsidence deceases with time, the difference in elevation across a fracture zone decreases with age.

A very narrow and deep rift valley is also associated with the transform. However, because of timedependent cooling, the oceanic lithosphere contracts horizontally. This contraction causes cracking of the lithospheric plate and widening and splaying of the rift associated with the fracture zone. Young fracture zones are characterized by an escarpment and a deep fissure along their length. The magnitude of the vertical offset decreases with age, and the fissure tends to widen. Down the lower flanks of the ridges, turbidities from the abyssal plains often finger into the fracture zone. The fracture zones form long curvilinear features throughout the oceans and are mappable because of morphology alone. *See* RIFT VALLEY.

The numerous strike-slip faults, which offset segments of the mid-oceanic ridge system, are classical examples of transform plate boundaries. The San Andreas system of California, which offsets portions of the East Pacific Rise, is probably the bestknown example. Transform faults show apparent lateral displacements of many tens or even hundreds of kilometers. Occasionally extension, and consequent igneous activity, may occur at a transform fault.

Transform faults are parallel to the direction of relative motion between the bounding plates. According to a theorem of spherical geometry, if plate 1 is moving with respect to plate 2 on a sphere, the instantaneous relative motion may be represented by the rotation of one plate with respect to the other about a single stationary pole. Circles drawn on the sphere concentric to the pole of relative motion will be parallel to the direction of instantaneous relative motion. Great circles drawn perpendicular to these concentric circles will intersect at the pole of instantaneous relative motion. The transform faults that separate block 1 from block 2 are parallel to the direction of present-day relative motion and therefore must lie on circles that are concentric to a pole of relative motion of 1 with respect to 2. Great circles drawn perpendicular to the transform faults will intersect at the pole of relative motion. In this way, the pole of relative motion may be found for any two plates whose common boundary consists in part of transform faults.

The instantaneous rate of relative motion is an angular rate of rotation of 1 with respect to 2 about an axis through the pole of relative motion and the center of the Earth. The linear rate of relative motion R (in centimeters per year) will then be proportional to the sine of the colatitude with respect to the instantaneous pole, as shown in the equation below,

$$R = a\sin\theta \frac{d\phi}{dt}$$

where a = the radius of the Earth, 2.4×10^8 in. (6 × 10^8 cm); $\theta =$ the colatitude of the point at which the rate is to be determined with respect to the instantaneous pole; and $d\phi/dt =$ the angular rate of motion of 1 with respect to 2.

In this way, spreading-rate data calculated from the distribution of magnetic lineations near the axis of a spreading ridge-transform system may be used to compute a pole of relative motion. Note that in this case the pole is not strictly instantaneous because magnetic lineations covering a finite period must be used. It has been found that the poles and rates of relative motion tend to remain nearly constant, perhaps changing quite slowly over long time intervals. Thus, poles and rates computed in the above fashion may often be used to describe the relative motion for several tens of millions of years.

The fracture zones may often be treated as fossil transform faults. To obtain paleopoles of relative motion, segments of fracture zones spanning the same age and from the same plate must be used. Perpendicular great circles drawn to these segments give a pole of relative motion for that time interval in a reference frame fixed with respect to that plate. The magnetic anomaly lineations abutting the fracture zone are used to determine the rate of motion. By calculating a time sequence of poles and rates using time sequential segments of fracture zones, the spreading history of an entire ocean basin may be determined. *See* FAULT AND FAULT STRUCTURES; TRANSFORM FAULT.

Triple junctions. The point where the boundaries between three pairs of plates join is known as a triple junction. Because the Earth's lithosphere shell is segmented into a mosaic, the boundary between any two plates must end in a triple junction. Quadruple junctions are theoretically possible, but except under very unusual circumstances they degenerate immediately into two triple junctions. Figure 1 shows a number of triple junctions that vary as to the types of boundaries that meet and as to the geometry.

When three bodies are moving with respect to each other across a spherical surface, if the relative motion of each of two pairs of the plates is known, the relative motion of the third pair may be determined. In the case of a spherical surface, the relative motion between two plates may be represented by a vector through the pole of instantaneous relative motion, with the length of the vector proportional to the rate of rotation. The relative motion of three plates or bodies on a sphere may be represented by three such vectors. Each vector gives the instantaneous relative motion of two of the three plates. Thus, three such vectors describe completely the relative motion of the three plates and must add to zero.

The orientation of the boundary between these plates will determine the type of interaction at the boundary (that is, spreading, subduction, or transform). There are a number of possible triple junctions. The Cocos, Pacific, and Nazca plates join at a ridge-ridge triple junction (Fig. 1). This is an example of a stable triple junction, that is, the triple junction moves in a constant direction and at a constant speed with respect to each of the plates. Other triple junctions are unstable or transitional, such as a ridge-ridge-transform triple junction which will change immediately to a transform-ridgetransform triple junction. The most geologically important triple junctions are those that migrate along a boundary. Consequently, the boundary along which the triple junction moves will experience a change in the direction of relative motion. If the change in relative motion is radical, there may be an accompanying change in tectonic style. This type of plate interaction occurred at the western boundary of the North American Plate during the Cenozoic, and has resulted in a time transgressive change, from south to north, from subduction to strike slip. This is reflected in the geologic record as a change from compressive tectonics with thrusting, folding, and Andean or island arc volcanism to strike slip, as occurs along the San Andreas Fault.

Paleogeography. Because the accretion of new lithosphere is bilaterally symmetric, the associated magnetic lineations must also be bilaterally symmetric across a ridge axis. Once formed, oceanic crust is rarely disturbed by horizontal shear, but each magnetic lineation preserves the shape of the midoceanic ridge axis and offsetting transforms at the time of its formation. The fact that lineations of the same age from opposite sides of a ridge axis may be fitted together is considered evidence of the undisturbed nature of the oceanic crust.

The North and South Atlantic oceans were formed by the rifting and drifting of the continents that now surround the Atlantic. These continents have been rifted as part of the separating lithospheric plates since the Triassic as the Atlantic Ocean slowly opened. The magnetic lineation pattern in the North Atlantic records this history of separation of Africa, Europe, and North America. When magnetic lineations of the same age from opposite sides of the ridge axis are fitted together, the relative positions of the continents at the time the lineations formed are obtained. This is equivalent to reversing the process of sea-floor spreading. In effect, the material younger than the two lineations is removed, and the older portions of the plates with the continents move toward each other until the two lineations fit together.

This technique is applicable to any ocean where separation at ridge axes has led to the passive rifting of the surrounding continents, such as the southeast and central Indian Ocean, the South and North Atlantic, the Labrador and Norwegian seas, and the Arctic Ocean.

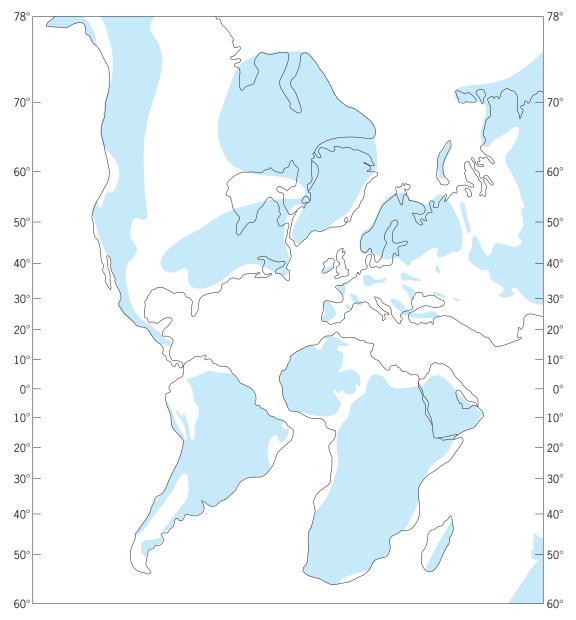


Fig. 7. Map illustrating the paleogeographic arrangement of the Atlantic continents during Coniacian time. Sea level may have been over 300 m (1000 ft) above present level at that time because of the volume of the mid-oceanic ridge system. Shaded areas are portions above sea level.

The above technique gives the position of the continental landmasses relative to each other, not their position relative to the rotational axis or paleomagnetic data. It is assumed that the Earth's magnetic field, now approximated by an axial geocentric dipole, has been similarly oriented throughout geologic time. Subaerial volcanic rocks acquire a remanent magnetism in the same way as oceanic crust, by cooling below the Curie point. Subaerial sedimentary rocks may also be remanently magnetized by the process of detrital remanent magnetization. As ironbearing particles settle from moving currents, their orientation is influenced by the prevailing magnetic field. By measuring the direction of the remanent magnetism of both igneous and sedimentary rocks whose ages can be determined, the position of the paleomagnetic pole for that age can be found. From

these data, the paleolatitude may be found for the set of continental landmasses whose positions relative to each other are known (**Fig. 7**). *See* PALEOGEOGRA-PHY; ROCK MAGNETISM.

Sea-level changes. Many important geologic phenomena may in part be related to plate tectonics processes, for example, changes in sea level and climate and changing patterns of evolution. The geologic record contains evidence of almost continuous eustatic sea-level changes throughout the Phanerozoic. Several large and rapid sea-level changes can be attributed to fluctuations in the volume of massive continental glaciers. However, it is apparent that large sea-level changes (greater than 300 ft or 100 m) also occurred during periods when continental glaciers were small or nonexistent.

Several factors affect the volume of the ocean

basins or the ocean waters and can thus alter eustatic sea level. One of the most important of these in terms of magnitude and rate of sea-level change is variation in the volume of the mid-oceanic ridge system. As explained above, once formed, the sea floor subsides systematically with time, with all oceanic crust following the same exponential subsidence curve. The volume of the mid-oceanic ridge system is quite large. If sea-floor spreading were to cease today, 70 m.y. from now the ridges would have subsided sufficiently to decrease the depth of the water over the abyssal area by about 1500 ft (450 m). The continental freeboard would increase by 1100 ft (320 m).

The volume of the mid-oceanic ridge system may be altered in several ways, one of which is to change the spreading rate. Because all ridges follow the same subsidence curve (which is a function of time only), age-versus-depth relationships are the same for all ridges. This means that if two ridges have been spreading at a constant but different rate for 70 m.y., the ratio of their volumes per unit length will equal the ratio of their spreading rates. If the spreading rate of the faster ridge is reduced to that of the slower ridge, volume/unit length of the larger ridge will gradually be reduced to that of the slower ridge. If spreading rates decrease, ridge volume also decreases, increasing the freeboard of the continents. The converse is also true: by increasing spreading rates, ridge volume will increase and the freeboard of the continents will be reduced. Several other ways of changing the volume of the mid-oceanic ridge system exist: ridges may be destroyed; segments of the ridge system may be subducted; and new ridges may be created by continental rifting and new rifting within ocean basins.

Both the length of the mid-oceanic ridge system and the spreading rate at its various segments have varied considerably during Phanerozoic time. These changes have caused large variations in sea level. For example, by using magnetic anomaly data to calculate spreading rates and ridge lengths back to the Upper Cretaceous, it has been estimated that the sea level then may have been as much as 900 ft (300 m) above present (**Fig. 8**).

Hot spots and mantle plumes. The existence of convective plumes originating in the deep mantle from below the level of the asthenosphere and rising to the bottom of the lithosphere has been proposed. About 14 major convective plumes are believed to exist today (Fig. 1). These plumes are believed to be nearly stationary with respect to each other, and hence they may be used as reference points with respect to which all the plates are moving. Mantle plumes, rather than convection currents, may also be the driving mechanism for plate motion. The latter two hypotheses are not necessary conditions for

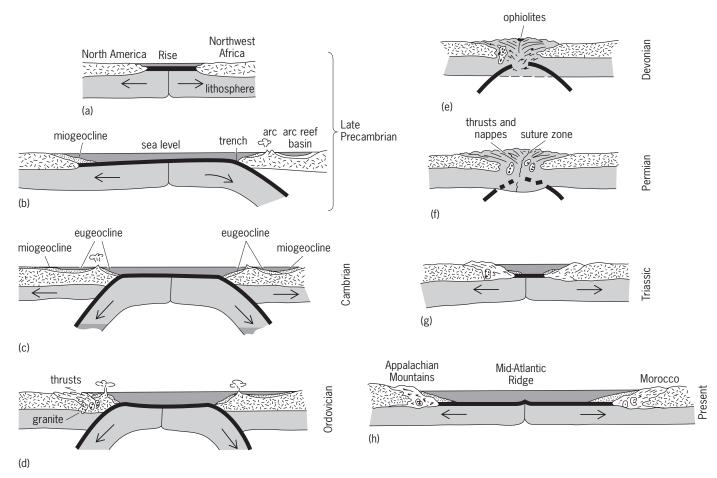


Fig. 8. Schematic representation of (a-h) the sequence of changes in the mid-oceanic ridge system in the plate tectonics history of the Appalachian orogenic belt. (After K. C. Condie, Plate Tectonics and Crustal Evolution, Pergamon, 1976)

the validity of the first. In other words, there may be mantle plumes that rise to the bottom of the lithosphere, but they may not be stationary with respect to each other and may not furnish the plate driving mechanism. The major clue to the existence of hot spots is the lines of intraplate volcanoes that are left as a trace of the passage.

As a plate passes over a hot spot, the hot spot burns its way through the plate. This releases volatile and eruptive magma on the plate surface. A classic example of the "volcanic" consequences of passage of a plate over a hot spot is the Hawaiian-Emperor Seamount chain. Hawaii and the islands immediate to it are volcanically active at present. The rest of the islands and seamounts of the Hawaiian and Emperor chains are inactive, but all are of volcanic origin. The islands and seamounts are found to be sequentially older westnorthwestward along the Hawaiian chains and are progressively older north-northwestward along the Emperor Seamount chain. As an explanation, it has been hypothesized that the Pacific Plate moved first north-northwesterly and then west-northwesterly over a single hot spot. Other seamount chains in the Pacific, such as the Line Islands, the Tuamotu Archipelago, and the Austral Seamount chain, are presumed to be of similar origin. See HOT SPOTS (GEOLOGY); MAGMA; OCEANIC ISLANDS.

In the Atlantic, features that are regarded as major hot-spot traces are Iceland and the Iceland Faroes Ridge, the New England Seamount chain, the Columbia Seamount chain, the Rio Grande Rise, and Walvis Ridge; and in the Indian Ocean, the Ninety-East Ridge is an analogous feature. The age progression along each of these features has been estimated; the theory that these hot spots have remained fixed in position relative to each other has been tested. The results indicate that some relative motion may occur, but it appears to be an order of magnitude less than the generally observed rate of interplate motion.

The hot spot hypothesis is a significant complement to plate tectonics. There is little doubt that hot spots or plumes of some type do occur. Intraplate volcanism often seems to be a manifestation of their existence. If, in addition, the hot spots prove to be essentially stable with respect to the rotational axis, they may be useful as a reference frame. They could serve as a latitudinal constraint in addition to paleomagnetic data. They could also be very useful in finding the relative motion between plates that have been separated by subduction zones or transform faults (that is, Pacific-North America relative motion).

Earth history. Not only does plate tectonics theory explain the present-day distribution of seismic and volcanic activity around the globe and physiographic features of the ocean basins such as trenches and mid-oceanic rises, but most Mesozoic and Cenozoic mountain belts appear to be related to the convergence of lithospheric plates. Two different varieties of modern mobile belts have been recognized: cordilleran type and collision type. The Cordilleran range, which forms the western rim of North and South America (the Rocky Mountains, Pacific Coast ranges, and the Andes), has for the most part been created by the underthrusting of an ocean lithospheric plate beneath a continental plate. Underthrusting along the Pacific margin of South America is causing the continued formation of the Andes. The Alpine-Himalayan belt, formed where the collision of continental blocks buckled intervening volcanic belts and sedimentary strata into tight folds and faults, is an analog of the present tectonic situation in the Mediterranean, where the collision of Africa and Europe has begun. *See* CORDILLERAN BELT; MOUNTAIN SYSTEMS; OROGENY.

Abundant evidence suggests that sea-floor spreading, continental drift, and plate tectonics have occurred for at least the past 600 m.y., during Phanerozoic time. Furthermore, it is probable that plate tectonics phenomena have dominated geologic processes for at least 2.5 billion years (b.y.). Presentday ocean basins are young (post-Paleozoic) features, and their origin is adequately explained by sea-floor spreading. The origin and evolution of the continental blocks, substantial portions of which are Precambrian in age, probably may be explained by accretion at subduction zones. Continental blocks, though largely composed of granite and granodiorite, appear to be mosaics of ancient mobile belts that progressively accreted laterally, vertically, or both, through time. Many late Precambrian and Paleozoic mobile belts, such as the Appalachian-Caledonian belt of the North Atlantic region, are interpretable in terms of plate tectonics mechanisms (Fig. 8). Initial rifting breaks a preexisting supercontinental landmass into two or more fragments that drift apart at the flanking margins of the growing ocean basin. At the margin of the growing basin, continental shelves, continental rises, and abyssal plains accumulate sediments. Cessation of sea-floor spreading and the development of new convergent margins along one or both sides of the ocean basin lead first to cordilleran-type orogeny and subsequently to continental collision. These orogenies suture or stitch together the original continental fragments and laterally buckle the intervening sediments and volcanic belts into folds and faults, initiate metamorphism, and generate volcanic and plutonic magmas. Belts may later be bisected by the development of new divergent plate margins that produce ocean basins whose axes may cut across the older mountain chains. This seems to be the case with the Appalachian-Caledonian belt, where the Mesozoic-Cenozoic development of the Atlantic Ocean has separated the chain into the Appalachians in North America and the Caledonides in Greenland, Great Britain, and Scandinavia. Most Paleozoic mobile belts, and many Precambrian belts at least 2.5 b.y. old, can be interpreted in an analogous fashion, for they contain the various petrogenetic associations of ocean opening and suture: ophiolites, blueschist terranes, mélanges, and a suite of sedimentary rocks whose composition, texture, and distribution suggest deposition in trenches and inland seas, on the abyssal plains of ocean basins,

and along continental shelves and rises. Because the Earth itself has evolved, chemical and petrologic aspects of these more ancient mobile belts often differ in detail from more modern analogs. Their organization leaves little doubt that they were formed by plate tectonics processes.

Plate tectonics is considered to have been operative as far back as 2.5 b.y. Prior to that interval, evidence suggests that the plate tectonics may have occurred, although in a markedly different manner, with higher rates of global heat flow producing smaller convective cells or more densely distributed mantle plumes which fragmented the Earth's surface into numerous small, rapidly moving plates. Repetitious collision of plates may have accreted continental blocks by welding primitive "greenstone island arc terranes" to small granitic subcontinental masses. *See* CONTINENTS, EVOLUTION OF; GEODY-NAMICS. Walter C. Pitman III

Details of Processes at Convergent Margins

Convergent continental and island arc margins encircle the Pacific Ocean and occur locally in all other oceans (**Fig. 9**). At convergent margins, two of the Earth's lithospheric plates are directed against each other. Normally, one plate consists of dense oceanic lithosphere which underthrusts, or subducts, beneath less dense continental or island arc lithosphere (**Fig. 10**).

Structure and tectonic units. The thrust fault between the plates is the major tectonic feature of a subduction zone, and opposing plate motion causes great earthquakes that begin at 10-20 km (6-12 mi) depths in the seismogenic zone. As the plates shear past one another, sediment and rock of the subducting plate may be added to the upper plate and upper-plate material may be eroded, introducing terrigeneous material into the subduction zone. The dominant material transfer is a basis for dividing convergent margins into accretionary and erosive types. In the absence of a standard nomenclature, the following is used here. The upper plate of convergent margins is commonly formed by four units, consisting of a frontal prism that is sometimes backed against an accreted prism, which in turn backs against the core rock framework and is overlain by a slope apron. The lower plate consists of igneous rock overlain by oceanic and trench fill sediment. Sediment that stays attached to the lower plate when introduced into the subduction zone forms the subduction channel that is capped by the interplate thrust. See MARINE GEOLOGY.

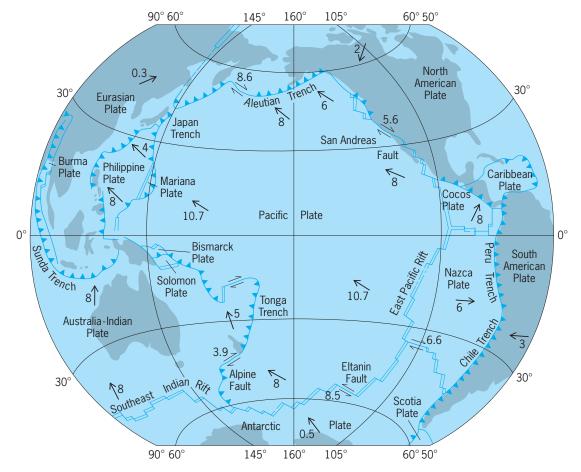


Fig. 9. Plate tectonic map of the circum-Pacific region showing major zones of plate convergence and spreading ridges. Arrows and numbers indicate the direction and the rate (cm/yr; 1 cm = 0.4 in.) of plate movement. The thrusts of the zones of convergence are marked with teeth on the upper plate. (*After G. W. Moore, Plate-Tectonic Map of the Circum-Pacific Region: Explanatory Notes, American Association of Petroleum Geolgists, 1982*)

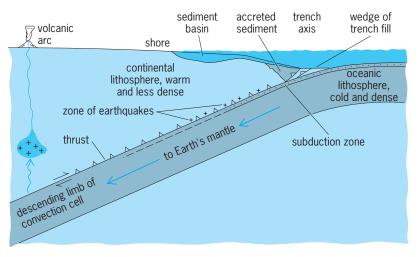


Fig. 10. Cross section through a typical convergent margin. The relatively dense oceanic lithosphere sinks beneath the relatively light continental lithosphere and forms the descending limb of a convection cell in the Earth's mantle.

The frontal prism begins at the initial contact between converging plates as represented by the first fault that displaces the trench axis sea floor, referred to as the deformation front. Rapid deformation characterizes the frontal prism. The prism extends to either a more rigid core framework or a well-consolidated accreted prism.

Compressional deformation tectonically thickens the prism in contrast to less prominent active tectonism landward. The frontal prism consists of material derived from off-scraped trench fill, slope apron sediment, or fragmented basement rock that was detached at the basement-prism contact. Along more than 50% of all convergent margins where trench fill is no thicker than 1 km (0.6 mi), frontal prisms are less than 20 km (12 mi) wide. Once they reach 25-30 km (15-20 mi) in width, material at the back of the prism has consolidated, and it becomes sufficiently rigid to become part of the accreted prism. Thus, the contact between frontal prism and accreted wedge migrates toward the trench axis. Frontal prisms widen continents only a small amount in contrast to accreted prisms.

The accreted prism, 30 to more than 100 km (20 to more than 60 mi) wide, is a body of oceanic and trench fill sediment that was transferred from the subducting lower plate at the frontal prism and became attached to the upper plate framework. Accreted prisms develop from the frontal prism and consist of consolidated and rigid sedimentary rock with porosities of about 15% or less. The accreted prism adds to the width of a continent or island arc and stores material for longer periods. Large accretionary prisms involve a margin configuration and plate convergence constant for tens of millions of years. They are common where sediment fill in the trench axis is more than 1 km (0.6 mi) thick and rates of orthogonal convergence are less than 5 km/m.y. Examples are found along the Lesser Antilles and the Nankai, Makran, and Indonesian margins.

The core framework is commonly referred to as

basement, and it includes diverse igneous as well as metamorphic and occasional sedimentary rock bodies. The framework basement exposed along coasts is a strong and rigid rock mass that can be followed tens of kilometers seaward with geophysical methods and sampling. Along more than half of the global length [about 41,000 km (25,000 mi)] of convergent margins, the framework extends to within 10–15 km (6–9 mi) of the trench axis. Many of the sedimentary rock bodies of the framework basement are fossil accreted wedges, as along the Japan and Alaskan margins.

The slope apron is a sediment sequence that overlies the margin framework and accreted prism. It extends from the edge of the shelf to the frontal prism and can be from a few meters to 5 km (3 mi) thick. The steepness of the continental slope causes gravity sliding and a general downslope migration of sediment. Seismic imaging of slope-apron stratification and high-resolution sea-floor bathymetry [100 m (330 ft) or less], show the lack of active deformation that indicates the tectonic stability of the margin framework and accreted prism relative to the frontal prism. Stability is also indicated by a sea-floor morphology with integrated canyons that are disrupted and disappear over the frontal prism.

The plate interface is underlain by a layer from about 200-1200 m (about 660-4000 ft) in thickness called the subduction channel. This layer consists of subducted trench sediment and eroded clastic material on the igneous ocean crust and is commonly detected as a zone of low seismic velocity, relative to its bounding rock masses. The low velocity is interpreted as indicating a fluid-rich unit. Subduction beneath the frontal prism inflicts a load on the porous sediment in the subduction channel, which elevates the pore fluid pressure because the prism's permeability does not allow the fluid to escape as quickly as the load is applied. Fluid pressure can approach the weight of the overlying material. Such pressures greatly reduce friction and allow less consolidated materials to be subducted rather than scraped off at the deformation front. Farther down, the subduction zone temperature increases to levels where some clay minerals alter and release chemically bound water into the subduction channel. Elevated pressure in a fluid-rich layer is an explanation for reduced seismicity, and the manner of slip along the plate interface thrust is considered stable. The unstable slip released during large earthquakes and tsunamis occurs at ~ 10 km (6 mi) or greater depths in the portion of the subduction zone called the seismogenic zone. See SEISMOLOGY; TSUNAMI.

The subducting lower plate commonly bends into the trench axis and breaks into extensional or normal faults. This faulting produces voids that fill with sediment carried down the subduction zone on the lower plate. Some normal faults are imaged seismically through the entire oceanic crust, and could form aquifers that conduct seawater into the lower crust. Seawater in contact with lower crustal rock can change the mineralogy to weaker serpentine minerals and affect the mechanical behavior and geochemical evolution of the subducting lithosphere. The cold, negatively buoyant subducting plate becomes a driving force for convection in the Earth's mantle, which is the layer between the lithosphere and the core. Ultimately, subduction of oceanic plates into the mantle is responsible for much of the tectonism expressed at the Earth's surface.

Tectonic processes. The dynamic character of convergent margins is evident from earthquakes and tsunamis. Convergent margins begin in the seismogenic zone at depths beyond the reach of direct observation even with scientific drilling. However, the mechanisms involved leave a geology to interpret that indicates the nature of dynamic processes involved. Accretionary mass addition has an erosional mass subtraction counterpart. The geology of material added to the upper plate is more easily observed than the record of massive material removed from a continent or island arc and carried out of sight. An important but difficult record to decipher from geologic evidence is the role of fluid in controlling tectonism and seismicity.

Erosional convergent margins. Subduction erosion is predominant along more than 60% of convergent margins globally. At erosional convergent margins, the net volume of the margin's upper plate decreases during subduction. Evidence of missing rock masses includes finding (1) the roots of ancient volcanic arcs along the coast or more than 100 km (60 mi) closer to the trench axis than where they formed; (2) largescale and long-term subsidence to trench depths of unconformities that formed in shallow-water environments; and (3) the migration of volcanic arcs hundreds of kilometers inland, indicating that the continental slope retreated a similar distance. Drill cores at seven Pacific convergent margins show 3-5 km (2-3 mi) of subsidence during the past 15-20 m.y. This subsidence requires upper-plate thinning from erosion along the plate interface because sediment deposition during subsidence rules out regional seafloor erosion.

Erosional margins characteristically occur where plate convergence is relatively rapid (more than 4 km/m.y.) and sediment in the trench axis is less than 1 km (0.6 mi) thick. The lower plate of igneous oceanic crust and sediment bends downward into the trench axis, causing extensional structure diagrammed as tilted blocks or half-grabens (depressed, faulted blocks). The wedge-shaped upper plate consists of a continental crustal framework covered by a slope sediment apron. The margin framework is as strong at the coast, but downslope it fractures and develops extensional faults that break up the plate beneath the middle slope. Disruption of the framework structure results in mass wasting of the slope apron and basement. Debris from mass wasting migrates downslope and collects in a frontal prism beneath the lower slope. This thin weak apex of the frontal prism displays ridges and valleys at the sea floor that mimic the buried lower plate relief. A thrust fault bounds the underside of the frontal prism, but landward the plate interface becomes a subduction channel. In seismic images, the channel has a lowfrequency-high-amplitude seismic reflectance that is 0.5-1.0 km (0.3-0.6 mi) wide, inferred to indicate a fluid-rich zone. Subduction beneath the frontal prism elevates fluid pressure in the lower plate and its sediment cover, and this pressure increases upward fluid migration along and across the plate interface. As fluid drains from the lower plate, its porosity is reduced, which consolidates and strengthens material below the interplate thrust, and at the same time fluid invading fractures in the upper plate weakens rock above the plate interface. Thus, the zone of minimum strength along the plate interface shifts upward. The shift results in a transfer of upper-plate material to the subducting lower plate, thereby thinning the plate. See GRABEN.

In the subduction erosional model, the upperplate framework looses strength downslope as the plate thins and fractures dilate from invasion by pressured fluid. As the lower plate subducts, it continues bending, which increases lower-plate relief. Relief promotes fracturing in the upper plate from jostling as lower-plate relief subducts beneath it. As hundreds of kilometers of lower plate subducts, the jostling and dilation of fractures by pressurized fluid ultimately fragments the base of the upper plate. Detached fragments are dragged piece by piece onto the underthrusting lower plate, which carries them into the seismogenic zone.

Accretionary margins. Accretionary margins preferentially occur where convergence is less than 40 km/m.y. (25 mi/m.y.) and where sediment in the trench axis is thicker than 1 km (0.6 mi). They are distinguished from erosional margins by the accreted prism that develops from the frontal prism. The sediment wedge filling the trench axis is deposited rapidly and has a high initial porosity (about 70%). As the wedge converges against the upper plate, horizontal force creates small compressional faults, which do not extend to the sea floor, in the lower part of the trench-sediment wedge, and pore fluid pressure increases. Before faults break the sea floor, fluid drains from the trench-sediment wedge through intergranular permeability. Closer to the deformation, front faults reaching the sea floor provide more efficient dewatering channels. The deformation front develops when upper sediment layers consolidate sufficiently to break in thrust faults. Faults near the deformation front may extend through most of the wedge sediment section, but a horizontal fault above the igneous basement and parallel to bedding (a decollement) that forms the interplate thrust fault intercepts them (Fig. 11). Sediment below the thrust fault subducts beneath the upper plate as part of the lower plate. Where it has been drilled, subducted sediment is little deformed and indicates low interplate friction, along with other observations. Once faulting landward of the deformation front begins, fluid drains rapidly through fracture permeability in the faults. The majority of original pore fluid in accreted sediment (80%) is expelled in approximately the first 5-10 km (3-6 mi) of the accretionary frontal prism. Accretionary frontal prisms thicken rapidly by

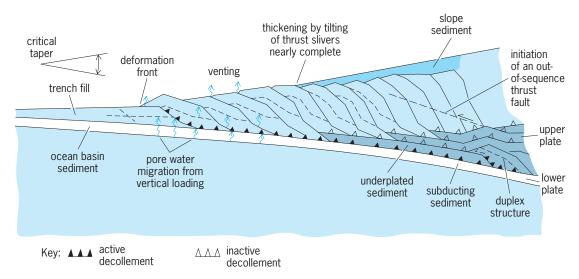


Fig. 11. Section across an accretionary convergent margin. The deformation front marks the partitioning of subducted and accreted sediment, a division that commonly occurs along the base of the trench fill. Rapid loading squeezes water from voids in the sediment, ultimately draining with the water from vents at the sea floor.

thrust faulting, and it is here that the greatest permanent strain deforms sedimentary bodies. In the subducted sediment and igneous crust, fluid is released farther down the subduction zone.

In many accretionary margins, the latest seismic and high-resolution bathymetric data show segmentation of tectonic activity not previously noted. A change in tectonic character separates the accretionary frontal prism from the accreted prism. The rapidly deforming frontal prism has a rough sea-floor morphology. The accreted prism deforms minimally, allowing deposition of a little deformed slope apron across it. The accreted prism is stronger and more rigid than the frontal prism, and the contact between them is a fault zone that commonly has sea-floor expression. This segmentation is not recognized along all accretionary margins, and it is difficult to resolve without the latest high-resolution geophysical data. Along the Nankai margin off Japan, the boundary between frontal prism and accreted prism is termed a splay fault, and similar features are resolved along the Sunda and Makran margins.

The concept of critical taper helps explain the thickening of an accretionary wedge. Taper refers to the angle between the upper and lower prism surfaces, and it depends on the strength of the material and the interplate friction. When the critical taper is exceeded, the slope becomes overcritical and fails by slumping to restore the optimum taper. When taper is undercritical, accretion restores it to the optimum taper for conditions at the time. To accomplish tectonic thickening, the accreted thrust packets are stacked and rotated, but only to a limiting angle. Once rotation and stacking of the detached packets can no longer thicken the wedge, thickening can be accomplished by the addition of material from below (underplating), by subsequent thrust faulting at a lower angle across previous structure (out-of-sequence thrusting), or by thrusting that is contained within a lower layer (duplexing). These

processes occur at depths where clear seismic imaging of structures is at the limits of the current technology. They are currently modeled to explain taper related to upper plate thickening landward of the frontal prism.

The mass flux at convergent margins is important for understanding arc volcanism and the origin of continental crust. Subduction zones return crust to the mantle from which is once originated, but in an altered state. The past tendency to view subduction zones as primarily areas of growth has diminished as the extent of subduction zone erosion becomes known. The volume of continental crust returned to the mantle is small compared to its entire volume, but sufficient to impart the distinct continental geochemistry observed in mantle rock. Erosion is one argument for the paucity of very old exposed continental crust, functioning as a balancing process to have kept the growth of continents at their approximate size during much of the Earth's history (since pre-Cambrian time, about 2.5 billion years ago). Roland von Huene

Bibliography. A. W. Bally (ed.), Seismic Expression of Structural Styles, Amer. Ass. Petrol. Geol. Stud. Geol. 15, 1983; N. L. Bangs et al., Evolution of the Nankai Trough decollement from the trench into the seismogenic zone: Inferences from threedimensional seismic reflection imaging, Geology, 32(4):273-276, 2004; K. C. Condie, Plate Tectonics and Crustal Evolution, 4th ed., 1997; Continents Adrift and Continents Aground: Readings from Scientific American, introduction by T. J. Wilson, 1976; R. von Huene, C. R. Ranero, and P. Vannucchi, Generic model of subduction erosion, Geology, 32(10):913-916, 2004; H. Kopp and N. Kukowski, Backstop geometry and accretionary mechanics of the Sunda margin, Tectonics, 22(6):1072, 2003; J. C. Moore et al., Tectonics and hydrogeology of the northern Barbados Ridge: Results from Ocean Drilling Program Leg 110, Geol.

Soc. Amer. Bull., 100(10):1578–1593, 1988; M. J. Selby, *Earth's Changing Surface*, 1985; C. Seyfert (ed.), *Encyclopedia of Structural Geology and Plate Tectonics*, 1987; B. F. Windley, *The Evolving Continents*, 3d ed., 1995.

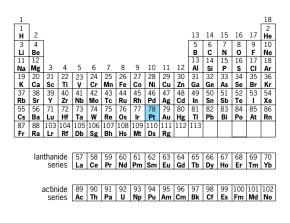
Plateau

Any elevated area of relatively smooth land. Usually the term is used more specifically to denote an upland of subdued relief that on at least one side drops off abruptly to adjacent lower lands. In most instances the upland is cut by deep but widely separated valleys or canyons. Small plateaus that stand above their surroundings on all sides are often called tables, tablelands, or mesas. The abrupt edge of a plateau is an escarpment or, especially in the western United States, a rim. In the study of landform development the word plateau is commonly used to refer to any elevated area, especially one underlain by nearly horizontal rock strata, that once had a smooth surface at high level, even though that surface may since have been largely destroyed by valley cutting. An example is the now-hilly Appalachian Plateau of western Pennsylvania, West Virginia, and eastern Kentucky. See ESCARPMENT; PLAINS.

Among the extensive plateau lands of the world are the Colorado and Columbia plateaus of the western United States, the plateau of southeastern Brazil, the Patagonian Plateau of southern South America, the Central Siberian Plateau, and the Deccan Plateau of peninsular India. Edwin H. Hammond

Platinum

A chemical element, Pt, atomic number 78, and atomic weight 195.09. Platinum is a soft, ductile, white noble metal. The platinum-group metals— platinum, palladium, iridium, rhodium, osmium, and ruthenium—are found widely distributed over the Earth. Their extreme dilution, however, precludes their recovery, except in special circumstances. For example, small amounts of the platinum metals, palladium in particular, are recovered during the electrolytic refining of copper. *See* IRIDIUM; OSMIUM; PALLADIUM; PERIODIC TABLE; RHODIUM; RUTHENIUM.



Properties	Value
Atomic weight (${}^{12}C = 12.00000$)	195.09
Naturally occurring isotopes and % abundance	190, 0.0127%
	192, 0.78%
	194, 32.9%
	195, 33.8%
	196, 25.3%
	198, 7.21%
Crystal structure	Face-centered cubic
Lattice constant <i>a</i> at 25°C, nm	0.39231
Thermal neutron capture cross	8.8
section, barns	
Common chemical valence	2, 4
Density at 25°C, g/cm ³	21.46
Melting point	1772°C (3222°F)
Boiling point	3800°C (6900°F)
Specific heat at 0°C, cal/g Thermal conductivity, 0–100°C,	0.0314 0.17
cal cm/cm ² s°C	0.17
Linear coefficient of thermal	9.1
expansion, 20–100°C, μin./in./°C	5.1
Electrical resistivity at 0° C,	9.85
microhm-cm	0.00
Temperature coefficient of electrical	0.003927
resistance, 0–100°C/°C	
Tensile strength, 1000 lb/in. ²	
Soft	18–24
Hard	30–35
Young's modulus at 20°C	
Ib/in. ² , static	24.8×10^{6}
lb/in.², dynamic	24.5×10^{6}
Hardness, Diamond Pyramid Number	
(DPN)	
Soft	37-42
Hard	90–95

The platinum-group metals have wide chemical use because of their catalytic activity and chemical inertness. As a catalyst, platinum is used in hydrogenation, dehydrogenation, isomerization, cyclization, dehydration, dehalogenation, and oxidation reactions. *See* CATALYSIS; ELECTROCHEMICAL PROCESS.

Platinum is not affected by atmospheric exposure, even in sulfur-bearing industrial atmospheres. Platinum remains bright and does not visually exhibit an oxide film when heated, although a thin, adherent film forms below 450° C (840° F). Platinum may be worked to fine wire and thin sheet, and by special processes, to extremely fine wire. Important physical properties are given in the **table**.

Platinum can be made into a spongy form by thermally decomposing ammonium chloroplatinate or by reducing it from an aqueous solution. In this form it exhibits a high absorptive power for gases, especially oxygen, hydrogen, and carbon monoxide. The high catalytic activity of platinum is related directly to this property. *See* CRACKING; HYDROGENATION.

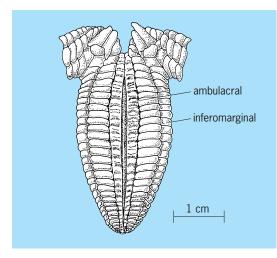
Platinum strongly tends to form coordination compounds. Platinum dioxide, PtO₂, is a dark-brown insoluble compound, commonly known as Adams catalyst. Platinum(II) chloride, PtCl₂, is an olive-green water-insoluble solid. Chloroplatinic acid, H₂PtCl₆, is the most important platinum compound. *See* CO-ORDINATION CHEMISTRY.

In the glass industry, platinum is used at high temperatures to contain, stir, and convey molten glass. In the electrical industry, platinum is used in contacts and resistance wires because of its low contact resistance and high reliability in contaminated atmospheres. Platinum is clad over tungsten for use in electron tube grid wires. In the medical field, the simple coordination compounds cisplatin and carboplatin are two of the most active clinical anticancer agents. In combination with other agents, cisplatin is potentially curative for all stages of testicular cancer. Both agents are used for advanced gynecologic malignancies, especially ovarian tumors, and for head and neck and lung cancers. Carboplatin was developed in attempts to alleviate the severe toxic side effects of the parent cisplatin, with which is shares a very similar spectrum of anticancer efficacy. See CANCER (MEDICINE); CHEMOTHERAPY.

Henry J. Albert; Nicholas Farrel Bibliography. L. Cerveny, *Catalytic Hydrogenation*, 1986; J. R. Davis (ed.), *Metals Handbook: Desk Edition*, 2d ed., ASM International, 1998; F. R. Hartley (ed.), *Chemistry of the Platinum Group Metals: Recent Developments*, 1991; L. R. Kelland and N. Farrell (eds.), *Platinum-Based Drugs in Cancer Ther apy*, 2000.

Platyasterida

A Paleozoic (Ordovician and Devonian) order of sea stars of the class Asteroidea; only two genera have been recognized. Platyasteridans have relatively elongate flat arms; broad, transversely aligned ventral arm plates; and paxilliform aboral plates (see **illus.)**. The morphology suggests living members of the Luidiidae, and based on those similarities, some authors contend that luidiids are surviving platyasteridans. Fundamental differences in ambulacral system construction have been found between modern asteroids and all of the well-known early Paleozoic



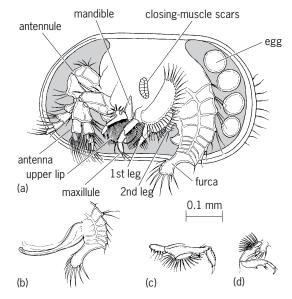
Ventral view of an arm and a portion of the disk of Platanaster ordovicus. Note lateral alignment of the arm plates. (After W. K. Spencer and C. W. Wright, Asterozoans, in R. C. Moore, Treatise on Invertebrate Paleontology, Pt. U, Echinodermata 3, Geological Society of America and University of Kansas Press, 1966)

asteroids, including the platyasteridans. Similarities between platyasteridans and modern asteroids are atributable to evolutionary convergence. *See* ASTER-OIDEA; ECHINODERMATA; PAXILLOSIDA.

Daniel B. Blake Bibliography. D. B. Blake, A classification and phylogeny of post-Paleozoic sea stars (Asteroidea: Echinodermata), *J. Nat. Hist.*, 21:481–528, 1987.

Platycopida

An order of the Podocopa (Ostracoda) containing a single family, Cytherellidae, and two extinct superorders. Platycopids are small marine or brackishwater ostracods with an asymmetrical shell that is



Cytherella abyssorum, an ostracod of the order Platycopida. (a) Female with left valve removed; eggs lie in posterior brood chamber outside body. (b) Posterior end of male body with paired furcae and long hemipenes. (c, d) Male first and second legs, both developed as claspers and very different from the corresponding appendages in the female. (After R. C. Moore, ed., Treatise on Invertebrate Paleontology, pt. Q, 1961, and D. L. McGregor and R. V. Kesling, Contributions from the Museum of Paleontology, University of Michigan, 1969)

oblong, nearly rectangular, and laterally compressed; their six pairs of appendages are adapted for burrowing and filter feeding. Both fossil and living species have the inner lamella of each valve calcified only in the distal rim, if at all; if a narrow duplicature is developed, it is continuous with the outer lamella and does not form a vestibule (troughlike cavity) between the lamellae. Recent platycopids are distinct from all other ostracods: The larger valve is rabbeted to overlap the smaller around the free edge; the carapace lacks vertical lobation; none of the postoral appendages serve as legs; and the posterior end of the body is more clearly delimited and tapering than in other ostracods.

Species are dimorphic. The female carapace is wider near the rear and longer than that of the

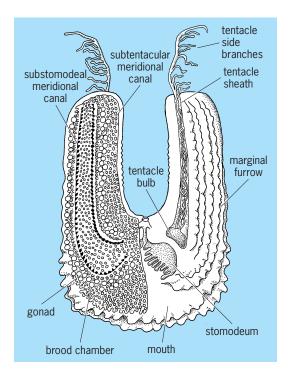
male, providing a brood chamber behind the body. Scars of adductor muscles are arranged in a biserial aggregate (see **illus**.). The biramous antenna has both endopod and exopod developed for locomotion, and the paired lamelliform furcae may assist. All living species are benthic, and fossil forms probably were also. Some species live at great depths in the Atlantic Ocean, but others have been found in the Arabian, Caribbean, and Mediterranean seas in shallower water.

Unquestioned Platycopida are known from Jurassic to Recent. It has been proposed that ancestors of the platycopids were nondimorphic Paleocopida. The Cytherellidea combine a primitive, albeit elaborate, soft-body anatomy with specialized filter feeding. Neoteny may have played a significant role in the evolution of platycopids. *See* CRUSTACEA; OSTRA-CODA; PODOCOPA. Patsy A. McLaughlin

Bibliography. L. G. Abele (ed.), *The Biology of Crustacea*, vol. 1: *Systematics, the Fossil Record, and Biogeography*, 1982; H. Kozur, Einige Bermerkungen zur Systematik der Ostracoden und Beschreibung neuer Platycopida aus der Trias Ungarns und der Slowakei, *Geol. Palaeontol. Mitt. Innsbruk*, vol. 2, 1972; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982; V. Pokorny, *Principles of Zoological Micropaleontology*, 2:219–237, 1965.

Platyctenida

An order of the phylum Ctenophora comprising four families (Ctenoplanidae, Coeloplanidae, Tjalfiellidae, Savangiidae) and six genera. All species are highly modified from the planktonic ctenophores.



Comb jelly, Lyrocteis imperatoris.

The platyctenes are fairly small (1-6 cm or 0.4-2.4 in.) and brightly colored. They have adopted a variety of swimming, creeping, and sessile habits, with concomitant morphological changes and loss of typical ctenophoran characteristics. The body is compressed in the oral-aboral axis, and the oral part of the stomodeum is everted to form a creeping sole (see illus.). Platyctenids have a fairly typical cydippid larva, and comb rows are retained in adult ctenoplanids, which are capable of swimming or creeping on a substrate. The remaining families all have a benthic habit and have lost the comb rows, and even the statocysts as adults in the Savangiidae. Sexual reproduction involves internal fertilization in many species, with retention of the developing cydippid larvae in brood pouches. Some species also reproduce asexually by fission. Most platyctenids are found in tropical coastal waters, where many are ectocommensals on benthic organ-Laurence P. Madin isms. See CTENOPHORA.

Platyhelminthes

A phylum of the invertebrates, commonly called the flatworms. They are bilaterally symmetrical, nonsegmented worms characterized by lack of coelom, anus, circulatory and respiratory systems, and exoand endoskeletons. Many species are dorsoventrally flattened. They possess a protonephridial (osmoregulatory-excretory) system, a complicated hermaphroditic reproductive system, and a solid mesenchyme which fills the interior of the body. Some parasitic species, that is, some trematodes, have secondarily acquired a lymphatic system resembling a true circulatory system. Some species of trematodes, the schistosomes, have separate sexes.

Traditionally, three classes were distinguished in the phylum: the Turbellaria, mainly free-living predacious worms; the Trematoda, or flukes, ecto- or endoparasites; and the Cestoda, or tapeworms, endoparasites found in the enteron (alimentary canal) of vertebrates, whose larvae are found in the tissues of invertebrates or vertebrates. However, recent cladistic analyses using morphology, including ultrastructure, as well as DNA analysis, have shown that the "Turbellaria" are an assemblage of taxa that are not monophyletic (that is, they are not a group containing all taxa with a common ancestor), and that the monogeneans, earlier included in the trematodes, do not belong to the trematodes. Most importantly, the Acoela do not belong to the Platyhelminthes, but are a very archaic group close to the base of the lower invertebrates; and all major groups of parasitic Platyhelminthes-that is, the Trematoda, Monogenea, and Cestoda-are monophyletic, constituting the Neodermata.

Morphology. Flatworm tissues and organs are derived from three germ layers. Most flatworms have adhesive or attachment devices. In turbellarians, these devices are glandular or muscular; in monogeneans, trematodes, and cestodes, they are suckers and/or hooks or spines. Muscular tissue occurs in mesenchymal layers and permits rapid change in body form. Numerous receptor cells (mechanoreceptors and chemoreceptors) and sense organs (tentacles, statocysts for balance, and ocelli for photore-

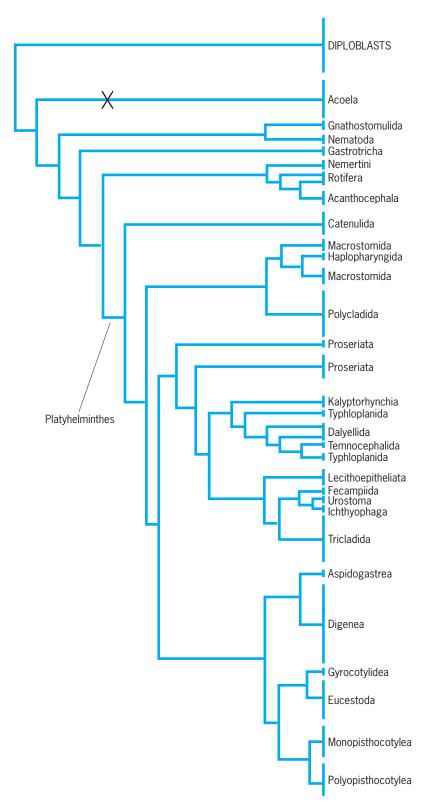


Fig. 1. Phylogeny of Platyhelminthes based on molecular (DNA) and morphological evidence. Note that the Accela do not belong to the Platyhelminthes (X). (*Strongly modified from D. T. J. Littlewood, K. Rohde, and K. A. Clough, Biol. J. Linnean Soc.*, 66:75–114, 1999)

ception) occur on the head and body. Some of these are reduced or absent in parasitic forms, although there is no general reduction in the complexity of the nervous system and sense receptors in the parasitic forms. Some parasitic species, in particular the Aspidogastrea, even have a nervous system of greater complexity and a greater variety of sensory receptors than free-living forms. The nervous system consists of longitudinal nerve cords (connectives) connected by transverse nerves (commissures). In some groups the nervous system has an irregular netlike appearance. Usually some of the anterior commissures are particularly well developed, forming a kind of brain (cephalization). *See* NERVOUS SYSTEM (INVER-TEBRATE).

A digestive system is present in turbellarians, monogeneans, and trematodes but is lacking in cestodes. A mouth, serving also for egestion, is situated either anteriorly or ventrally. It is often provided with an aspirating sucker or pharynx by which host tissues or other food particles are ingested. The gut is saclike or branched and lined with a single layer of cells often packed with granules. The protonephridial system is of uniform construction in the phylum. It consists of ramifying blind tubes capped with large cells, called flame cells, each bearing a tuft of cilia which projects into the lumen. The tubes course through the mesenchyme and discharge at the surface by means of one or more openings. The function of this system is not well understood but contributes to osmoregulation and excretion. See DI-GESTION (INVERTEBRATE).

The reproductive system reaches very great morphological complexity. In each reproductive unit which is the entire organism in the Turbellaria, Monogenea, Trematoda, Amphilinidea, and Gyrocotylidea, and which is the proglottid (segment) in the Eucestoda—one to several testes and ovaries occur as well as various accessory reproductive organs.

Reproduction. The Platyhelminthes reproduce both sexually and asexually. In sexual reproduction, fertilization is internal, following copulation or hypodermic insemination. Cross-fertilization is the rule but self-fertilization may occur. In most flatworms, the eggs are ectolecithal; that is, the ova are invested with yolk cells and the egg mass is enclosed in a capsule. In turbellarians, capsules are often deposited in cocoons, in which the juveniles develop. Eggs of monogeneans, trematodes, and cestodes emerge from the hosts. In the monogeneans, larvae infect the final vertebrate hosts. In the other two, they either are eaten or hatch in water as free-swimming larvae; then they either actively penetrate, or are passively ingested by, intermediate hosts. Alternatively, many species among the trematodes encyst on a substratum outside a host.

Asexual reproduction is a frequent occurrence in this phylum. Many turbellarians reproduce by fragmentation or binary fission; in some, chains of individuals are temporarily produced. The Digenea reproduce asexually in their larval stages, as do some tapeworms, such as *Echinococcus* and *Multiceps*. Formation of proglottids is an asexual process, resulting from activity of a proliferating zone in a neck region. *See* REPRODUCTION (ANIMAL).

Economic and biological importance. Turbellaria are widespread in freshwater and the littoral zones (the biogeographic zones between the high- and lowwater marks) of the sea, while one group of triclads occurs on land in moist habitats. In the meiofauna (small benthic animals) along the coasts, huge numbers of small turbellarians of many species occur, but only a very small proportion of species has been described. Turbellaria have been used to study regeneration, including the effects of chemicals and radiation upon the process. They have also been used in axial gradient research. Adults of parasitic flatworms occur on or in many tissues and cavities of the vertebrates on which they feed. They are responsible for troublesome diseases in humans and animals. Monogeneans are of foremost importance as agents of fish disease, especially in aquaculture, whereas digeneans are important agents of disease in domestic animals and humans (schistosomiasis). Larval flukes are frequent in mollusks, mainly gastropods, and occasionally occur in pelecypods (bivalves). Vector hosts, such as insects and fish, are often interpolated between the mollusk and vertebrate. Adult tapeworms, living in the gut or the biliary ducts, compete with the host for food and accessory food factors such as vitamins. Larval tapeworms reside chiefly in arthropods, but larvae of the Cyclophyllidea develop in mammals, which may be severely impaired or even killed by the infection. Investigations utilizing flukes and tapeworms have given clearer insight into the host-parasite relationship. See PARASITOLOGY.

Classification. Figures 1 and **2** show the interrelationships of the major groups of Platyhelminthes according to cladistic analyses based on DNA analysis and morphology (gross morphology and ultra-

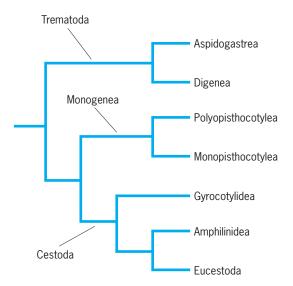


Fig. 2. Phylogeny of the Neodermata, the major groups of parasitic Platyhelminthes. Note that the Aspidogastrea plus Digenea, the Polyopisthocotylea plus Monopisthocotylea, and the Gyrocotylidea plus Amphilinidea plus Eucestoda form monophyletic groups, that is, the Trematoda, Monogenea, and Cestoda.

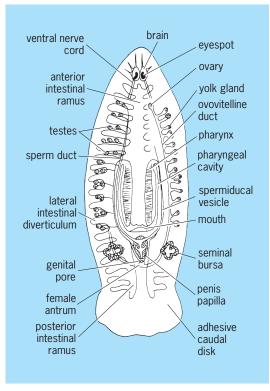


Fig. 3. *Bdelloura candida* (Tricladida), ectocommensal on the king crab, *Limulus*. Complete digestive and male systems are shown on the left, female systems on the right.

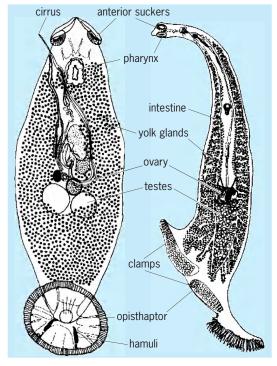


Fig. 4. A monopisthocotylean monogenean on the left and a polyopisthocotylean monogenean on the right. (*Based on K. Rohde, Ecology of Marine Parasites, 2d ed., CABI, Wallingford, Oxon, U.K., 1993*)

structure). Note that the Acoela do not belong to the Platyhelminthes and that all major groups of parasitic flatworms, the Neodermata, form a monophylum including all groups derived from one common ancestor. The Catenulida is the sister group of all other Platyhelminthes, and there are many taxa of equal taxonomic rank.

"Turbellaria." This group is an assemblage of taxa which are not monophyletic. The body is cylindrical or flattened and may be less than 1 mm to over 50 cm long (0.04 to 20 in.). The members are chiefly free-living, but commensal and parasitic species occur. A ciliated epidermis may uniformly cover the body or may be restricted to the ventral or other surfaces. Peculiar rodlike rhabdoids are embedded in the epidermis. Characteristic color patterns are imparted to some species by chromatophores, while other species contain symbiotic algae and are green or brown. Species of Temnocephalida occur as ecto-

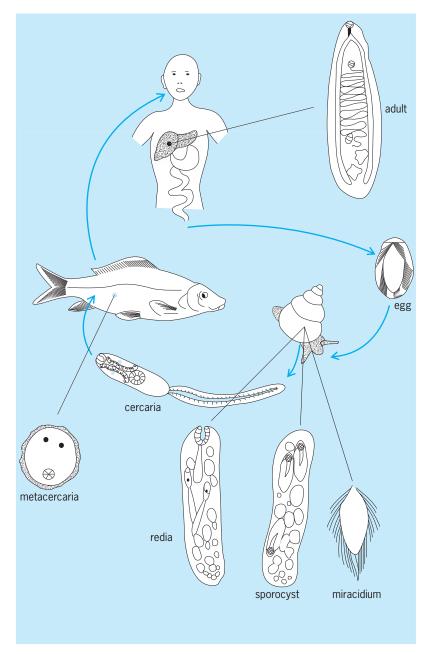


Fig. 5. Life cycle of a liver fluke. (Based on K. Rohde, Ecology of Marine Parasites, 2d ed., CABI, Wallingford, Oxon, U.K., 1993)

commensals (organisms living on the outer surface of the body of another organism, without affecting the host) mainly on freshwater crayfish. Tricladida consists of larger forms which are marine, freshwater, or terrestrial and have a three-branched intestine (**Fig. 3**). Polycladida contains marine forms of elongate or leaflike shape with few to many eyes near the tentacles and on the body margin. Many intestinal rami branch from a central digestive cavity above the pharynx. Many of the other taxa, in earlier taxonomic systems often grouped in the Rhabdocoela (taxa with a simple saclike gut), are either marine or freshwater forms. The Fecampiida and *Ichtbyophaga* are entirely endoparasitic. *See* POLYCLADIDA; TRICLA-DIDA; TURBELLARIA.

Neodermata. This large taxon consists of parasites covered with a surface syncytium (tegument) formed by underlying cells. It includes the Monogenea, Trematoda, and Cestoda.

Monogenea. The monogeneans are found mainly as ectoparasites on cold-blooded vertebrates, with direct and simple life cycles. They consist of two groups, the Polyopisthocotylea and Monopisthocotylea. The former have a posterior attachment organ (opisthaptor) with clamplike attachment sclerites and hamuli (hooks), whereas the latter have a posterior attachment organ consisting of a large sucker and/or hamuli (hooks) [**Fig. 4**]. Adult monogeneans produce eggs from which typically a ciliated larva (oncomiracidium) hatches. It attaches itself to the surface of the final host, usually the skin, gills, or fins of fish, and develops to the adult. *See* MONOGENEA.

Trematoda. The flukes consist of two groups, the Aspidogastrea (=Aspidobothria) and Digenea. The Aspidogastrea are endoparasites of invertebrates and vertebrates, characterized by their ventral attachment disk consisting of a row of suckers or many suckerlets, and by a simple life cycle which may, however, include two hosts. The Digenea are endoparasites of vertebrates, with two or more hosts involved in their life cycle (Figs. 5 and 6), and the intermediate host is almost always a mollusk. In the Aspidogastrea, there never is secondary multiplication of larval stages-that is, the larva is transformed into the adult. In the Digenea, larval stages multiply-that is, a miracidium (ciliated first larva) is formed in the egg, it develops to a sporocyst which gives rise to several or many daughter sporocysts or rediae, and they in turn give rise to many cercariae, which may encyst outside a host or in a second intermediate host, forming the metacercariae. Each cercaria is transformed to the adult after infection of the final (=definitive) host. See ASPIDOGASTREA; DIGENEA; TREMATODA.

Cestoda. These highly specialized animals are parasites of vertebrates and consist of the "segmented" Eucestoda and the nonsegmented Amphilinidea and Gyrocotylidea, formerly grouped in the Cestodaria, which, however, is unlikely to be a monophyletic group. A typical tapeworm in the Eucestoda consists of an anterior attachment organ, the scolex, and a series of similar reproductive segments, the proglottids. *See* CESTODA; EUCESTODA.

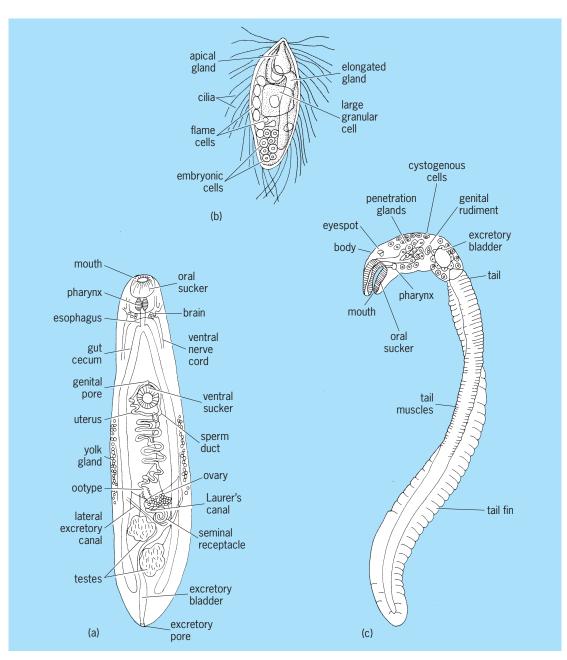


Fig. 6. Some of the stages in the life cycle of *Opisthorchis felineus*. (a) Adult from the bile ducts of a cat. (b) Swimming miracidium hatched from an egg (this stage penetrates the snail host). (c) Cercaria, view from the side (this stage encysts in a fish which is then eaten by the cat). (*After H. Vogel, Zoologica, 33:1–103, 1934*)

Amphilinidea infect the body cavity of teleost fishes and freshwater turtles; their larvae occur in crustaceans. The Gyrocotylidea infect chimaerid fishes; their life cycle is not known. *See* AMPHILIN-IDEA; GYROCOTYLIDEA.

In Eucestoda the ripe eggs or hatched ciliated larvae (coracidia) are ingested by intermediate hosts in which six-hooked embryos, the oncospheres, transform into larval forms such as procercoids, plerocercoids, cysticercoids, cysticerci, and other specialized types. Final hosts become infested by ingesting these larvae. Klaus Rohde; C. G. Goodchild

Bibliography. R. H. Barth and R. E. Broshears, *The Invertebrate World*, Sanders College Publishing, Philadelphia, 1982; K. Rohde (ed.), *Marine Para*- *sitology*, CSIRO Publishing, Melbourne, and CABI Publishing, Wallingford, Oxon, U.K., 2005; G. D. Schmidt and L. S. Roberts, *Foundations of Parasitology*, 6th ed., McGraw-Hill, New York, 1999.

Platypus

The single species *Ornithorhynchus anatinus* of the family Ornithorhynchidae in the order Monotremata, which occurs in eastern Australia and Tasmania. This mammal is a monotreme, which lays eggs and incubates them in a manner similar to birds. The platypus is also known as the duckbill, duckmole, and water mole. One of the most primitive mammals,



Platypus, a primitive aquatic mammal. (Photograph by Paddy Ryan/Ryan Photographic)

it retains some reptilian characteristics. The female lacks a marsupium; however, marsupial or epipubic bones are well developed.

On the male's hindleg is a poison spur whose function is unknown, but which occurs in other monotremes. The skeletal system possesses several reptilianlike characteristics. In addition, as an oviparous species, reptilian features such as a cloaca, double uterus, and primitive reproductive organs are retained.

Ornithorhynchus is well adapted to aquatic life. It ranges from the tropical, sea-level streams to cold lakes at altitudes of 6000 ft (1800 m). It is covered with short dense fur, typical of many aquatic mammals; the external ears are lacking; and the tail is broad and flattened, resembling that of the beaver (see **illus.**). The rubbery bill is covered with a highly sensitive skin. There are no teeth, but horny ridges are used for crushing food and grubbing on the bottom of rivers. Food is principally crustaceans, worms, and tadpoles.

The adult is about 2 ft (0.6 m) long and weighs approximately 4 lb (1.8 kg). The duckbill is a burrowing animal, and digs a chamber in the side of a riverbank. The female lays and hatches two leathery shelled eggs in the chamber, which is lined with dry grass. Incubation takes about 2 weeks. The young are born naked and helpless, and the female feeds them on lacteal secretions produced by mammary glands which more closely resemble sweat glands. The young have temporary molar, deciduous teeth that are lost in the adult. *See* MONOTREMATA. Charles B. Curtin

Bibliography. S. Jackson, M. Serena, and D. Middleton, Platypus, in S. Jackson (ed.), *Australian Mammals: Biology and Captive Management*, pp. 1–32, CSIRO, Collingwood, Victoria, Australia, 2003; R. M. Nowak, *Walker's Mammals of the World*, Johns Hopkins University Press, 1999; M. Serena, Duck-billed platypus: Australia's urban oddity, *Nat. Geog.*, 197(4):118–129, 2000.

Playa

A nearly level, generally dry surface in the lowest part of a desert basin with internal drainage (see **illus**.). When its surface is covered by a shallow sheet of water, it is a playa lake. Playas and playa lakes are also called dry lakes, alkali flats, mud flats, saline lakes, salt pans, inland sabkhas, ephemeral lakes, salinas, and sinks. The original meaning in Spanish of shore or beach has been lost. An example is Death Valley in California.

A playa surface is built up by sandy mud that settles from floodwater when a playa is inundated by downslope runoff during a rainstorm. A smooth, hard playa occurs where ground-water discharge is small or lacking and the surface is flooded frequently. These mud surfaces are cut by extensive desiccation polygons caused by shrinkage of the drying clay. Puffyground playas form by crystallization of minerals as ground water evaporates in muds near the surface. Ground water moves upward because artesian conditions exist on many playas. Artesian conditions require that subsurface flow be confined within an aquifer, usually a sandy layer, so the piezometric surface created by the hydrologic head is near or above the ground surface. See AQUIFER; ARTESIAN SYSTEMS.

Subsurface brine is present beneath many playas. Brines evolve from surface waters that sink into alluvial fans adjacent to the playa basin. As ground water flows through the fans, through the adjoining sandflats, and under the playa, various minerals are precipitated progressively in the sediments, increasing the ionic concentration of the water and modifying the proportions of the ions. The type of brine depends on the original composition of the surface water and reflects the lithology of the rocks weathered in the surrounding mountains. *See* GROUND-WATER HYDROLOGY.

Numerous playas in the southwestern United States yield commercial quantities of evaporite minerals, commonly at shallow depths. Important are salt (NaCl) and the borates, particularly borax (Na₂B₄O₇·10H₂O), kernite (Na₂B₄O₇·4H₂O), ulexite



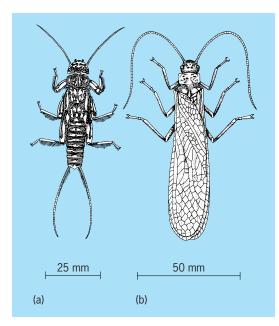
Light-colored playa in lowest part of Sarcobatus Flat in southern Nevada.

(NaCaB₅O₉·8H₂O), probertite (NaCaB₅O₉·5H₂O), and colemanite (Ca₂B₆O₁₁·5H₂O). Soda ash (sodium carbonate; Na₂CO₃) is obtained from trona (Na₃H(CO₃)₂·2H₂O) and gaylussite (Na₂Ca(CO₃)₂· 5H₂O). Lithium and bromine are produced from brine waters. *See* DESERT EROSION FEATURES; SALINE EVAPORITES. John F. Hubert

Bibliography. J. L. Melvin (ed.), Evaporites, Petroleum and Mineral Resources, *Dev. Sedimentol.*, no. 50, 1991; H. G. Reading (ed.), *Sedimentary Environments: Processes, Facies and Stratigraphy*, 3d. ed., 1996; R. W. Renault and W. M. Last (eds.), Sedimentology and Geochemistry of Modern and Ancient Saline Lakes, *Soc. Sed. Geol. Spec. Publ.*, no. 50, 1994.

Plecoptera

An order of primitive insects known as the stoneflies. The order comprises 16 families and more than 2000 species distributed on all continents except Antarctica. Stoneflies spend the majority of their lives aquatically as nymphs (immature larval forms); they exit the water and live only briefly as adults on land. There are relatively slight differences between the aquatic nymphal stage and the terrestrial adult form, except for wings and tracheal gills. In stoneflies, the soft and somewhat flattened or cylindrical body ending in two tail filaments, the strong legs with paired tarsal claws, the chewing mouthparts, and the rusty blacks, dull yellows, and browns are characteristic of both the nymph and adult (see illustration). The name Plecoptera literally means pleated wing, referring to the hindwings that are folded and hidden under the forewings, which the adult holds close to the abdomen when at rest or walking. See INSECTA.



Plecoptera. (a) Nymph of *Paragnetina*. (b) Adult of *Pteronarcys*, one of the largest stoneflies. (After A. H. Morgan, Field Book of Ponds and Streams, Putnam, 1930)

Characteristics. Stonefly nymphs live in the rapid stony parts of clean swift streams, although a few species occur along the rocky shores of large temperate lakes. Filamentous tracheal gills are characteristic of all nymphs except those that live in highly aerated water and respire through the body wall. In various species, gills are attached at the neck, on the side of the thorax, and on the sides and top of the abdomen. In some species, such nymphal structures persist in the adults but do not function.

Most stonely species require 1 year for development, but a few may take as long as 3 years. In an observed complete life history of one species, the nymphs went through 22 instars, the form assumed by an insect between molts; the number of known instars ranges from 12 to 36. Nymphs have diverse feeding habits, including species that are carnivores and others that are herbivores and/or detritivores. Their diet can be determined by examining the stomach contents in newly collected specimens.

The adult life stage begins when full-grown nymphs climb onto rocks or plants, molt, and emerge as adults. In various parts of North America stonefly adults emerge during every month of the year. Numerous species of "winter stoneflies" reach their peaks of abundance from late winter to early spring. For these species, emergence typically takes place after openings appear in ice-covered streams. Adults are poor fliers and are often found creeping about on stream banks and shrubbery.

Reproductive behavior. Adults live from one to a few weeks, and the time is mainly spent engaged in reproductive activities. Mating occurs near the ground, never in flight. For most North American species of stoneflies, males and females locate each other through a complex system of vibrational communication called drumming. The male taps or rubs his abdomen on various substrates and waits for a receptive female to respond, whereby they are able to locate each other. Each species has its own unique pattern of drumming that varies in amplitude, frequency, and duration. Females deposit egg masses into the water using various means. Stoneflies that inhabit large streams deposit their eggs directly into the swirling water in clutches of more than 1000; other species place the eggs at the water's edge, drop them from the air, or even crawl underwater and deposit eggs onto the steam bottom. See INSECT PHYSIOLOGY. Andrew K. Rasmussen; Lewis Berner

Bibliography. E. Leiser and R. H. Boyle, *Stone-flies for the Angler*, Knopf, New York, 1982; W. P. McCafferty, *Aquatic Entomology: The Fishermen's and Ecologists' Illustrated Guide to Insects and Their Relatives*, Jones and Bartlett, Sudbury, MA, 1981; R. W. Merritt and K. W. Cummins (eds.), *An Introduction to the Aquatic Insects of North America*, 3d ed., Kendall Hunt, Dubuque, Iowa, 1996; K. W. Stewart and B. P. Stark, *Nymphs of North American Stonefly Genera (Plecoptera)*, 2d ed., Caddis Press, Columbus, Ohio, 2002; K. W. Stewart, S. W. Szczytko, and B. P. Start, The language of stoneflies, *Bioscience*, 33:117-118, 1983.

Pleiades

A typical open cluster of stars (see **illus.**), and the best example in the sky of an open cluster for the casual observer. (Open clusters are one of the two basic types of star clusters, distinguished from globular clusters by their smaller total populations and central densities, and by their younger age. Besides being a focus of astronomical research in the present era, the Pleiades has been considered important since ancient times in various cultures, including ancient civilizations of Pre-Columbian Mexico, Greece, Persia, India, and China. *See* STAR CLUSTERS.

The Pleiades, located in the constellation Taurus, contains several hundred stars within a radius of about 1° , centered on its brightest star, Alcyone. (In comparison, the Moon is 0.5° across.) The cluster is permeated with a faint diffuse nebulosity. Though the Pleiades is always referred to as having seven bright stars, only six are seen today. The "lost Pleiad" is likely to be Pleione, which is presently a 5th-magnitude variable star. The configuration of the brighter stars resembles a very small dipper. *See* TAU-RUS; VARIABLE STAR.

Several characteristics of the Pleiades make the cluster useful for research. It is a young cluster, about 100 million years old, and still contains bright hot blue stars (type B) on the main sequence (hydrogenburning stars) with abundances of elements similar to that of the Sun. It is compact, and close enough to see its smaller fainter members (including brown dwarfs). Finally, although it has stars of widely varying masses and brightnesses, they are all approxi-



The Pleiades. (Lick Observatory photograph)

mately the same age and distance from Earth, about 130 parsecs (425 light-years). *See* LIGHT-YEAR; PAR-SEC; STAR; STELLAR EVOLUTION.

Until the Hipparcos satellite was launched in 1989, the Pleiades was just beyond the distance at which trigonometric parallaxes could be measured. While the distances measured by Hipparcos for the Hyades and other clusters were generally near the values expected from previous measurements, the Pleiades cluster was found to be closer than its expected distance. This implied that the actual light given off by the stars was only about threequarters of what had been assumed from the pre-Hipparcos measured distance. After several years of investigation, the most likely resolution to the problem seems to be that the Hipparcos measurements for the Pleiades cluster contain unexpected systematic errors, in the complex solution of the extremely precise positional measurements obtained within a limited period of time, and the Pleiades is actually at its pre-Hipparcos measured distance. See PARALLAX (ASTRONOMY).

Because it is relatively nearby, the Pleiades is frequently studied. Brown dwarfs, whose masses are smaller than that of the faintest true star but larger than those of planets, have been found in the Pleiades. While they may make up a quarter of the objects in the cluster, they probably account for no more than 1% of the cluster mass. The Pleiades is also a useful basis for comparison with other clusters in order to obtain the distances of clusters much farther away. Other topics for research include the effects of rapid rotation (stars slow down as they age) and the fitting of theoretical models to the Pleiades. *See* BROWN DWARF; STELLAR ROTATION.

Nancy Remage Evans; Ronald Pitts Bibliography. G. Basri, The discovery of brown dwarfs, *Sci. Amer.*, 282(4):76, April 2000; A. G. A. Brown, The Hyades—So close, and now, so familiar, *Mercury*, 27(3)17–21, May/June 1998; R. Burnham, Jr., *Burnham's Celestial Handbook*, vol. III, Dover Publications, New York, 1978; B. Paczynski, Astronomy: A problem of distance, *Nature*, 427:299– 300, 2004; L. Kun Tan, The Pleiades, in Gallery, *Sky Telesc.*, 99(2):136, February, 2003.

Pleistocene

The older of the two epochs of the Quaternary Period. The Pleistocene Epoch represents the interval of geological time (and rocks accumulated during that time) extending from the end of the Pliocene Epoch (and the end of the Tertiary Period) to the start of the Holocene Epoch. Most recent time scales show the Pleistocene Epoch spanning the interval from 1.8 million years before present (m.y. B.P.) to 10,000 years B.P. The Pleistocene is commonly characterized as an epoch when the Earth entered its most recent phase of widespread glaciation. *See* QUATERNARY.

British naturalist Charles Lyell introduced the term Pleistocene in 1839. Earlier, he had subdivided the

	QUATERNARY	Holocene
CENOZOIC		Pleistocene
	TERTIARY	Pliocene
		Miocene
		Oligocene
		Eocene
		Paleocene
MESOZOIC	CRETACEOUS	

Tertiary Period on the basis of the degree of faunal similarity of the epochs with modern molluscan fauna. He identified an older Pliocene and a younger Pliocene, with 50% and 70% faunal similarity with modern fauna, respectively. In 1839 he introduced the concept of Pleistocene as a separate epoch to replace the younger Pliocene. Later he attempted to abandon this term in favor of a generic post-Pliocene, but this revisionism was largely ignored by the geological community as the use of the term Pleistocene had become popular in stratigraphic literature. The Holocene Epoch was proposed by the French naturalist P. Gervais in 1867 to cap the Pleistocene and the geological column. The Holocene represents the youngest of the geological epochs and spans the last 10,000 years of geological time. Its fossil content is characterized by 100% faunal similarity to modern assemblages. See HOLOCENE; PLIOCENE; TERTIARY.

Following the publication in 1840 of the work by Swiss zoologist Louis Agassiz, who envisioned a climatically induced continental ice sheet all the way from the North Pole to Central Asia and across the Alps, geologists for the first time began associating the Pleistocene as a geological epoch with widespread glaciation. Though Agassiz's notion was correct in its basic tenet, it was based on several ill-founded assumptions and was not taken seriously. By 1850, however, the concept of the Pleistocene ice age took hold, largely through the documentation of glacial till sediments by Scandinavian and Scottish geologists. The glacial deposits normally lack contemporaneous fossils (they may contain reworked older fossils) and are thus difficult to date. Nonetheless, the fossils contained in sediments below and above the glacial deposits showed considerable affinity to modern assemblages, thereby indicating to the early workers that these sediments were relatively recent, that is, of the Quaternary age. *See* TILL.

The early workers envisioned a single large glacial event in the Pleistocene, but observations soon led to evidence of more than one glaciation, separated by more equitable interglacial periods. As evidence has accumulated in recent decades from both the land and the sea, it clearly shows numerous glacial advances and retreats. At least 17 glacial events have been shown to have occurred during the Pleistocene. Modern paleoclimatic studies also indicate that the Pleistocene does not characterize the first glaciation. The first significant glaciation in the most recent phase of Earth's glacial history began as far back as the middle Miocene, some 15 m.y. ago. Nevertheless, extensive and high-frequency glacials events are characteristic of the Pleistocene Epoch. See CLIMATE HISTORY; GLACIAL EPOCH.

While geologists and glaciologists think widespread glaciation typifies the Pleistocene Epoch, in the past anthropologists considered the appearance of modern humans as a defining criterion, and vertebrate paleontologists used the fossiliferous occurrence of the elephant, modern horse, and cattle in their definition of this interval. Some of these events were later shown to be of greater antiquity and not restricted to the Pleistocene.

Subdivisions. In modern geological time scales, the Pleistocene is subdivided into a lower and an upper series. In Europe, the lower series is considered equivalent to the Calabrian Stage, while the upper series is equated with the Sicilian and Tyrrehenian stages (**Fig. 1**). The exact boundaries between these stages are, however, still a matter of debate among stratigraphers.

The Calabrian Stage was established by the French geologist M. Gignoux in 1913 based on shallow marine sediments in Calabria, the toe of Italy. The occurrence of the cold-water mollusk species *Arctica islandica* and the benthic forminifer *Hyalinea balthica* characterizes the Calabrian beds, which were assumed to correspond to the latest Pliocene cold period. The latter was elevated to represent the basal

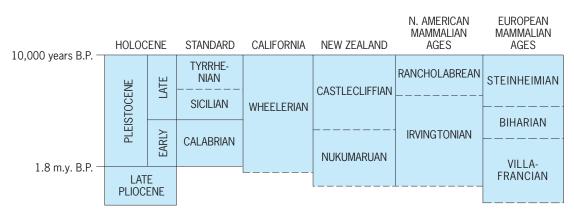


Fig. 1. Pleistocene stages and their temporal equivalents. (After B. U. Haq and F. W. B. van Eysinga, Geological Time Table, Elsevier, Amsterdam, 1998)

Pleistocene at the International Geological Congress in 1948. However, the base of the Pleistocene (or the Pliocene-Pleistocene boundary) has remained controversial. At present the Calabrian sections at La Castella and Vrica and near the village of Santa Maria di Catanzaro are variously used by stratigraphers to typify the Pliocene-Pleistocene boundary and the Calabrian Stage. In 1977 B. Haq and coworkers, in an attempt to reconcile the paleoclimatic and chronologic data, used multiple, fossil, radiometric and paleomagnetic criteria to define the boundary at the Calabrian sections. They equated the base of the Pleistocene with the top of the Olduvai magnetic event, and considered it as coeval with the first of the four major climatic deteriorations of the Pleistocene. This finding also reconciled the boundary with the arrival of the cold molluscan and foraminiferal fauna in the Mediterranean and a dramatic reduction in temperatures. More recently, Italian stratigraphers have preferred the Vrica section as the type section for the Pliocene-Pleistocene boundary. It is a thicker section, containing much the same events as the La Castella section but a better ash-bed chronology. Nevertheless, their placement of the boundary yielded the same conclusions as those reached by Haq and coworkers. See DATING METHODS; PALEOCLIMATOL-OGY; PALEOMAGNETISM.

The marine Calabrian is replaced locally in Italy by fresh-water deposits. These were designated by L. Pareto in 1856 as the Villafranchian Stage based on a section near Villafranca di Asti that is rich in mammalian fauna. The occurrence of *Equus* and *Elephas* in these beds was considered to be their earliest, and to mark the beginning of the Quaternary according to vertebrate paleontologists. However, the time equivalency of the Villafranchian with Calabrian has been questioned, and the former has been shown to be at least in part older than the Calabrian, and therefore late Pliocene. The first appearances of modern horses and elephants have also been shown to date from the late Pliocene.

The Calabrian deposits are overlain by shallow marine sediments assigned to the Sicilian Stage, which was considered to characterize the second glacial episode marked by the return of *A. islandica* to the Mediterranean area. In the later part of the Sicilian, the climate once again ameliorated and warmer faunas returned. This was followed by the deposition of the sediments of the Tyrrhenian Stage, which represents another regressive-transgressive cycle. The Tyrrhenian sediments represent overall warming conditions with several smaller, colder intervening episodes.

Local Pleistocene stage equivalents (Fig. 1) in Italy include the Emilian (which straddles the early-late Pleistocene boundary) and the Milazzian (which is equivalent to the later part of Sicilian and early Tyrrhenian stages). In California, the Wheelarian Stage is considered largely to be equivalent to the Pleistocene, though its lower boundary is somewhat older and occurs in the latest Pliocene. In Japan the upper half of the Kechienjian Stage and the Yuzanjian Stages are considered to span the early Pleistocene. The Australian Werrikooian Stage also spans only the early Pleistocene. In New Zealand, stratigraphers recognize the Nukumaruan Stage as overlapping the latest Pliocene and the earlier Pleistocene, and the Castlecliffian as representing the later Pleistocene. They divide the Castlecliffian further into a lower, Okehuan substage, and an upper, Putikian substage. Students of vertebrate fossils, who often name their own stages based on mammalian fauna, equate the upper part of Villafrancian, the Biharian, and the Steinheimian (Oldenburgian) with the Pleistocene in the Old World, and use Irvingtonian and Rancholabrean in North America to subdivide the Pleistocene. Each of these ages is characterized by its own typical mammalian assemblage. *See* INDEX FOSSIL.

Glaciations and climate. Contrary to the belief of early workers who equated the Pleistocene with the ice age, evidence accumulated in recent years indicates that glaciation in the Northern Hemisphere actually began in the late Pliocene. Data from the stable oxygen isotopes from the marine record suggests that both Iceland and Greenland were glaciated at least 3.2 m.y. B.P. An especially cold interval also occurred around 3.0 m.y. B.P., when glacial ice reached the sea in both polar regions as indicated by substantial ice-rafted debris on the sea floor. The remainder of the Pliocene shows cyclic waxing and waning of the ice sheets. The cyclic nature of the glacialinterglacial variations becomes even more characteristic of the Pleistocene. Research in the 1970s and 1980s firmly tied this cyclicity to variations in astronomical parameters that control the distance between the Earth and the Sun (Milankovitch cyclicity, discussed below). The onset of the Pleistocene brought glaciations that were more widespread than those in the Pliocene. Mountain glaciers expanded and continental ice fields covered large areas of the temperate latitudes. Sea ice also became more widespread. Where there were no glaciers, such as in the lower latitudes, considerable cooling occurred.

Four or five major glacial events have been identified in the Pleistocene continental record of the temperate and high latitudes, and each of these episodes shows a higher frequency of advances and retreats of glacial ice. In the Alps, five major glacial episodes have been identified and are known as (from the oldest to the youngest) the Donau, Günz, Mindel, Riss, and Würm. On the North American continent, four major glacial advances have been identified in the last million years of the continental record. These are known as (from the oldest to the youngest) the Nebraskan, Kansan, Illinoian, and Wisconsinan. During each of these episodes, ice sheets covered much of the North American continent (Fig. 2). The synchroneity of the younger four glaciations in the Alps with those of North American is, however, not fully established due to difficulty in direct dating of glacial deposits. In New Zealand, four major Pleistocene glaciations are also evident when highland glaciers advanced to the lowlands, but their correlation with Northern Hemisphere glacial episodes have not been clearly established.

The most complete record of the Pleistocene glacial and climatic variations that can be tied to a

meaningful chronology comes from the deep sea. Deep-sea sediment cores, especially those of the North Atlantic, contain a record of climatic cyclicity that has been shown to be forced by Milankovitch modulation. In the early twentieth century, M. Milankovitch established curves for solar insolation (the amount of solar radiation impinging on the Earth's surface) as a factor of temporal variations in the Earth's distance from the Sun. This distance is controlled by three astronomical parameters with different periodicities. These include variations in (1) the stretch of the Earth's orbit around the Sun, or eccentricity, with a periodicity of about 95,000 years; (2) the tilt of the Earth's axis of rotation, or obliquity, with a periodicity of about 40,000 years; and (3) the wobble in the Earth's rotation, or precession, that has a periodicity of about 21,000 years. Eccentricity and precession directly define the distance of the Earth from the Sun, whereas obliquity influences the distribution of solar radiation in different latitudes. Milankovitch's curves implied that the variation in the solar insolation was the most important factor in modulating the Pleistocene climate. These timedependent astronomical parameters have been directly tied to the observed periodicities in the stableisotopic and faunal and floral record from the deep sea. See EARTH ROTATION AND ORBITAL MOTION; INSOLATION; MARINE SEDIMENTS.

Although the paucity of accurate time correlations remains a limiting factor, climatic data on a regional basis suggest that Pleistocene climates differed in different areas, and cooling and warming trends were not uniform. Generally, cooler climates followed the onset of glaciations, and warmer climates trailed glacial retreats. The interglacial periods, however, have been inferred to have become progressively cooler through the Pleistocene, with each successive interglacial as much as $1^{\circ}C$ (1.8°F) cooler than the previous one. Vegetation and faunal distributions also followed these climatic fluctuations. For example, in North America, tundralike vegetation spread several thousand kilometers to the south during the peak of the Wisconsinan glaciation when compared to its present-day distribution. See GEOLOGIC THER-MOMETRY; TUNDRA.

Oceans. The marine sedimentary record shows that in the oceans the equatorial regions seem to have suffered less cooling during glacial episodes than the temperate to higher latitudes. For example, near the end of the last glacial maximum, some 18,000 years ago, the Caribbean and Central Atlantic waters cooled only slightly, while the North Atlantic at 40°N cooled by as much as 10°C (18°F). This was a consequence of a significant southward shift of the paleo-Gulf Stream during the glacial maximum. The average temperature of the ocean waters was, however, only 2.3° C (4.1° F) cooler than at present. On the continents the temperatures were some 5°C (9°F) cooler than today. Temperature increases during the interglacials were also of greater magnitude in the continental interiors, where the average summer temperature may have increased by as much as $15^{\circ}C(27^{\circ}F).$

The expansion and decay of the ice sheets had a

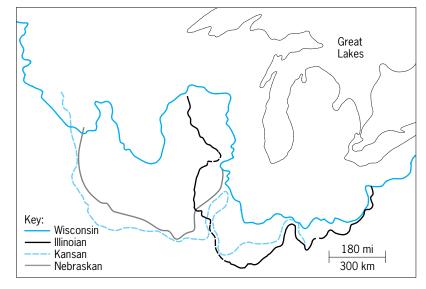


Fig. 2. Spread of the continental ice sheets during various Pleistocene ice ages over North America. (After R. F. Flint, Glacial and Quaternary Geology, Wiley, New York, 1971)

direct effect on the global sea level. Global sea-level fluctuations of 50-150 m (170-500 ft) have been estimated for various glacial-interglacial episodes during the Pleistocene, following the sequestering of water as glacial ice on the continents and its subsequent melting. Since the last deglaciation, which began some 17,000 years ago, the sea level has risen by about 110 m (360 ft) worldwide, drowning all of the ancient lowstand shorelines. One important product of the sea-level drops was the migration of large river deltas to the edges of the continental shelves and to the deeper parts of the basins. These coarse-grained sandy sediment bodies constitute ideal reservoirs for the accumulation of hydrocarbons. Conversely, the last marine transgression that started in the late Pleistocene after rapid deglaciation and ended in the Holocene (6000 to 7000 years ago) resulted in new deltas that formed at the mouths of modern rivers. All present-day river deltas date from the cessation of the rapid sea-level rise and the attainment of relative eustatic stability in the mid-Holocene. See DELTA; PALEOCEANOGRAPHY; PETROLEUM GEOLOGY.

Life. The Miocene three-toed horse genus Hipparion survived into the latest Pliocene part of the Villafranchian. These and other survivor species, such as mastodonts, and premodern rhinos became largely extinct at the close of the Pliocene. The onset of cooler climate and Pleistocene glaciation is also approximated with a wave of mammalian migration from the east to the west. A relatively modernlooking fauna that included the first true oxen, elephants, and the first one-toed horse first appeared at the beginning of Pleistocene. The modern horse, E. caballus, made its first appearance some 250,000 years ago in the late Pleistocene in North America. From North America it migrated to Asia and then west to Europe. However, during the last glacial maximum, some 18,000 years ago, it became extinct in North America when it was unable to cross the deserts to migrate to South America. Oxen, deer and reindeer, large cats, mammoth, great elk, wolf,

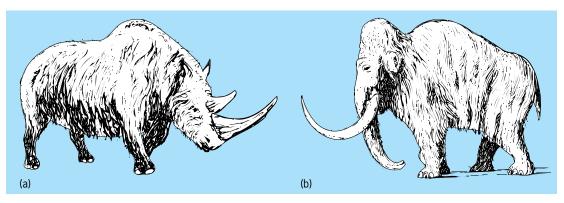


Fig. 3. Two European mammals of the Pleistocene that ranged widely and died out before the close of the epoch. (a) Woolly rhino (Coelodonta antiquitatis) (after B. Kurten, Pleistocene Mammals of Europe, Weidenfeld and Nicolson, 1968). (b) Woolly mammoth (after O. Abel, Die vorzeitlichen Säugetiere, Gustav Fischer, 1914)

hyena, and woolly rhinos (**Fig. 3**) proliferated during the middle and late Pleistocene. Mammoths, which have been found preserved nearly intact in frozen soils in Siberia, ranged over much of Europe during the glacial times. Thick fur and a substantial layer of protective fat gave these species an advantage in the cold climates. Their reign ended at the close of the glacial period. *See* GEOLOGIC TIME SCALE; MAM-MALIA.

The unique Pleistocene mammalian faunas of some of the isolated islands, such as Madagascar, the Philippines, Taiwan, and the Japanese Archipelago, indicate restriction in the dispersal of species during the Pleistocene.

Before the onset of the Pleistocene glaciation, the Pliocene Epoch had already seen the evolution of a whole lineage of hominids in Africa that had given rise to the human precursor, Homo habilis, around 2 m.y. B.P. This was followed by the appearance of H. erectus, which came on the scene almost at the Plio-Pleistocene boundary, around 1.8 m.y. B.P. These species were followed by more recent human ancestors, and the Pleistocene Epoch is characterized by the final evolution of the Man. The first archaic H. sapiens are now considered to have arrived on the scene earlier than previously believed, around about a million years ago. The appearance of H. neanderthalensis, or the Neanderthal Man, whose tenure of the tropics and temperate regions partly overlapped with the archaic man, is now dated at least as far back as 250,000 years B.P. This species became extinct at approximately the same time as the rise of modern man. Recent datings have also pushed back the appearance of the first true modern H. sapiens (the Cro-Magnon Man) to around 100,000 years B.P. Even these dates are, however, quite likely to be revised (to greater antiquity) as new age data accumulate. See FOSSIL HUMANS; GEOLOGIC TIME SCALE. Bilal U. Hag

Bibliography. B. U. Haq et al., The corrected age of the Pliocene-Pleistocene boundary, *Nature*, 269:483-488, 1977; B. U. Haq and F. W. B. van Eyinga, *Geological Time Table*, 6th ed., 2007; T. Nilsson, *The Pleistocene: Geology and Life in the Quaternary Ice Age*, 1982; C. Pomerol, *The Cenozoic Era: Tertiary and Quaternary*, 1982.

Pleochroic halos

Spherical or elliptical regions up to 40 micrometers in diameter in which there is a change in color from the surrounding mineral when viewed with a petrographic microscope. Pleochroic halos are found around small inclusions of radioactive mineralsfor example, zircon, monazite, allanite, xenotime, and apatite-and in rock-forming minerals, principally quartz, micas, amphiboles, and pyroxenes. Halos have also been identified in coalified wood preserved in deposits on the Colorado Plateau. These small regions of discoloration were recognized by H. Rosenbusch as early as 1873 in cordierite, but it was not until the early twentieth century that J. Joly ascribed pleochroic halos to the effect of irradiation originating from the inclusions. See CORDIERITE; PETROFABRIC ANALYSIS.

The change in color is a result of radiation damage caused by alpha particles emitted during the radioactive decay of nuclides in the decay chains of uranium-238, uranium-235, and thorium-232. Typically, the alpha particle has an energy of 4.5-9 MeV and a range of 10-20 μ m, dissipating most of its energy by ionization with only a few hundred atomic displacements at the end of its trajectory. The alpha recoil has an energy of 0.07-0.09 MeV and a range of only 0.02 μ m. Thus, it is the range of the alpha particle and ionization effects that account for the size and color of the halos. The halos have a distinctive ring structure with varying degrees of discoloration between the rings: the coloration in the halos increases, saturates, and finally diminishes with increasing ion dose. A total dose over time of 10¹³ alpha particles per square centimeter is required to initiate discoloration in laboratory experiments. Natural fluences must be four to seven times greater in order to initiate the same effect in micas, suggesting annealing of radiation damage over geologic periods. Thus, the size of a halo is limited first by the range of the alpha particle and then by the dissipation of the fluence (inversely proportional to distance) below a critical threshold value required for discoloration. See ALPHA PARTICLES; METAMICT STATE.

The early interest in pleochroic halos was in their use for geologic age dating. Careful attempts were made to correlate the halo color with the alpha-irradiation dose in order to estimate the age of the enclosing mineral. Additionally, the constant size of the rings of the uranium and thorium halos for minerals of different ages was taken as evidence that the decay constants for radionuclides used in age dating had remained constant throughout geologic time. Thermal annealing of the halos has been used to model the thermal histories of rock units. *See* FISSION TRACK DATING.

Of great interest are the giant halos (50–100 μ m in radius) or the Y halos, which lack the distinctive ring structure of uranium and thorium halos. These halos were attributed to extinct or nearly extinct nuclides, such as plutonium-244. The much larger halos (up to 100 μ m) might also have been caused by highly energetic alpha particles that accompany the spontaneous fission of plutonium-244 or by alpha decay of heavy elements (atomic number 126, 127, 116, or 124). Various interpretations suggest mechanical and chemical mechanisms for the enlargement of the discolored areas of the halo. *See* ATOMIC NUMBER; MICA; PLEOCHROISM; RADIOACTIVE MINERALS. Rod Ewing

Bibliography. R. V. Gentry, Radioactive halos, *Annu. Rev. Nucl. Sci.*, 23:347–362, 1973; C. H. Holbrow, Haloes from plutonium minerals and fission alphas?, *Nature*, 265:504–508, 1977; M. A. K. Lodhi (ed.), *Superbeavy Elements*, 1978; C. Moazed, R. M. Spector, and R. F. Ward, Polonium radiohalos: An alternative interpretation, *Science*, 180:1272–1274, 1973.

Pleochroism

In some colored transparent crystals, the effect wherein the color is quite different in different directions through the crystals. In such a crystal the absorption of light is different for different polarization directions. Tourmaline offers one of the best-known examples of this phenomenon. In colored transparent tourmaline the effect may be so strong that one polarized component of a light beam is wholly absorbed, and the crystal can be used as a polarizer. For a fuller discussion of the effect *see* DICHROISM; POLARIZED LIGHT; TRICHROISM. Bruce H. Billings

Plesiosauria

An order of extinct aquatic diapsid reptiles within the infraclass Sauroptrygia, common throughout the world during the Jurassic and Cretaceous periods. These large carnivores are characterized by long, paddle-shaped limbs and short, dorsoventrally compressed bodies (see illus.). In contrast with nothosaurs (more primitive members of the Sauropterygia), plesiosaurs have greatly expanded ventral portions of both pectoral and pelvic girdles to provide large areas for the attachment of muscles to move the limbs anteriorly and posteriorly. Lateral undulation of the trunk was severely restricted by elaboration of the ventral scales. The great length of the limbs indicates that they must have been used symmetrically, rather than alternatively as in primitive reptiles. The configuration of the skeleton suggests that plesiosaurs swam in the manner of sea lions by horizontal retraction of the limbs, rather than like sea turtles and penguins which use the limbs like wings to fly through the water.

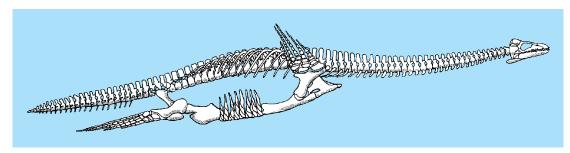
Approximately 40 genera of plesiosaurs are recognized. They are divided into two large groups: The plesiosauroids include the most primitive genera and others that have small heads and very long necks. Among the plesiosauroids, the elasmosaurids had as many as 76 cervical vertebrae. The pliosauroids had shorter necks but larger skulls. The skull of *Kronosaurus* exceeded 9 ft (2.7 m) in length. Both elasmosaurs and pliosaurs persisted until the end of the Mesozoic.

The common attributes of the loss of the lower temporal bar and the superficial position of the scapula relative to the clavicle support relationship between the plesiosaurs and the Triassic nothosaurs, but known members of these two groups differ significantly in limb proportions and the structure of the girdles. Intermediate forms might be expected in the latest Triassic, but few remains of marine reptiles are known from this time interval. *See* DIAPSIDA; NOTHOSAURIA; REPTILIA; SAUROPTERYGIA. Robert L. Carroll

Bibliography. R. L. Carroll, Vertebrate Paleontology and Evolution, 1987.

Pleuronectiformes

A very distinctive order of actinopterygian fishes, also known as the Heterosomata or commonly the flatfishes, which in turn are known by a variety of common names: halibut, turbots, plaice, flounders, soles, tonguefishes, and others. All adults are characterized by a flattened (laterally compressed) body with both eyes located on the same side of the head. Larval flatfishes are bilaterally symmetrical,



Skeleton of the plesiosauroid Cryptocleidus, about 10 ft (3 m) long.

but early in development one eye migrates across the top of the head to the other side. The asymmetry is reflected in complex modifications of the skull bones and associated muscles and nerves, as well as dentition, squamation (arrangement of scales), visceral anatomy, and paired fins. A given family is usually either dextral (eyes on the right side) or sinistral (eyes on the left side); however, some families have both dextral and sinistral members even at the species level. Flatfishes are further characterized by pigmented upper side; blind side usually lacking pigment; cycloid or ctenoid scales; tubercles in some species; long-based dorsal and anal fins, with the dorsal fin base extending to or beyond the eyes (except in the family Psettodidae); body cavity quite reduced in size; usually six or seven branchiostegal (gill bone) rays; and a swim bladder in adults of only a few species. See ACTINOPTERYGII; ANIMAL SYMME-TRY; SWIM BLADDER.

Flatfishes are invariably benthic (bottom-dwelling) as postlarvae, spending their entire adult lives swimming on their side (except *Psettodes*, which occasionally swims upright), with the eyeless side down. They are carnivorous predators, keeping much to the bottom. The eyes are protrusible, allowing these fishes to see while buried in the substrate awaiting a wandering prey. In pursuit of prey, the flatfishes move far from the substrate. Worldwide many species of flatfish are valued as food and are important to the commercial fisheries industry. *See* MARINE FISH-ERIES.

The order Pleuronectiformes comprises about 678 extant species in about 134 genera, 14 families, and two suborders. About 10 species of soles are thought to be limited to freshwater habitats; a few more, primarily freshwater species enter estuaries or marine water, and another 20 that are normally marine species occasionally enter freshwater.

The families below were selected to represent the diversity of the pleuronectiforms.

Suborder Psettodoidei. The family Psettodidae (spiny turbots) is represented by a single genus, *Psettodes* (Fig. 1), and three species. It occurs in marine waters of tropical western Africa, the eastern Atlantic, eastern Africa and the Red Sea, and the western Pacific. It is distinguished from other flatfishes by a combination of the following characters: dorsal fin not extending onto the head to or past the eye; spinous anterior dorsal and anal rays; nearly symmet-



Fig. 1. Indian spiny turbot (Psettodes erumei) from the Indo-West Pacific. (Photo by Jack E. Randall)



Fig. 2. Ocellated flounder (*Pseudorhombus* dupliciocellatus) from the Indo-West Pacific. (*Photo by Jack E. Randall*)

rical pelvic fins, with one spine and five soft rays; 15 branched caudal fin rays; sinistral or dextral eyes; very large mouth, with maxilla extending far beyond the eyes; large supramaxilla; barbed jaw teeth; gill arches with groups of teeth; palatine with teeth; and distinct preopercular margin, not covered with skin. Maximum length is about 60 cm (24 in.).

Suborder Pleuronectoidei. In this group, the dorsal fin extends onto the head at least to the eyes, and the dorsal and anal fins lack spines.

Family Paralichthyidae (sand flounders). The paralichthyids (**Fig. 2**) are primarily marine fishes of the Atlantic, Indian, and Pacific oceans, and rarely enter freshwater. They can be distinguished from other flatfishes by the following combination of characters: sinistral eyes in most species; short and nearly symmetrical pelvic fin bases (but position of bases is variable between species); and branched pectoral rays. Maximum length is 75 to 152 cm (30 to 60 in.). There are about 16 genera and about 105 species.

Family Pleuronectidae (righteye flounders). These are marine (occasionally brackish-water, rarely freshwater) fishes of the Arctic, Atlantic, Indian, and Pacific oceans that are distinguished from other flatfishes by the following characters: eyes almost always dextral; origin of dorsal fin above the eyes; symmetrical pelvic fins; usually no oil globule in the egg yolk; and a well-developed lateral line on both sides. There are 23 genera with about 60 species. The commercially important and large halibuts belong to this family, including *Hippoglossus bippoglossus* (Atlantic halibut), with a maximum published weight of 320 kg (700 lb), and *H. stenolepis* (Pacific halibut), weighing in at 363 kg (800 lb).

Family Bothidae (lefteye flounders). Known from the tropical and temperate waters of the Atlantic, Indian, and Pacific oceans, these flatfishes are distinguished by sinistral eyes; pelvic fin base on eye side longer than on eyeless side, on midventral line, and origin anterior to the base on eyeless side; unbranched pectoral and pelvic fin rays; pelvic fins without a spine; and eggs with a single oil globule in the yolk. Twenty genera are known, with about 140 species, some of which are commercially important. Maximum total length is 3.5 to 46 cm (1.4 to 18 in.), depending on the species.

Sole families. Three families, Achiridae, Soleidae, and Cynoglossidae, form a monophyletic group, with

Achiridae (American soles) includes shore fishes in the eastern central Pacific and western Atlantic, and freshwater fishes of temperate and tropical waters of South America. They are characterized by the following combination of characters: dextral eyes; margin of preoperculum (another gill cover bone) detectable by a shallow groove; dorsal and anal fins free from the caudal fin; right pelvic fin joined to the anal fin; and eyed side of the lower lip having a conspicuous, variously fringed fleshy rim. American soles are carnivores, feeding primarily on fishes and benthic invertebrates. Maximum total length is 4 to 28 cm (1.6 to 11 in.), depending on the species. Seven genera are known, with about 33 species.

Soleidae (soles) occurs widely in cold temperate to tropical waters of the eastern Atlantic, the Indo-West Pacific, and Australia. The following combination of characters readily identifies the soles: dextral eyes; preoperculum completely concealed by skin and scales; dorsal fin origin above or before the eyes; dorsal and anal fins free from or united with the caudal fin; pectoral fins present or absent; and pelvic fins free from the anal fin. Maximum total length is about 2.5 to 110 cm (1 to 43 in.), depending on the species. Thirty-five genera have been identified, with 135 species.

Cynoglossidae (tonguefishes; also called tonguesoles) is primarily a marine family of the tropical and subtropical seas of the world. It is distinguished from other soles by the following characters: very small sinistral eyes that are usually close together; preopercle margin covered with skin and scales; dorsal and anal fins confluent with the pointed caudal fin; pelvic fin of eyeless side having four rays attached along midventral line, linked to the anal fin in some species; pelvic girdle and fin on the eyed side absent in some species; pectoral fins absent; and asymmetrical mouth. Most species occur in shallow water or estuaries; however, some species of the genus Symphurus may be found at depths to 1900 m (6240 ft). Five species are known primarily from rivers. Maximum total length of most species is from 4 to 30 cm (1.6 to 12 in.), rarely 40 to 48 cm (16 to 19 in.). Three genera are known, with about 127 species. Herbert Boschung

Bibliography. F. Chapleau, Pleuronectiform relationships: A cladistic reassessment, *Bull. Mar. Sci.*, 52(1):516-540, 1993; R. Froese and D. Pauly (eds.), FishBase, World Wide Web electronic publication, <www.fishbase.org>, version (05/2005); T. A. Munroe, Bothidae (pp. 1885-1895), Scophthalmidae (1896-1897), Paralichthyidae (1898-1921), Poecilopsettidae (1922-1923), Achiridae (1925-1933), and Cynoglossidae (1934-1959), in K. E. Carpenter (ed.), *The Living Marine Resources of the Western Central Atlantic: FAO Species Identification Guide for Fishery Purposes*, vol. 3, pt. 2, FAO, Rome, 2003 (dated 2002); T. A. Munroe, Systematic diversity of the Pleuronectiformes (pp. 10-41), Distributions and biogeography (pp. 42–67), in R. N. Gibson (ed.), *Flatfishes: Biology and Exploitation*, Fish and Aquatic Resources Series 9, Blackwell Publishing, Oxford, England, 2005; J. S. Nelson, *Fishes of the World*, 4th ed., Wiley, New York, 2006; R. T. C. Ramos, Achiridae (American soles), pp. 666–669 in R. E. Reis, S. O. Kullander, and C. J. Ferraris, Jr. (eds.), *Checklist of the Freshwater Fishes of South and Central America*, EDIPUCRS, Porto Alegre, Brasil, 2003.

Pliocene

The youngest of the five geological epochs of the Tertiary Period. The Pliocene represents the interval of geological time (and rocks deposited during that time) extending from the end of the Miocene Epoch to the beginning of the Pleistocene Epoch of the Quaternary Period. Modern time scales assign the duration of 5.0 to 1.8 million years ago (Ma) to the Pliocene Epoch. *See* MIOCENE; PLEISTOCENE; QUATERNARY; TERTIARY.

Charles Lyell introduced Pliocene in 1833 to characterize the youngest of the Tertiary strata. He based his subdivision of geological eras on the degree of faunal similarity of those times with the modern molluscan assemblages. For the Pliocene he identified an older Pliocene (with more than 50% extant species) and a newer Pliocene (with up to 95% modern species). In 1839 he substituted the term Pleistocene for the newer Pliocene, which he later tried to abandon in favor of the term post-Pliocene; but this suggestion was largely ignored, and the use of the term Pleistocene became widespread in the geological literature. Deposits in the hills of the lesser Apennines of northern Italy were shown to characterize the Pliocene type assemblages during Lyell's time.

Pliocene marine sediments are commonly distributed along relatively restricted areas of the continental margins and in the deep-sea basins. Continental margin sediments are most often terrigenous and range from coarser-grained sandstone to finergrained mudstone and clay. Major rivers of the world, such as the Amazon, Indus, and Ganges, contain thick piles of Pliocene terrigenous sediments in their offshore fans. The Pliocene deep-sea sediments are carbonate-rich (commonly biogenic oozes) and are often very thick (up to 5000 m or 16,400 ft). *See* BASIN; CONTINENTAL MARGIN.

			Holocene
		QUATERNARY	Pleistocene
	CENOZOIC		Pliocene
			Miocene
		TERTIARY	Oligocene
			Eocene
			Paleocene
	MESOZOIC	CRETACEOUS	

				CALIFORNIA	JAPAN	AUSTRALIA	NEW ZEALAND
1.8 Ma			STANDARD	WHEELERIAN	KECHIENJIAN		NUKUMARUAN
		ATE	PIACENZIAN	VENTURIAN	SUCHIAN		MANGAPANIAN
	ENE	LA	(ASTIAN)	-		YATALAN	MANGAFANIAN
	E			REPETTIAN	TOTOMIAN		WAIPIPIAN
	PLIOCE	ARLY	ZANCLEAN	DELMONTIAN		KALIMNAN	OPOITIAN
5.0 Ma		EA	(TABIANIAN)	DELIVIONTIAN	YUIAN	CHELTENHAM- IAN	OFOTHAN
J.U IVIA				1			-

Fig. 1. Pliocene stages and their temporal equivalents in different regions.

Pliocene subdivisions. Modern stratigraphic usage subdivides the Pliocene Epoch into two standard stages, the lower, Zanclean stage and the upper, Piacenzian stage (**Fig. 1**).

In 1868 G. Seguenza proposed the Zanclean stage as the lower stage of the Pliocene, characterized by the deposits of Trubi Marl in Sicily that contained a diverse marine microfauna. The Zanclean is considered an equivalent of the Tabianian stage, introduced by K. Mayer-Eymer in that same year, with a type area in northern Italy. The Tabianian contained a less diverse and more restricted cooler-water fauna, less suitable for broader correlations, thus leading to the use of Zanclean as the standard stage for the lower Pliocene over the Tabianian. In 1973 M. Cita proposed a type section for the Zanclean in the Trubi Marl at Capo Rosello in Sicily.

Mayer-Eymer proposed the Piacenzian in 1858 with the type area in northern Italy. It was first proposed as a lower subdivision of the Astian stage that had been introduced by P. de Rouville in 1853 as an equivalent of Lyell's older Pliocene. The Piacenzian was later upgraded to a stage level by E. Renevier, with a type area designated near Castell' Arquato by M. Pareto in 1865. The sands and blue marls of type Piacenzian contain diverse benthic marine fauna as well as several planktonic foraminifer species. The equivalent interfingering strata of the Astian (type area near Asti, northern Italy) are largely calcareous sand, conglomerate, and marl, containing shallow-water benthos and only minor planktonic constituents. *See* MARL.

The Romanian and Dacian stages are sometimes used for the Pliocene in the Paratethys region (eastern and southeastern Europe) and are considered time-equivalent to Zanclean and Piacenzian, respectively. The Kimmerian and Akchagylian stages (Moldavian and Khaprovian substages for the latter) are often used as approximate equivalents for the lower and upper Pliocene in the former Soviet republics. Other regionally used stage terms of the Pliocene include the lower brackish-water Sarmatian stage and the upper nonmarine Pontian stage of the Vienna basin and the western Paratethys; the Piacenzian equivalent Redonian stage in Brittany; and the Anversian stage for the younger Pliocene in Belgium. In Japan, the Yuian, Totomian, Suchian, and younger part of the Kechienjian stages span the Pliocene, and in Australia it is subdivided into the Cheltenhamnian, Kalimnan, and Yatalan stages. New Zealand stratigraphers assign the Opoitian, Wapipian, Mangapanian, and the lower part of Nakumaruan stages to the Pliocene. In California, the regional Pliocene stages include the Delmontian, Repettian, Venturian, and the younger part of the Wheelarian.

Paleontologists working with vertebrate fauna often use their own age and stage terminology to characterize particular mammalian assemblages. In European Pliocene they identify the Ruscinian and the younger part of Villafranchian ages; while in North America they consider the oldest part of Hemphillian, the Blancan, and the youngest part of Irvingtonian ages as characteristic of the mammalian ages of the Pliocene.

Tectonics, seas, and climate. The most notable tectonic events in the Pliocene include the beginning of the third and last phase of the Himalayan uplift, the Attican orogeny that began in the late Miocene and continued into the Pliocene, and the Rhodanian and Walachian orogenies that occurred during the later Pliocene.

The latest Miocene is marked by a global cooling period that continued into the earliest Pliocene, and there is evidence that the East Antarctic ice sheet had reached the continental margins at this time. The Miocene Epoch terminated with the isolation of, and a salinity crisis in, the Mediterranean. At this time, the Mediterranean was cut off from the open ocean due to the emergence of the sill blocking the Gibraltar Straits, due to uplift or falling global sea level. The desiccated Mediterranean developed high salinities, and a large volume of evaporites was deposited. The onset of the Pliocene is marked by the reconnection with the Atlantic, when normal marine conditions returned to the Mediterranean. *See* SALINE EVAPORITES.

The global sea level had been falling through the late Miocene, and with the exception of a marked rise in the mid-Zanclean, the trend toward lowered sea levels continued through the Pliocene and Pleistocene. The mid-Zanclean sea-level rise (3.5–3 Ma) was also accompanied by a significant global warming event, which is evidenced by the paleontological and glacial records of both the Arctic and Antarctic regions and New Zealand. The oxygen isotopic data, which record the prevailing sea surface temperatures and total ice volume on the ice caps, show little variations in the Equatorial Pacific during the middle Pliocene. *See* GEOLOGIC THERMOMETRY.

By early Pliocene time, the major surface circulation patterns of the world ocean and the sources of supply of bottom waters were essentially similar to their modern counterparts. The only significant difference was the continuing connection of the Central Atlantic with the Pacific at the Isthmus of Panama, which did not close until the mid-Pliocene at about 3.2 Ma. The closure of the isthmus had a profound effect on the North Atlantic circulation patterns. From the marine oxygen isotopic record, the initiation of the Arctic glaciation has been dated at 3.5 Ma, but the continental record shows that land glaciation on North America did not start until 3.2 Ma. This time-lag effect may be linked to the closing of the Isthmus of Panama and the strengthening of the North Atlantic Gulf Stream, which would then be displaced farther north. Until the closure of the isthmus, the terrestrial and marine ice mass generation systems were not coupled. But the combined effect of the northward displacement of the Gulf Stream and the initial development of the Arctic ice cap, which resulted in increased availability of moisture to northern high latitudes and lowered temperatures, led to the development of the first terrestrial ice sheet. From then on, the terrestrial and marine ice mass systems became coupled. A significant ice cap on the Arctic region, however, did not develop until 2.5 Ma. In the Southern Hemisphere, two major glacial advances of the East Antarctic ice sheet have been documented for the periods 3.2-2.7 Ma and 2.1-1.6 Ma.

The third phase of the uplift of the Himalayas that began in the Pliocene accentuated the monsoontype circulation in the Arabian Sea and North Indian Ocean. The monsoonal system in this region had already been activated after the Miocene (second) phase of the Himalayan orogeny.

Thus, by mid-Pliocene time the conditions were in place that would lead to glacially dominated climates of the Quaternary Period.

Life. By Pliocene time, much of the marine and terrestrial biota had essentially evolved its modern characteristics. The late Pliocene cooling led to the expansion of cooler-water marine assemblages of the higher latitudes into lower latitudes, particularly the foraminifers, bivalves, and gastropods. At the onset of cooling, the warm-water-preferring calcareous nannoplankton group of discoasters began waning in the late Pliocene and became extinct at the close of the epoch. *See* BIVALVIA; COCCOLITHOPHORIDA; FORAMINIFERIDA; GASTROPODA.

The widespread grasslands of the Pliocene were conducive to the proliferation of mammals and increase in their average size. The mid-Zanclean sealevel rise led to the geographic isolation of many groups of mammals and the increase in endemism. But the late Pliocene-Pleistocene lowering of sea level facilitated land connections and allowed extensive mammalian migration between continents. The closing of the Panama isthmus in the late Pliocene also allowed interchanges between North and South America. Armadillos, capybaras, ground sloths, and porcupines, among others, migrated from south to north, and horses, elephants, cats, dogs, deer, and northern rodents went south. The arrival of the North American mammals led to increased competitive pressure and extinction of many typically South American groups, including the notoungulates and litopterns. Horses evolved and spread widely in the Pliocene (Fig. 2). The late Miocene North American horse genus, Hipparion, became extinct in the mid-Pliocene after spreading into the Old World during the latest Miocene and early Pliocene. The modern horse genus, Equus, had appeared before the demise of Hipparion. See MAMMALIA.

The Pliocene Epoch also saw the appearance of several hominid species that are considered to be directly related to modern human ancestry. The

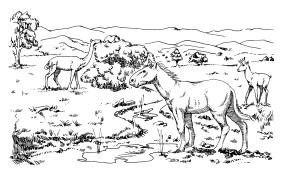


Fig. 2. Pliocene scene in the San Joaquin Valley, California, with ancestral species of camel (*Parocamelus*), horse (*Pliohippus*), and pronghorn (*Sphenophalos*). The spread of grasslands in the Pliocene was conducive to the proliferation of grazing mammalian species. (*After R. A. Stirton, Time, Life and Man, John Wiley*, 1959)

earliest hominid bones have been discovered from Baringo, Kenya, in sediments that are dated to be of earliest Pliocene age. After this first occurrence, a whole suite of australopithicine species made their appearance in the Pliocene. These include *Ardipithecus ramidus*; and *Australopithecus anamensis*, *A. afarensis*, *A. babrelgbazali*, *A. africanus*, *A. boisei*, and *A. robustus*. Before the beginning of the Pleistocene glaciation, the first *Homo* species, *Homo habilis*, had already evolved in Africa, first appearing around 2 Ma, and *Homo erectus* appeared on the scene near the Pliocene-Pleistocene boundary. Both of these species are considered to be the precursor of modern humans. *See* AUSTRALOPITHECINE; FOSSIL HUMANS; PALEOBOTANY; PALEONTOLOGY.

Bilal U. Haq

Bibliography. W. A. Berggren and J. A. van Couvering, The late Neogene, *Palaeogeog. Palaeoclimat. Palaeoecol.*, 16:1-216, 1974; B. U. Haq and F. W. B. van Eysinga, *Geological Time Table*, Elsevier, Amsterdam 1998.

Plum

Any of the smooth-skinned stone fruits grown on shrubs or small trees. Plums are widely distributed in all land areas of the North Temperate Zone, where many species and varieties are adapted to different climatic and soil conditions.

Types and varieties. There are four principal groups: (1) Domestica (*Prunus domestica*) of



Fig. 1. Branch of a Damson plum.



Fig. 2. Cluster of prunes decayed by brown rot. Fungus spore masses are on the fruit, and there is an exudation of gum at the base of the stems where the pathogen has entered the shoot. (*Courtesy of J. M. Ogawa*)

European or Southwest Asian origin, (2) Japanese or Salicina (*P. salicina*) of Chinese origin, (3) Insititia or Damson (*P. insititia*) of Eurasian origin (**Fig. 1**), and (4) American (*P. americana* and *P. hortulana*). The Domesticas are large, meaty, prune-type plums, the principal varieties being Agen or French prune (grown in California) and Italian or Italian hybrids (grown primarily in Oregon, Washington, Idaho, and Michigan).

A prune is a plum which dries without spoiling. Commonly but incorrectly the term "prune" is applied to all Domestica-type plums, whereas the

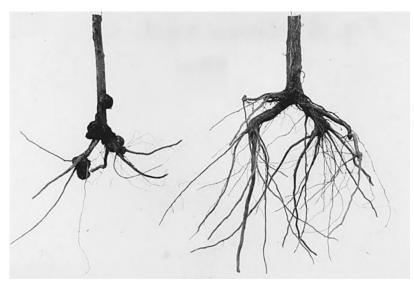


Fig. 3. Crown gall on plum rootstock at left. Tree at right was protected against infection by using the biocontrol agent Agrobacterium radiobacter. (Courtesy of W. J. Moller)

term "plum" is used only for Japanese-type plums. Japanese plums are typically round, reddish or yellow, and juicy, and are represented by such varieties as Beauty, Santa Rosa, Duarte, Burbank, Wickson, and Satsuma. Insititias (Damson), represented by Shropshire, are small, meaty fruits grown sparingly for jam. American plums are small, watery fruits of low quality, represented by De Soto, Pottawattomi, and Golden Beauty, and valued chiefly for cold hardiness of the tree.

Propagation and culture. Plums are propagated by budding on seedlings of the myrobalan plum (*P. cerasifera*) and less commonly on the peach and certain strains of *P. domestica. See* PEACH; PLANT PROPAGATION.

The Domestica and Institutia plums are slightly hardier than the peach; the Japanese plums have about the same hardiness as the peach; and native American plums are considerably hardier than the other groups.

Trees of the Domestica and Damson types may grow 20-25 ft (6-7.5 m) high unless kept lower by training and pruning. They are usually planted about 20 ft (6 m) apart, and are adapted to a wide range of soils, from light to heavy. Trees of the Japanese type are smaller (15-20 ft or 4.5-6 m) and are best adapted to lighter soils. They are normally planted 18-20 ft (5.4-6 m) apart, but in the hands of skilled horticulturists they are adaptable to close spacings and controlled shapes. Most plums require cross-pollination. *See* REPRODUCTION (PLANT).

Production and economic importance. California is the leading producer (90%) of the prune-type Domestica plums, followed by Oregon, Washington, Michigan, and Idaho. In states other than California, where 100% are dried as prunes, 52% of the prunetype plums are sold fresh, 40% canned, and 8% dried. R. Paul Larsen

Plum and prune diseases. Brown rot, a fungus disease caused by *Monilinia fructicola* and *M. laxa*, is probably the most widespread and serious disease of these crops (**Fig. 2**). Symptoms include blossom and twig blight and pre- and postharvest fruit rot. Injuries caused by the plum curculio in eastern states and twig borer in western states play an important role in fruit infection. Control of blossom blight is based on the application of a dormant eradicant spray and one or more protective sprays during bloom. Fruit decay can be prevented only by controlling the involved insects and the pathogens.

Other important fungus diseases include *Armillaria* root rot (caused by *A. mellea*), controlled mainly by soil fumigation and use of resistant root-stocks; black knot (caused by *Dibotryon morbosum*), a serious disease in eastern states that is controlled by gall removal and spraying; Valsa (Cytospora) canker (caused by *Valsa* ssp.), favored by freeze injury to the North and sunburning in the Southwest; and *Ceratocystis* canker (caused by *C. fimbriata*, an insect-vectored pathogen that invades mechanical-harvesting injuries).

Bacterial diseases of importance on plums and prunes are bacterial spot (caused by *Xanthomonas pruni*), bacterial canker (caused by *Pseudomonas* *syringae*), and crown gall caused by *Agrobacterium tumefaciens* (Fig. 3). Bacterial spot occurs only east of the Rocky Mountains, where it is especially damaging to Japanese plums. Bacterial canker is serious in the Pacific coast states on all types of plums and prunes. It is partially controlled through soil fumigation to kill nematodes that predispose trees to the disease, and by the use of peach rootstocks. Crown gall is biologically controlled in nurseries, where it is of greatest concern, by the use of *A. radiobacter*, which produces a bacteriostatic bacteriocin specific for the pathogen. *See* CROWN GALL.

Although plums and prunes are affected by several virus diseases—Prunus ring spot, line pattern, stem pitting, and diamond canker, to mention a few—these disorders mostly are of minor importance. The serious plum pox virus disease of Europe has not yet been found in North America. *See* FRUIT, TREE; PLANT PATHOLOGY. Harley English

Bibliography. W. H. Chandler, *Deciduous Orchards*, 1957; N. F. Childers, *Modern Fruit Science*, 1973; U.S. Department of Agriculture, *Virus Diseases and Noninfectious Disorders of Stone Fruits in North America*, Agr. Handb. 437, 1976; E. E. Wilson and J. M. Ogawa, *Fungal, Bacterial, and Certain Nonparasitic Disease of Fruit and Nut Crops in California*, Division of Agricultural Sciences, University of California, 1979.

Plumbaginales

An order of flowering plants, division Magnoliophyta (Angiospermae), in the subclass Caryophyllidae of the class Magnoliopsida (dicotyledons). The order consists of only the family Plumbaginaceae, with about 400 species. The plants are herbs or less often shrubs. The flowers are strictly pentamerous (that is, each floral whorl has five members); the petals are fused (sympetalous condition) and all alike in shape and size; the pollen is trinucleate; and there is a single basal ovule located in a compound ovary that has a single locule or cell. The Plumbaginaceae differ from most families of their subclass in their straight embryo, copious endosperm, absence of perisperm, and the presence of anthocyanin pigments instead of betalains. The family contains a few garden ornamentals, such as species of Armeria, known as thrift or sea-pink. See CARYOPHYLLIDAE; MAGNOLIO-PHYTA; MAGNOLIOPSIDA; PLANT KINGDOM; POLLEN. Arthur Cronquist; T. M. Barkley

Pluto

Formerly considered the outermost planet, tiny Pluto is now recognized to be one of the largest and closest members of a disk of icy planetesimals that surrounds the solar system beyond the orbit of Neptune. For this reason, the International Astronomical Union resolved to reclassify Pluto in a new category, dwarf planet, in 2006.

Pluto was discovered on February 18, 1930, by C. W. Tombaugh at the Lowell Observatory, Flagstaff, Arizona, on photographic plates taken with a special astronomical camera as part of a systematic search. The presence of a planet beyond Neptune's orbit had been predicted independently by P. Lowell and W. H. Pickering (among others) on the basis of analyses similar to those that had led U. J. Leverrier and J. C. Adams to the prediction of Neptune. It was thought that there were perturbations in the motion of Uranus that Neptune alone could not explain. Pluto was found surprisingly near its predicted position, since it is now known that the mass of Pluto is far too small to have caused the (evidently spurious) perturbations. *See* NEPTUNE; PLANET.

Charon, Pluto's largest satellite, was discovered in 1978 by J. W. Christy and R. Harrington at the nearby Flagstaff station of the U.S. Naval Observatory. H. Weaver and a team of astronomers using the *Hubble Space Telescope* discovered two additional satellites, Hydra and Nix, in 2005 (**Fig. 1**).

Orbit. Pluto's orbit (**Fig. 2**) has a semimajor axis (mean distance to the Sun) of 5.9×10^9 km (3.7×10^9 mi), an eccentricity of 0.25, and an inclination of 17.1° . The inclination and eccentricity are larger than those of any of the planets. At a mean orbital velocity of 2.95 mi/s (4.7 km/s), it takes Pluto 248.0 years to make one revolution around the Sun. The large orbital eccentricity means that at perihelion Pluto is closer to the Sun than Neptune, but the orbits of the two bodies do not intersect.

Telescopic appearance. Pluto is visible only through fairly large telescopes since even at perihelion its visual magnitude is fainter than 13.5. This is thousands of times below the sensitivity of the unaided eye. Its visual magnitude at mean opposition is 15.1, thousands of times fainter than our visual threshold. Periodic variations in its brightness demonstrate that the surface of Pluto is covered with bright and dark markings and indicate a period of rotation of 6.4 days. This is longer than the rotational period of any of the planets except Mercury and Venus. *See* MAGNITUDE (ASTRONOMY).

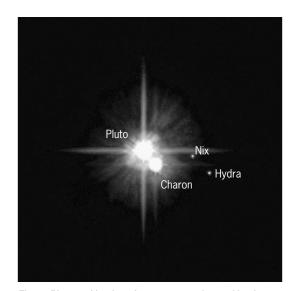


Fig. 1. Pluto and its three known moons, imaged by the Hubble Space Telescope on February 15, 2006. (NASA, ESA, H. Weaver (JHUAPL), A. Stern (SwRI), HST Pluto Companion Search Team)

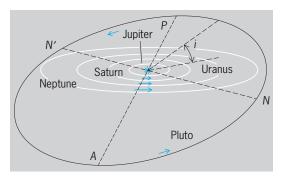


Fig. 2. Orbit of Pluto; a perspective view to show the inclination *I* and eccentricity of the orbit. *A*, aphelion; *P*, perihelion; *NN'*, line of nodes. (*After L. Rudaux and G. de Vaucouleurs, Astronomie, Larousse, 1948*)

Size and mass. The radius of Pluto is 1153 ± 10 km $(716\pm6 \text{ mi})$, making Pluto significantly smaller than the Earth's Moon (whose radius is 1738 km or 1080 mi), while Charon's radius is 603.5±1.5 km (375.0±0.9 mi). (The uncertainty in Pluto's radius includes the possibility that a lower atmosphere haze prevents the true surface of this dwarf planet from being measured by the occulation observation on which it is based.) Charon is thus about half the size of Pluto itself, making the most closely matched pair in the solar system. The density of Pluto is 2.0 g/cm⁻³, while that of Charon is 1.65 g/cm⁻³. These densities suggest compositions of ice and rock within the range exhibited by Saturn's regular satellites. Nix and Hydra are too small to permit determinations of size and mass.

Surface and atmosphere. The near-infrared spectrum of Pluto reveals absorption features of solid nitrogen (N₂), carbon monoxide (CO), and methane (CH₄) ice. From these absorptions it is possible to conclude that nitrogen is the dominant ice on Pluto's surface (some 20-100 times more abundant than the other ices), which means it must also be the major constituent of the planet's tenuous atmosphere. The nitrogen absorption in the spectrum indicates a surface temperature near 38 K (-391°F), consistent with an atmospheric surface pressure roughly 10^{-5} times that on Earth. Such a pressure was deduced directly from observations of a stellar occultation in 1988, which argues against the possible presence of a thick haze in Pluto's lower atmosphere. (If such a haze exists, it could affect the radius measurement and hence the deduction of the planet's mean density.) Its size, surface temperature, composition, and thin nitrogen atmosphere make Pluto very similar to Triton, Neptune's largest satellite. However,

there are important differences such as the absence of frozen carbon dioxide (CO_2) on Pluto. *See* OCCUL-TATION.

Seasons on Pluto are modulated by the large eccentricity of the orbit. Pluto reached perihelion in 1989, which brought it closer to the Sun than Neptune. This means that during the 1988 occultation Pluto was observed at the warmest time in its seasonal cycle, when the atmospheric pressure might be expected to be at its maximum value.

However, observations of another stellar occultation by Pluto in 2002 revealed that the atmosphere was actually twice as massive as it had been in 1988. This result had been anticipated by some scientists, who had predicted that the changing orientation of Pluto as it moved around its orbit would expose polar deposits of frozen atmospheric gases to the Sun, if such deposits existed. The atmospheric pressure would then increase as the polar deposits sublimed. The same seasonal change in atmospheric pressure is observed on Mars, where the dominant gas is carbon dioxide (CO₂), which condenses out at the winter polar cap. On Pluto, this seasonal effect will soon be overwhelmed by the decrease in temperature as the dwarf planet moves farther from the Sun in its eccentric orbit, and the atmospheric pressure will gradually diminish as nitrogen again freezes out on Pluto's surface. Thus, to observe these processes it would be advantageous to explore the planet with a spacecraft as soon as possible, and such a mission is underway. Another stellar occultation, in 2006, found that the atmosphere had not changed further since 2002. See MARS.

Satellites. Pluto has three small satellites: Charon, Nix, and Hydra, with Charon far larger than the other two (see table; Fig. 3). Demonstrating another striking difference with Pluto, Charon apparently has little if any frozen methane on its surface, which instead exhibits the spectral signature of water ice. The orbit of Charon is unique in the solar system in that the satellite's period of revolution is identical to the rotational period of the planet. Thus an inhabitant of Pluto who lived on the appropriate hemisphere would see Charon hanging motionless in the sky. From the opposite hemisphere the satellite would be invisible. (On Earth, artificial satellites are put into such geostationary orbits for communications and television broadcasting.) See COMMUNI-CATIONS SATELLITE; SATELLITE (ASTRONOMY); SATEL-LITE (SPACECRAFT).

Pluto and the Kuiper Belt. In 1992, the first of what is proving to be a huge number of icy planetesimals

Name	Orbit radius, km (mi)	Period, days	Eccentricity	Inclination*	Radius, km (mi)	Magnitude at mean opposition
Charon	19,571 (12,161)	6.4	~0	\sim 0 $^{\circ}$	603.5±1.5 (375.0±0.9)	17.3
Nix	48,675 (30,245)	24.9	\sim 0	\sim 0 $^{\circ}$	~45 (28)?†	24.6
Hydra	64,780 (40,252)	38.2	\sim 0	\sim 0 $^{\circ}$	~45 (28)?†	24.4

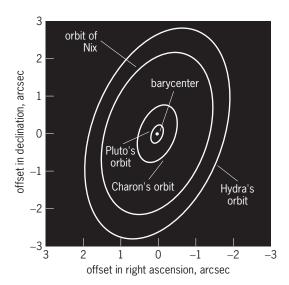


Fig. 3. Orbits of four bodies in the Pluto system (Pluto and its satellites, Charon, Nix, and Hydra) about the barycenter (center of mass) of the system, as seen from Earth. (NASA; ESA; W. J. Merline, Southwest Research Institute; Pluto Companion Search Team)

(200 km or 125 mi in diameter or less) was discovered orbiting the Sun at distances comparable to or greater than the orbit of Pluto. These are members of the Kuiper Belt, a disk-shaped distribution of material left over from the formation of the solar system. These objects appear to be representative of the solid components required to initiate the growth of giant planets. Held in the cosmic deep freeze of trans-Neptunian space, they must be among the most primitive, unprocessed objects still circulating in the solar system. Collisional fragments occasionally make their way into the inner solar system, where they become the short-period comets. In the course of this journey, they lose some of their most volatile constituents but still represent a unique resource for studying the early days of the solar system. The European Space Agency (ESA) mission Rosetta is scheduled to carry out a detailed investigation of a comet in 2012. See COMET; SOLAR SYSTEM; SPACE PROBE.

The existence of these small bodies (Kuiper Belt objects) finally provides a proper context for Pluto itself. Because of its small size and the high inclination and eccentricity of its orbit, Pluto has been an "ugly duckling" among the planets. As one of the largest members of the Kuiper Belt, however, it becomes a majestic "swan," swimming serenely at the boundary between the planets and this disk of debris left over from the formation of the giant planets.

By the middle of 2006, astronomers had discovered six Kuiper Belt objects with diameters of approximately 1000 km (625 mi), comparable in size to Charon, and at least one, Eris, that may be slightly larger than Pluto. They have also found that several of the Kuiper Belt objects, including Eris, have satellites. Of the larger bodies, Sedna and Eris are particularly interesting. The diameter of Sedna is 1000–1500 km (625–940 mi), and is still poorly known as it depends on an assumed reflectivity. (The more reflective the surface, the smaller the object needs to be to produce the same brightness.) Sedna's orbit has a very high eccentricity of 0.85 and an inclination of 12° , but it is the perihelion distance of 76 astronomical units that is most remarkable: Sedna's closest approach to the Sun is more than twice as distant as that of Pluto and is the farthest of any known Kuiper Belt object. This orbit is well beyond the gravitational influence of Neptune, and raises the questions of how Sedna got there and whether there are still more large bodies in such distant orbits. Eris is indeed larger, but its orbit lies somewhat closer to the Sun. Current models of solar system formation do not allow enough material at these enormous distances to form objects the size of these two Kuiper Belt objects, so they must have been scattered from closer orbits. Perhaps other large icy bodies suffered the same fate and ongoing searches will reveal them.

In 2003, the National Aeronautics and Space Administration (NASA) authorized the *New Horizons* Pluto-Kuiper Belt mission. A single spacecraft was launched on January 19, 2006, and is scheduled to fly past Pluto and Charon in 2015, making measurements and recording images. The spacecraft will then venture forth into the Kuiper Belt and attempt to fly past one or more of these small distant worlds. *See* KUIPER BELT. Tobias C. Owen

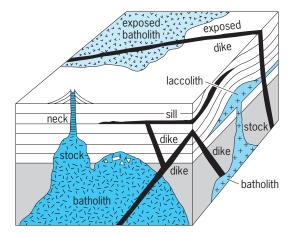
Bibliography. J. K. Beatty, C. C. Petersen, and A. Chaikin (eds.), *The New Solar System*, 4th ed., Cambridge University Press, 1999; W. G. Hoyt, *Planets X and Pluto*, University of Arizona Press, 1980; D. Morrison and T. Owen, *The Planetary System*, 3d ed., Addison Wesley, 2003; S. A. Stern and J. Mitton, *Pluto and Charon: Ice Worlds on the Ragged Edge of the Solar System*, Wiley, 1998; S. A. Stern and D. J. Tholen (eds.), *Pluto and Charon*, University of Arizona Press, 1997; D. A. Weintraub, *Is Pluto a Planet?: A Historical Journey Through the Solar System*, Princeton University Press, 2006.

Pluton

A solid rock body that formed by cooling and crystallization of molten rock (magma) within the Earth. Most plutons, or plutonic bodies, are regarded as the product of crystallization of magma intruded into surrounding "country rocks" within the Earth (principally within the crust). During the middle part of the twentieth century some geologists believed that many igneous or plutonic rock bodies were the result of ultrametamorphism (granitization). However, that model has largely been abandoned.

Igneous rock bodies are referred to generally as either extrusive or volcanic on one hand, or as intrusive or plutonic on the other, although the term volcanic is sometimes also used to refer to small, shallow intrusive bodies associated with volcanoes. *See* IGNEOUS ROCKS; MAGMA.

Types. Plutons occur in a nearly infinite variety of shapes and sizes, so that definition of types is arbitrary in many cases. In general, two modes of emplacement can be recognized with regard to the country rock. Concordant plutons are intruded between layers of stratified rock, whereas the more



Schematic block diagram showing the occurrence of the more common types of plutons.

common discordant plutons are characterized by boundaries that cut across preexisting structures or layers in the country rock. The principal types of concordant plutons are sills, laccoliths, and lopoliths; the principal types of discordant plutons are dikes, volcanic necks or plugs, stocks, and batholiths. In complex plutons, characteristics of more than one type may be found from place to place. The more common types are shown in the **illustration**.

Dike. A dike is a tabular, discordant pluton. Dikes can range in size from a few millimeters wide by a few centimeters long to major structures several hundred meters wide and several hundred kilometers long. Many dikes tend to be sheetlike and nearly planar, whereas others can follow complex pathways through zones of weakness in country rocks. The concentric fracture systems found in many volcanic regions can also give rise to ring dikes.

Sill. While they are also tabular plutons, sills are by definition concordant with the stratification or layering in the surrounding country rocks. True sills are much less common than dikes.

Laccolith. Larger concordant plutons, particularly more viscous (granitic) ones, may tend to bulge upward, creating a domal shape. Such bodies are called laccoliths and may show well-defined feeder dikes or sills. Laccoliths tend to be several kilometers in diameter and several hundred meters thick in the center (1 km = 0.6 mi; 1 m = 3.3 ft).

Lopolith. A second type of large concordant pluton is the lopolith, which is characterized by a saucerlike or funnellike shape. They are associated primarily with large concordant intrusions of mafic (gabbroic) rock, and true lopoliths are relatively rare.

Plug (neck). The central conduits to most volcanoes become filled, generally, with solidified magma, and the resulting pipelike igneous bodies are referred to as volcanic plugs or necks. They tend to possess steep, discordant sides and roughly equidimensional horizontal sections; they may range from several meters to hundreds of meters in diameter.

Batholiths and stocks. A complete range exists in sizes of discordant plutons, and the designation of a particular pluton as stock or batholith may depend upon

degree of exposure due to uplift and erosion. Plutons smaller than $40 \text{ mi}^2 (100 \text{ km}^2)$ and 6 mi (10 km)across are referred to generally as stocks, and larger ones are referred to as batholiths. Batholiths may be considerably larger than 40 mi² (100 km²), and the term batholith has also come to refer to very large plutonic masses (up to several hundred kilometers in largest dimension) that are made up of many smaller, compositionally distinct individual plutons (stocks or batholiths). Batholiths do not extend indefinitely into the Earth, but possess a finite thickness, which ranges from 6 mi (10 km) or less to 18 mi (30 km). In some cases they show well-defined lower boundaries (floors), but in others they may gradually grade downward into root zones where the original magma was derived by melting of older crustal rocks. Unfortunately the deep zones of most batholiths are not exposed, so details of their deep structure must be inferred from petrologic and geophysical studies.

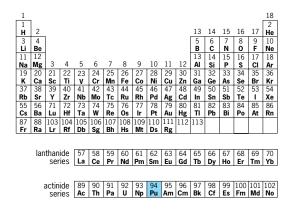
Origin. The modern view of most plutons is that they are the result of intrusion of molten rock (magma) into surrounding country rock, with subsequent cooling of the magma to form plutonic igneous rock. Several mechanisms of intrusion are known or proposed. The most simple ones, pertaining to smaller plutonic bodies, are forceful injection or passive migration into fractures. Larger plutons may form by several processes. For example, less dense magma may migrate upward along a myriad of channelways to accumulate as a large molten body within the upper crust. Further migration could occur by forceful injection, by stoping (a process where the magma rises as blocks of the roof of the magma chamber break off and sink), and by diapiric rise, where country rocks flow around the upward-moving magma body. The variations and interactions of such processes are complex, so that no two large plutons formed in exactly the same way. See PETROLOGY. W. Randall Van Schmus

Bibliography. D. S. Barker, *Igneous Rocks*, 1983; M. G. Best, *Igneous and Metamorphic Petrology*, 1982; H. Blatt and R. Tracy, *Petrology (Igneous, Sedimentary, and Metamorphic)*, 2d ed., 1996; E. G. Ehlers and H. Blatt, *Petrology: Igneous, Sedimentary, and Metamorphic*, 1982; A. Hall, *Igneous Petrology*, 2d ed., 1996; F. Press and R. Siever, *Earth*, 4th ed., 1986; F. Press and R. Siever, *Understanding Earth*, 3d ed., 2001.

Plutonium

A chemical element, Pu, atomic number 94. Plutonium is a reactive, silvery metal in the actinide series of element elements. The principal isotope of chemical interest is ²³⁹Pu, with a half-life of 24,131 years. It is formed in nuclear reactors by the process shown in the following reaction. Plutonium-239 is fissionable,

$$^{238}\text{U} + n \longrightarrow ^{239}\text{U} \xrightarrow{\beta^-} ^{239}\text{Np} \xrightarrow{\beta^-} ^{239}\text{Pu} \xrightarrow{\beta^-} ^{239}\text{Pu}$$



but may also capture neutrons to form higher plutonium isotopes. *See* PERIODIC TABLE.

Plutonium-238, with a half-life of 87.7 years, is utilized in heat sources for space application, and has been used for heart pacemakers. Plutonium-239 is used as a nuclear fuel, in the production of radioactive isotopes for research, and as the fissile agent in nuclear weapons.

Plutonium exhibits a variety of valence states in solution and in the solid state. Plutonium metal is highly electropositive. Numerous alloys of plutonium have been prepared, and a large number of intermetallic compounds have been characterized.

Reaction of the metal with hydrogen yields two hydrides. The hydrides are formed at temperatures as low as $150^{\circ}C$ ($300^{\circ}F$). Their decomposition above $750^{\circ}C$ ($1400^{\circ}F$) may be used to prepare reactive plutonium powder. The most common oxide is PuO₂, which is formed by ignition of hydroxides, oxalates, peroxides, and nitrates of any oxidation state in air of $870-1200^{\circ}C$ ($1600-2200^{\circ}F$). A very important class of plutonium compounds are the halides and oxyhalides. Plutonium hexafluoride, the most volatile plutonium compound known, is a strong fluorinating agent. A number of other binary compounds are known. Among these are the carbides, silicides, sulfides, and selenides, which are of particular interest because of their refractory nature.

Because of its radiotoxicity, plutonium and its compounds require special handling techniques to prevent ingestion or inhalation. Therefore, all work with plutonium and its compounds must be carried out inside glove boxes. For work with plutonium and its alloys, which are attacked by moisture and by atmospheric gases, these boxes may be filled with helium or argon. *See* ACTINIDE ELEMENTS; NEPTUNIUM; NU-CLEAR CHEMISTRY; TRANSURANIUM ELEMENTS; URA-NIUM. Fritz Weigel

Bibliography. W. T. Carnall and G. R. Choppin (eds.), *Plutonium Chemistry*, 1983; R. Guillaumont et al. (eds.), *Update on the Chemical Thermodynamics of Uranium*, *Neptunium*, *Plutonium*, *Americium and Technetium*, 2003; S. Hofmann, *On Beyond Uranium: Journey to the End of the Periodic Table*, 2002; J. J. Katz, G. T. Seaborg, and L. R. Morss (eds.), *The Chemistry of the Actinide Elements*, 2 vols., 2d ed., 1986.

Plywood

A wood product in which thin sheets of wood are glued together, grains of adjacent sheets being at right angles to each other in the principal plane. Because of this cross-grained orientation, mechanical properties are less directional than those of natural lumber and more dimensionally stable. Originally developed to replace wide wood boards that had been sawed directly from large logs, plywood was manufactured from logs not suitable for other purposes. It has since found such widespread use that tree farms are now cultivated specifically to yield logs suitable for processing into sheets for plywood. The softwood plywood capacity in the United States in 1997 was 12.5 million cubic meters. Southern pine represents the largest segment of this market.

Grades. The American Plywood Association identifies several grades of product. Plywood is designated group 1 when made from northern-grown Douglas-fir, western larch, and such southern pines as loblolly and longleaf, or other woods noted for their strength. Plywoods in groups 2, 3, and 4 are made from woods of successively lower strengths. Consequently, group 1 plywood offers the greatest stiffness, group 4 the least.

Plywood with waterproof glue is designated exterior type; it is also used interiorly where moisture is present. Plywood with nonmoisture-resistant glue is designated interior type; it can withstand an occasional soaking but neither repeated soakings nor continuous high humidity.

Veneer grades A through D extend from a smooth surface to a surface with occasional knotholes and limited splits. If the outer face of the plywood is cut from only heartwood or sapwood, free from open defects, the plywood is assigned veneer grade N, indicating that it will take a natural finish. *See* STEM.

Most commonly used plywoods are 1/4-in. (0.6-cm) sanded interior paneling or 1/2-in. (1.3-cm) exterior grade plywood sheeting. Other standard thicknesses extend to 1 in. (2.5 cm) for interior types and to $1^{1}/_{8}$ in. (2.8 cm) for exterior types. The most common panel size is 4×8 ft (1.2 × 2.4 m); larger sizes are manufactured for such special purposes as boat hulls.

Manufacture. In modern plywood plants, as described in *Forest Industries* magazine, logs are debarked and graded, and those of suitable diameter are trimmed and delivered to a lathe where they are turned into plys of required thickness. Automatic handling machines then stack the sheets into sandwiches, glue being spread as the sheets are moved into position. If a thin hardwood veneer face is called for, it may be conveyed by vacuum rollers. The sandwiches are either stacked in a hot press or placed in a multiopening press for adhesive curing. *See* VENEER.

Finished plywood may be unsanded, sanded, or overlayed with several types of coatings for decorative and specialty uses.

If other than a flat board is required, the plywood is built up ply by ply over or inside a form and molded to the finished shape under heat and pressure as required by the glue that is used.

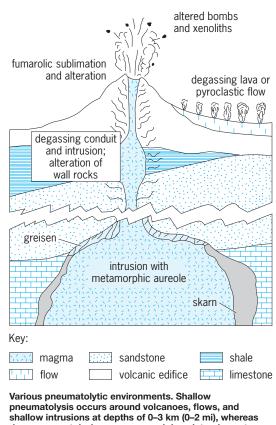
Plywood in appropriate grades is used in many different applications, such as furniture, wall facings, shelving, containers, crates, fences, forms, subflooring, and roof decking. *See* WOOD PRODUCTS. Frank H. Rockett; Roger M. Rowell

Bibliography. H. N. Spelter, D. B. McKeever, and I. A. Durbak, *Review of Wood-based Panel Sector in United States and Canada*, Gen. Tech. Rep. FPL-GTR-99, USDA Forest Service, Forest Products Laboratory, 1997.

Pneumatolysis

The alteration of rocks or crystallization of minerals by gases or supercritical fluids (generically termed magmatic fluids) derived from solidifying magma. At surface conditions, magmatic fluids sampled from active volcanoes contain steam with lesser amounts of carbon dioxide, sulfur dioxide, hydrogen sulfide, hydrogen chloride, and hydrogen fluoride, and trace amounts of many other volatile constituents. Magmatic fluids may contain relatively high concentrations of light and heavy elements, particularly metals, that do not crystallize readily in common rockforming silicates constituting most of the solidifying magma; thus, valuable rare minerals and ores are sometimes deposited in rocks subjected to pneumatolysis. Magmatic fluids are acidic and may react extensively with rocks in the volcanic edifice or with wall rocks surrounding intrusions. Penetration of magmatic fluids into adjacent rocks is greatly aided by faults, fractures, and cracks developed during intrusion and eruption or created by earlier geologic events. See MAGMA; METAMORPHISM; METASO-MATISM; ORE AND MINERAL DEPOSITS; VOLCANO.

Pneumatolysis describes specific mechanisms of mineral deposition, hydrothermal alteration, or metasomatism in which magmatic fluids play an extremely significant role (see illus.). For example, lavas and ejecta at volcanoes may contain blocks (xenoliths) of wall rock that react with magmatic fluids to form pneumatolytic minerals such as vesuvianite (idocrase). Many silicic lavas and pyroclastic flows, particularly rhyolites, contain pneumatolytic tridymite, cristobalite, alkali feldspar, and rarely, topaz, garnet, rutile, fluorite, cassiterite, hematite, among others. Gases streaming from volcanic fumaroles deposit sublimates of sulfur, sulfates, chlorides, fluorides, and oxides of many metals. Wall rocks surrounding volcanic conduits may be thoroughly altered to mixtures of quartz, alunite, anhydrite, pyrite, diaspore, kaolin, as well as other minerals by acidic fluids degassed from magma. Rarely, gold, silver, base-metal sulfides, arsenides, and tellurides are deposited by the fluids, making valuable ores. Physical conditions during shallow pneumatolysis range from 100 to 1000°C (200 to 1850°F) at relatively low pressure (≤ 1 kbar or $\leq 14,000$ lb/in.²), depending on eruption temperature, cooling history, depth to magma, and wall rock configuration. In such



shallow intrusions at depths of 0–3 km (0–2 mi), where deep pneumatolysis occurs around deep intrusions at depths of 3–15 km (2–10 mi).

environments, magmatic fluids are commonly subcritical, and near-surface ground waters and precipitation usually mix with degassing magmatic fluids. *See* LAVA; METASOMATISM; PYROCLASTIC ROCKS; RHY-OLITE; SUBLIMATION; VESUVIANITE; XENOLITH.

When magma bodies intrude rocks at greater depths, lithostatic pressures are high (about 1-5 kbar or 14,000-70,000 lb/in.2), and magmatic temperatures (650-1200°C or 1200-2200°F) are sustained for long periods of time ($\leq 10^6$ years). Released fluids are supercritical brines. Wall rocks undergo contact metamorphism, and a metamorphic aureole surrounds the intrusion. If highly reactive rocks such as limestones and dolomites are intruded, replacement deposits called skarns are formed that contain quartz, calcite, and a variety of iron, magnesium, and calcium silicates and oxides. Pneumatolytic minerals including ores of precious and base metals may be deposited from magmatic fluids during contact metamorphism. Silicic intrusions, especially granites, release substantial quantities of residual magmatic fluids that may alter wall rocks to form banded deposits and veinlike bodies called greisens. These deposits contain quartz, lithium-rich micas, and topaz, usually with accessory tourmaline, axinite, fluorite, apatite, cassiterite, wolframite, and others. An intrusion may undergo similar pneumatolytic alteration near contacts, displaying pegmatite-like mineralization. In these environments, pneumatolysis generally occurs at 400-600°C (750-1100°F). See CONTACT AUREOLE; GRANITE; GREISEN; PEGMATITE; SKARN. Fraser Goff

Bibliography. M. R. Carroll and J. R. Holloway (eds.), Volatiles in Magmas, Reviews in Mineralogy, vol. 30, Mineralogical Society of America, 1994; J. S. Cline and R. J. Bodnar, Direct evolution of brine from a crystallizing silicic melt at the Questa, New Mexico, molybdenum deposit, Econ. Geol., 89:1780-1802, 1994; W. D. Duffield et al., Origin of rhyolite-hosted tin mineralization: Evidence from the Taylor Creek rhyolite, New Mexico, Econ. Geol., 85:392-398, 1990; F. Goff et al., Gold degassing and deposition at Galeras Volcano, Colombia, GSA Today, September 1994; D. M. Kerrick (ed.), Contact Metamorphism, Reviews in Mineralogy, vol. 26, Mineralogical Society of America, 1991; H. Sigurdsson et al. (eds.), Encyclopedia of Volcanoes, Academic Press, 2000.

Pneumonia

An acute or chronic inflammatory disease of the lungs. More specifically when inflammation is caused by an infectious agent, the condition is called pneumonia; and when the inflammatory process in the lung is not related to an infectious organism, it is called pneumonitis.

An estimated 45 million cases of infectious pneumonia occur annually in the United States, with up to 50,000 deaths directly attributable to it. Pneumonia is a common immediate cause of death in persons with a variety of underlying diseases. With the use of immunosuppressive and chemotherapeutic agents for treating transplant and cancer patients, pneumonia caused by infectious agents that usually do not cause infections in healthy persons (that is, pneumonia as an opportunistic infection) has become commonplace. Moreover, individuals with acquired immune deficiency syndrome (AIDS) usually die from an opportunistic infection, such as pneumocystis pneumonia or cytomegalovirus pneumonia. Concurrent with the variable and expanding etiology of pneumonia and the more frequent occurrence of opportunistic infections is the development of new antibiotics and other drugs used in the treatment of pneumonia. See ACQUIRED IMMUNE DEFICIENCY SYN-DROME (AIDS); OPPORTUNISTIC INFECTIONS.

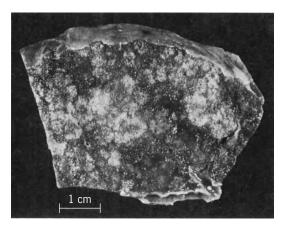
The incidence of noninfectious inflammatory conditions of the lung also seems to be increasing, or at least they are more frequently recognized. The number of drugs capable of inducing pneumonitis appears to be increasing as well. For example, cocaine derivatives have been shown to cause an inflammatory condition of the lung that may sometimes result in death.

Infectious pneumonia. Bacteria, as a group, are the most common cause of infectious pneumonia, although influenza virus has replaced *Streptococcus pneumoniae* (*Diplococcus pneumoniae*) as the most common single agent. Common bacterial agents causing pneumonia are *Streptococcus pneumoniae* (diplococci), other species of *Streptococcus. Staphylococcus* species, *Hemophilus influenzae, Klebsiella* species, *Pseudomonas aerugi-* nosa, Legionella pneumophila, Mycobacterium tuberculosis, and Mycoplasma pneumoniae. Some of these bacteria are normal inhabitants of the body and proliferate to cause disease only under certain conditions. Other bacteria are contaminants of food or water. *See* STEPTOCOCCUS PENUMONIAE (PNEUMO-COCCUS).

Most bacteria cause one of two main morphologic forms of inflammation in the lung. Streptococcus pneumoniae causes a type of pneumonia referred to as lobar pneumonia, in which an entire lobe of a lung or a large portion of a lobe, usually the lower lobe, becomes consolidated (firm, dense) and nonfunctional secondary to an influx of fluid and acute inflammatory cells (polymorphonuclear leukocytes) that represent a reaction to the bacteria. This type of pneumonia is uncommon today, usually occurring in people who have poor hygiene and are debilitated. If lobar pneumonia is treated adequately, the inflammatory process may entirely disappear, although in some instances it undergoes a process called organization, in which the inflammatory tissue changes into fibrous tissue. This results in a permanent obliteration of the normal lung architecture, usually rendering that portion of the lung nonfunctional.

The other morphologic form of pneumonia, which is caused by the majority of bacteria, is called bronchopneumonia. In this form there is patchy consolidation of lung tissue, usually around the small bronchi and bronchioles, again most frequent in the lower lobes (see **illus**.). Bronchopneumonia usually begins as an infection in the lumina of the small airways of the lower lobe and progresses to involve the surrounding alveoli, with an associated acute inflammatory response. This type of pneumonia may also undergo complete resolution if there is adequate treatment, although rarely it organizes.

Among the many viruses known to cause pneumonia are influenza A, influenza B, adenovirus, respiratory syncytial virus, cytomegalovirus, and herpes simplex virus. Viral pneumonia is usually a diffuse process throughout the lung and produces a different type of inflammatory reaction than is seen



Representative region of lung tissue from the right lower lobe. The nodular whitish areas represent acute bronchopneumonia and appear white due to the influx of white blood cells, which represent an inflammatory reaction to the bacteria.

in bronchopneumonia or lobar pneumonia. Mycoplasma pneumonia, caused by *Mycoplasma pneumoniae*, is referred to as primary atypical pneumonia and causes an inflammatory reaction similar to that of viral pneumonia. Mycoplasma pneumonia usually occurs in families and groups and may become chronic and difficult to eradicate.

Fungal organisms relatively frequently causing pneumonia include *Coccidioides immitus*, *Cryptococcus neoformans*, and *Histoplasma capsulatum*. Pneumonia can be caused by a variety of other fungal organisms, especially in debilitated persons such as those with cancer or AIDS. *Mycobacterium tuberculosis*, the causative agent of pulmonary tuberculosis, produces an inflammatory reaction similar to fungal organisms. *See* MYCOBACTERIAL DISEASES; TUBERCU-LOSIS.

Legionella pneumonia. Legionella pneumonia, initially called Legionnaire's disease, is caused by bacteria of the genus Legionella. The condition is frequently referred to under the broader name of legionellosis. Although the illness was originally described as being caused by Legionella pneumophila, the organism that was isolated from patients who succumbed during the 1976 American Legion convention in Philadelphia, many other Legionella species are now recognized. Legionellosis is not uncommon, and two clinical syndromes are manifested: an acute pneumonia with a high mortality rate, and a self-limited flulike illness. Legionella pneumonia may also occur in compromised hosts as an opportunistic infection. The disease may be preceded by several days of fever, chills, malaise, and muscle aches, and it may be associated with nonpulmonary symptoms such as diarrhea and encephalopathy (brain dysfunction). The legionella organism is a coccobacillus that does not strain easily with the conventional (Gram) stain used to identify bacteria, and requires a special type of medium to grow in culture. See LEGIONNAIRES' DISEASE.

Pneumocystis pneumonia. Pneumocystis pneumonia is caused by Pneumocystis carinii, a putative protozoan parasite. Before the AIDS epidemic, this organism was a rare cause of pneumonia, occurring mainly in infants and young children and also in immunocompromised cancer patients. Between October 1980 and May 1981, five men in California were treated for pneumocystis pneumonia, and all five also had cytomegalovirus pneumonia. They were subsequently found to be immunodeficient, and pneumocystis pneumonia was quickly recognized as the primary cause of death in patients with AIDS. The pneumocystis organism, which is about the size of a red blood cell, causes a rapid consolidation of the lung. It is best identified with a silver stain in tissue sections or in special preparations made of fluid obtained from the lungs. The pneumonia can be effectively treated with erythromycin and pentamidine

Hypersensitivity pneumonitis. Some persons are allergic to certain organic antigens, often the spores of various types of bacteria. When these antigens are inhaled into the lungs of susceptible persons,

an inflammatory reaction occurs that can progress from acute to a more chronic disease. The most frequently recognized hypersensitivity pneumonitis is called farmer's lung: susceptible farmers or other persons dealing with such things as moldy hay inhale the spores of the bacteria *Thermoactinomyces vulgaris*. About 4–6 h later, they develop fever, cough, shortness of breath, and chest pain. X-ray shows a patchy infiltrate in their lungs. The disease can rapidly clear when the person is removed from the offending antigen or is treated with cortisone-type medicines.

Drug-induced pneumonitis. Many drugs can induce inflammation in the lungs. The mechanisms by which various drugs cause pneumonitis can either depend on the dose or be secondary to a hypersensitivity reaction. Alkylating agents such as busulfan, which is used in treating various types of solid tumors, are the drugs most frequently recognized to cause pneumonitis. The pneumonitis may occur weeks to months after the drug has been started and usually appears to depend on the dose. Gold sodium thiomalate, a drug used to treat rheumatoid arthritis, and amiodarone, a drug used in treating heartbeat irregularities, are agents that can cause a pneumonitis that may progress to fibrosis and that represents a hypersensitivity reaction. See CHEMOTHERAPY AND OTHER ANTINEOPLASTIC DRUGS.

Cocaine and its derivatives have been associated with the development of a nonspecific type of pneumonitis and fibrosis, as well as pulmonary hemorrhage. The exact mechanism by which cocaine causes these types of change is uncertain but may be related to its effect on the contraction and relaxation of blood vessels.

Other types of pneumonitis and agents. Idiopathic interstitial pneumonitis, also referred to as idiopathic pulmonary fibrosis, is characterized by a chronic inflammation in the lung in association with the formation of fibrous tissue. The cause of most cases is uncertain, although some have been associated with diseases such as systemic lupus erythematosus. Bronchiolitis obliterans or organizing pneumonitis is characterized by inflammation of the bronchioles as well as lobar fibrosis. It usually is associated with cough, fever, and chest pain. Its cause is usually unknown, and most cases are reversible when treated with cortisone-type drugs.

Oxygen, when delivered at concentrations greater than 40% for more than 24 h, can cause inflammation in the lung. This usually begins as acute damage to the alveoli and can progress to pulmonary fibrosis. Radiation, used in the treatment of thoracic tumors, can also cause inflammation of the lung with subsequent fibrosis. Aspiration of blood or gastric contents can produce a localized region of severe inflammation and cellular death in the lung. The aspiration of an oily substance, such as oily nose drops, can result in the accumulation of lipid material in the lung, referred to as lipoid pneumonitis.

Signs, symptoms, and diagnosis. The signs and symptoms of pneumonia and pneumonitis are usually nonspecific, consisting of fever, chills, shortness of breath, and chest pain. Fever and chills are

more frequently associated with infectious pneumonias but may also be seen in pneumonitis. The physical examination of a person with pneumonia or pneumonitis may reveal abnormal lung sounds indicative of regions of consolidation of lung tissue. A chest x-ray also shows the consolidation, which appears as an area of increased opacity (white area). Cultures of sputum or bronchial secretions may identify an infectious organism capable of causing the pneumonia. In AIDS patients, fluid is frequently instilled into the pulmonary bronchial tree and then rapidly removed; examination of special preparations of this fluid may show such microorganisms as Pneumocystis carinii. In some cases of pneumonia or pneumonitis in which no definite cause can be determined from routine studies, the chest cavity must be opened surgically and the lung biopsied.

Treatment. The treatment of pneumonia and pneumonitis depends on the cause. Bacterial pneumonias are treated with antimicrobial agents such as penicillin. If the organisms can be cultured, the sensitivity of the organism to a specific antibiotic can be determined. Viral pneumonia is difficult to treat, as most drugs only help control the symptoms. The treatment of pneumonitis depends on identifying its cause; many cases are treated with cortisone-type medicines. Samuel P. Hammar

Bibliography. D. H. Dail and S. P. Hammar (eds.), *Pulmonary Pathology*, 2d ed., 1993; R. B. Fick and H. Y. Reynolds, The changing spectrum of pneumonia: News media creation or clinical reality?, *Amer. J. Med.*, 74:1-8, 1983; A. P. Fishman (ed.), *Update: Pulmonary Disorders and Diseases*, 1983.

Poales

An order of flowering plants (angiosperms) that includes the bromeliads, grasses, restios, rushes, and sedges. It comprises approximately 18 families and more than 18,000 species, over half of which are grasses (Poaceae). These families in general include plants without showy flowers (except for Bromeliaceae, Xyridaceae, and Rapateaceae) and with a grasslike form. These plants dominate vast areas of the Earth's surface and include some important grain crops, such as corn (Zea mays), rice (Oryza sativa), and wheat (Triticum aestivum), as well as rushes, sedges, grasses used for thatch, and bamboos used for construction. See DUNE VEG-ETATION; GRAIN CROPS; GRASS CROPS; PLANT EVOLU-TION. Mark W. Chase; Michael F. Fay

Podicipediformes

A small order of aquatic birds that contains only a single living family, the Podicipedidae (grebes), with 20 species found throughout the world. They breed on freshwater, but some species winter along the coasts. Grebes have a number of convergent similarities with loons, but are not related to them or to the fossil hesperornithiforms. The close affinities of the grebes are still unresolved. *See* GAVIIFORMES; HESPERORNITHIFORMES.

Grebes are small to medium-sized birds that are specialized for aquatic life and can swim and dive skillfully. Their legs are placed far posteriorly with lobed-not webbed-toes that are used to propel the birds through the water, but their terrestrial locomotion is awkward. Their wings are small, and they rarely fly, preferring to dive to escape enemies; yet they migrate over long distances and have colonized many remote islands. Rollandia microptera, living on Lake Titicaca in the Andes, is flightless, as is Podiceps taczanowskii, living on the high-altitude Lake Junin in Peru; the latter species is endangered. Grebe bodies are compact with dense waterproof plumage. The neck is medium to long, the head is large, and the bill is short to medium and pointed. Grebes feed on fish and other aquatic animals that they pursue underwater. The sexes are alike in their plumage, which is dull except on the head, where it is bright and patterned and used in courtship displays. Grebes are monogamous, with an elaborate courtship and strong pair bond. The nest consists of a floating mat of vegetation and can be solitary or part of a large colony. The clutch of three to nine eggs is incubated by both parents. The downy young leave the nest after hatching and are cared for by both parents, which often carry the small young on their backs when swimming.

Grebes are found throughout the world except on some remote oceanic islands. They have a poor fossil record in spite of being aquatic birds. The earliest fossil is early Miocene, with only a handful of younger forms. Some researchers have suggested a gruiform affinity for grebes, but that remains speculative. Note that some taxonomists have enlarged the Ciconiiformes to include the podicipediform grebes, as in the Sibley-Ahlquist taxonomy a radical new approach to taxonomy, based on DNA hybridization studies. *See* CICONIIFORMES.

The western grebe (*Aechmophorus occidentalis*; see **illustration**) and Clark's grebe (*A. clarkii*) constitute an interesting pair of sibling species that are similar, not only in their nearly identical plumage coloration but also in their ecology and distribution. The giant pied-billed grebe (*Podilymbus gigas*),



Aechmophorus occidentalis, the western grebe. (© California Academy of Sciences, photo by Dr. Lloyd Glenn Ingles)

found only on Guatemala's Lake Atitlan, probably became extinct during the past decade as a result of interbreeding with the closely related pied-billed grebe (*P. podiceps*). *See* AVES. Walter J. Bock

Bibliography. J. Fjeldä, *Grebes: Podicipedidae*, Oxford University Press, 2004.

Podocopa

One of two Recent subclasses of Ostracoda. The two (Podocopa and Myodocopa) may share many characteristics but differ significantly in others. Shared characteristics include a carapace (shell) with a straightto-concave ventral margin and somewhat convex dorsal margin; adductor-muscle-scar patterns varying from numerous to few; the absence of an anterior notch in the bivalve shell; and the absence of lateral eyes, heart, and frontal organ (Bellonci organ). Podocopid ostracods are small, rarely exceeding 0.25 in. (7 mm). The carapace valves are unequal, weakly or strongly calcified, and may be heavily ornamented. The hinge varies from simple to complex. Six or seven pairs of appendages are present; the seventh is adapted for walking, for cleaning, or as a clasping leg in males. It is never annulated and wormlike, as in the Myodocopa. The caudal furca, when present, is lamelliform, flagellate, or footlike and ventral to the anus. In one podocopid genus the posterior part of the body is pronounced and has a pair of setae-bearing lobes, structures that have been interpreted by some researchers as representing a telson and uropods, respectively.

The exopod of the antenna, which consists of only one or two segments, is never adapted for swimming; it may be modified as a spinning structure or reduced to a simple bristle or scale. The maxillule is usually provided with a branchial epipod or plate. The first postoral appendage is lamelliform or sticklike.

The sexes are separate, and the male genital organs are large, highly specialized, and usually paired. Females of some fresh-water species are parthenogenetic and produce eggs capable of with-standing years of desiccation. Many podocopids brood their eggs; eight—or less frequently, seven—juvenile instars make up the developmental sequence. Podocopid ostracods are found in fresh-water, marine, and brackish-water environments, and a few are terrestrial. The majority are burrowing or crawling inhabitants of the benthos and are filter or detritus feeders or herbivores. Some are capable of swimming, but none appear to ever lead a planktonic existence.

The subclass is represented by the two extant orders, Platycopida and Podocopida. Although morphologic differences between these two taxa are sufficiently distinct to have prompted some researchers to recommend taxonomic hierarchical separation, evidence supports the hypothesis of common ancestry. Despite differences in filter-feeding adaptations, some older representatives exhibit converging morphologic characteristics, and taxa from extinct lineages provide possible links between the modern representatives of the two orders. *See* CRUSTACEA; OSTRACODA; PLATYCOPIDA; PODOCOP-IDA. Patsy A. McLaughlin

Bibliography. T. E. Bowman, The case of the nonubiquitous telson and the fraudulent furca, *Crustaceana*, 21, 1971; A. C. Cohen, Ostracoda, in S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, vol. 2, 1982; R. F. Maddocks, Evolution within the Crustacea, pt. 4: Ostracoda, in L. G. Abele (ed.), *The Biology of Crustacea*, vol. 1: *Systematics, the Fossil Record, and Biogeography*, 1982; R. C. Moore (ed.), *Treatise on Invertebrate Paleontology*, pt. Q: *Arthropoda*, 1961; H. K. Schminke, The ubiquitous telson and deceptive furca, *Crustaceana*, vol. 30, 1976.

Podocopida

An order of the subclass Podocopa, class Ostracoda, that consists of the extant suborder Podocopina and the extinct Metacopina. There is agreement on the assignment of the Cypridoidea, Bairdioidea, Cytheroidea, and Darwinuloidea to the Podocopida; however, some researchers also include the Sigillioidea, and others consider the Terrestrichtheroidea a podocopid taxon of equivalent rank. In all podocopids the two valves fit firmly, hermetically sealing the animal inside when closed. Although variously sculptured, none of the podocopids have strong vertical lobation like that found in the Paleocopa.

Appendages include antennules, antennae, mandibles, maxillules, and three pairs of thoracic legs. Antennules and antennae are both used in locomotion; swimming forms are provided with long, feathered setae, whereas crawling forms have only short, stout setae. Mandibles are strongly constructed and, with rare exceptions, used for mastication. The maxilla has setiferous endites and bears a large branchial plate to circulate water alongside the body. The furca is variously developed but never lamelliform. If eyes are present, the two lateral eyes and the median eye are joined into one central structure. No heart is developed.

Principal superfamilies. Differences in appendage specializations, male copulatory structures (Zenker's organs), furcae, and shell shape make a brief review of the four principal superfamilies appropriate.

Cypridoidea. These ecologically versatile ostracods are mostly fresh-water, but a few are marine; they are adapted to extremes in temperature, oxygen, pH, and postoral salinity. The legs are all different (**Fig. 1***a*). The first legs are accessory feeding structures, the second legs are for walking, and the third are modified for cleaning the interior of the carapace. Some species are parthenogenetic; if males are present, the Zenker's organs are large and well developed (Fig. 1*b*). The furca is typically an elongate ramus and two terminal claws (Fig. 1*c*), but in *Cypridopsis* the end is flagelliform. Basically, the adductor muscle scars consist of one elongate scar above and five others clustered below. The hinge is relatively

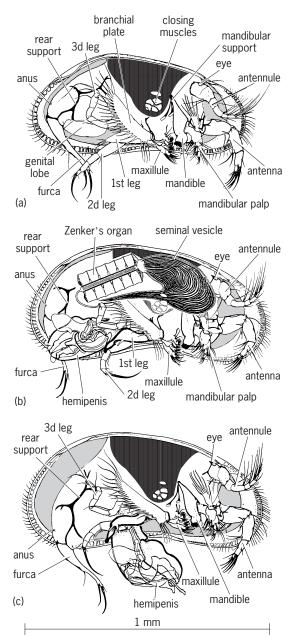


Fig. 1. Candona suburbana, an ostracode of the Cypridoidea. Animals with right valves removed. (a) Female; note dimorphic differences in carapace outline and first leg. (b) Male with hemipenis rotated, folded, and retracted; part of the body wall is removed to show right Zenker's organ and seminal vesicle, the latter distended with spermatozoa. (c) Male with hemipenis erect. (After D. J. McGregor and R. V. Kesling, Contributions from the Museum of Paleontology of the University of Michigan, 1969)

weak and simple, and teeth rarely occur.

Carapace shape varies from ovate to elongate and somewhat subtriangular; the carapace surface is usually smooth and weakly calcified, but may be strongly calcified and sculptured.

The first leg of the male has a pincerlike clasper. Ontogeny is brief, and some fresh-water species reach maturity in a month. In many species the eggs are encased and can withstand drying over periods of years, hatching only when returned to water; hence, many fresh-water species have spread over vast areas. *Bairdioidea.* The Bairdioidea are exclusively marine ostracods with the three postoral appendages adapted for walking. The furca is rod-shaped, with several distal bristles. Males lack Zenker's organs. Muscle scars are irregular and loosely clustered. Hingement is relatively simple, although small teeth occur. The round-backed carapace is acuminate at the ends in many genera, and this lemon-seed form is known as the *Bairdia* shape. These ostracods are not as numerous as the Cypridoidea or Cytheroidea, but they are found throughout the world.

Cytheroidea. The superfamily contains the most genera and species of any ostracode group; in some faunas the cytheraceans are twice as numerous as all other ostracodes combined. Only a few live in fresh water. Most are crawling and digging forms on the ocean floor. Postoral legs are all alike, adapted for walking (Fig. 2). The furca is very small, without claws, and Zenker's organ is absent. The four closing-muscle scars are horizontal, arranged in a vertical row, although one or more scars may be interrupted. The hinge is strong, with complex interlocking structures in many genera. The carapace tends to be nearly straight-backed because of the long hinge; many are distinctly ornamented and colored. These are the only ostracodes with spinning glands, which are located in the antennae. The commensal ostracodes Entocythere and Sphaeromicola belong to this superfamily.

Darwinuloidea. This small superfamily, containing only two genera, is primarily fresh-water but widespread. The ostracodes crawl on bottom sediments. Males are very rare; most darwinuloids are parthenogenetic. The eggs and young animals are protected in a brood space behind the body (**Fig. 3**). The two postoral pairs of legs are alike and are used in walking. The rear end of the body is curiously tapered; the furca is rudimentary. The several muscle

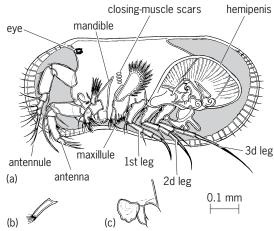


Fig. 2. Limnocythere sanctipatricii, a podocopid ostracode of the Cytheroidea. (a) Male with left valve removed; the paired hemipenes occupy much of the space at the rear of the carapace. (b) Male left "brush-shaped organ," an unusual structure found only in adult males between the postoral legs, presumed to assist in copulation. (c) Female furca and genitalia. (After R. C. Moore, ed., Treatise on Invertebrate Paleontology, pt. Q, 1961, and D. L. McGregor and R. V. Kesling, Contributions from the Museum of Paleontology of the University of Michigan, 1969)

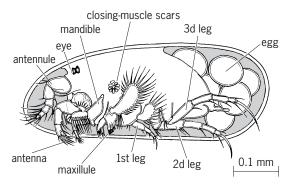


Fig. 3. Darwinula stevensoni, a podocopid ostracode of the Darwinuloidea. In this parthenogenetic species, eggs and young are protected in a brood space behind the body. (After R. C. Moore, ed., Treatise on Invertebrate Paleontology, pt. Q, 1961)

scars form a vertically elongate rosette. The hinge is relatively simple. The carapace is smooth and notably elongate.

Other groups. The Podocopida (Ordovician to Recent) have many fossil representatives, both marine and fresh-water. The superfamily Cytheroidea, in particular, is abundant in the Mesozoic and Cenozoic, and has many divisions. Nearly half of all fossil species described belong to this taxon. The Sigillioidea includes two extant species of *Saipanetta*; however, there is no universal agreement regarding the placement of the genus, or of the family in the Podocopida. Both are regarded by some as correctly being retained in the suborder Metacopina.

The Terrestrictheroidea is represented by a single genus (*Terrestricythere*) of terrestrial ostracods from the Kuril Islands of the northwestern Pacific Ocean. Other terrestrial ostracodes are assigned to the superfamily Cypridoidea, and it is the opinion of some specialists that the Terrestricytheroidea are no more than acutely specialized cypridoids.

Evolution. Within the Podocopida there appears to be two principal evolutionary pathways, one including the Cypridoidea and Darwinuloidea (also Sigillioidea), and the other the Bairdioidea and Cytheroidea. It is believed that the oldest podocopids may belong to the Bairdioidea and that those may approach the ancestral podocopid stem. The Cytheroidea are then assumed to have been descended from the Bairdioidea. The Darwinuloidea (and Sigilloidea) have retained many primitive characters, although that may be the result of the fact that most living species reproduce parthenogenetically. Although the Cypridoidea cannot be considered primitive, they do share several significant characteristics, such as appendage structure and genitalia, with the former taxa. See CRUSTACEA; PODOCOPA. Patsy A. McLaughlin

Bibliography. A. C. Cohen, Ostracoda, in S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, vol. 2, 1982; P. D. Deckker, Terrestrial ostracods in Australia, in J. K. Lowry (ed.), *Papers from the Conference on the Biology and Evolution of Crustacea*, Austral. Mus. Mem. 18, 1983; R. F. Maddocks, Evolution within the Crustacea, pt. 4: Ostracoda, in L. G. Abele (ed.), *The Biology of Crustacea*, vol. 1: *Systematics, the Fossil Record, and Biogeography,* 1982; K. G. McKenzie, Saipanetta and the classification of podocopid Ostracoda: A reply to Schornikov and Gramm (1974), *Crustaceana*, vol. 29, 1975; R. C. Moore (ed.), *Treatise on Invertebrate Paleontology,* pt. Q, 1961; G. O. Sars, *Ostracoda: An Account of the Crustacea of Norway*, Bergen Museum, vol. 9, 1922–1928.

Podostemales

An order of flowering plants, division Magnoliophyta (Angiospermae), in the subclass Rosidae of the class Magnoliopsida (dicotyledons). The order consists of only the family Podostemaceae, with about 200 species, the greatest number occurring in tropical America. They are submerged aquatics with modified, branching, often thalluslike shoots; small, perfect flowers with a much reduced perianth; a superior ovary with united carpels and numerous small seeds without endosperm. *See* MAGNOLIOP-SIDA; PLANT KINGDOM; ROSIDAE. Arthur Cronquist

Poecilosclerida

An order of sponges of the class Demospongiae in which the skeleton includes two or more types of megascleres, each localized in a particular part of the sponge colony. Frequently one type of megasclere is restricted to the dermis and another type occurs in the interior of the sponge (**Fig. 1**). Sometimes one category is embedded in spongin fibers while a second category, usually spinose, protrudes from the fibers at right angles. Spongin is always present but varies in amount from species to species. Microscleres are usually present; often several types occur in one species. A wide variety of microsclere categories is found in the order but asters are never present.

In shape, poecilosclerid sponges are encrusting, massive, lobate, or branching. Deep-sea species often have bizarre shapes (**Fig. 2**). Sponges of this order are

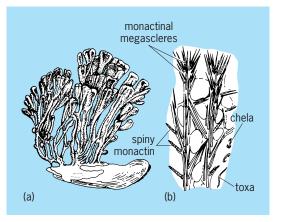


Fig. 1. *Microciona prolifera*. (a) Whole organism. (b) Spiculation.

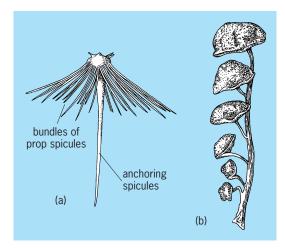


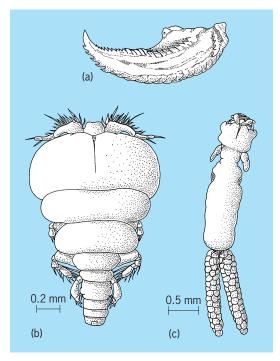
Fig. 2. Deep-sea poecilosclerids. (a) Cladorhiza longipinna, found at depth of 1800 ft (5500 m). (b) Esperiopsis challengeri, found at depth of 500 ft (1500 m).

found in all seas and range from tidal waters to depths of at least 1800 ft (5500 m). Undoubted poecilosclerid spicules are known from Tertiary deposits. *See* DEMOSPONGIAE. Willard D. Hartman

Poecilostomatoida

One of two major orders of parasitic Copepoda that were previously included in the Cyclopoida. The classification of parasitic copepods has been established on the basis of the structure of the mouth. In poecilostomatoids the mouth is represented by a transverse slit, partially covered by the overhanging labrum, which resembles an upper lip. Although there is variability in the form of the mandible among poecilostomatoids, it can be generalized as being falcate (see illus.). Body segmentation is typically podoplean, having prosome-urosome articulation between the fifth and sixth thoracic somites; however, this segmentation is often lost with the molt to adulthood. The antennules frequently are reduced in size and the antennae modified to terminate in small hooks or claws that are used in attachment to host organisms.

Most poecilostomatoid copepods (see illus.) are ectoparasites of marine fishes or invertebrates, usually attaching to the external surface of the host or in the branchial cavity on the walls or gill surfaces. Representatives of one family, however, have successfully made the transition to fresh-water habitats, and a second family has evolved an endoparasitic mode of life. The latter inhabit the mucous ducts, body wall, or viscera of the host teleost fish. The body structure of the parasites is reduced in females to a vermiform, saccular, or spherical mass usually lacking observable appendages; body segmentation in males is obsolete, but appendages are retained. Among ectoparasitic poecilostomatoids, external body segmentation is obsolete in those species that attach permanently to their hosts. More mobile species retain much of their body segmentation and attach to their hosts by large, subchelate antennae or by antennae pro-



Representatives of poecilostomatoid copepods. (a) Poecilostomatoid mandible (after L. Margolis and Z. Kabata, Guide to the Parasites of Fishes of Canada. pt. II: Crustacea, Can. Spec. Publ. Fisher. Aquat. Sci. 101, 1988). (b) Bomolchus concinnus (after W. Vervoort, Caribbean Bomolochidae (Copepoda: Cyclopoida), Studies on the Fauna of Curaçao and Other Caribbean Islands, vol. 28, 1969). (c) Heterochondria atypica (after J.-S. Ho, Four new parasitic copepods of the family Chondracanthidae from California inshore fishes, Proc. Biol. Soc. Wash., vol. 85, 1972).

vided with rows of small spinules or denticles. As a group, poecilostomatoids are usually less deleterious to their hosts than other parasitic copepods, since most species are small, tend to be less mobile, and thus are less likely to cause extensive tissue damage.

Poecilostomatoids are thought to have evolved rather recently from the Cyclopoida lineage. Perhaps the strongest argument for that hypothesis is found in the similarity of nauplii in the two orders. A possible link has been suggested through the genus *Hemicyclops*, which has a broad labrum that covers only the anterior edge of the mouth. Subsequent evolutionary steps might then have included hypertrophy and paedomorphosis, which led to the reduction in appendages and adult body segmentation. *See* COPEPODA; CRUSTACEA; CY-CLOPOIDA. Patsy A. McLaughlin

Bibliography. Z. Kabata, *Parasitic Copepoda of British Fishes*, Ray Society, London, vol. 152, 1979; B. M. Marcotte, Evolution within the Crustacea, pt. 2: Copepoda, in L. G. Abele (ed.), *The Biology of Crustacea*, vol. 1: *Systematics, the Fossil Record, and Biogeography*, 1982; L. Margolis and Z. Kabata, *Guide to the Parasites of Fishes of Canada*, pt. II: *Crustacea*, Can. Spec. Publ. Fisher. Aquat. Sci. 101, 1988; W. Vervoort, *Caribbean Bomolochidae (Copepoda: Cyclopoida*), Studies on the Fauna of Curacao and Other Caribbean Islands, vol. 28, 1969.

Pogonophora (Siboglinidae)

A family of sedentary marine worms (known as beard worms) that live in all the world's oceans, generally at depths between 330 and 13,200 ft (100 and 4000 m), shallower at higher latitudes and deeper in trenches. The first specimens of the Pogonophora, previously considered to be a phylum, but now known more accurately as Siboglinidae (a family of polychaete annelids), were dredged up late in the nineteenth century, but first investigated in the 1950s by the Russian zoologist A. V. Ivanov. Siboglinids construct a chitinous tube and have no mouth, gut, or anus in their postembryonic anatomy. Instead of actively feeding, they derive their nutrition from symbiotic bacteria hosted in specialized tissues or via absorption. With new technologies in deep-sea exploration, such as crewed and remotely operated submersibles, many new species of Siboglinidae have been discovered in unexpected habitats over the past three decades, notably around hydrothermal vents and cold-water hydrocarbon seeps, and on decaying whale bones. See DEEP-SEA FAUNA; HYDROTHERMAL VENT; MARINE ECOLOGY.

Anatomy. The body is generally divided into four regions (see illus.). The anterior region (cephalic lobe) bears from one to thousands of tentacles or branchial filaments and is followed by a short collar region known as the forepart or vestimentum. The largest section is usually an extended trunk region, which can have regional specializations such as papillae, cilia, and chaetae. Except in the whalebone-inhabiting Osedax, this region hosts symbiotic bacteria. In the frenulates, vestimentiferans, and moniliferans, the terminal region (opisthosoma) has the internal septa and external chaetae typical of polychaete annelids. No opisthosoma has been found in Osedax; instead this worm forms an extensive root system that reaches deeply into the whale bones and hosts the female gonads and symbiotic bacteria.

Physiology. Frenulates and moniliferans can absorb dissolved organic matter and other nutrients through their body wall, but symbiotic bacteria supplement their nutrition. Vestimentiferans seem to depend completely on their symbionts for nutrition. The bacteria in frenulates, moniliferans, and vestimentiferans are usually chemoautotrophic sulfuroxidizers, but one frenulate species is known to host methanotrophic bacteria. The symbionts of *Osedax* are heterotrophic: they are capable of extracting and degrading complex carbohydrates from the whale bones that this genus inhabits. *See* FEEDING MECHANISMS (INVERTEBRATE).

Vestimentiferans have a unique type of hemoglobin which can bind sulfide and oxygen simultaneously. Sulfide and oxygen are absorbed through the bright red branchial filaments and transported to the symbionts, where they are released. The sulfidebinding capacity of vestimentiferan hemoglobin has several advantages: it prevents toxic effects of sulfide in the body as well as its spontaneous reaction with oxygen, while channeling the energy of the oxida-

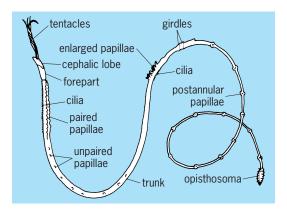


Diagram of a frenulate siboglinid; length has been greatly reduced. (After J. D. George and E. C. Southward, A comparative study of the setae of Pogonophora and polychaetous annelida, J. Mar. Biol. Ass. U.K., 53(2):403–424, 1973)

tion to the symbionts. *See* HEMOGLOBIN; VESTIMEN-TIFERA.

Reproduction and development. The sexes are separate in most siboglinids; only one hermaphroditic species has been reported. Osedax shows an extreme degree of sexual dimorphism in which only the females reach full size, while the males are essentially sexually mature larvae that live in the female tube, where they fertilize the eggs. Larval development has been studied in very few siboglinid species. Some frenulates brood their larvae in their tubes; in others, larvae have never been observed. Vestimentiferan larvae are planktonic (free-floating) and have a transient digestive tract just before settlement. They may spend several weeks in the water column, subsisting on yolk supplied with the egg, before metamorphosing into a juvenile. The larvae are symbiont-free, and the worms must acquire new symbionts from the environment with each generation. See SEXUAL DIMORPHISM.

Taxonomy. The phylogenetic position of the Siboglinidae has been the subject of much debate. The earliest specimens examined were incomplete. With the opisthosoma missing, the most obvious clue for their relationship with the segmented annelids was lacking. Due to the absence of the mouth and gut, some early researchers erroneously considered the nerve cord to be dorsal, suggesting deuterostome affinities. The frenulates and vestimentiferans were once considered separate phyla and were named Perviata and Obturata, respectively. However, recent molecular studies, supported by similarities in embryonic and larval development, the segmentation of the opisthosoma, and the ultrastructure of chaetae formation, clearly indicate that siboglinids are polychaete annelids. See ANNELIDA.

Anja Schulze; Edward B. Cutler Bibliography. A. D. Nussbaumer et al., Horizontal endosymbiont transmission in hydrothermal vent tubeworms, *Nature*, 441:345–348, 2006; G. W. Rouse et al., *Osedax:* Bone-eating marine worms with dwarf males, *Science*, 305:668–671, 2004; A. Schulze and K. M. Halanych, Siboglinid evolution shaped by habitat preference and sulfide tolerance, *Hydrobiologia*, 496:199–205, 2003; E. C. Southward et al., Pogonophora (Annelida): Form and function, *Hydrobiologia*, 535/536:227–251, 2005.

Poison

A substance which by chemical action and at low dosage can kill or injure living organisms. Broadly defined, poisons include chemicals toxic for any living form: microbes, plants, or animals. For example, antibiotics such as penicillin, although nontoxic for mammals, are poisons for bacteria. In common usage the word is limited to substances toxic for humans and mammals, particularly where toxicity is a substance's major property of medical interest. Because of their diversity in origin, chemistry, and toxic action, poisons defy any simple classification. Almost all chemicals with recognized physiological effects are toxic at sufficient dosage. The same compound may be considered a drug or a poison, depending on dosage, effect, or intended use.

Origin and chemistry. Many poisons are of natural origin. Some bacteria secrete toxic proteins (for example, botulinus, diphtheria, and tetanus toxins) that are among the most poisonous compounds known. Lower plants notorious for poisonous properties are ergot (Claviceps purpurea) and a variety of toxic mushrooms. Ergot, a fungal parasite of rye, has been the source of numerous epidemics of poisoning from the use of contaminated rye flour. The fungus contains many different alkaloids, some of which are also useful drugs. Among the best-known toxic mushrooms are the fly agaric (Amanita muscaria) containing muscarine, and the destroying angel (A. phalloides) whose toxic agents are phalloidin and aminitin. See ERGOT AND ERGOTISM; MUSHROOM; TOXIN

Higher plants, which constitute the major natural source of drugs, contain a great variety of poisonous substances. Many of the plant alkaloids double as drugs or poisons, depending on dose. These include curare, quinine, atropine, mescaline, morphine, nicotine, cocaine, picrotoxin, strychnine, lysergic acid, and many others. Some of the alkaloids were used in classical antiquity (coniine was the toxic agent of the extract of spotted hemlock, Conium maculatum, drunk by Socrates). Some are of prehistoric antiquity (quinine and curare were used by South American Indians before the advent of Europeans), and some date from the earliest records of humans (the opium poppy is believed to have been cultivated in the Stone Age). See ATROPINE; COCAINE; MORPHINE ALKALOIDS; QUININE.

Poisons of animal origin (venoms) are similarly diverse. Toxic marine animals alone include examples of every phylum from Protozoa (dinoflagellates) to Chordata (a number of fishes). Insects and snakes represent the best-known venomous land animals, but on land, too, all phyla include poison-producing species. Among mammalian examples are certain shrews with poison-producing salivary glands. *See* POISON GLAND.

Poisons of nonliving origin vary in chemical complexity from the toxic elements, for example, the heavy metals, to complex synthetic organic molecules. Most of the heavy metals (gold, silver, mercury, arsenic, and lead) are poisons of high potency in the form of their soluble salts. Strong acids or bases are toxic largely because of corrosive local tissue injury; ingestion, for example, sometimes results in fatal perforation of the gastrointestinal tract. Many other elemental substances are toxic at low concentrations: selenium, beryllium, cadmium, manganese, phosphorus, and zinc. In general, it may be anticipated that any chemically reactive element is likely to be toxic to humans unless it represents one of the bulk elements in the body, in the ionic form in which it exists in the body. Na⁺, K⁺, Ca²⁺, and Cl⁻, for example, are relatively nontoxic.

The chemically reactive gases hydrogen sulfide, hydrocyanic acid, chlorine, bromine, and ammonia are also toxic, even at low concentration, both because of their corrosiveness and because of more subtle chemical interaction with enzymes or other cell constituents. An example is the toxic action of cyanide ion, probably explained by its inactivation of cytochrome oxidase.

Many organic substances of synthetic origin are highly toxic and represent a major source of industrial hazard. Most organic solvents are more or less toxic on ingestion or inhalation. Ethanol is a relatively innocuous exception, in part because available enzymatic reactions channel it into major normal metabolic pathways along which it can be oxidized to products familiar to mammalian metabolism, such as acetate. Many other alcohols, such as methanol, are much more toxic. Many solvents (for example, carbon tetrachloride, tetrachloroethane, dioxane, and ethylene glycol) produce severe chemical injury to the liver and other viscera, sometimes from rather low dosage. With certain poisons of this group, a high degree of variation in individual susceptibility exists.

Chemical correlations. Since poisons represent all chemical classes from the elements to complex alkaloids and large proteins, general chemical constitution has no defining value for toxicity. However, in limited instances chemical features correlate with toxic action. Many of the chlorinated hydrocarbons, for example (carbon tetrachloride), have similar toxicity for liver, heart, and kidneys. A number of alkyl phosphates (diisopropylfluorophosphate, tetraethylpyrophosphate, and related compounds) are very potent inhibitors of the enzyme acetylcholinesterase and produce a consistent set of physiological changes arising from the inactivation of this enzyme. A number of tertiary amino esters of aromatic acids (such as procaine and cocaine) are local anesthetics and also share marked cardiac and nervous system toxicity. The digitalis glycosides, all represented by steroids condensed with sugars and containing an additional lactone ring, share many pharmacological properties, notably characteristic effects on cardiac function which make these drugs cardiotoxins at low concentration.

Physiological actions. The action of poisons is generally described by the physiological or biochemical changes which they produce. For most poisons, a descriptive account can be given which indicates what organic system (for example, heart, kidney, liver, brain, and bone marrow) appears to be most critically involved and contributes most to seriously disordered body function or death. In many cases, however, organ effects are multiple, or functional derangements so generalized that a cause of death cannot be localized.

Although a comprehensive list cannot be presented here, some illustrations may be given. Anoxia, lack of oxygen for cellular respiration, can result from poisons acting at different sites. Phosgene or other inhaled corrosive vapors may produce massive flooding of the lungs with edema fluid and thus cause mechanical suffocation. Carbon monoxide acts by binding to hemoglobin with a much greater affinity than oxygen and so produces anoxia by interference with oxygen transport. Cyanide prevents tissue respiration at a terminal intracellular site by poisoning cytochrome oxidase which catalyzes the final step of oxygen utilization. Central nervous system depression, with coma and ultimate respiratory and circulatory failure, results from a large group of drugs: the general anesthetics (ether, chloroform, cyclopropane, and ethylene), the barbiturates, the opium alkaloids (morphine, codeine, and related compounds), and less regularly from many other poisons. Liver injury with jaundice (toxic hepatitis) is a prominent result of many poisons, characteristically the chlorinated hydrocarbons, elemental phosphorus, certain heavy metals (arsenic, antimony, barium, and copper), and a large diversity of organic compounds. Kidney damage results from many toxic chemicals, but is common after exposure to carbon tetrachloride, mercuric salts, and ethylene glycol. Convulsions, sometimes fatal, are a common response to acute poisoning by diverse agents, but are a particularly characteristic effect of strychnine, atropine, ergot, alkyl phosphates, picrotoxin, cocaine, and certain snake venoms. Among the medically important toxic agents are those that reduce or exaggerate effective transmission of the nerve impulse across synaptic junctions, especially in the autonomic nervous system. All such compounds, sometimes called autonomic agents, act so as to mimic or to block the physiological effects of either of the two known neurohumoral transmitter substances of the motor branches of autonomic nerves, acetylcholine and norepinephrine (Fig. 1). See ACETYLCHOLINE.

Poisons that act to exaggerate those physiological effects normally produced by acetylcholine release include muscarine and pilocarpine. The anticholinesterases (neostigmine, physostigmine, and the alkyl phosphates) have the same general physiological effects, although by a different mechanism. These compounds slow the heart, lower blood pressure, increase secretion of fluids into the respiratory tract, and narrow the respiratory passages. Toxic doses may produce death by suffocation or by failure

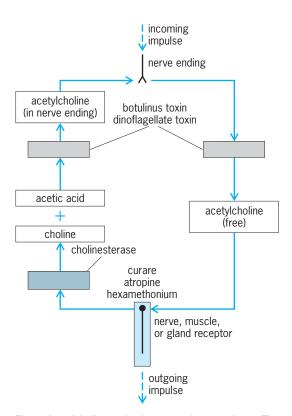


Fig. 1. Acetylcholine cycle, the target of many poisons. The "anticholinesterase" poisons prevent breakdown of acetylcholine by inactivating cholinesterase. Botulinus and dinoflagellate toxins hinder synthesis or release of acetylcholine. Curare and atropine make the receptor less sensitive to the chemical stimulus.

of the circulation. Poisons that prevent acetylcholine from producing skeletal muscle contraction are curare and a number of related synthetic compounds used in anesthesia (succinylcholine, decamethonium). These drugs result in weakness of all voluntary muscle and may cause death through paralysis of respiratory muscles. Exaggeration of the normal effects of norepinephrine and epinephrine (adrenergic drugs) may produce dangerously high blood pressure, rapid heart action, and sometimes fatal disturbances of cardiac rhythm. *See* EPINEPHRINE.

Many normal functions (glandular secretion and contraction of voluntary and smooth muscle) depend on the cyclic release, breakdown, and resynthesis of acetylcholine at the endings of cholinergic (acetylcholine-secreting) nerves. Figure 1 indicates the different ways in which several distinct poisons may interfere with the normal operation of the acetylcholine cycle. In the acetylcholine cycle, an impulse reaching a nerve ending liberates acetylcholine, which stimulates a receptor. The receptor is freed for further impulses by the enzyme cholinesterase, which breaks down acetylcholine into acetic acid and choline. These are resynthesized by other enzymes into new acetylcholine.

Mechanism of action. More precise understanding of the mechanism of poisons requires detailed knowledge of their action in chemical terms. Information of this kind is available for only a few compounds, and then in only fragmentary detail. Poisons that inhibit acetylcholinesterase have toxic actions traceable to a single blocked enzyme reaction, hydrolysis of normally secreted acetylcholine. Detailed understanding of the mechanism of chemical inhibition of cholinesterase is not complete, but allows some prediction of chemical structures likely to act as inhibitors.

Carbon monoxide toxicity is also partly understood in chemical terms, since formation of carboxyhemoglobin, a form incapable of oxygen transport, is sufficient to explain the anoxic features of toxicity.

Heavy metal poisoning in many cases is thought to involve inhibition of enzymes by formation of metal mercaptides with enzyme sulfhydryl groups, the unsubstituted form of which is necessary for enzyme action. This is a general reaction that may occur with a variety of sulfhydryl-containing enzymes in the body. Specific susceptible enzymes whose inhibition explains toxicity have not yet been well documented.

Metabolic antagonists active as poisons function by competitive blocking of normal metabolic reactions. Some antagonists may act directly as enzyme inhibitors, others may be enzymatically altered to form derivatives which are even more potent inhibitors at a later metabolic step. An example of the latter is the biosynthetic incorporation of metabolite analogs into much more complex molecules, particularly the incorporation into nucleic acids of altered purine or pyrimidine bases, such as 8-azaguanine and 5-bromouracil, or the incorporation into proteins of altered amino acids, such as *p*-fluorophenylalanine or 7-azatryptophan. *See* ENZYME INHIBITION. A well-studied example of biosynthetic production of a highly toxic metabolic analog arose from studies of fluoroacetate, a toxic component of plants of the genus *Dichapetalum*. Fluoroacetate is enzymatically converted to fluorocitrate, through reactions analogous to those normal steps leading from acetate to citrate. The fluorocitrate formed from fluoracetate is a potent inhibitor of aconitase, a citrateutilizing enzyme in the tricarboxylic acid cycle, and this inhibition can explain the toxicity of fluoroacetate.

Where poison mechanisms are relatively well understood it has sometimes been possible to employ rationally selected antidotes. Thus the clarification of arsenic as a sulfhydryl poison led to the introduction of the effective metal-poisoning antidote, dimercaptopropanol (British antilewisite). Knowledge of the physiology of anticholinesterase poisons has permitted the use of rational physiological antagonists of excessive cholinergic stimulation, such as atropine and curare. On a more chemical basis, studies of the enzymatic mechanism of cholinesterase have led to more direct antidotes (pyridine aldoxime methiodide and other oximes) for the anticholinesterase poisons. These antidotes act by removing the inhibitor from the enzyme.

It should be noted that for the great majority of poisons the mechanism of toxic action is not as well understood as in the selected examples above. In only a few cases are poisons known to act directly by altering enzyme function. Other poisons, probably much more numerous, may involve reactions with less well-defined cellular components.

Toxicity rating	practically nontoxic	slightly toxic	moderately toxic	very toxic	extremely toxic	supertoxic	supremely toxic
Examples	glycerin, graphite, lanolin	ethyl alcohol, lysol, castor oil, soaps	methyl (wood) alcohol, kerosene, ether	tobacco, aspirin, boric acid, phenol, carbon tetrachloride	morphine, bichloride of mercury	potassium cyanide, heroin, atropine	botulinus toxin, some snake venoms
Probable lethal dose, mg/kg	more than 15,000	5000 to 15,000	500 to 5000	50 to 500	5 to 50	less than 5	less than 0.5
Probable lethal dose, for a 70-kg (155-lb) human	more than 1 qt	1 pt-1 qt	۴ 1 oz–1 pt	1 tsp-1 oz	7 drops- 1 tsp	a taste (less than seven drops)	يď⊲

Fig. 2. Toxicity-rate scale of substances according to the size of the probable lethal dose. Many drugs fall into the highly toxic categories. Drawings suggest the fatal doses of water, whiskey, ether, aspirin, morphine, and cyanide, the last three depicted in terms of aspirin-sized tablets. Scale suggested by M. N. Gleason, R. E. Gosselin, and H. C. Hodge.

Potency. The strength or potency of poisons is most frequently measured by the lethal dose, potency being inversely proportional to lethal dose (Fig. 2). Although killing dose represents a useful end point for potency appraisal, much smaller doses of a given poison may produce toxic changes in an organism. In estimating potency, a large test population should be used because of inherent individual variation in susceptibility. The relationship between percent of a population killed and dose of poison sometimes, but not often, conforms to a symmetrical sigmoid curve representing the integrated or cumulative form of the bell-shaped gaussian distribution curve. From statistically treated dose-response data, the dose killing 50% of the sample population can be determined, and is usually designated the MLD (median lethal dose) or LD₅₀. This is the commonest measure of toxic potency, although for special purposes the lethal dose for other fractions of the population may also be used (for example, LD_{10} and LD_{90}). The spread of the distribution (for example, dose difference between LD_1 and LD_{99}) is highly variable for different poisons and is of course also greatly influenced by the relative heterogeneity or homogeneity of the test population. See LETHAL DOSE 50.

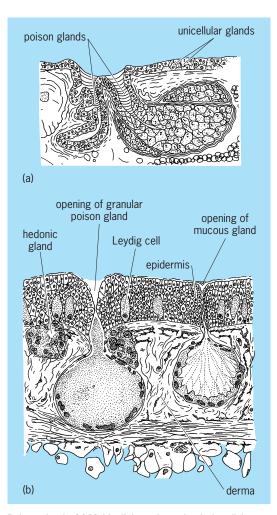
Health aspects. The environment of modern industrial society abounds in toxic chemicals: drugs, industrial products, fuels, pesticides, and many common household products such as cleaning agents and paint products. Acute and chronic poisoning represent an appreciable public health hazard.

The multiplication of toxic chemicals in everyday use led, in 1953, to the establishment of information centers known as Poison Control Centers, which provide physicians with emergency telephone information concerning toxicity and antidotes for common poisons as well as for brand-named chemical formulations. These services were greatly assisted by the publication in 1957 of a compilation of brand-named products with information as to ingredients and quantitative toxicity. *See* TOXICOLOGY. Elijah Adams

Bibliography. J. M. Arena, *Poisoning: Toxicology, Symptoms, Treatment*, 4th ed., 1979; R. H. Dreisbach, *Handbook of Poisoning: Prevention, Diagnosis, and Treatment*, 10th ed., 1982; M. N. Gleason, R. E. Gosselin, and H. C. Hodge, *Clinical Toxicology of Commercial Products*, 4th ed., 1976.

Poison gland

The specialized gland of certain fishes (**illus**. *a*), as well as the granular and some mucous glands of many aquatic and terrestrial Amphibia (illus. *b*). The poison glands of fishes are simple or slightly branched acinous structures which use the holocrine method of secreting a mucuslike substance. The poison glands of snakes are modified oral or salivary glands. Amphibian glands are simple, acinous, holocrine, with granular secretion. In some cases these amphibi



Poison glands. (a) Multicellular poison gland of catfish. (b) Amphibian skin, generalized.

ian poison glands produce mucus by a merocrine method of secretion. These glands function as protective devices. *See* EPITHELIUM; GLAND.

Olin E. Nelsen

Poison ivy

A general name applied to certain species of the genus *Toxicodendron*, previously known as *Rhus*, in the sumac family (Anacardiaceae). *Toxicodendron radicans* is the poison ivy of eastern North America;



Three-foliolate leaf of poison ivy.

henbane

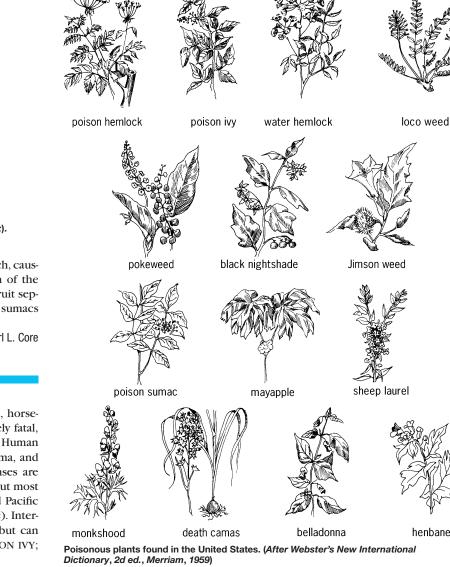
T. diversiloba is the poison oak of California. These plants are natives of North America. Both cause ivy poisoning, an annoying and often painful dermatitis.

Toxicodendron radicans, the most widespread species, is extremely variable. It has a bushy or climbing habit and three-foliolate leaves which are smooth and glossy or hairy and are entire, toothed, or lobed (see illus.). Poison ivy bears white fruits whereas the nonpoisonous sumacs bear red fruits. See SAPIN-DALES. Perry D. Strausbaugh; Earl L. Core

Poison sumac

The plant Toxicodendron vernix (previously in the genus Rhus), a member of the sumac family (Anacardiaceae). It is an inhabitant of swamps ranging from Quebec to Minnesota, and southward to Florida, Louisiana, and Texas. It is a tall bush or small tree bearing pinnately compound leaves with 7-13 entire (without marginal teeth) leaflets, and drooping, axillary clusters of persisting white fruits (see illus.).

Glycoside-containing plants. Glycosides are common compounds in plants. They decompose to form one or more sugars, but sometimes the remaining compounds, aglycones, can be quite poisonous. Cyanogenic glycosides, which produce hydrocyanic acid, are found worldwide in many plant families; the best known are in the rose family (Rosaceae) and in the pea family (Fabaceae). Leaves, bark, and seeds of stone fruits such as cultivated and wild cherries, plums, peaches, bitter almonds, and apricots contain the glycoside amygdalin, which hydrolyzes to form hydrocyanic acid that can be fatally toxic to humans or animals. The same toxic substance is found in apple and pear seeds. Animals have been killed by high amounts of hydrocyanic acid in wilted leaves of wild species of cherry (Prunus). After ready absorption into the bloodstream, the acid causes death of body cells by blocking use of oxygen. Cattle poisoned by hydrocyanic acid develop convulsions, rapid breathing, coma, and a red froth at the muzzle. Cardiac glycosides are found in many unrelated species of



Poison sumac fruits and leaf (Toxicodendron vernix).

Like poison ivy, this plant is poisonous to touch, causing in many persons a severe inflammation of the skin, or dermatitis. The presence of white fruit separates this species from the nonpoisonous sumacs with their red fruits. See SAPINDALES.

Perry D. Strausbaugh; Earl L. Core

Poisonous plants

More than 700 species of seed plants, ferns, horsetails, and fungi that cause toxic, though rarely fatal, reactions in humans and animals (see illus.). Human allergic responses, including hay fever, asthma, and dermatitis, are widespread. Allergic responses are produced by many different plant species, but most common are poison ivy, poison sumac, and Pacific poison oak (all are species of Toxicodendron). Internal injury by toxic plants is less common but can be detrimental or lethal. See ASTHMA; POISON IVY; POISON SUMAC.

plants. Those of the foxglove (*Digitalis purpurea*) contain a number of these glycosides used medicinally to slow and strengthen the heartbeat. Accidental poisoning or an overdose results in blurred vision, nausea, abdominal pain, and vomiting, and can be fatal. Oleander (*Nerium oleander*), which is cultivated in the warmer parts of the United States, contains a toxic glycoside that has an action similar to that of digitalis. The leaves have a very bitter taste that usually prevents their ingestion, but livestock have been poisoned from eating dried leaves mixed with edible feeds. *See* CYANIDE; DIGITALIS; GLYCOSIDE.

Alkaloid-containing plants. Alkaloids, compounds containing a nitrogen atom, have specific pharmacological effects on both humans and animals. Found in many different plant families, they have been used in drug therapy since ancient times, but misuse of these plants can produce poisonings. The potato family (Solanaceae) has many species that contain a number of alkaloids. Hyoscyamine and atropine are the alkaloids occurring in belladonna or deadly nightshade (Atropa belladonna), black henbane (Hyoscyamus niger), thornapples and jimsonweed (Datura), and tree daturas or angel's-trumpets (Brugmansia). Ingestion of these can produce pupil dilation, nausea, muscular incoordination, delirium, and finally death from respiratory paralysis and heart failure. The black nightshades, which comprise a number of species of Solanum, are common weeds in the United States. Plants of Solanum contain glycoalkaloids that break down to aglycones such as solanine and solanidine. Eating green berries of Solanum plants can result in dizziness, vomiting, diarrhea, and changes in heartbeat, but rarely causes death.

Plants of tobacco, *Nicotiana*, contain numerous alkaloids, principally the very toxic nicotine or its isomer anabasine. Humans can smoke tobacco because nicotine is detoxified in the body. Nicotine poisoning produces heart palpitations, clammy skin, nausea, vomiting, diarrhea, labored breathing, and even respiratory paralysis. In severe poisoning death may be rapid. Plants of poison hemlock (*Conium macalatum*) have several alkaloids similar to nicotine, which affect the central nervous system.

Plants of rattlebox (*Crotalaria*), groundsel (*Senecio*), and fiddleneck (*Amsinckia*) have alkaloids of similar molecular structure. Livestock losses have occurred from ingestion of hay containing seeds or dried plants of these genera. The alkaloids cause liver damage by constricting the smaller blood vessels. Symptoms, chronic or acute, vary in different animals. Pigs, cattle, horses, and poultry are particularly susceptible.

Anagyrine is a toxic alkaloid found in several species of lupine (*Lupinus*) in the western United States. It is especially abundant in young plants in spring, and in seeds and fruits during late summer. Although lupine poisoning is most common in sheep, other animals such as cattle, horses, swine, and deer are also affected. Acute poisoning produces nervousness, depression, loss of muscular control, convulsions, and coma. Anagyrine can move through the

fetal membranes, producing deformed calves without affecting the pregnant dam.

Alkaloids present in species of monkshood (*Aconi-tum*) are extremely toxic. Ingestion of any part of the plant, but particularly roots and seeds, produces cardiac rhythmical disturbances and may be fatal to humans or livestock in 12 to 6 h. Symptoms include lowering of blood pressure and convulsions. Livestock may develop bloat. Larkspur plants (*Delphinium*) have similar toxic alkaloids affecting the central nervous system, causing excitability and muscular spasms.

Plants of false hellebores (Veratrum) and death camas (Zigadenus) have complex alkaloids of similar structure and cause livestock deaths in the western United States. Sheep that eat false hellebore plants in early pregnancy may produce deformed lambs. Other livestock and humans have also been poisoned by false hellebores, developing lowered blood pressure and slowed cardiac action. Nausea and vomiting may also occur. Although normally not eaten, death camas plants, the first to grow in the spring, may cause extensive loss of sheep. Dried death camas plants in hay have also caused death of livestock. Human poisonings have occurred by bulbs of this species mistakenly eaten for onions or edible camas (Camassia). Symptoms include salivation and vomiting, slowed heart action, lowered blood pressure, convulsions, and coma. See ALKALOID.

Heath plants. Toxic resins, andromedotoxins, formed by members of the heath family (Ericaceae) are derived from diterpenes. The most toxic species are mountain laurel, sheep laurel, and bog laurel, all in the genus *Kalmia*. Livestock develop symptoms within 6 h of ingestion, including copious salivation, vomiting, bloat, diarrhea, loss of muscular control, slowed heartbeat, coma, and death. Human toxicity has also been recorded.

Pokeweed. Pokeweed (*Pbytolacca americana*) is a garden weed throughout the United States. but it is native to the eastern and central areas. The entire plant, especially the seeds and the large root, is poisonous. Young shoots are edible after boiling in three separate waters. Human poisonings have resulted from inadvertently including parts of the root along with the shoots. Symptoms include an immediate burning bitter taste, salivation, vomiting, diarrhea, and possible shock. Recovery is usually within 1 day.

Waterhemlock. The waterhemlocks (*Cicuta*) are widespread in North America. The large underground tubers, mistakenly considered edible, have caused human poisonings and death. Livestock usually do not eat waterhemlock, but have been fatally poisoned. The highly toxic principle, an unsaturated aliphatic alcohol which acts directly on the central nervous system, causes violent convulsions and in less than 1 h may be followed by paralysis and death.

Oxalate poisoning. Oxalic acid, as oxalate salts, accumulates in large amounts in some species of plants such as those of the genera *Halogeton*, *Bassia*, *Rumex*, and *Oxalis*. It reacts with calcium, removing it from the blood. Crystals of insoluble calcium

oxalate cause mechanical damage to the kidneys and other organs. Livestock may eat plants with large oxalate content; death of sheep and cattle can occur quickly as a result of calcium deficiency. Chronic poisoning produces kidney damage which may result in death accompanied by paralysis. Horses show bone degeneration from calcium deficiency.

Nitrate poisoning. Nitrate poisoning is widespread and results in many cattle deaths yearly. Any disruption of the normal synthesis of nitrates into amino acids and proteins causes large accumulations of nitrates in various species of plants, particularly in the goosefoot family (Chenopodiaceae). A nitrate content of 1.5% or more of the plant's dry weight is considered lethally toxic. Individual plants in the same pasture show great variations of nitrate content, but some plants accumulate as much as 10% or more. In the rumen of cattle, nitrates are converted to nitrites which are absorbed into the bloodstream and change hemoglobin into chocolate-colored methemoglobin which cannot carry oxygen. Animals die manifesting severe lethargy with increased cardiac and respiratory rates.

Allelopathic toxins. Allelopathic phytotoxins are chemical compounds produced by vascular plants that inhibit the growth of other vascular plants. Residues of grain sorghum (Sorghum bicolor) can markedly reduce the following year's growth of wheat and some weedy grass seedlings. Seeds of jimsonweed (Datura stramonium) will not germinate until the alkaloids are washed from the seed coat. However, these same alkaloids prevent the growth of seedlings of other plant species in the vicinity. Locoweeds, belonging to the genera Oxytropis and Astragalus, produce loss of livestock that become addicted to eating these unpalatable plants. The unknown toxin causes the animals to develop incoordination, erratic gaits, and excitability followed by depression. Other species of Astragalus produce toxic aliphatic nitro compounds. Acute nitro poisoning results in weakness and labored breathing with death within 20 h. Chronic poisoning is manifested as weakness in the hindquarters, a peculiar goosestepping gait, and blurred vision. Poisoned animals that are forced to move may collapse and die. Still other species of Astragalus accumulate toxic quantities of selenium which usually causes chronic poisoning of livestock. They may develop excitability followed by depression, emaciation, loss of hooves, and death. See ALLELOPATHY.

Fungi. Every year human fatalities occur from ingestion of wild poisonous mushrooms. Those gathered in the wild require individual identification since toxic species may grow alongside edible ones.

Humans were poisoned by ergot fungal bodies in grain in past centuries, but modern graincleaning processes have eliminated this problem. Ergot fungus, infecting many species of grasses, causes widespread poisoning of livestock. Normal grass grains are replaced by hardened masses of ergot fungus (*Claviceps purpurea*) containing a number of toxic alkaloids and amines. Cattle show acute stimulation of the central nervous system, including stiff gait and lateral jumping. Chronic ergot poisoning produces contractions of the smaller blood vessels, leading to gangrene and sloughing of the tail, hooves, and ear tips. *See* ERGOT AND ERGOTISM; TOXICOL-OGY. Thomas C. Fuller

Bibliography. J. W. Hardin and J. M. Arena, *Human Poisoning from Native and Cultivated Plants*, 2d ed., 1974; R. F. Keeler, K. R. Van Kampen, and L. F. James (eds.), *Effects of Poisonous Plants on Livestock: Symposium on Poisonous Plants, Utab State University, 1977, 1978; J. M. Kingsbury, Poisonous Plants of the United States and Canada*, 1964; W. H. Lewis and M. P. F. Elvin-Lewis, *Medical Botany: Plants Affecting Man's Healtb*, 1977.

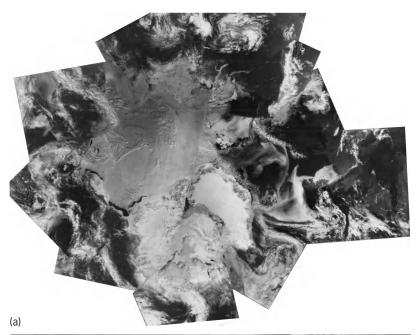
Polar meteorology

The science of weather and climate in the high latitudes of the Earth. In the polar regions the Sun never rises far above the horizon and remains below it continuously for part of the year, so that snow and ice can persist for long periods even at low elevations. The meteorological processes that result have distinctive local and large-scale characteristics in both polar regions.

The polar regions are shown in **Fig. 1** as mosaics of satellite photos taken in late spring. In winter the effective area of the Arctic region is enlarged three-fold by snow covers that average 2.7×10^6 mi² (7×10^6 km²) of the land surrounding the Arctic Ocean, and by an expansion of the area of the sea ice from 2×10^6 mi² (5×10^6 km²) to 4.5×10^6 mi² (11.5×10^6 km²). Similarly sea ice forming on the Southern Ocean each winter more than doubles the total ice-covered area of Antarctica, from 5.8×10^6 mi² (1.5×10^7 km²). *See* ANTARCTIC CIRCLE; ANTARCTICA; ARCTIC CIRCLE.

Energy budgets and temperatures. A general feature of polar energy budgets is their small radiative income. Even in midsummer, when the height of the Sun above the horizon in latitude L varies in the range $23.5 \pm (90 - L)$ degrees, the average daily energy intake of a horizontal unit surface at the top of the atmosphere is reduced by about $60\% (= 1 - \sin 23.5)$ to 550 W/m² from the 1370 W/m² received where the Sun is directly overhead. After further losses during its long slanted passage through the polar atmosphere, up to 80% of the remaining solar radiation is reflected before it can warm a surface covered by snow. In that case, the surface will become extremely cold unless it can draw upon the energy stored above it in warmer air advected from lower latitudes, and below it in the ground or ocean.

Representative energy budgets of typical polar surfaces are shown in the **table**. All have an excess of radiation in summer, but they use it differently. On the Arctic tundra the entire snow pack is melted; temperatures can then rise well above freezing, and so a great deal of the surplus energy is transferred to the atmosphere. Melting accounts for an even larger part of the radiational surplus on Arctic sea ice, but less



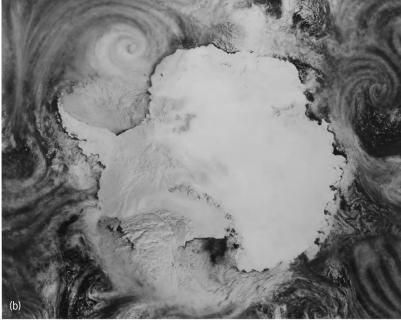


Fig. 1. Composites of polar regions as photographed by meteorological satellites. (a) Arctic, May 1989. (b) Antarctic, November 4–25, 1986. (From U.S. Air Force Defense Meteorological Satellite Program film, NOAA/NESDIS, CIRES/National Snow and Ice Data Center. Boulder)

heat goes into the atmosphere since the temperature of the ice surface cannot rise above $32^{\circ}F(0^{\circ}C)$. In Antarctica the various budget items have smaller values, and melting occurs only near the coast, where even midsummer monthly mean temperatures remain below freezing. In winter, radiational deficits prevail and produce temperatures that increase with height in the lower atmosphere (temperature inversions), which indicate that heat is being transferred from the air to the surface. The strongest inversions occur on the Antarctic plateau, where Vostok (elevation 11,440 ft or 3488 m) has a midwinter monthly mean temperature of $-76^{\circ}F(-60^{\circ}C)$, and a lowest recorded temperature of $-129.3^{\circ}F(-89.6^{\circ}C)$; even in midsummer the monthly mean temperature there is only -26° F (-32° C). *See* AIR MASS; SEA ICE; TEM-PERATURE INVERSION; TUNDRA.

Fuller details of the temperatures and other meteorological parameters on the large ice sheets of Antarctica and Greenland are provided by automatic weather stations reporting via satellites (**Fig. 2**). The annual mean surface temperatures of the ice sheets lie between 32°F (0°C) and 14°F (-10°C) at their edges and decrease to -26°F (-32°C) in central Greenland and to below -76°F (-60°C) in the central Antarctic; they have been determined independently as snow temperatures some 33 ft (10 m) below the surface, where the seasonal air temperature changes are smoothed out by the heat transfer through the snow. *See* METEOROLOGICAL SATELLITES.

On the global scale the overall energy deficit of the polar regions can be seen in satellite measurements made during the Earth Radiation Budget (ERB) project. These have shown that the effective temperature (proportional to the fourth root of the long-wave radiation loss into space) of the Arctic is $-25.6^{\circ}F(-32^{\circ}C)$, and that of the Antarctic, $-65.2^{\circ}F$ $(-54^{\circ}C)$, compared with $-9.4^{\circ}F(-23^{\circ}C)$ for the Earth as a whole. These so-called planetary temperature differences between low and high latitudes find expression in the atmospheric circulations of the polar regions. *See* ATMOSPHERIC GENERAL CIRCU-LATION; HEAT BALANCE, TERRESTRIAL ATMOSPHERIC; INSOLATION; TERRESTRIAL RADIATION.

Pressure and wind. Frequent storms form along the boundary between polar and middle-latitude air masses (the polar front), and they create belts of low surface pressure (polar troughs) around both polar regions. In the north these storms move freely across the Arctic Basin, which they enter mainly via the Norwegian Sea; other small, intense polar lows often form rapidly in polar air outbreaks and accompany them into lower latitudes. By contrast, due to the Antarctic ice sheet [which has an average elevation exceeding 6560 ft (2 km) and a central plateau rising above 13,200 ft (4 km)], most Antarctic storms and their associated very strong westerly winds remain over the surrounding Southern Ocean, although a few storms move on to the ice sheet by way of the low-level Ross and Filchner-Ronne ice shelves. The atmospheric pressure along the ice sheet surface therefore depends primarily on elevation; however, it undergoes a general increase by 10-15 hectopascals (10-15 millibars) from October to December, when radiative heating expands the lower atmospheric layers around the continent. That permits some of the raised air to flood the ice sheet; the accompanying reduction in surface pressure over the rest of the globe was discovered a century ago and was later used to estimate correctly the average height of the ice sheet before its inner regions were explored. See FRONT; GLACIOLOGY; STORM.

The average circulation over the Antarctic ice sheet consists of inflow in the middle and upper troposphere and outflow in a shallow surface layer descending along the ice sheet; a similar katabatic drainage of cold surface air toward the coast occurs

Area	Net radiation, gain (+) or loss (–) by surface	Energy from above, gain (+) or loss (-) by surface	Energy from below, gain (+) or loss (-) [also energy used for melt]
Arctic tundra			
Summer	104	-75	-29
Winter	-17	15	2
Arctic sea ice			
Summer	74	-13	-61
Winter	-29	15	14
Antarctic coast			
Summer	31	-6	-25
Winter	-41	7	34
Antarctic plateau			
Summer	20	-18	-2
Winter	-19	17	2

also in Greenland. The Antarctic katabatic winds become extraordinarily strong in some coastal regions. Thus, at Port Martin, Adelie Land, in 1951 there was recorded a 24-h mean wind of 108.6 mi/h (48.5 m/s); the strongest monthly mean wind reached 65 mi/h (29.1 m/s). Terra Nova Bay, Victoria Land, is another strong-wind location where the katabatic outflow can actually be seen in infrared satellite photos as a warmer plume over the colder sea ice. The katabatic winds are sufficiently strong and persistent, even inland, to cut up the snow surface into sharp ridges. Known as sastrugi, from the Russian word



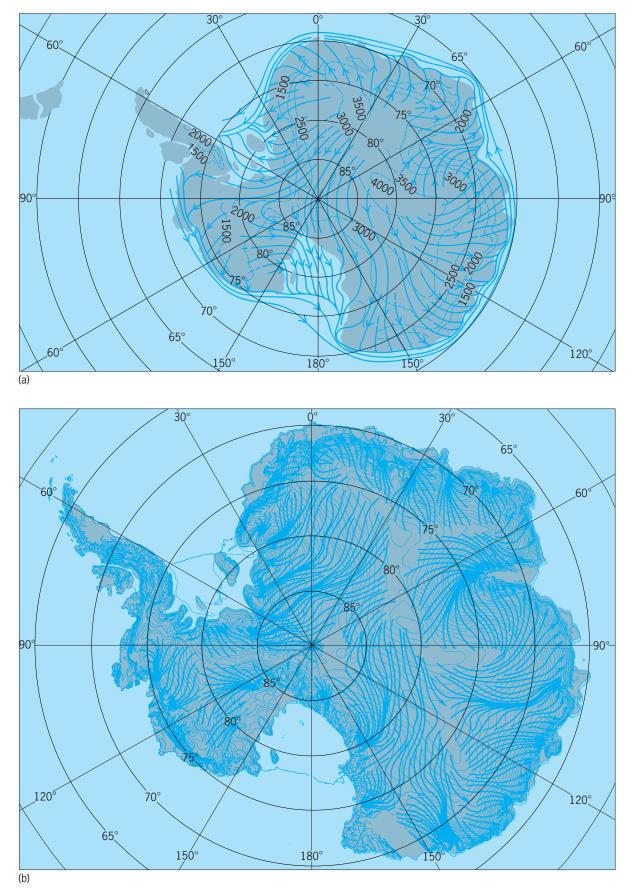
Fig. 2. Automatic weather station on the Antarctic ice sheet. (Courtesy of G. Wendler)

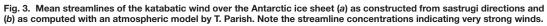
for a wood plane, these winds have been mapped from the air and in satellite photos to establish the broad directional pattern of the surface flow (**Fig.** 3*a*). Computer model simulations of the katabatic flow have refined these patterns and identified the major outflow concentrations (Fig. 3*b*). See TRO-POSPHERE; WIND.

In the long polar winter night the upper layers of the atmosphere become very cold, and a persistent stratospheric jet stream is created at the boundary between the warmer sunlit and colder dark parts of the atmosphere. Inside the stable polar vortex thus formed, the stratospheric ozone previously created by solar radiation decays until the return of the Sun in spring destroys the vortex. Over Antarctica, that normal decay has become enhanced, apparently through reactions of the ozone with chemicals produced by human activities; if this process continues, this ozone hole could ultimately weaken the global ozone shield. *See* ATMOSPHERE; STRATO-SPHERIC OZONE; JET STREAM; STRATOSPHERE.

Precipitation. Polar precipitation, like solar radiation, faces a physical limitation: the atmosphere's saturation limit for water vapor, which decreases from about 0.38% of atmospheric mass (6.1 hPa or 6.1 mbar vapor pressure) at 32° F (0°C) to 2.5×10^{-3} % (0.04 hPa or 0.04 mbar) at $-58^{\circ}F(-50^{\circ}C)$. Both polar regions receive around 8 in. (20 cm) of water as annual precipitation, one-fifth of the average global precipitation rate of 37.4 in. (95 cm) per year. Since the area inside the two polar circles represents 8.3% of the global surface, the polar precipitation accounts for a mere 1.7% of the total global precipitation of 115,000 mi³ (480,000 km³) per year. Special interest attaches to the snow, which over millions of years has created the Antarctic and Greenland ice sheets; together they presently contain 6.5×10^6 mi³ (2.7 × 10^7 km³) of ice, which if melted would raise sea level by 203 ft (62 m); this figure takes into account a large portion of the Antarctic ice presently afloat or below sea level.

Since the yearly global precipitation equals 35 times the average atmospheric water-vapor content of some 3130 mi^3 (13,000 km³) of water, the latter has





an atmospheric residence time of only 11 days. By contrast, the small fraction deposited as snow on the ice sheets consolidates to ice and remains stored for thousands of years until it is discharged in icebergs. During that time the ice acquires a wealth of information on changes in polar and global climate, in the atmospheric composition, and even in the height of the ice sheet surface. The climatic information is provided by the isotopic composition of the ice, which contains fewer (more) heavy oxygen (¹⁸O) and hydrogen (²H) atoms in periods of lower (higher) temperatures; furthermore, the dust content of the ice has changed markedly with the changing strength of the prevailing winds and extent of deserts. Air bubbles trapped in the ice have revealed past changes in atmospheric composition [in particular, lower amounts of carbon dioxide (CO2) and methane (CH₄)] during glacial periods. Moreover, the amounts of air enclosed have provided the atmospheric pressure (and hence the elevation) of the ice sheet surface at the time that its snow was compacted to ice. See GLACIAL EPOCH; PALEOCLIMATOLOGY.

The snow accumulating in the course of a year provides the best estimate of the current precipitation rate, which is difficult to measure directly in the windy conditions prevailing on the ice sheets. The winds are accompanied by almost incessant snow drift that has been calculated to carry during a year as much as 1×10^8 tons of snow across each 0.6 mi (1 km) of coastline at one of the coastal strongwind sites. Inland transports are one to two orders of magnitude smaller and directed more nearly along the height contours (Fig. 3); consequently only about 5% of the annual accumulation on the Antarctic ice sheet and 1% of that on the Greenland ice sheet leave the ice sheets as drifting snow. The total amount of snow gained by the ice sheets each year and the ice losses due to melting and iceberg calving determine whether the ice sheets are tending to grow or shrink in the current polar climate; no firm answer has yet been reached to this question that has a strong bearing on future sea levels. See ICE-BERG; PRECIPITATION (METEOROLOGY); SNOW.

Physiological aspects of polar climate. The windblown snow is a feature of the polar climate that can cause serious problems. It can markedly restrict low-level visibility and create a so-called white-out in which all contrasts are eliminated for the eye by a uniform scattering of light in all directions. In combination with low temperatures, the strong polar winds produce large body-heat losses. The windchill temperature in a 45 mi/h (20-m/s) wind at -40° F (-40°C) is -121°F (-85°C), and at -112°F (-80°C) even a mere 11.2-mi/h (5-m/s) wind feels like its temperature is $-116^{\circ}F$ ($-110^{\circ}C$). A further effect is a loss of body moisture to the very dry polar air; just breathing that air, with a moisture content of around 10 oz/ft³ (0.1 g/m), has been estimated to involve a body loss of 0.7 fluid oz (0.3 liter) of water or more in 24 h. See BIOMETEOROLOGY: COMFORT TEMPERA-TURES Llwe Radok

Bibliography. Atmospheric sciences in Antarctica: Special section, *Rev. Geophys.*, vol. 26, 1988; J. O. Fletcher and U. Radok, The role of polar regions in climate dynamics, Appendix A of Polar Research Board, *The Polar Regions and Climatic Change*, 1984; Y. Herman (ed.), *The Arctic Seas: Climatology, Oceanography, Biology, and Geology*, 1990; S. Orvig (ed.), *Climates of the Polar Regions*, World Survey of Climatology, vol. 14, 1971; Polar Group, Polar atmosphere-ice-ocean processes: A review of polar problems in climate research, *Rev. Geophys. Space Phys.*, 18:525-543, 1980; W. Schwerdtfeger, *Weather and Climate of the Antarctic*, 1984.

Polar molecule

A molecule possessing a permanent electric dipole moment. Molecules containing atoms of more than one element are polar except where forbidden by symmetry; molecules formed from atoms of a single element are nonpolar (except ozone). The dipole moments of polar molecules result in stronger intermolecular attraction, increased viscosities, higher melting and boiling points, and greater solubility in polar solvents than in nonpolar molecules. *See* DIPOLE MOMENT.

The electrical response of polar molecules depends in part on their partial alignment in an electric field, the alignment being opposed by thermal agitation forces. This orientation polarization is strongly temperature-dependent, in contrast to the induced polarization of nonpolar molecules. *See* FERROELECTRICS; MOLECULAR STRUCTURE AND SPECTRA; PERMITTIVITY; POLARIZATION OF DIELECTRICS. Robert D. Waldron

Polar navigation

The complex of navigational techniques modified from those used in other areas to suit the distinctive regional character of polar areas. Although polar navigation has become routine to a rising number of navigators operating in and through such highlatitude parts of the word, their success continues to be based on a sound grasp of the regional differences and the developing adaptations of navigational principles and aids to suit these peculiar area needs. This article therefore singles out salient physical differences, from those of other areas, which demand navigational modifications and relates them to the best-known special techniques and applications.

Coordinates, directions, and charts. In polar regions the meridians radiate outward from the poles, and parallels are concentric circles. Thus the rectangular coordinates familiar to the navigator accustomed to using the Mercator projection are replaced by polar coordinates. A rhumb line is of little use in high latitudes. The usual significance of direction decreases as latitude increases, disappearing at the poles, where all directions are south (or north).

Polar navigational grid. For navigation, it is convenient to restore rectangular coordinates. This is done by placing a series of parallel lines on the chart, and

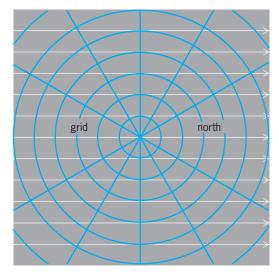


Fig. 1. Polar stereographic graticule with polar grid.

letting them represent fictitious meridians. This is called a navigational grid. Generally, these lines are parallel to the Greenwich meridian, and grid north is taken as north along this meridian (**Fig. 1**). In this way, grid and true directions near the poles are interconverted by the addition or subtraction of longitude. Variation of the compass (magnetic declination) is replaced by grid variation, or grivation, which is the angle between a magnetic meridian and a grid meridian. Thus grid and magnetic directions are interconverted by the addition or subtraction of grid variation.

Selection of chart projection. As latitude increases, the Mercator chart so familiar to the marine navigator of lower latitudes becomes of decreased utility. In subpolar regions—roughly in the area between the 60° and 70° parallels of latitude—the Lambert conformal projection is used increasingly for nautical charts. In higher latitudes both nautical and aeronautical charts are generally on a polar projection. The polar stereographic and transverse Mercator projections are used most commonly, but the modified Lambert conformal (Ney's), polar gnomonic, and polar azimuthal equidistant projections are also used. *See* MAP PRO-JECTIONS.

Light conditions. At the poles the day and year become synonymous as regards the proportion of daylight to darkness. The duration of twilight is lengthened proportionately, lasting several weeks. About 32 h elapse from the time the lower limb of the Sun touches the horizon until the upper limb disappears at sunset, during which time the Sun travels $1^{1}/_{3}$ times (480°) around the horizon. The period of several days that occurs each spring and autumn when the Sun is below the horizon but too close to it for other celestial bodies (except the Moon) to be visible is a critical time for navigation, because of the great dependence upon celestial bodies for both position and direction determination.

Celestial sphere and time patterns. The diurnal motion of celestial bodies is essentially horizontal. Half of the celestial sphere is always visible at each pole, and the other half is always below the horizon. Time loses its usual significance in regions where all time zones come together. The time of any zone would be about equally satisfactory. It is common practice to keep Greenwich mean time. *See* TIME.

Meteorological factors. Temperatures are generally low, but not as low as might be supposed. It is true that the coldest atmospheric temperatures recorded have been on the Antarctic continent, but temperatures at the North Pole are never as low as some recorded in Yellowstone National Park. Although few places in Antarctica ever get above freezing, much of the Arctic does so regularly each summer. A temperature of 100°F (38°C) has been recorded on the Arctic Circle. *See* POLAR METEOROLOGY.

The polar air is relatively dry, but because of the low prevailing temperatures, fog and clouds are common. Visibility may also be limited by blowing snow. In most parts of the polar regions, precipitation is so light that these areas are sometimes classed as deserts. Nevertheless, over much of both polar areas the ice does not completely melt during the summer, and large quantities of ice are left both on the land and in the sea. When there is no fog, the atmosphere is exceptionally clear. In the clear, cold air of polar regions sounds travel great distances. It is not unusual to hear the bark of a dog 10 mi (16 km) away. *See* ATMOSPHERIC ACOUSTICS.

The strongest surface winds in the world are probably encountered in certain regions of Antarctica, where speeds of more than 200 knots (100 m/s) are not unusual. Over the Arctic Ocean, however, strong winds are not encountered except in some regions near land.

The conditions described are primarily responsible for the differences between navigation in polar regions and elsewhere. There is no well-defined line of demarcation between polar regions and subpolar regions, nor is there universal agreement as to the definition of polar regions. For the purposes of this discussion, however, the parallel of latitude 70° can conveniently be used as a general dividing line. Polar grid navigation is usually limited to the latitudes poleward of this parallel.

Piloting hazards. Piloting in polar regions is strongly affected by the absence of any great number of aids to navigation. Also, natural landmarks may not be shown on the chart, or may be difficult to identify. The appearance of some landmarks changes markedly under different ice conditions. When snow covers both the land and a wide ice foot attached to the shore and extending for miles seaward, even the shoreline is difficult to locate. Adjacent islands sometimes merge together as the straits between them fill up completely with ice. Along a rugged coast such as that of much of Greenland, snow-covered headlands may look alike. *See* PILOTING.

Unreliability of charts. Charts of polar regions are less reliable than those of other regions, because relatively little surveying has been done in the polar areas. Attempts to fix the position of a craft can be discouraging when the various landmarks used are not charted in correct position relative to each other. Marine navigators sometimes plot positions relative to land known to be shown in the wrong place, rather than in the correct geographical positions, because it is the land and its attendant shoals that constitute danger to their vessels.

Because relatively few soundings are shown on charts, ships entering harbors often send small boats ahead to determine the depth of water available. The polar marine navigator finds a knowledge of geology useful in predicting safe areas and those areas in which rocks and shoals might be encountered. Ice concentrations and movement may also be an indication of the relative depth of the water.

The reliability of charts of polar regions has steadily improved as additional information has become available.

Limitations of electronic applications. Coverage of electronic navigation systems has also improved, but is still somewhat limited. Omega signals are available throughout both polar regions. However, there are several sources of error in the propagation of the very low frequencies utilized by the Omega system. Some of these are predictable, permitting corrections to be applied. But temporary sources of anomalous propagation identified as sudden ionospheric disturbances caused by influx of x-rays from the Sun and polar cap disturbances associated with proton enhancement by the Sun are more troublesome. In the Arctic, rapid attenuation of very low frequency signals is experienced when the signals cross the Greenland ice cap. Additionally, user equipment-installation problems can be severe, notably in aircraft, where electromagnetic interference problems are particularly troublesome. For maximum effective utilization of Omega signals, accurate absolute time is available, resulting in dependence on time signals from an external source. Loran C sky waves are available throughout the Arctic, and ground waves extend to some parts of this area, but neither ground waves nor sky waves are available in the Antarctic. See LORAN.

Radar is useful, but experience in interpretation of the scope in polar regions is essential for reliable results. This is particularly true in aircraft, where the relative appearance of water and land areas often reverse in winter and summer. Hummocked ice presents a different appearance from unbroken ice. A radio direction finder is useful, when radio signals are available. Except along the northern coast of Alaska and Russia, few radio beacons are available. The use of electronics in polar regions is further restricted by magnetic storms, which are particularly severe in the auroral zones. *See* ELECTRONIC NAVIGA-TION SYSTEMS; RADAR.

Difficulties of dead reckoning. Reliable dead reckoning depends upon the availability of accurate measurement of direction and distance (or speed and time). There are difficulties in meeting these requirements in polar regions. *See* DEAD RECKONING.

Compass limitations. Direction is measured largely by a compass. The compasses in common use are the magnetic, in which the directive element attempts to align itself with the horizontal component of the Earth's magnetic field, and the north-seeking gyro, in which the directive element attempts to align itself with the Earth's axis of rotation. The magnetic compass becomes unreliable in the vicinity of the magnetic poles of the Earth, and the north-seeking gyrocompass becomes unreliable in the vicinity of the geographical poles of the Earth. The magnetic and geographical poles are both in high latitudes, and the areas in which the compass errors are large and somewhat erratic overlap. In these areas frequent checks and comparisons are needed.

One solution to the directional problem in high latitudes is to use a directional gyro, a gyroscopic device that maintains its axis in a set direction but must be reset at frequent intervals, as every 15 min, because of gyroscopic drift. One of the devices discussed below for celestial direction determination can provide the directional data needed for setting or resetting a directional gyro. Directional gyros are not generally used aboard ship, and in aircraft they have been largely replaced by inertial navigators, which are also used by submarines and to a limited extent by surface ships. However, because inertial systems are self-contained, without reference to external signals, the insertion of incorrect data is always a possibility. Proposals to remedy this situation by means of geostationary satellites have encountered the problem that such satellites do not provide coverage beyond 65° latitude. See AIRCRAFT COMPASS SYSTEM; GYRO-SCOPE; INERTIAL GUIDANCE SYSTEM.

Celestial determination of direction. Direction can be determined by means of celestial bodies, but this is usually an instantaneous indication used to check the compasses, unless a device equipped with a clock mechanism is used to provide more continuous information. Several types of devices have been developed to facilitate the use of celestial bodies for determination of direction. The oldest is the sun compass, which utilizes the shadow of a shadow pin, or gnomon, and a suitable dial. This, of course, will not operate unless the Sun is visible. A sky compass indicates direction of the Sun by means of polarized light in the sky when the Sun is near the horizon, even though it may be below the horizon or otherwise obscured. This device may offer the only means of determining direction during the brighter part of the long polar twilight. A Canadian version is called a twilight compass because of the principal period of its use. An astrocompass (Fig. 2) can be set for the coordinates of any celestial body and the latitude of the observer, and then gives an indication of azimuth, true north, and heading.

Distance or speed problems. Such measurement in polar regions presents no problems in aircraft. When ships operate in ice, however, the sensing element in the water may be adversely affected or damaged by the ice. A method of determining speed or distance that has proved successful has been to track an iceberg or other prominent feature, either visually or by radar. If the feature is stationary, as a grounded iceberg, the result is speed or distance over the ground.

At best, dead reckoning is difficult aboard a ship operating in the ice, not only because of the difficulty encountered in measuring course and speed, but also because neither of these may be constant for very long. Land ice which flows down to the sea and breaks off in the form of icebergs or ice islands is not usually a problem, and may even prove beneficial.

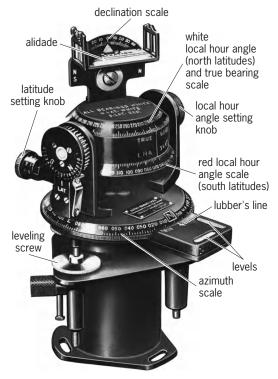


Fig. 2. An astrocompass.

Its use as an aid in the determination of speed has been mentioned. Individual pieces are usually so large that they move with deep-water currents, often in a direction differing from that of the sea ice, which moves mainly in the direction the wind blows. Thus an iceberg might clear a path in the desired direction of motion. One precaution is essential, however. It is dangerous to approach close to an iceberg because of possible underwater rams which might extend out for some distance from the berg and also because it is not uncommon for an iceberg to acquire unstable equilibrium, because of uneven melting, and capsize. *See* ICEBERG.

Ice formed at sea, called sea ice, is seldom an unbroken sheet over any very large area. The unequal pressure exerted by tides, currents, winds, and temperature changes produce stresses that break the ice and move different parts of it relative to each other, producing leads, long cracks that have opened wide enough to permit passage of a ship; polynyi, large areas, other than leads, relatively free from ice; or pressure ridges, ridges of ice piled up where two floes have come together under pressure. *See* SEA ICE.

A large field of floating pieces of ice which have drifted together is called a pack. If this is relatively loose (**Fig. 3**), a ship, skillfully handled, can negotiate it. If it is packed tightly under pressure, however, it is best avoided.

Successful negotiation of pack ice is the result of working with, not against, the pack, seeking out weak spots, ramming when appropriate, and retreating at other times, taking advantage of leads and polynyi, and avoiding heavy pressure that might stop the vessel or crush it. It is usually considered better to keep moving in the general direction desired than to risk complete stoppage with possible damage by staying exactly on the desired course. All this, plus a lack of detailed information on currents, with little opportunity to acquire it, makes marine dead reckoning difficult in polar regions.

Celestial navigation problems. Celestial navigation is of great importance in polar regions, sometimes providing the only means of determining position accurately, or establishing a directional reference.

In aircraft the use of a sextant is the same as it is in other latitudes. Aboard ship a marine sextant can be used when celestial bodies and a horizon are both available, an occurrence more frequent in most ocean areas at lower latitudes. Since the pack smooths the sea, resulting in very little rolling or pitching of a ship operating in ice, some marine navigators have found an artificial-horizon sextant of the type carried in aircraft useful. If the acceleration error is too great aboard ship, observations from a nearby ice cake may be possible. *See* SEXTANT.

When operating in lower latitudes, navigators generally avoid observation of bodies near the horizon because of the uncertainty of the refraction correction there. In polar regions, even though refraction is more uncertain, navigators often have no choice. Near the equinoxes the Sun may be the only body available for several weeks, and it remains close to the horizon. Under these conditions, the navigator observes the Sun and makes the most of the available information, usually with satisfactory results.

During the polar summer the Sun is often the only celestial body available. The Moon is above the horizon half the time, but is in a favorable position relative to the Sun during relatively few days each month. When only one body is available, it is observed at frequent intervals, perhaps hourly, and a series of running fixes is plotted.

Timing of celestial observations is less critical in polar regions than in lower latitudes because apparent motion of the bodies is nearly horizontal. A relatively large error in longitude, the effect of an error in time, is a relatively small error in miles in an area where the meridians are close together.

Sight reduction in polar regions is not attended by any unusual problems. The same methods in common use in lower latitudes are used in polar regions. Simplified tables might be prepared, but they are generally avoided in favor of familiar methods. *See* CELESTIAL NAVIGATION.



Fig. 3. Ship negotiating pack ice.

Impact of modern technology. Advances in technology have contributed significantly to the safety and reliability of navigation in polar regions. The availability of modern echo sounders has made possible a continuous plot of the bottom profile beneath a vessel while under way. Sonar, operating on the same principle as an echo sounder, indicates the presence of an underwater obstruction. Reliable inertial navigators having an error of less than 1 mi/h (0.5 m/s) are available to make aircraft navigation in polar regions almost routine, providing both positional and directional data. Ship inertial navigation systems are available and have made practicable submarine operations under sea ice. *See* ECHO SOUNDER; SONAR.

The U.S. Navy Navigation Satellite System (NNSS) provides means for intermittent position determination, with an error of a small fraction of a mile, for craft equipped with suitable instrumentation. Because of the need for knowledge of craft velocity vector to achieve such accuracy and intermittent availability, ships have benefitted from satellites more than aircraft. However, because all satellites of the system transmit similar signals on the same frequencies (150 and 400 MHz), both ships and aircraft encounter mutual interference between satellites when more than one is above the horizon at the same time, a frequent occurrence in high latitudes, where satellite orbits converge.

The NAVSTAR Global Positioning System (GPS) of the U.S. Department of Defense, when fully implemented operationally, will eliminate the limitations of the Navy Navigation Satellite System. Although all GPS satellites also operate on the same frequencies (1227 and 1575 MHz), signals are modulated with two codes, one of which provides for identification of and lock-on to the desired signal. The former Soviet Union advised member nations of the International Telecommunications Union (ITU) that it was developing a navigation satellite system similar to GPS, to be known as GLONASS. *See* SATELLITE NAVI-GATION SYSTEMS.

The establishment of all eight Omega navigation system transmitters on an operational basis has provided an electronic means of continuously fixing the position of a craft anywhere in the world, including both polar regions.

The overall impact of modern technology has been to remove much of the uncertainty and hazard of polar navigation, which approaches the high reliability attained previously in lower latitudes. *See* NAV-IGATION. Alton B Moody

Bibliography. N. Bowditch, American Practical Navigator, Publ. 9, vol. 1, 1977; L. Logsdon, The Navstar Global Positioning System, 1992; A. B Moody, Navigation Afloat—A Manual for the Seaman, 1980.

Polarimetric analysis

A method of chemical analysis based on the optical activity of the substance being determined. Optically active materials are asymmetric; that is, their molecules or crystals have no plane or center of symmetry. These asymmetric molecules can occur in either of two forms, *d*- and *l*-, called optical isomers. Often a third, optically inactive form, called meso, also exists. Asymmetric substances possess the power of rotating the plane of polarization of planepolarized light. Measurement of the extent of this rotation is called polarimetry. Polarimetry is applied to both organic and inorganic materials. *See* OPTICAL ACTIVITY.

The extent of the rotation depends on the character of the substance, the length of the light path, the temperature of the solution, the wavelength of the light which is being used, the solvent (if there is one), and the concentration of the substance. In most work, the yellow light of the D line of the sodium spectrum (589.3 nanometers) is used to determine the specific rotation, according to the equation below. Here α is the measured angle of rotation,

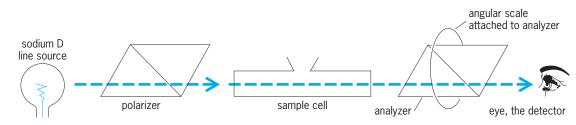
Specific rotation =
$$\left[\alpha\right]_{\rm D}^{20} = \frac{\alpha}{l\rho}$$

l is the length of the column of liquid in decimeters, and ρ the density of the solution. In other words, the specific rotation is the rotation in degrees which this plane-polarized light of the sodium D line undergoes in passing through a 10-cm-long (4-in.) sample tube containing a solution of 1 g/ml concentration at 20°C (68°F).

In the **illustration**, light from the sodium lamp is polarized by the polarizer (a fixed nicol prism) before it passes through the cell containing the material being analyzed. After the light passes through the cell, it passes through the analyzer (another nicol prism) and then is detected by the eye or a photocell. A comparison of the angular orientation of the analyzer as measured on the scale with the cell empty and with the cell filled with solution serves to measure the rotation of the polarized light by the sample. This rotation may be either clockwise (+) or counterclockwise (-), depending on the substance in question.

Polarimetry may be used for either qualitative or quantitative analytical work. In qualitative applications, the presence of an optically active material is shown, and then a calculation of specific rotation often leads to the identification of the unknown. In quantitative work, the concentration of a given optically active material is determined either from a calibration curve of percentage of the constituent versus angular rotation or from the specific rotation, assuming the angular rotation to be a linear function of concentration. For this method of analysis to be useful, it is necessary that only one optically active material be present in solution.

Polarimetry is used in carbohydrate chemistry, especially in the analysis of sugar solutions. Polarimeters used for this work are specially designed and are called saccharimeters. Other materials often determined by polarimetry are tartaric acid, Rochelle salt (potassium sodium tartrate), various terpenes such as *d*- and *l*-pinene, many steroids, and other compounds of biological and biochemical importance. Since there is great difference between the biological



Simplified diagram of a polarimeter.

activities of the different optical forms of organic compounds, polarimetry is used in biochemical research to identify the molecular configurations.

Optical rotatory dispersion is the measurement of the specific rotation as a function of wavelength. To accomplish this, the sodium lamp is replaced by a monochromator, and a source of continuous radiation. A photocell circuit is substituted for the eye as a detector. In this way, the specific rotation may be determined in the ultraviolet or near-infrared, as well as in the visible region of the spectrum.

The information obtained by optical rotatory dispersion has shown that minor changes in configuration of a molecule have a marked effect on its dispersion properties. By using the properties of compounds of known configuration, it has been possible to determine the absolute configurations of many other molecules and to identify various isomers. Most of the applications have been to steroids, sugars, and other natural products, including amino acids, proteins, and polypeptides. *See* COT-TON EFFECT; OPTICAL ROTATORY DISPERSION; POLAR-IZED LIGHT. Robert F. Goddu; James N. Little

Bibliography. E. Charney, *The Molecular Basis of Optical Activity*, 1979, reprint 1985; I. M. Koltoff and P. J. Elving, *Treatise on Analytical Chemistry*, pt. 1, vol. 6, 1965; S. F. Mason, *Molecular Optical Activity and the Chiral Discriminations*, 1982; G. Snatzke (ed.), *Optical Rotatory Dispersion and Circular Dicbroism in Organic Chemistry*, 1976.

Polarimetry

The science of determining the polarization state of electromagnetic radiation (x-rays, light or radio waves). Radiation is said to be linearly polarized when the electric vector oscillates in only one plane. It is circularly polarized when the x-plane component of the electric vector oscillates 90° out of phase with the *y*-plane component. In experimental work, the polarization state is usually expressed in terms of the Stokes parameters I (the total intensity of the radiation), Q (the preference for light to be linearly polarized in the x plane or at 0° in the instrument reference system), U (the preference for linear polarization at 45°), and V (the preference for left-circular polarization). When Q/I = 1, for example, the radiation is 100% linearly polarized along an axis at 0° . Similarly, V/I = -1 means that the light is 100% rightcircularly polarized.

To completely specify the polarization state, it is necessary to make six intensity measurements of the light passed by a quarter-wave retarder and a rotatable linear polarizer, such as a Polaroid or a Nicol prism.

The retarder converts circular light into linear light. It is placed upstream of the polarizer and is moved into the beam to determine Stokes *V*. Then *I*, *Q*, *U*, and *V* can be calculated from the six intensity measurements of the table. For fully polarized light, $I^2 = Q^2 + U^2 + V^2$. If the light is only partially polarized, the degree of polarization *P* is given by the equation below.

$$P = \sqrt{\frac{Q^2 + U^2 + V^2}{I^2}}$$

Polarization of radiation. An ensemble of atoms in a hot gas, such as a stellar atmosphere, usually radiates in no preferred state of polarization. Most starlight is unpolarized. However, atoms in the presence of a magnetic field align themselves at fixed, quantized angles to the field direction. Then the spectral lines they emit are circularly polarized when the magnetic field is parallel to the line of sight and linearly polarized when the field is perpendicular. In 1908, G. E. Hale used this property in solar spectral lines to prove that sunspots have magnetic fields of 1000-3000 gauss (0.1-0.3 tesla). For comparison, the Earth's magnetic field is 0.3 gauss (30 microteslas). Fields as high as 100 megagauss (10,000 T) have been measured in white dwarf stars. See SOLAR MAG-NETIC FIELD; SUN; WHITE DWARF STAR; ZEEMAN EF-FECT

The light from sunspots is polarized because the magnetic fields impose some direction in the emitting gas. Other phenomena also remove isotropy and produce polarization. Sunlight scattered by electrons in the solar corona is partially linearly polarized because the radiation from the solar surface is much greater than the radiation into the surface (anisotropic radiation). Similarly, beams of electrons or protons will excite partially polarized light when they strike a gaseous cloud. Polarimetry can reveal the presence of the atomic beams.

Synchrotron emission and scattering are important polarizers of light. Lord Rayleigh showed that scattering by the barbell-shaped molecules of the Earth's atmosphere produces partial linear polarization of the blue sky. Similarly, needle-shaped grains in interstellar dust clouds polarize the starlight that passes through them whenever magnetic fields have aligned the grains. By measuring the angle and degree of polarization in many stars, astronomers have been able to map the magnetic fields of the Milky Way Galaxy and many other galaxies. Galactic magnetic fields are typically only a few microgauss (a fraction of a nanotesla). *See* INTERSTELLAR MATTER; MILKY WAY GALAXY.

Synchrotron radiation is produced when electrons moving at nearly the speed of light spiral rapidly in a magnetic field. Because of the centripetal acceleration, the electrons emit linearly polarized light or radio waves. The Crab Nebula is a well-known source of highly polarized synchrotron radiation from the ultraviolet to the radio regime. The inferred field strength in the nebula is approximately 100 microgauss (10 nT). Radio galaxies, quasars, and active galactic nuclei also produce polarized synchrotron radiation. Sometimes the degree of polarization *P* is as high as 50%. *See* CRAB NEBULA; GALAXY, EXTERNAL; QUASAR; RADIO ASTRONOMY.

The Wilkinson Microwave Anisotropy Probe (WMAP) mission mapped the polarization of the cosmic microwave background radiation over the full sky. Polarization has been detected that arises from the anisotropic scattering of the photons off free electrons. This occurred both at the decoupling epoch and when the first stars reionized the universe. WMAP and other experiments are searching for a fainter polarization pattern that is predicted by inflation theory, arising not from scattering but from the gravitational waves produced during the sudden expansion of the universe. See COSMIC BACKGROUND RADIATION; GRAVITATIONAL RADIATION; INFLATION-ARY UNIVERSE COSMOLOGY; WILKINSON MICROWAVE ANISOTROPY PROBE.

Magnetic fields penetrate every corner of the universe. They are generated by dynamos deep inside stars and planets and possibly in galaxies, but the dynamo mechanism is not well understood. And magnetic fields must have played some role in the evolution of the universe. These are good reasons to continue to develop more efficient and accurate polarimetry. *See* MAGNETOHYDRODYNAMICS.

Measurement of polarization. Electrooptical devices are rapidly replacing rotating polarizers and fixed retarders. To make daily maps of the weak magnetic fields on the Sun, H. W. Babcock invented the magnetograph, which consists of a spectrograph to isolate the atomic spectral line for study; a Pockels cell, an electrooptic crystal whose retardance depends on an applied voltage; a polarizing prism to isolate the polarization state passed by the retarder; a pair of photocells to detect the transmitted light; and a scanning mechanism to sweep the solar image across the spectrograph entrance slit. Two photocells are needed to simultaneously measure left- and right-circular polarization. *See* SPECTROGRAPH.

Since Babcock's magnetograph measured only circular polarization, it recorded only the line-of-sight, or longitudinal, component of the field. The sensitivity was about 5 gauss (500 μ T). A magnetograph can be made sensitive to linear polarization, but the signal levels are about 100 times weaker for the in-

ferred transverse fields than for longitudinal fields of comparable strength. In this case, longer measurement times are required to obtain useful signal-tonoise levels.

To improve signal-to-noise levels while measuring weak linear or circular polarization over a wide field, the spectrograph can be replaced with an optical filter having a narrow passband, and the photocells can be replaced with an array of photosensitive picture elements (pixels). In this manner, the narrow portion of the spectrum most sensitive to the effect (Zeeman effect) of magnetic fields can be isolated and recorded at many points simultaneously.

Most classical polarizing optical devices are effective over only narrow fields of view, sometimes as little as 1°. This characteristic severely limits the amount of light that can be analyzed at once. The introduction of polarizing dichroic films (Polaroids) was a great improvement over classical polarizing prisms (for example, the Nicol prism) because of their wide acceptance angle. It is more difficult to make a wide-angle retarder. The widest-angle device is the photoelastic modulator, invented by J. Kemp. The photoelastic modulator can be made to alternately pass right- and left-circular light over a field with a half-angle of 20° . The photoelastic modulator aperture is generally limited to about 5 cm (2 in.), but this dimension is about twice that of the aperture of classical prisms.

Some devices for detecting linear and circular polarization depend on scattering. To detect circular polarization in sunlight at the wavelength of the sodium line (589 nanometers), a cell filled with sodium vapor is sometimes used. Such a resonance cell is placed inside a laboratory magnet. Then the atoms preferentially absorb and reemit the circularly polarized component of the incident light. The resonance cell is very stable and accepts light from a very large angle.

In the infrared, glass containing tiny aligned silver needles can discriminate efficiently between the two senses of linear polarization because the needles scatter electromagnetic waves aligned with their long axis far more efficiently than they scatter nonaligned waves. David M. Rust

Bibliography. A. Adamson et al., *Astronomical Polarimetry: Current Status and Future Directions*, 2005; J. O. Stenflo, *Solar Magnetic Fields: Polarized Radiation Diagnostics*, 1994; J. Tinbergen, *Astronomical Polarimetry*, 1996, paper 2005; J. C. del Toro Iniesta, *Astrophysical Spectropolarimetry*, 2003.

Polaris

The star α Ursae Minoris, also known as the North Star or Pole Star. It is perhaps the best-known star in the northern sky. Its location only 1 degree of arc from the north celestial pole, the point where the Earth's rotation axis intersects the celestial sphere, has made it a very useful reference point for navigation. It may easily be found by following the line joining the two bright stars at the end of the bowl of the Big Dipper. *See* URSA MAJOR; URSA MINOR.

Polaris (apparent magnitude 1.99) is a supergiant of spectral type F7 or F8, with an intrinsic brightness about 1500 times that of the Sun. Its effective temperature is approximately 6000 K (10,300°F). It is accompanied by a 9th-magnitude main-sequence star that has the same proper motion, at an angular separation of 18". In addition, its spectrum shows evidence of another, much closer companion in an eccentric orbit with a period of 30 years. *See* SPEC-TRAL TYPE; SUPERGIANT STAR.

Polaris is a variable star, displaying slight changes in brightness with a period close to 4 days. The radial velocity also changes with the same period indicating that the diameter of the star is changing periodically. Polaris is a member of an important group of stars known as the Cepheid variables, which are typically high-luminosity stars that pulsate because of internal energy imbalances. Cepheids serve as an important tool for measuring distances in the Milky Way Galaxy and beyond because they are easily recognized by their characteristic patterns of light variation. Polaris is the brightest and nearest of the stars of this class, at a distance of approximately 130 parsecs from the Sun (4 \times 10¹⁵ km or 2.5 \times 10¹⁵ mi). However, it is atypical in that the amplitude of the variations is very small compared to other Cepheids and has decreased steadily over 100 years to the point where the pulsation of the star has virtually stopped. See CEPHEIDS; STAR; VARIABLE STAR. David W. Latham

Polarization of dielectrics

A vector quantity representing the electric dipole moment per unit volume of a dielectric material. *See* DIELECTRIC MATERIALS; DIPOLE MOMENT.

The polarization \mathbf{P} , measured in coulombs/m², is related to the macroscopic electric parameters by the equation shown below, where \mathbf{D} is the electric

 $\mathbf{P} = \mathbf{D} - \epsilon_0 \mathbf{E} = \chi \epsilon_0 \mathbf{E} = \epsilon_0 (\epsilon_r - 1) \mathbf{E}$

displacement, **E** is the electric field strength, ϵ_0 is the permittivity of free space (8.854 × 10⁻¹² farad/m), ϵ_r is the relative permittivity, and χ is the electric susceptibility of the material. *See* ELECTRICAL UNITS AND STANDARDS.

Dielectric polarization arises from the electrical response of individual molecules of a medium and may be classified as electronic, atomic, orientation, and space-charge or interfacial polarization, according to the mechanism involved.

Electronic polarization represents the distortion of the electron distribution or motion about the nuclei in an electric field. This polarization occurs for all materials and is nearly independent of temperature and frequency up to about 10^{14} Hz for insulators.

Atomic polarization arises from the change in dipole moment accompanying the stretching of chemical bonds between unlike atoms in molecules. This mechanism contributes to polarization at frequencies below those of the vibrational modes of molecules (about 10^{12} to 10^{14} Hz). For a discussion of molecular vibrations *see* MOLECULAR STRUCTURE AND SPECTRA

Orientation polarization is caused by the partial alignment of polar molecules (molecules possessing permanent dipole moments) in an electric field. This mechanism results in a temperature-dependent component of polarization at lower frequencies.

Space-charge or interfacial polarization occurs when charge carriers are present which can migrate an appreciable distance through a dielectric but which become trapped or cannot discharge at an electrode. This process always results in a distortion of a macroscopic field and is important only at low frequencies. *See* ELECTRIC FIELD; ELECTRIC SUS-CEPTIBILITY; PERMITTIVITY. Robert D. Waldron

Polarization of waves

The directional dependence of certain wave phenomena, at right angles to the propagation direction of the wave. In particular, ordinary light may be regarded as composed of two such asymmetrical components, referred to as its two states of linear polarization.

These two components are refracted differently by doubly refracting crystals, such as calcite, or Iceland spar. Each state of linear polarization is refracted according to its own separate refractive index. On a subsequent refraction by the same crystal, but now rotated through an angle θ about the direction of the beam, each component appears as a mixture of the original two polarization components, according to the proportions $\cos^2 \theta : \sin^2 \theta$. See BIREFRINGENCE; CRYSTAL OPTICS; REFRACTION OF WAVES.

The most familiar effect of light polarization occurs with polaroid material, such as in polaroid sunglasses. Polaroid absorbs one component of the light and allows the other component to pass through. For a perfect polaroid, the above-mentioned mixing effect could be seen by holding two such pieces against one another. For an angle θ between their polarizing directions, the second polaroid reduces the total intensity transmitted by a factor $\cos^2 \theta$.

Nature of light. In the early nineteenth century, T. Young suggested that light polarization arises from transverse oscillations. In J. C. Maxwell's theory of light as electromagnetic waves, visible light—and also other types of electromagnetic radiation such as radio waves, microwaves, and x-rays (distinguished from visible light only by wavelength)—consists of electric and magnetic fields, each oscillating in directions perpendicular to the propagation direction, the electric and magnetic field vectors being perpendicular to each other (**Fig. 1**). The plane of polarization of the wave contains the electric vector (or magnetic vector; there is no general agreement which) and the propagation direction. *See* ELECTROMAGNETIC RADI-ATION; LIGHT; MAXWELL'S EQUATIONS.

Circular polarization. If the plane of polarization remains constant along the wave (as in the case of each light component in a doubly refracting medium), the

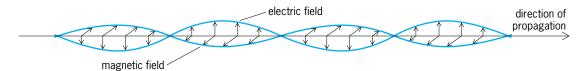


Fig. 1. Plane-polarized electromagnetic wave. (After R. Penrose, The Emperor's New Mind, Oxford University Press, 1989)

wave has linear (or plane) polarization. However, the plane of polarization can also rotate. If the rotation rate is constant, the intensity of the wave being also constant, a circularly polarized wave results (**Fig. 2**). These are of two types: right-handed and left-handed.

Any electromagnetic wave can be considered to be composed of monochromatic components, and each monochromatic component can be decomposed into a left-handed and a right-handed circularly polarized part. The states of linear polarization are each made up of equal magnitudes of the two circularly polarized parts, with differing phase relations to provide the different possible directions of plane polarization. Monochromatic waves composed of unequal magnitudes of the two circularly polarized parts are called elliptically polarized. This refers to the fact that the electric and magnetic vectors trace out ellipses in the plane perpendicular to the direction of motion.

Relation to spin. These descriptions refer to classical electromagnetic waves but apply equally well to the individual wave functions of single photons. Photons have quantum-mechanical spin, which refers to the angular momentum of the photon, necessarily about its direction of motion. A photon's spin has magnitude 1, in fundamental units. This spin can point along the direction of motion (positive helicity, right-handed spin) or opposite to it (negative helicity, left-handed spin), and this corresponds (depending on conventions used) to a classical electromagnetic wave of right- or left-handed circular polarization. *See* HELICITY (QUANTUM MECHANICS); PHOTON.

The wave functions of general, massless quantum particles of nonzero spin are similar, except that the magnitude of the spin can be some other integral or half-integral value, in fundamental units. Massless particles of half-integral spin do not give rise to classical fields, in the ordinary sense, since they are fermions (particles satisfying Pauli's exclusion principle). Particles of integral spin are the quanta of classical fields. In the half-integral case, there is the curious property that a rotation through 360° does not leave the wave function unaltered, but changes its sign. This occurs because this wave function is a spinor quantity, which indeed changes sign under a single complete rotation (about any axis). *See* BOSE-EINSTEIN STATISTICS; EXCLUSION PRINCIPLE; FERMI-DIRAC STATISTICS; SPIN (QUANTUM MECHANICS).

Classical waves. For a classical wave, corresponding to a massless quantum particle, with an integral spin, the symmetry of the field is the same as for the corresponding wave function. In particular, a gravitational wave (of Einstein's general relativity) has the same symmetry as the wave function of a graviton, a massless particle (as yet unobserved) of spin 2. For a gravitational wave traveling in one direction, the quantity describing the strength of the wave (namely the Riemann curvature tensor, describing tidal distortion) has the symmetry that a rotation through 180° about the direction of motion leaves it unaltered. If there were a classical field of spin 3, then for a wave of such a field propagating in one direction, a rotation through 120° about the propagation direction would leave the field unaltered. See GRAVITATIONAL RADIATION; GRAVITON.

In the case of zero spin, a rotation through any angle about the propagation direction will leave the field unaltered. This is the situation of a scalar wave, and it differs from all the other cases in that polarized

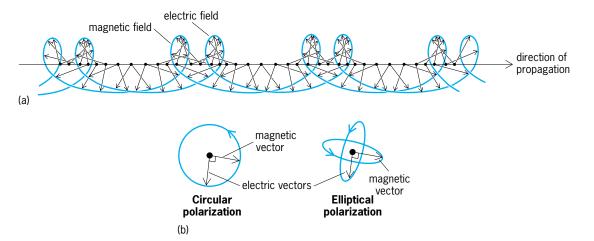


Fig. 2. Circularly polarized electromagnetic wave. (a) View along propagation direction. (b) Plane perpendicular to propagation direction for circular polarization and elliptical polarization (which is intermediate between plane polarization and circular polarization). (After R. Penrose, The Emperor's New Mind, Oxford University Press, 1989)

waves do not exist, and there is but a single state of polarization rather than two.

Orthogonal states of polarization. For a state of polarization of a general wave, changing the sign of the field (or wave function) leaves it unchanged. This applies to states of linear polarization as well as circular polarization. Thus, for an electromagnetic wave, a rotation about the propagation direction through 180° leaves a state of linear polarization unaltered. To pass to the opposite, or orthogonal state of linear polarization, it is necessary to rotate through half this angle, namely 90° . Thus, the two orthogonal states of linear polarization have polarization planes that are at right angles to each other. The plane of oscillation of the electric vector, for a particular state of linear polarization, would be at right angles to that for the orthogonal state of linear polarization.

For a gravitational wave, however, to pass to the orthogonal linear polarization, a rotation through 45° is needed. This can be understood in terms of the physical effect of a gravitational wave: inducing a tidal distortion on whatever matter it encounters. For a gravitational wave in a state of linear polarization, this tidal distortion can be pictured in terms of an ellipse which is perpendicular to the direction of propagation, this ellipse describing the distortion that is experienced by a circular pattern of particles in the path of the wave and parallel to its plane. For an oscillatory linearly polarized wave, the ellipse oscillates so that the axes of the ellipse remain constant, but the major and minor axes continually interchange with one another. The orthogonal state of linear polarization can be pictured in terms of an oscillating ellipse whose axes are at 45° to those of the original ellipse.

Elastic waves. Electromagnetic and gravitational waves both have the specific property that they are entirely transverse in character, which is a consequence of their speed of propagation being the absolute speed of relativity theory (the speed of light). This corresponds to the fact that their respective quanta, namely photons and gravitons, are massless particles. In the case of waves that travel at a smaller speed, as with fields whose quanta are massive rather than massless, there can be (unpolarized) longitudinal as well as transverse effects. Seismic waves traveling through the Earth's material, for example, can be transverse (polarized sideways oscillations) or longitudinal (unpolarized pressure waves). These waves travel at different speeds. Wave motion in other types of elastic material exhibits similar behavior. See SEIS-MOLOGY; SOUND.

Polarized and unpolarized light. In most situations encountered in practice, light (or gravitational waves) consists of an incoherent mixture of different polarization states, and is referred to as unpolarized. However, light reflected off a refracting surface (for example, glass or water) is polarized to some extent; that is, there is a certain preponderence of one state of linear polarization over the orthogonal possibility. Complete polarization occurs for a particular angle of incidence, known as the Brewster angle. Light from the sky is slightly polarized. Light from distant celestial objects is sometimes polarized (when produced by synchrotron radiation). This plane of polarization may become rotated en route to the Earth by an effect known as Faraday rotation, and this gives valuable astronomical information about the intervening magnetic fields. *See* FARADAY EFFECT; PO-LARIZED LIGHT; RADIO ASTRONOMY; REFLECTION OF ELECTROMAGNETIC RADIATION; SYNCHROTRON RADI-ATION; WAVE MOTION. Roger Penrose

Polarized light

Light which has its electric vector oriented in a predictable fashion with respect to the propagation direction. In unpolarized light, the vector is oriented in a random, unpredictable fashion. Even in short time intervals, it appears to be oriented in all directions with equal probability. Most light sources seem to be partially polarized so that some fraction of the light is polarized and the remainder unpolarized. It is actually more difficult to produce a completely unpolarized beam of light than one which is completely polarized.

The polarization of light differs from its other properties in that human sense organs are essentially unable to detect the presence of polarization. The Polaroid Corporation with its polarizing sunglasses and camera filters has made millions of people conscious of phenomena associated with polarization. Light from a rainbow is completely linearly polarized; that is, the electric vector lies in a plane. The possessor of polarizing sunglasses discovers that with such glasses, the light from a section of the rainbow is extinguished.

According to all available theoretical and experimental evidence, it is the electric vector rather than the magnetic vector of a light wave that is responsible for all the effects of polarization and other observed phenomena associated with light. Therefore, the electric vector of a light wave, for all practical purposes, can be identified as the light vector. *See* CRYSTAL OPTICS; ELECTROMAGNETIC RADIATION; LIGHT; POLARIZATION OF WAVES.

Types of polarized light. Polarized light is classified according to the orientation of the electric vector. In linearly polarized light, the electric vector remains in a plane containing the propagation direction. For monochromatic light, the amplitude of the vector changes sinusoidally with time. In circularly polarized light, the tip of the electric vector describes a circular helix about the propagation direction. The amplitude of the vector is constant. The frequency of rotation is equal to the frequency of the light. In elliptically polarized light, the vector also rotates about the propagation direction, but the amplitude of the vector on a plane at right angles to the propagation direction describes an ellipse.

These different types of polarized light can all be broken down into two linear components at right angles to each other. These are defined by Eqs. (1)

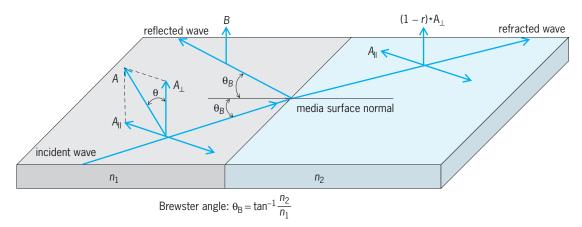


Fig. 1. Illustration of the law of Malus.

and (2), where A_x and A_y are the amplitudes, φ_x and

$$E_r = A_r \sin(\omega t + \varphi_r) \tag{1}$$

$$E_{\nu} = A_{\nu} \sin\left(\omega t + \varphi_{\nu}\right) \tag{2}$$

 φ_y are the phases, ω is 2π times the frequency, and *t* is the time. For linearly polarized light Eqs. (3) hold. For circularly polarized light Eqs. (4) hold. For elliptically polarized light Eqs. (5) hold.

$$\varphi_x = \varphi_y \qquad A_x \neq A_y \tag{3}$$

$$\varphi_x = \varphi_y \pm \frac{\pi}{2} \qquad A_x = A_y \tag{4}$$

$$\varphi_x \neq \varphi_y \qquad A_x \neq A_y \tag{5}$$

In the last case, it is always possible to find a set of orthogonal axes inclined at an angle α to *x* and *y* along which the components will be E'_x and E'_y , such that Eqs. (6) hold.

$$\varphi'_x = \varphi'_y \pm \frac{\pi}{2} \qquad A'_x \neq A'_y \tag{6}$$

In this new system, the x' and y' amplitudes will be the major and minor axes a and b of the ellipse described by the light vector and α will be the angle of orientation of the ellipse axes with respect to the original coordinate system. The relationships between the different quantities can be written as in Eqs. (7) and (8). The terms are defined by Eqs. (9)-(12).

$$\tan 2\alpha = \tan 2\gamma \cos \varphi \tag{7}$$

$$\sin 2\beta = \sin 2\gamma \sin \varphi \tag{8}$$

$$\tan \gamma = \frac{A_y}{A_x} \tag{9}$$

$$\varphi = \varphi_x - \varphi_y \tag{10}$$

$$\tan \beta = \pm \frac{b}{a} \tag{11}$$

$$A_x^2 + A_y^2 = a^2 + b^2 \tag{12}$$

These same types of polarized light can also be broken down into right and left circular components or into two orthogonal elliptical components. These different vector bases are useful in different physical situations.

One of the simplest ways of producing linearly polarized light is by reflection from a dielectric surface. At a particular angle of incidence, the reflectivity for light whose electric vector is in the plane of incidence becomes zero. The reflected light is thus linearly polarized at right angles to the plane of incidence. This fact was discovered by E. Malus in 1808. Brewster's law shows that at the polarizing angle the refracted ray makes an angle of 90° with the reflected ray. By combining this relationship with Snell's law of refraction, it is found that Eq. (13) holds, where

$$\tan \theta_B = \frac{n_2}{n_1} \tag{13}$$

 θ_B is the Brewster angle of incidence and n_1 and n_2 are the refractive indices of the two interfaced media (**Fig. 1**). This provides a simple way of measuring refractive indices. *See* REFRACTION OF WAVES.

Law of Malus. If linearly polarized light is incident on a dielectric surface at Brewster's angle (the polarizing angle), then the reflectivity of the surface will depend on the angle between the incident electric vector and the plane of incidence. When the vector is in the plane of incidence, the reflectivity will vanish. To compute the complete relationship, the incident light vector **A** is broken into components, one vibrating in the plane of incidence and one at right angles to the plane, as in Eqs. (14) and (15),

$$A_{\parallel} = A\sin\theta \tag{14}$$

$$A_{\perp} = A\cos\theta \tag{15}$$

where θ is the angle between the light vector and a plane perpendicular to the plane of incidence. Since the component in the plane of incidence is not reflected, the reflected ray can be written as Eq. (16),

$$B = Ar\cos\theta \tag{16}$$

where r is the reflectivity at Brewster's angle. The intensity is given by Eq. (17). This is the mathematical

$$I = B^2 = A^2 r^2 \cos^2 \theta \tag{17}$$

statement of the law of Malus (Fig. 1).

Linear polarizing devices. The angle θ can be considered as the angle between the transmitting axes of a pair of linear polarizers. When the polarizers are parallel, they are transparent. When they are crossed, the combination is opaque. The first polarizers were glass plates inclined so that the incident light was at Brewster's angle. Such polarizers are quite inefficient since only a small percentage of the incident light is reflected as polarized light. More efficient polarizers can be constructed.

Dichroic crystals. Certain natural materials absorb linearly polarized light of one vibration direction much more strongly than light vibrating at right angles. Such materials are termed dichroic. For a description of them. *See* DICHROISM.

Tourmaline is one of the best-known dichroic crystals, and tourmaline plates were used as polarizers for many years. A pair was usually mounted in what were known as tourmaline tongs.

Birefringent crystals. Other natural materials exist in which the velocity of light depends on the vibration direction. These materials are called birefringent. The simplest of these structures are crystals in which there is one direction of propagation for which the light velocity is independent of its state of polarization. These are termed uniaxial crystals, and the propagation direction mentioned is called the optic axis. For all other propagation directions, the electric vector can be split into two components, one lying in a plane containing the optic axis and the other at right angles. The light velocity or refractive index for these two waves is different. *See* BIREFRINGENCE.

One of the best-known of these birefringent crystals is transparent calcite (Iceland spar), and a series of polarizers have been made from this substance. W. Nicol (1829) invented the Nicol prism, which is made of two pieces of calcite cemented together as in **Fig. 2**. The cement is Canada balsam, in which the wave velocity is intermediate between the velocity in calcite for the fast and the slow ray. The angle at which the light strikes the boundary is such that for one ray the angle of incidence is greater than the critical angle for total reflection. Thus the rhomb is transparent for only one polarization direction.

Canada balsam is not completely transparent in the ultraviolet at wavelengths shorter than 400 nanometers. Furthermore, large pieces of calcite material are exceedingly rare. A series of polarizers has been

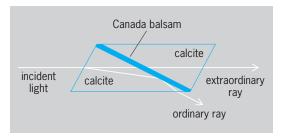


Fig. 2. Nicol prism. The ray for which Snell's law holds is called the ordinary ray.

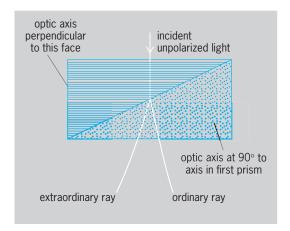


Fig. 3. Wollaston prism.

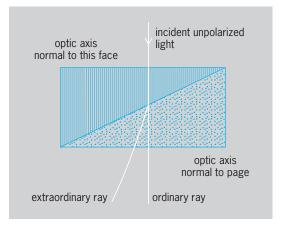


Fig. 4. Rochon prism.

made using quartz, which is transparent in the ultraviolet and which is more commonly available in large pieces. Because of the small difference between the refractive indices of quartz and Canada balsam, a Nicol prism of quartz would be tremendously long for a given linear aperture.

A different type of polarizer, made of quartz, was invented by W. H. Wollaston and is shown in **Fig. 3**. Here the vibration directions are different in the two pieces so that the two rays are deviated as they pass through the material. The incoming light beam is thus separated into two oppositely linearly polarized beams which have an angular separation between them. By using appropriate optical stops (obstacles that restrict light rays) in the system, it is possible to select either beam.

In the Wollaston prism, both beams are deviated; since the quartz produces dispersion, each beam is spread into a spectrum. This is not the case of a prism which was invented by A. Rochon. Here the two pieces are arranged as in **Fig. 4**. One beam proceeds undeviated through the device and is thus achromatic. *See* OPTICAL PRISM.

Sheet polarizers. A third mechanism for obtaining polarized light is the Polaroid sheet polarizer invented by E. H. Land. Sheet polarizers fall into three types. The first is a microcrystalline polarizer in which small crystals of a dichroic material are oriented parallel to each other in a plastic medium. Typical microcrystals, such as needle-shaped quinine iodosulfate, are embedded in a viscous plastic and are oriented by extruding the material through a slit.

The second type depends for its dichroism on a property of an iodine-in-water solution. The iodine appears to form a linear high polymer. If the iodine is put on a transparent oriented sheet of material such as polyvinyl alcohol (PVA), the iodine chains apparently line themselves parallel to the PVA molecules and the resulting dyed sheet is strongly dichroic. A third type of sheet polarizer depends for its dichroism directly on the molecules of the plastic itself. This plastic consists of oriented polyvinylene. Because these polarizers are commercially available and can be obtained in large sheets, many experiments involving polarized light have been performed which would have been quite difficult with the reflection polarizers or the birefringent crystal polarizer.

Characteristics. There are several characteristics of linear polarizers which are of interest to the experimenter. First is the transmission for light polarized parallel and perpendicular to the axis of the polarizer; second is the angular field; and third is the linear aperture. A typical sheet polarizer has a transmittance of 48% for light parallel to the axis and $2 \times 10^{-4\%}$ for light perpendicular to the axis at a wavelength of 550 nm. The angular field is 60° , and sheets can be many feet in diameter. The transmittance perpendicular to the axis varies over the angular field.

The Nicol prism has transmittance similar to that of the Polaroid sheeting but a much reduced linear and angular aperture.

Polarization by scattering. When an unpolarized light beam is scattered by molecules or small particles, the light observed at right angles to the original beam is polarized. The light vector in the original beam can be considered as driving the equivalent oscillators (nuclei and electrons) in the molecules. There is no longitudinal component in the original light beam. Accordingly, the scattered light observed at right angles to the beam can only be polarized with the electric vector at right angles to the propagation direction of the original beam. In most situations, the scattered light is only partially polarized because of multiple scattering. The best-known example of polarization by scattering is the light of the north sky. The percentage polarization can be quite high in clean country air. A technique was invented for using measurements of sky polarization to determine the position of the Sun when it is below the horizon. See SCATTERING OF ELECTROMAGNETIC RADIATION.

Production of polarized light. Linear polarizers have already been discussed. Circularly and elliptically polarized light are normally produced by combining a linear polarizer with a wave plate. A Fresnel rhomb can be used to produce circularly polarized light.

Wave plate. A plate of material which is linearly birefringent is called a wave plate or retardation sheet. Wave plates have a pair of orthogonal axes which are designated fast and slow. Polarized light with its electric vector parallel to the fast axis travels faster than light polarized parallel to the slow axis. The thickness of the material can be chosen so that for light traversing the plate, there is a definite phase shift between the fast component and the slow component. A plate with a 90° phase shift is termed a quarter-wave plate. The retardation in waves is given by Eq. (18), where $n_s - n_f$ is the birefringence; n_s is

$$\delta = \frac{(n_s - n_f)d}{\lambda} \tag{18}$$

the slow index at wavelength λ ; n_f is the fast index; and d is the plate thickness.

Wave plates can be made by preparing X-cut sections of quartz, calcite, or other birefringent crystals. For retardations of less than a few waves, it is easiest to use sheets of oriented plastics or of split mica. A quarter-wave plate for the visible or infrared is easy to fabricate from mica. The plastic wrappers from many American cigarette packages seem to have almost exactly a half-wave retardation from green light. Since mica is not transparent in the ultraviolet, a small retardation in this region is most easily achieved by crossing two quartz plates which differ by the requisite thickness.

Linearly polarized light incident normally on a quarter-wave plate and oriented at 45° to the fast axis can be split into two equal components parallel to the fast and slow axes. These can be represented, before passing through the plate, by Eqs. (19) and (20), where x and y are parallel to the wave-plate

$$E_x = A_x \sin\left(\omega t + \varphi_x\right) \tag{19}$$

$$E_y = A_x \sin\left(\omega t + \varphi_x\right) \tag{20}$$

axes. After passing through the plate, the two components can be written as Eqs. (21) and (22), where

$$E_x = A_x \sin\left(\omega t + \varphi_x + \frac{\pi}{2}\right) \tag{21}$$

$$E_y = A_x \sin(\omega t + \varphi_x) \tag{22}$$

 E_x is now advanced one quarter-wave with respect to E_y .

It is possible to visualize the behavior of the light by studying the sketches in Fig. 5, which show the projection on a plane z = 0 at various times. It is apparent that the light vector is of constant amplitude, and that the projection on a plane normal to the propagation direction is a circle. If the linearly polarized light is oriented at -45° to the fast axis, the light vector will revolve in the opposite direction. Thus it is possible with a quarter-wave plate and a linear polarizer to make either right or left circularly polarized light. If the linearly polarized light is at an angle other than 45° to the fastest axis, the transmitted radiation will be elliptically polarized. When circularly polarized light is incident on a quarter-wave plate, the transmitted light is linearly polarized at an angle of 45° to the wave-plate axes. This polarization is independent of the orientation of the wave-plate axis. For elliptically polarized light, the behavior of the quarter-wave plate is much more complicated. However, as was mentioned earlier, the elliptically polarized light can be considered as composed of two

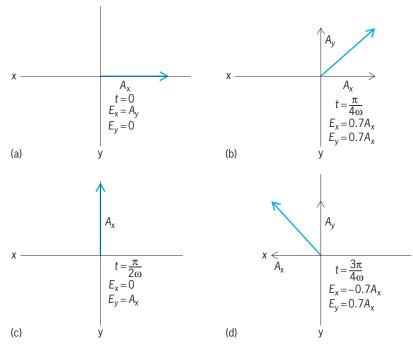


Fig. 5. Projection of light vector of constant amplitude on plane z = 0 for circularly polarized light. (a) t = 0. (b) $t = \pi/4\omega$. (c) $t = \pi/2\omega$. (d) $t = 3\pi/4\omega$.

linear components parallel to the major and minor axes of the ellipse and with a quarter-wave phase difference between them. If the quarter-wave plate is oriented parallel to the axes of the ellipse, the two transmitted components will either have zero phase difference or a 180° phase difference and will be linearly polarized. At other angles, the transmitted light will still be elliptically polarized, but with different major and minor axes. Similar treatment for a half-wave plate shows that linearly polarized light oriented at an angle of θ to the fast axis is transmitted as linearly polarized light oriented at an angle $-\theta$ to the fast axis.

Wave plates all possess a different retardation at each wavelength. This appears immediately from Eq. (18). It is conceivable that a substance could have dispersion of birefringence, such as to make the retardation of a plate independent of wavelength. However, no material having such a characteristic has as yet been found.

Fresnel rhomb. A quarter-wave retardation can be provided achromatically by the Fresnel rhomb. This device depends on the phase shift which occurs at total internal reflection. When linearly polarized light is totally internally reflected, it experiences a phase shift which depends on the angle of reflection, the refractive index of the material, and the orientation of the plane of polarization. Light polarized in the plane of incidence experiences a phase shift which is different from that of light polarized at right angles to the plane of incidence. Light polarized at an intermediate angle can be split into two components, parallel and at right angles to the plane of incidence, and the two components mathematically combined after reflection.

The phase shifts can be written Eqs. (23) and (24),

$$\tan\frac{\phi_{\parallel}}{2} = \frac{\sqrt{n^2 \sin^2 i - 1}}{\cos i} \tag{23}$$

$$\tan\frac{\varphi_{\perp}}{2} = \frac{\sqrt{n^2 \sin^2 i - 1}}{n \cos i} \tag{24}$$

where φ_{\parallel} is the phase shift parallel to the plane of incidence; φ_{\perp} is the phase shift at right angles to the plane of incidence; *i* is the angle of incidence on the totally reflecting internal surface; and n is the refractive index. The difference $\varphi_{\parallel} - \varphi_{\perp}$ reaches a value of about $\pi/4$ at an angle of 52° for n = 1.50. Two such reflections give a retardation of $\pi/2$. The Fresnel rhomb shown in Fig. 6 is cut so that the incident light is reflected twice at 52°. Accordingly, light polarized at 45° to the principal plane will be split into two equal components which will be shifted a quarterwave with respect to each other, and the transmitted light will be circularly polarized. Nearly achromatic wave plates can be made by using a series of wave plates in series with their axes oriented at different specific angles with respect to a coordinate system.

Analyzing devices. Polarized light is one of the most useful tools for studying the characteristics of materials. The absorption constant and refractive index of a metal can be calculated by measuring the effect of the metal on polarized light reflected from its surface. *See* REFLECTION OF ELECTROMAGNETIC RADIATION.

The analysis of polarized light can be performed with a variety of different devices. If the light is linearly polarized, it can be extinguished by a linear polarizer and the direction of polarization of the light determined directly from the orientation of the polarizer. If the light is elliptically polarized, it can be analyzed with the combination of a quarter-wave plate and a linear polarizer. Any such combination of polarizer and analyzer is called a polariscope. As explained previously, a quarter-wave plate oriented parallel to one of the axes of the ellipse will transform elliptically polarized light to linearly polarized light. Accordingly, the quarter-wave plate is rotated until the beam is extinguished by the linear polarizer. At this point, the orientation of the quarter-wave plate gives the orientation of the ellipse and the orientation of the polarizer gives the ratio of the major to

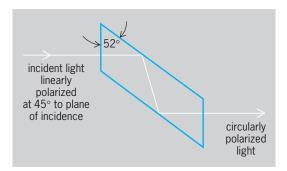


Fig. 6. Fresnel rhomb.

minor axis. Knowledge of the origin of the elliptically polarized light usually gives the orientation of the components which produced it, and from these various items, the phase shifts and attenuations produced by the experiment can be deduced.

One of the best-known tools for working with polarized light is the Babinet compensator. This device is normally made of two quartz prisms put together in a rhomb. One prism is cut with the optic axis in the plane of incidence of the prism, and the other with the optic axis perpendicular to the plane of incidence. The retardation is a function of distance along the rhomb; it will be zero at the center, varying to positive and negative values in opposite directions along the rhomb. It can be used to cancel the known or unknown retardation of any wave plate.

Retardation theory. It is difficult to see intuitively the effect of a series of retardation plates or even of a single plate of general retardation δ on light which is normally incident on the plate and which is polarized in a general fashion. This problem is most easily solved algebraically. The single-wave plate is assumed to be oriented normal to the direction of propagation of the light, which is taken to be the *z* direction of a set of cartesian coordinates. Its fast axis is at an angle α to the *x* axis. The incident light can be represented by Eqs. (25) and (26). A first step is to break

$$E_x = A_x \sin\left(\omega t + \varphi_x\right) \tag{25}$$

$$E_y = A_y \sin(\omega t + \varphi_y) \tag{26}$$

the light up into components $E_{x'}$ and $E_{y'}$ parallel to the axes of the plate. It is possible to write Eqs. (27) and (28). These components can also be written as Eqs. (29) and (30). After passing through the plate, the components become Eqs. (31) and (32).

$$E_{x'} = E_x \cos \alpha - E_y \sin \alpha \tag{27}$$

$$E_{y'} = E_x \sin \alpha + E_y \cos \alpha \tag{28}$$

$$E_{x'} = A_{x'} \sin\left(\omega t + \varphi_{x'}\right) \tag{29}$$

$$E_{\gamma'} = A_{\gamma'} \sin\left(\omega t + \varphi_{\gamma'}\right) \tag{30}$$

$$E_{x''} = A_{x'} \sin\left(\omega t + \varphi_{x'} + \delta\right) \tag{31}$$

$$E_{y''} = A_{y'} \sin\left(\omega t + \varphi_{y'}\right) \tag{32}$$

In general, it is of interest to compare the output with the input. The transmitted light is thus broken down into components along the original axes. This results in Eqs. (33) and (34).

$$E_{x'''} = E_{x''} \cos \alpha + E_{v''} \sin \alpha \tag{33}$$

$$E_{\nu'''} = -E_{x''} \sin \alpha + E_{\nu''} \cos \alpha \tag{34}$$

With this set of equations, it is possible to compute the effect of a wave plate on any form of polarized light. *Jones calculus.* Equations (33) and (34) still become overwhelmingly complicated in any system involving several optical elements. Various methods have been developed to simplify the problem and to make possible some generalizations about systems of elements. One of the most straightforward, proposed by R. C. Jones, involves reducing Eqs. (33) and (34) to matrix form. The Jones calculus for optical systems involves the polarized electric components of the light vector and is distinguished from other methods in that it takes cognizance of the absolute phase of the light wave.

The Jones calculus writes the light vector in the complex form as in Eqs. (35) and (36). Matrix op-

$$Z_x = A_x e^{i(\omega t + \varphi_x)}$$

$$Z_y = A_x e^{i(\varphi t + \varphi_y)}$$
(35)

$$E = \begin{vmatrix} A_x e^{i\varphi_x} \\ A_y e^{i\varphi_y} \end{vmatrix} e^{iwt}$$
(36)

erators are developed for different optical elements. From Eqs. (25)-(34), the operator for a wave plate can be derived directly. *See* MATRIX THEORY.

1

The Jones calculus is ordinarily used in a normalized form which simplifies the matrices to a considerable extent. In this form, the terms involving the actual amplitude and absolute phase of the vectors and operators are factored out of the expressions. The intensity of the light beam is reduced to unity in the normalized vector so that Eq. (37) holds. Under

$$4_x^2 + A_y^2 = 1 \tag{37}$$

this arrangement, the matrices for various types of operations can be written as Eq. (38). This is the

$$G(\delta) = \begin{vmatrix} e^{i(\delta/2)} & 0\\ 0 & e^{-i(\delta/2)} \end{vmatrix}$$
(38)

operator for a wave plate of retardation δ and with axes along *x* and *y*. Equation (39) gives the operator

$$S(\alpha) = \begin{vmatrix} \cos \alpha & -\sin \alpha \\ \sin \alpha & \cos \alpha \end{vmatrix}$$
(39)

for a rotator which rotates linearly polarized light through an angle α . Equation (40) gives the operator

$$P_b = \begin{vmatrix} 1 & 0 \\ 0 & 0 \end{vmatrix} \tag{40}$$

for a perfect linear polarizer parallel to the *x* axis. A wave plate at an angle α can be represented by Eq. (41). A series of optical elements can be repre-

$$G(\delta, \alpha) = S(\alpha)G(\delta)S(-\alpha) \tag{41}$$

sented by the product of a series of matrices. This simplifies enormously the task of computing the effect of many elements. It is also possible with the Jones calculus to derive a series of general theorems concerning combinations of optical elements. Jones has described three of these, all of which apply only for monochromatic light.

1. An optical system consisting of any number of retardation plates and rotators is optically equivalent to a system containing only two elements, a retardation plate and a rotator.

2. An optical system containing any number of partial polarizers and rotators is optically equivalent to a system containing only two elements-one a partial polarizer and the other a rotator.

3. An optical system containing any number of retardation plates, partial polarizers, and rotators is optically equivalent to a system containing four elements-two retardation plates, one partial polarizer, and one rotator.

As an example of the power of the calculus, a rather specific theorem can be proved. A rotator of any given angle α can be formed by a sequence of three retardation plates, a quarter-wave plate, a retardation plate at 45° to the quarter-wave plate, and a second quarter-wave plate crossed with the first, as in Eq. (42), where β is the angle between the axis

$$S(\alpha) = S(\beta)G\left(-\frac{\pi}{2}\right)S(-\beta)S\left(\beta + \frac{\pi}{4}\right)$$
$$\cdot G(\delta)S\left(-\beta - \frac{\pi}{4}\right)S(\beta)G\left(\frac{\pi}{2}\right)S(-\beta) \quad (42)$$

of the first quarter-wave plate and the x axis, and δ is the retardation of the plate in the middle of the sandwich.

The first simplification arises from the fact that the axis rotations can be done in any order. This reduces Eq. (43) to Eq. (44). Now Eqs. (45) and (46) hold. When the multiplication is carried through, Eq. (47) is obtained.

$$S\left(\beta + \frac{\pi}{4}\right) = S(\beta)S\left(\frac{\pi}{4}\right) = S\left(\frac{\pi}{4}\right)S(\beta) \qquad (43)$$

 $S(\alpha) = S(\beta)G\left(-\frac{\pi}{2}\right)S\left(\frac{\pi}{4}\right)$ $\cdot G(\delta)S\left(-\frac{\pi}{4}\right)G\left(\frac{\pi}{2}\right)$ (44)

$$S\left(-\frac{\pi}{4}\right)G\left(\frac{\pi}{2}\right) = \frac{1}{\sqrt{2}} \begin{vmatrix} 1 & -i \\ -i & 1 \end{vmatrix}$$
(45)

$$G\left(-\frac{\pi}{2}\right)S\left(\frac{\pi}{4}\right) = -\frac{1}{\sqrt{2}} \begin{vmatrix} 1 & i \\ i & 1 \end{vmatrix}$$
(46)

$$S(\alpha) = S(\beta)$$

-
$$\frac{1}{2} \begin{vmatrix} e^{i\delta/2} + e^{-\delta/2} & -ie^{i\delta/2} + ie^{-i\delta/2} \\ ie^{i\delta/2} - ie^{-i\delta/2} & e^{i\delta/2} + e^{-i\delta/2} \end{vmatrix} S(-\beta)$$

SCON

$$S(\alpha) = S(\beta) - \frac{1}{2} \begin{vmatrix} e^{i\delta/2} + e^{-\delta/2} & -ie^{i\delta/2} + ie^{-i\delta/2} \\ ie^{i\delta/2} - ie^{-i\delta/2} & e^{i\delta/2} + e^{-i\delta/2} \end{vmatrix} S(-\beta)$$
(47)

The rotation angle is therefore equal to one-half

the phase angle of the retardation. This combination is a true rotator in that the rotation is independent of the azimuth angle of the incident polarized light.

A variable rotator can be made by using a Soleil compensator for the central element. This consists of two quartz wedges joined to form a plane parallel quartz plate.

Mueller matrices. In the Jones calculus, the intensity of the light passing through the system must be obtained by calculation from the components of the light vector. A second calculus is frequently used in which the light vector is split into four components. This also uses matrix operators which are termed Mueller matrices. In this calculus, the intensity I of the light is one component of the vector and thus is automatically calculated. The other components of the vector are given by Eqs. (48)-(50). The matrix

$$M = A_x^2 - A_y^2 \tag{48}$$

$$C = 2A_x A_y \cos(\varphi_x - \varphi_y) \tag{49}$$

$$S = 2A_x A_y \sin(\varphi_x - \varphi_y) \tag{50}$$

of a perfect polarizer parallel to the x axis can be written as Eq. (51).

The calculus can treat unpolarized light directly. Such a light vector is given by Eq. (52). The vector for

$$\begin{vmatrix} I\\M\\C\\S \end{vmatrix} = \begin{vmatrix} 1\\0\\0\\0 \end{vmatrix}$$
(52)

light polarized parallel to the x axis is written as expression (53). In the same manner as in the Jones cal-

$$\begin{vmatrix} 1\\1\\0\\0\end{vmatrix}$$
(53)

culus, matrices can be derived for retardation plates, rotators, and partial polarizers. This calculus can also be used to derive various general theorems about various optical systems. See FARADAY EFFECT; INTER-FERENCE OF WAVES; OPTICAL ACTIVITY; OPTICAL RO-TATORY DISPERSION. Bruce H. Billings; Hong Hua

Bibliography. R. M. A. Azzam and N. M. Bashara, Ellipsometry and Polarized Light, 1977, paper 1987; B. Billings (ed.), Selected Papers on Applications of Polarized Light, 1992; M. Born and E. Wolf, Principles of Optics, 7th ed., 1999; E. Collett, Polarized Light, 1993; F. A. Jenkins and H. E. White, Fundamentals of Optics, 4th ed., 1976; D. S. Kliger, J. W. Lewis, and C. E. Randall, Polarized Light in Optics and Spectroscopy, 1990.

Polarized light microscope

A microscope that utilizes polarized light to form a highly magnified image of an object. Polarizing microscopes play an important role in crystallography, petrography, microchemistry, and biology. Although all light microscopes compare poorly with electron microscopes with respect to image resolution, polarized light microscopes have the unique ability to deliver information about the submicroscopic structure of the objects being examined. They also have the advantage of being relatively nondestructive, and may be used safely with living cells. Polarized light interactions with electromagnetically anisotropic structures, down to atomic dimensions, can be measured by polarized light microscopy. The sensitivity of polarized light microscopy as well as its importance to biology have been enhanced by the use of video technology. See POLARIZED LIGHT.

Optical anisotropy. Conventional light microscopes develop image contrast primarily as a result of light absorption in the specimen. Phase-contrast and interference microscopes develop contrast for nonabsorbing objects whose optical thickness differs from that of their surroundings. In turn, polarizing microscopes develop contrast for objects whose optical thickness has more than one value, depending on the orientation of the electric vector of the light relative to the axes of the object. *See* INTERFERENCE MICROSCOPE; PHASE-CONTRAST MICROSCOPE.

Just as materials such as wood have different mechanical properties in different directions, optically anisotropic materials have different electrical properties in different directions. A consistent directional difference in the alignment of chemical bonds, ions, or molecules results in materials that have different dielectric constants for electric fields applied in different directions. Because the velocity of light in a material is proportional to the square root of the dielectric constant in the direction of the electric-field displacement (electric-vector direction), the light velocity-and hence refractive index-in these materials differs for light polarized in different directions. For transparent materials, the refractive index is equal to the square root of the dielectric constant at optical frequencies.

Crystals, oriented polymers, and complex macromolecules tend to be optically anisotropic and can have up to three different refractive indices, one for each of the three orthogonal directions. The majority of crystals have two or three refractive indices; only regular cubic crystals and amorphous materials show the same refractive index in all directions. Light traveling through anisotropic materials experiences at most two refractive indices because its electricvector direction is always normal to the direction of propagation. Hence, optically anisotropic materials are often referred to as birefringent, that is, having two refractive indices. The birefringence of such materials is a manifestation of their unique internal electrical structure and is independent of their environment. This is known as intrinsic birefringence. Stress birefringence and strain birefringence demonstrate the electric displacements, in initially isotropic solids, that result from elastic and plastic deformation, respectively. Stress birefringence is reversible, disappearing when the stress is removed, whereas strain birefringence is permanent. Stress and strain birefringences are also intrinsic in that they are strictly the result of internal electrical anisotropy.

Form birefringence, also called structural birefringence, occurs when asymmetric particles of one refractive index are aligned more or less parallel to each other in a surrounding medium of different refractive index. Form birefringence is dependent on the interparticle spacing, in at least one direction, being much less than the wavelength of light. Typical biological examples of form birefringence are the bundles of microtubules and microfilaments of the cytoskeleton. Flow birefringence is a special case of form birefringence in which submicroscopic particles in a fluid matrix are aligned by shear during flow. A fluid suspension of submicroscopic rodlets, such as tobacco mosaic virus particles in water, shows strong flow birefringence when squirted out of a pipette.

Form birefringence requires different refractive indices for the oriented structures and the intervening medium. The magnitude, but not the sign, of birefringence is a function of this difference. Consequently, form birefringence is indicated when changing the refractive index of the medium changes the birefringence. *See* BIREFRINGENCE; REFRACTION OF WAVES.

Polarizing microscopes. A polarizing microscope differs from a conventional light microscope in a number of ways (see **illus.**).

Polarizing devices. A polarizing microscope has a pair of polars (polarizing devices) in the optical train. The first polar (polarizer) defines the initial plane of polarization for light entering the microscope and is located between the illuminator and the condenser. The other polar (analyzer) is usually placed between the objective and the ocular tube and defines the plane of polarization of the light reaching the ocular. One or both must be accurately rotatable about the optical axis of the instrument. Usually the analyzer is also removable from the optical path.

A real polarizer is never totally efficient in rejecting light whose electric vector is at 90° to the desired direction. Thus, when two polars are placed in sequence in the optical path with their electricvector transmission directions mutually crossed at 90°, most of the incident light is blocked. However, a small amount of unpolarized light I_{\perp} leaks through. When the polars are oriented with their electricvector transmission directions parallel, the intensity of light transmitted I_{\parallel} is half the intensity of the incident unpolarized light minus any light lost by reflection or absorption in the polars, plus I_{\perp} . The figure of merit for polars is the extinction factor (EF), as given in Eq. (1).

$$\mathrm{EF} = \frac{I_{\parallel}}{I_{\perp}} \tag{1}$$

The most frequently used type of polar is a dichroic

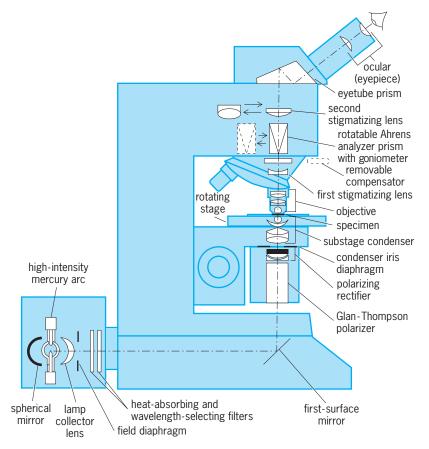


Diagram of a high-sensitivity biological polarizing microscope. Polarizing microscopes for petrography may not require the use of the expensive prism polarizer and analyzer, which can be replaced by dichroic polarizing filters. Some petrographic applications require the addition of a universal stage, which permits specimen rotation about all three axes. In all other respects the petrographic polarizing microscope has a similar optical layout. The parts of the optical train are indicated.

sheet polarizing filter. It can have an extinction factor as high as 500,000. Filter polars have wide acceptance angles, are easily accommodated without major changes in the optical system, and are significantly less expensive than prism polars. However, they do absorb as much as 56% of the light that should be transmitted. When used for both polarizer and analyzer, the maximum amount of light that filter polars can pass through the system is only 10% of the light input to the polarizer.

The sensitivity of a polarizing microscope is ultimately limited by the radiance of the illumination source and the extinction of the optical system. A high-quality prism polarizer can have an extinction factor in excess of 10⁶, while transmitting virtually all of the properly polarized light. Other factors being equal, a microscope with high-quality filter polars requires a light source with more than five times the radiance needed to achieve the same sensitivity as a microscope using high-quality prism polars.

Unfortunately, the use of prism polarizers complicates the microscope design. Additionally, the polarizing prisms available share the common characteristic of transmitting the extraordinary component whose velocity varies with its angle relative to the crystal axes. This results in a slightly astigmatic image. *See* ABERRATION (OPTICS). Thus, for petrographic, crystallographic, and most microchemical applications, dichroic filter polars are the better choice. For critical biological applications, such as investigating the weak birefringence of cytoskeletal structures in living cells, the expense and complication attendant to the use of prism polars can be justified by the attainment of sensitivity unobtainable by other means.

Other requirements and features. In addition to the polars, all polarizing microscopes need rotatable specimen stages and one or more removable birefringence compensators. Slots for insertion of the compensators are located near an aperture plane either between objective and analyzer or between polarizer and condenser.

The compensators are birefringent devices used to measure magnitude and sign of retardation due to specimen birefringence, to enhance specimen image contrast, and to manipulate the state of polarization of light passing through any point in the specimen. In general, compensators provide either a fixed amount of birefringent retardation, to which the specimen identifiably adds or subtracts, or a quantitatively variable amount of retardation that is added to or subtracted from the specimen retardation to bring the combination to an identifiable state.

Polarizing microscopes for different applications have some differences in construction. For petrography and crystallography, the microscope should be able to accept a universal specimen stage capable of rotating the specimen about three axes. A polarizing microscope for biological use has less rigid requirements for angular orientation but needs rectified optics, the extremely high extinction polars noted above, and a sensitive elliptic compensator because the birefringence of the typical biological specimen is much smaller. An ocular telescope or a Bertrand lens (a built-in lens that optionally converts the ocular into a telescope) is essential for some applications and useful for all.

Rectified optics. Even optics free of any form of birefringence have been found to depolarize light, degrading the extinction factor when large numerical apertures are involved. The cause of lens depolarization is the result of differential reflection at lens surfaces. Improved designs include polarization rectifiers consisting of a simple zero-power lens to duplicate the rotation caused by the lens to be rectified, and a half-wave retardation plate to turn the rotation into its mirror image at every point over the lens aperture. These eliminate or greatly reduce depolarization at high numerical apertures. It has also been shown that an image-degrading diffraction artifact is present in polarized light images produced with unrectified optics, and it has been demonstrated that rectification eliminates the artifact. Thus, for objects requiring both high extinction and high resolution, such as living cells, rectified optics greatly extend the power of polarization microscopy.

Modes of use. There are two traditional modes of use for the polarizing microscope, the orthoscopic mode and the conoscopic mode.

Orthoscopic mode. In the orthoscopic mode, the ocular projects an image of the specimen, as in conventional microscopy. Rotation of the specimen stage reveals the location and orientation of any anisotropic features in the specimen.

When rotated 360° between crossed polars, a birefringent object will go from dark to light four times, and all isotropic features will remain dark. A birefringent object is darkest whenever its slow axis is parallel to either polar, and brightest when the slow axis is at 45° to the polars. (The slow axis is the direction in the material for which the dielectric constant is highest.) This brightening is the result of the conversion by the specimen birefringence of the linearly polarized light from the polarizer to elliptically polarized light that has a component in the direction of the analyzer. The light intensity emerging from the analyzer I_a is related to the light intensity emerging from the analyzer when the polars are parallel. I_{\parallel} by Eq. (2), where δ is the specimen retardation

$$I_a = I_{\parallel} \sin^2\left(\frac{\delta}{2}\right) \sin^2(2\theta) + I_{\perp}$$
(2)

(expressed as the angular fraction of a wavelength) between ordinary and extraordinary waves, θ is the angle between the specimen slow axis and the polarizer, and I_{\perp} is the unpolarized light leaking through the crossed polars.

Although the relative brightness of the specimen, placed with its axes at 45° to the polarizer, is a function of the specimen birefringence, it is not a linear function. Also birefringence is not the exclusive source of light between crossed polars. A compensator is therefore used to measure the birefringence retardation introduced by the specimen. For weakly retarding specimens, an elliptic compensator provides the greatest sensitivity. This compensator consists of a thin mica plate, providing one-tenth-wave or less retardation, mounted perpendicular to the microscope's axis and rotatable about that axis to 0.1° accuracy.

A measurement is made by setting the compensator to introduce the same ellipticity as the specimen but of opposite sign so that the image of the specimen is extinguished (maximally darkened). The procedure is as follows: (1) The compensator slow axis is set parallel to the polarizer (indicated by extinction of the nonbirefringent parts of the field of view). (2) The specimen is rotated to its extinction position, aligning its axes with the polars, and then it is rotated 45° to its brightest position (which the eye can less readily determine than its darkest position). (3) The compensator is rechecked for background extinction, in case the slide or cover glass has any residual birefringence, and the compensator angle is recorded. (4) The compensator is then rotated a small amount and the change in specimen brightness observed. If the specimen is brighter, the compensator is adding to the specimen birefringence and should be rotated in the opposite direction for compensation. If the specimen darkens, the compensator is compensating the specimen and should be rotated to the point of specimen extinction. (5) The compensator angle is recorded, and the difference between the two readings is used as the compensator azimuth angle in calculating the specimen retardation in degrees or in nanometers.

This procedure yields the value for the amount of the differential in light path through the specimen, but it does not yield directly either the specimen's extraordinary and ordinary refractive indices or its coefficient of birefringence. Because specimen retardation is the product of the thickness of the specimen times its coefficient of birefringence, any means that can yield the specimen thickness will give the coefficient of birefringence, but the actual value of either refractive index must come from some form of refractometry. In many cases, knowledge of specimen symmetry will allow a good estimate for specimen thickness. Estimates based on use of the microscope's fine focus calibration have poor accuracy but are at times the only available alternative.

A dichroic object is also birefringent and will show up as such in this test. To determine which birefringent objects are also dichroic, the analyzer is removed from the optical path, and the stage is rotated again. In the absence of the analyzer, a dichroic object darkens twice per revolution, being darkest when its slow axis is parallel to the polarizer. Optically active objects and scattering objects pass at least some light when placed between crossed polars, and do not darken with specimen rotation. An optically active object may be distinguished from a scattering object by rotating one polar relative to the other. At some offset from the crossed position between the polars, the optically active object darkens differentially relative to the background, while the scatterer does not darken.

Conoscopic mode. The conoscopic mode is used to characterize crystalline specimens. Here the ocular is not used to project an image of the specimen, but rather—with the aid of a special lens—the image of the objective exit pupil is examined to reveal the relative retardation experienced by polarized light as a function of its angle of incidence on the specimen.

For effective use of this mode, the specimen is mounted in a universal stage that allows it to be rotated about all three axes relative to the optical axis of the microscope. This permits the location of the crystal's optic axis (or axes) and the axial ratios of its index ellipsoid (indicatrix) to be determined, as well as the optic axial angle, in the case of biaxial crystals. Because the optic axis is the direction of travel through the crystal for which it appears isotropic, rotation of the specimen about the viewing axis will reveal the optic axes as spots that remain dark regardless of their orientation relative to the crossed polarizer and analyzer. *See* CRYSTAL OPTICS; MICRO-SCOPE. Gordon W. Ellis

Bibliography. P. J. Duke and A. G. Michette (eds.), *Modern Microscopy: Techniques and Applications*, 1989; M. Pluta, *Advanced Light Microscopy: Specialized Methods*, vol. 2, 1989; E. M. Slayter and H. S. Slayter, *Light and Electron Microscopy*, 1993.

Polarographic analysis

An electrochemical technique used in analytical chemistry. Polarography involves measurements of current-voltage curves obtained when voltage is applied to electrodes (usually two) immersed in the solution being investigated. One of these electrodes is a reference electrode: its potential remains constant during the measurement. The second electrode is an indicator electrode. Its potential varies in the course of measurement of the current-voltage curve, because of the change of the applied voltage. In the simplest version, so-called dc polarography, the indicator electrode is a dropping-mercury electrode, consisting of a mercury drop hanging at the orifice of a fine-bore glass capillary (usually about 0.08 mm inner diameter). The capillary is connected to a mercury reservoir so that mercury flows through it at the rate of a few milligrams per second. The outflowing mercury forms a drop at the orifice, which grows until it falls off. The lifetime of each drop is several seconds (usually 2 to 5). Each drop forms a new electrode; its surface is practically unaffected by processes taking place on the previous drop. Hence each drop represents a well-reproducible electrode with a fresh, clean surface. See ELECTRODE; REFER-ENCE ELECTRODE.

Apparatus. The dropping-mercury electrode is immersed in the solution to be investigated and placed in a cell containing the reference electrode (**Fig. 1**). Polarographic current-voltage curves can be recorded with a simple instrument consisting of a potentiometer or another source of voltage and a current-measuring device (such as a sensitive galvanometer). The voltage can be varied by manually changing the applied voltage in finite increments, measuring current at each, and plotting current as a function of the voltage. Alternatively, commercial in-

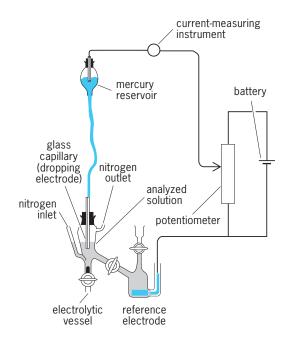


Fig. 1. Diagram showing components of a polarographic circuit.

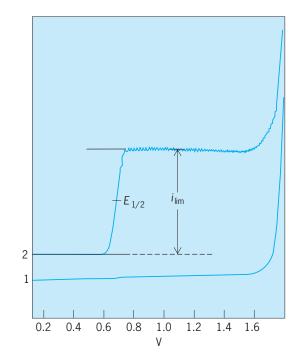


Fig. 2. Polarographic current-voltage curves: 1 = supporting electrolyte; 2 = curve in the presence of an electroactive (reducible) compound; $E_{1/2} =$ half-wave potential; $i_{lim} =$ limiting current (wave height).

struments are available in which voltage is increased linearly with time (a voltage ramp), and current variations are recorded automatically (**Fig. 2**).

Polarographic curves. Oscillations of current result from growth of the individual mercury drops. Their mean value is usually measured. Three portions can be observed on a typical polarographic curve. At sufficiently positive potentials, only a small current flows. Then, in a region characteristic of the particular species and solution, an S-shaped current rise is observed. Finally, at sufficiently negative potentials, the current is again independent of applied voltage over a potential range. This current is called the limiting current and is usually a few microamperes or a fraction of a microampere. The limiting current (also called the height of the polarographic wave) increases, usually with concentration of the investigated species in the solution. The wave height can be thus used for determination of concentration (that is, how much of the species is present in the solution). The potential at the point of the polarographic wave, where the current reaches half of the limiting value, is called the half-wave potential (Fig. 2). This potential is characteristic of the species studied (that is, it can be used to confirm which species are present).

Solutions. The solutions investigated polarographically must contain the species to be studied in 10^{-3} to 10^{-6} *M* concentration. In addition, the solution must contain a large excess (50-fold or greater) of a supporting electrolyte which does not react at the electrode in the potential region of interest and which can be a neutral salt, acid, base, or a mixture (such as a buffer). The function of the supporting electrolyte is to reduce the resistance of the solution, to ensure that species are transported to the

surface of the electrode by diffusion rather than by migration in the electric field, to keep conditions at the electrode surface unchanged in the course of polarographic electrolysis, and to keep or convert the species into a form most suitable for electrolysis (for example, by complex formation or protonation).

Evaluations. The heights of polarographic waves are measured and compared with heights obtained with standard solutions of the species, the concentrations of which are known. Accuracy varies from 2 to 5%, according to the potential range, shape of polarographic waves, and presence of other components. Under strictly controlled conditions, even 1% accuracy can be achieved.

Polarographically active species. To be studied polarographically, a species must undergo reduction or oxidation at the surface of the mercury drop, must form compounds with mercury, or must catalytically affect electrode processes. From changes in the shape of polarographic curves, it is also possible to determine various surfactants.

To be able to follow a given reducible species, its reduction must occur at more positive potentials than reduction of the component of the supporting electrolyte or solvent. Oxidations can be followed if they occur at potentials more negative than that of oxidation of the solvent, supporting electrolyte, or electrode material. Reduction processes are indicated by cathodic waves (above the zero current line), and oxidation processes by anodic waves (below the zero current line). *See* OXIDATION-REDUCTION.

Polarography has been used for the determination of most metals (typically lead, copper, zinc, cadmium, iron, uranium, cobalt, nickel, manganese, potassium, sodium, and so forth), some inorganic acids and their anions (for example, iodic and periodic acids, nitrates, and nitrites), and some gases (for example, sulfur dioxide and oxygen). The wave of oxygen forms the basis of a convenient oxygen determination in technical gases, waters, or biological material. Alternatively the reducibility of oxygen means that oxygen must be removed (usually by purging with nitrogen) from solutions to be analyzed for other constituents. Reduction of inorganic species results in the lowering of the oxidation state, which may be accompanied by amalgam formation.

Electrode processes involving organic compounds at the dropping-mercury electrode result in a cleavage or formation of a chemical bond in the investigated organic molecule. Polarographic behavior depends primarily on the nature of the bond involved, even when the molecular environment (that is, the presence and kind of neighboring groups and other substituents, the type of the molecular frame to which the electroactive group is bound, and the spatial arrangement) also affects the electrode process. The reducibility of multiple bonds (for example, C=O, C=N, C=C, and N=N) or of strongly polarizable groups (such as NO₂, NO, and C-halogen), especially when conjugated, was recognized early. Recently the reducibility of some single bonds such as C—O, C—N, C—S, or C—P when the molecule contains activating groups was proved.

Because mercury ions react with various anions, the anodic current corresponding to the dissolution of mercury can be used, for example, for the determination of halides, thiocyanate, cyanide, and sulfide among inorganic species; and of mercaptans, urea and thiourea derivatives, and dithiocarbamates among organic compounds.

Most important among catalytic waves are those which correspond to catalytic hydrogen evolution. Among compounds giving catalytic waves in buffered ammoniacal solutions of cobalt salts are some types of proteins.

The protein molecule must contain a suitably situated thiol or disulfide group and must be able to form a complex with cobalt and ammonia. It is assumed that the proton transferred by this complex undergoes reduction more easily than the hydronium ion (H_3O^+) . See PROTEIN.

When blood serum is alkali-denatured and the proteins with largest molecular weight are separated by precipitation with sulfosalicylic acid, the remaining filtrate contains proteins and their fragments, the contents of which differ in the serum of healthy and pathological individuals. Increase of the catalytic wave is observed for cancer and inflammatory diseases, and decrease of the catalytic wave for hepatitis and other liver diseases. In Europe the polarographic test is used in clinical analysis as a general screening test, in connection with nine other tests as a proof of malignancy, and in controlling cure of cancer and measuring the effectiveness of treatments such as surgery or irradiation. *See* CLINICAL PATHOLOGY.

Applications. Polarographic studies can be applied to investigation of electrochemical problems, to elucidation of some fundamental problems of inorganic and organic chemistry, and to solution of practical problems.

In electrochemistry polarography allows measurement of potentials, and yields information about the rate of the electrode process, adsorption-desorption phenomena, and fast chemical reactions accompanying the electron transfer. Since polarographic experiments are simple and not too time-consuming, polarography can be used in a preliminary test designed to find the most suitable model compound for detailed electrochemical investigations.

In fundamental applications, polarography makes it possible to distinguish the form and charge of the species (for example, inorganic complex or organic ion) in the solution. Polarography also permits the study of equilibria (complex formation, acidbase, tautomeric), rates, and mechanisms. For equilibria established in the bulk of the solution in more than 15 s, measurement of wave heights of individual components makes possible the evaluation of equilibrium constants. For equilibria that are very rapidly established at the electrode surface, equilibrium constants can be determined from shifts of half-wave potentials. Finally, for some equilibria between these two extremes, which are established in times comparable to the drop time (3 s), calculation of rate constants is possible (for example, for dehydration of hydrated aldehydes, or for protonation of anions derived from organic acids, such as $C_6H_5COCH_2COCH_3$). In this way, rate constants of very fast reactions of the order 10^5-10^{10} liters mol⁻¹ s⁻¹ can be determined.

For slower reactions, rate constants can be found from changes of wave heights with time. Moreover, as some reaction intermediates giving separate waves can be detected, identified, and followed polarographically (if their half-lives are longer than 15 s), polarography can prove useful in mechanistic studies. Elimination of Mannich bases, hydration of multiple bonds in unsaturated ketones, and aldolization are examples studied. *See* ORGANIC REACTION MECHANISM.

Finally, polarography can be used for investigation of the relationship between electrochemical data and structure. Reduction of most organic systems, particularly in aqueous media, involves steps with a high activation energy and is therefore irreversible. Half-wave potentials of such systems are a function of the rate constant of the electrode process. This heterogeneous rate constant is frequently influenced by structural effects, as are, in a similar manner, rate constants of homogeneous reactions. Therefore, effects of substituents on half-wave potentials in aromatic systems and in aliphatic systems can be treated by the Hammett relations and the Taft substituent constants, respectively. Among steric effects, for example, steric hindrance of coplanarity or effects of cis-trans isomerisms affect polarographic curves. Polarography also makes it possible to distinguish between some epimers, for example, bearing axial or equatorial halogen. See MOLECULAR ISOMERISM.

Practical applications are predominantly analytical procedures. In inorganic analysis, polarography is used predominantly for trace-metal analysis (with increased sensitivity of differential pulse polarography and stripping analysis). In organic analysis, it is possible in principle to use polarography in elemental analysis, but such applications are deservedly infrequent. More frequent are applications in functional group analysis. Either the reaction product (for example, semicarbazone in the determination of carbonyl compounds, or N-nitrosoamine in determinations of secondary amines) is measured, or a decrease in concentration of a reagent (for example, chromic acid in determination of alcohols, or mercuric or silver ions in determination of thiols) is followed. Or, alternatively, the polarographic limiting current can be taken as a measure of the amount of a given electroactive group (for example, quinones, aliphatic nitro compounds, phenazines, or thiols).

More frequent is determination of individual compounds. If the species is electroactive, the analysis frequently consists of dissolving the material in a proper supporting electrolyte, recording the waves, and evaluating them. Electroinactive species can be determined indirectly, by converting them into electroactive by a suitable chemical reaction (for example, aromatics by nitration, secondary amines by nitrosation). The most important fields of application of inorganic determinations are in metallurgy, environmental analysis (air, water, and seawater contaminants), food analysis, toxicology, and clinical analysis. The possibility of being able to determine vitamins, alkaloids, hormones, terpenoid substances, and natural coloring substances has made polarography useful in analysis of biological systems, analysis of drugs and pharmaceutical preparations, determination of pesticide or herbicide residues in foods, and so forth. Polarography also makes possible determination of monomers, catalysts, and some reactive groupings in polymers.

Other polarographic techniques. To eliminate unwanted charging current and to increase the sensitivity of polarography, the voltage is applied in regular pulses instead of gradually. When the current is measured only during the second half of the pulse, the technique is called pulse or differential pulse polarography. This technique is more sensitive by two orders of magnitude than dc polarography, and in inorganic trace analysis competes with atomic absorption and neutron activation analysis. With smaller dosage of more effective drugs, the sensitivity of differential pulse polarography has found application in drug analysis.

When an alternating voltage of small amplitude (a few millivolts) is superimposed on the dc voltage ramp, the technique is called ac polarography. It is particularly useful for obtaining information on adsorption-desorption processes at the surface of the dropping-mercury electrode. An increase in accuracy of polarographic methods can be achieved when the dropping-mercury electrode (or some other indicator electrode) is used in titrations. At constant applied voltage the current is measured as a function of volume of added titrant in amperometric titrations. The amperometric titration curve frequently has a shape of two linear sections. Their intersection corresponds to the titration end point.

Related methods. Methods in which a potential sweep is used (instead of the practically constant potential of the electrode during the life of a single drop in polarography) are frequently called voltammetry. Either an electrode with nonrenewed surface-solid electrode, mercury pool, or mercury drop-is used, or the whole scan is carried out during the life of a single drop when the dropping-mercury electrode is used. The use of solid electrodes (for example, platinum, gold, or various forms of graphite) makes it possible to extend the voltage range to more positive potentials, so that numerous substances which cannot be oxidized when mercury is used can be oxidized. Solid electrodes can be used when they are stationary, rotating, or vibrating, but the change at the electrode surface in the course of electrolysis often decreases the reproducibility of the current-voltage curve compared with polarographic curves.

When extremely small concentrations of amalgamforming metals (such as lead, cadmium, or zinc) are to be determined, a preconcentration can be carried out. The metal is deposited into a hanging mercury drop or into mercury plated on surfaces of solid electrodes. After a chosen time interval, a voltage sweep is applied, and the dissolution current of the metal ion from the amalgam is measured. Such anodic stripping, particularly when combined with the differential pulse technique, allows determination up to 1 part in 10^{12} parts of solution. Similarly, anions forming compounds with mercury can be preconcentrated by electrodeposition and dissolved in cathodic stripping. Alternatively, compounds absorbed at the mercury surface can be stripped.

When the applied voltage sweep is first increased and then decreased and the plot of dependence of voltage on time resembles a triangle, the technique is called cyclic voltammetry and offers, in particular, information about products and intermediates of electrode processes and their reactions.

Finally, when current is applied and the potential of an electrode measured, another group of techniques is developed. Most important among them is chronopotentiometry. In this technique, a constant current is applied to the indicator electrode in an unstirred solution, and its potential change is measured as a function of time. The time at which a sudden potential change is observed is known as the transition time, and its square root is proportional to the concentration of the reacting species. The method has found only limited practical application but has proven useful for the study of reactions of products of electrolysis. *See* ELECTROCHEMICAL TECHNIQUES. Petr Zuman

Bibliography. A. Bard, *Electroanalytical Chemistry*, 1993; A. J. Bard (ed.), *Electroanalytical Chemistry*, vols. 1-16, 1966-1988; A. J. Bard, *Electrochemical Methods: Fundamentals and Applications*, 2d ed., 2000; A. M. Bond, *Modern Polarographic Methods in Analytical Chemistry*, 1980; Z. Brainina and E. Neyman, *Electroanalytical Stripping Methods*, 1993; G. W. Ewing, *Instrumental Methods of Chemical Analysis*, 5th ed., 1985; L. Meites and P. Zuman, *Handbook Series in Inorganic Electrochemistry*, vols. 1-8, 1979-1988; L. Meites and P. Zuman, *Handbook Series on Organic Electrochemistry*, vols. 1-6, 1977-1983; J. T. Stock and M. V. Orna (eds.), *Electrochemistry*, *Past and Present*, 1989; B. H. Vassos and G. W. Ewing, *Electroanalytical Chemistry*, 1983.

Polaron

A quasiparticle that forms from an electronic charge carrier interacting with a solid's vibrating displaceable atoms. In particular, a polaron comprises an electronic charge carrier together with the carrier-induced altered motions of surrounding atoms. Different classes of material have distinct types of polarons. Polarons may be divided into weak-coupling polarons and strong-coupling polarons. Strong-coupling polarons comprise large selftrapped polarons and small self-trapped polarons.

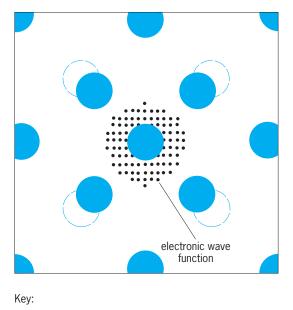
All types of polarons form because of the electronlattice interaction. The electron-lattice, or electronphonon, interaction describes the dependence of the energy of an electronic charge carrier on the positions of the surrounding atoms. As a result of this interaction neither electronic charge carriers nor atoms can move without affecting the other. *See* LAT-TICE VIBRATIONS; PHONON.

The electron-phonon interaction can be divided into two components. The long-range component of the electron-lattice interaction describes the dependence of the energy of an electronic charge carrier in an ionic (or polar) material on the positions of the surrounding ions. This interaction, like the underlying Coulomb interaction, is of long range. The short-range component of the electron-lattice interaction describes the dependence of the energy of an electronic charge carrier on the positions of nearby atoms with which the electronic carrier forms bonding or antibonding states. The short-range interaction predominates in covalent semiconductors. The term polaron was adopted in recognition of the electronlattice interaction typically being especially strong in polar (ionic) materials. See COHESION (PHYSICS); COULOMB'S LAW; IONIC CRYSTALS; SEMICONDUCTOR.

Electrons are lighter and more rapidly moving than are the relatively heavy and sluggish atoms with which they interact. Thus, electronic carriers tend to follow atoms' motions. Polaron effects result as electronic charge carriers adjust to and modify atomic motion.

Weak-coupling polaron. In some instances the electron-lattice interaction only slightly alters the motions of an electronic carrier and the atoms with which it interacts. Then the range over which an electronic carrier and the surrounding atoms affect one another is termed the weak-coupling polaron radius. This radius can be viewed as the distance that an electronic carrier can diffuse in the time that it takes for atoms to execute a single vibration. In this weak-coupling regime the electron-lattice interaction simply acts to slightly lower the net energy and to slightly enhance the effective mass of an electronic charge carrier.

Strong coupling: self-trapping. More dramatic effects occur when the electron-lattice interaction is strong enough and the electronic carrier moves slowly enough for it to become self-trapped. To understand the physical origin of self-trapping, it is useful to consider an electronic charge held fixed in a condensed medium. The presence of a localized electronic charge forces a shift of the equilibrium positions of the surrounding atoms (Fig. 1). This change of atoms' equilibrium positions lowers the electron's potential energy. If the potential well enveloping the electron is deep enough, the electron will be bound within it. Then, without the imposition of external forces, the electron cannot escape the potential well unless it is altered by atoms changing their positions. This situation will be energetically stable if the lowering of the electron's energy due to its being bound exceeds the strain energy associated with displacing atoms' equilibrium positions. This situation will be dynamically stable if the bound electron can complete an orbit in its well before the surrounding atoms can move appreciably.



displaced atomic equilibrium position
 equilibrium position without the electron

Fig. 1. An electron self-trapped by the displacements of equilibrium positions of atoms that surround it.

When these two stability conditions are fulfilled, an electronic carrier is said to be self-trapped. The term self-trapping emphasizes that the electronic carrier is bound within a potential well that its very presence induces. The associated polaron comprises the self-trapped carrier and the associated pattern of displaced atomic equilibrium positions.

Energies of self-trapped polarons lie below those for carriers that are not self-trapped (**Fig. 2**). The width of these polarons' energy band is less than that of a characteristic phonon, bv, where b is Planck's constant and v is the characteristic atomic vibration frequency. The narrowness of the polaron band is indicative of the slowness of polaron motion. Selftrapped carriers cannot move without atoms altering their positions.

Polarons are also associated with the absorption of electromagnetic radiation. These absorption bands arise when self-trapped electrons are excited from their ground state. The electrons may thereby be freed from the self-trapped potential well or raised to higher states within it.

Polarons may also pair as bipolarons. Pairing is fostered because placing two carriers proximate to one another tends to deepen the self-trapping potential

conduction band

polaron band

Fig. 2. Energy levels of quasifree and self-trapped states when the polaron is energetically stable.

well that both carriers experience. This tendency of polarons to amalgamate competes with the carriers' mutual Coulomb repulsion. This Coulomb repulsion is reduced and bipolaron formation aided if the carriers can avoid one another while moving within a common potential well. The coordination of the two carriers' motion is termed an electronic correlation effect. A bipolaron forms a singlet when the electronic spins of its two electronic carriers are aligned in opposition to one another. Thus, a singlet bipolaron has no net electronic spin but twice the electronic charge of a single electron. These properties may be used to identify singlet bipolarons.

Polaron formation is a nonlinear process. That is, self-trapping may be regarded as a feedback phenomenon. In particular, confining a charge carrier induces atomic displacements that deepen the potential well that confines the carrier, thereby enhancing the charge carrier's confinement. These feedback effects are subtle. Changes of the electron-lattice interaction qualitatively change the type of self-trapped polaron that is formed. *See* NONLINEAR PHYSICS.

Large (self-trapped) polaron. The self-trapping polaron that is formed with only the long-range component of the electron-lattice interaction is termed a large polaron since its self-trapped carrier generally extends over multiple atomic sites. The spatial extent of the self-trapped carrier decreases continuously as the strength of the electron-lattice interaction is increased.

Large polarons can move coherently through a solid. In particular, the centroid of a large polaron's self-trapped carrier shifts when the atoms responsible for the carrier's self-trapping alter their positions appropriately. Since atomic motion is required for the self-trapped carrier to move, large polarons move slowly with very large effective masses. Estimated effective masses are as high as several hundred times the rest mass of an electron. *See* BAND THEORY OF SOLIDS.

Large polarons' very high effective masses cause them to move with exceptionally large momenta. As a result, large polarons are not easily scattered. In particular, the rate with which a large polaron is scattered by atomic vibrations is often inversely proportional to its effective mass. Then a large polaron's mobility, its velocity divided by the electric field that drives its motion, is independent of its effective mass. The mobility of a large polaron also falls as its scattering by vibrating atoms increases upon raising the temperature. The large polaron mobility often varies inversely with temperature over a significant temperature regime (Fig. 3). Near room temperature, a large polaron's mobility is typically comparable to that of a conventional charge carrier that is not selftrapped. Thus, it is the relatively small scattering rate, not the mobility, that best distinguishes large-polaron motion from that of a conventional charge carrier, a weak-coupling polaron.

Small (self-trapped) polaron. The self-trapped carrier that is formed with just the short-range component of the electron-lattice interaction is termed a small polaron since its self-trapped carrier shrinks to

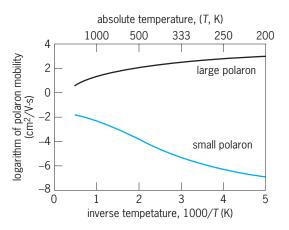


Fig. 3. Polaron mobilities versus inverse temperature.

the smallest relevant structural unit (such as an atom, bond, or molecule). Alternatively, if the electronlattice interaction is too weak to induce a collapse into a small polaron, the charge carrier does not selftrap at all. The carrier's interaction with vibrating atoms then produces only a weak-coupling polaron. This dichotomy between a severely localized selftrapped carrier and a carrier that is not self-trapped is a manifestation of the nonlinear character of selftrapping.

In the presence of both long-range and short-range components of the electron-lattice interaction, a selftrapped carrier may form a large polaron or a small polaron, or both types of self-trapped polaron may coexist. Since imposition of disorder fosters a carrier's localization, disorder tends to trigger smallpolaron formation. Thus small-polaron formation is more easily achieved in a disordered material than in a crystal.

The motion of a small polaron is generally described as proceeding by a succession of thermally assisted jumps between adjacent sites. Thus, a small polaron has an extremely low mobility that increases with rising temperature (Fig. 3). The temperature dependence of the small-polaron mobility becomes progressively milder as the temperature is lowered. This weakening of the mobility's temperature dependence results from atoms' motion being increasingly dominated by quantum-mechanical effects upon cooling. All told, small-polaron motion is qualitatively different from large-polaron motion. A large polaron has its free motion occasionally impeded by a scattering event. By contrast, a small polaron only occasionally departs from being static to jump to an adjacent site.

A magnetic field deflects a current of small polarons, thereby producing a Hall effect. However, the small-polaron Hall effect is unlike that of conventional charge carriers. In particular, the magnitude and temperature dependence of the mobility determined in a Hall-effect experiment, the Hall mobility, differs from the mobility that enters into the electrical conductivity (Fig. 3). The Hall mobility is generally larger and less temperature dependent than is the conductivity mobility. Most strikingly, the sign of the small-polaron Hall effect is often anomalous. Then, *p*-type small polarons in a semiconductor produce an *n*-type Hall effect, and similarly, *n*-type small polarons can generate a *p*-type Hall effect. These distinctive features provide a means of identifying the presence of small polarons. *See* HALL EFFECT. David Emin

Bibliography. C. L. Chien and C. R. Westgate (eds.), *The Hall Effect and Its Applications*, Plenum Press, New York, 1980; D. Emin, Small polarons, *Phys. Today*, 35(6):34–40, June 1982; C. Kittel, *Quantum Theory of Solids*, Wiley, New York, 1963; C. G. Kuper and G. D. Whitfield (eds.), *Polarons and Excitons*, Plenum, New York, 1963; E. K. H. Salje, A. S. Alexandrov, and W. Y. Liang (eds.), *Polarons and Bipolarons in High-T_c Superconductors and Related Materials*, Cambridge University Press, 1995, paper 2005.

Poliomyelitis

An acute infectious viral disease which in its serious form affects the central nervous system and, by destruction of motor neurons in the spinal cord, produces flaccid paralysis. However, about 99% of infections are either inapparent or very mild. *See* CENTRAL NERVOUS SYSTEM.

Infectious agent. Polioviruses are no longer thought to attack only neural tissue, because (1) they multiply in cultures of many nonnervous tissues; (2) their presence in the blood and antibody formation appear before the paralytic phase is reached and in cases where even transient signs of central nervous system involvement do not occur; and (3) the virus is regularly found in the throat and stools before the onset of disease, and after onset it is found for a week in the throat, and for several weeks in the stools. The infective virus particle is about 28 nanometers in diameter. Freezing preserves it for long periods. In contrast to the arboviruses, it is destroyed only slowly by alcohol and not at all by ether and deoxycholate. Poliovirus has a very restricted host range; most strains are limited to primates when cultured in the animal body and to primate tissue when in tissue culture. Three antigenic types of poliovirus are known. See ANIMAL VIRUS; ARBOVIRAL ENCEPHALITIDES; ENTEROVIRUS; TISSUE CULTURE.

Pathogenesis. The virus probably enters the body through the mouth; primary multiplication occurs in the throat and intestine. Transitory viremia occurs; the blood seems to be the most likely route to the central nervous system. The severity of the infection may range from a completely inapparent through minor influenzalike illness, or an aseptic meningitis syndrome (nonparalytic poliomyelitis) with stiff and painful back and neck, to the severe forms of paralytic and bulbar poliomyelitis. In all clinical types, virus is regularly present in the enteric tract. In paralytic poliomyelitis the usual course begins as a minor illness but progresses, sometimes with an intervening recession of symptoms (hence biphasic), to flaccid paralysis of varying degree and persistence. When the motor neurons affected are those of the diaphragm or of the intercostal muscles, respiratory paralysis occurs. Bulbar poliomyelitis results from viral attack on the medulla (bulb of the brain) or higher brain centers, with respiratory, vasomotor, facial, palatal, or pharyngeal disturbances.

Diagnosis. Laboratory diagnosis is by isolation, usually from stools inoculated into tissue cultures, and subsequent identification by neutralization with specific antiserums in a test tube, and by complement-fixing and neutralizing serum antibody rises. Isolation of a virus which is cytopathogenic in tissue cultures, but which is not antigenically identified, is not sufficient for diagnosis, for many other cytopathogenic enteroviruses also inhabit the enteric tract and produce syndromes similar to mild or early poliomyelitis. *See* NEUTRALIZING ANTIBODY.

Epidemiology. Poliomyelitis occurs throughout the world. In temperate zones it appears chiefly in summer and fall, although winter outbreaks have been known. It occurs in all age groups, but less frequently in adults because of their acquired immunity. In crowded underdeveloped areas and in tropical countries, where conditions favor a constant widespread dissemination of virus, poliomyelitis continues to be a disease of infancy, and a high percentage of children over 4 years old are already immune. Before the introduction of vaccine, in some areas of the temperate zones the age incidence had tended to change in marked parallel with improving sanitation and hygiene; exposure had been postponed, sometimes so long that parents were without immunity and succumbed to infections transmitted from their children. In 1952, the year of maximum incidence in the United States, more than 57,000 cases were reported; more than half involved paralysis.

The virus is spread by human contact; the nature of the contact is not clear, but it appears to be associated with familial contact and with interfamily contact among young children. The virus may be present in flies.

Prevention and control. Inactivated poliovirus vaccine (Salk; IPV), prepared from virus grown in monkey kidney cultures, was developed and first used in the United States, but oral poliovirus vaccine (Sabin; OPV) is now generally used throughout the world. The oral vaccine is a living, attenuated virus, and like most viruses it is unstable except when held at very low temperatures in the frozen state. The use of magnesium chloride (MgCl₂) as a stabilizing agent for poliovirus overcomes this problem; in molar concentrations, MgCl₂ protects live poliovirus vaccines, so that they may be stored in an ordinary refrigerator for over a year with no loss in immunizing potency. Vaccines stabilized with MgCl₂ are apt to maintain potency over periods of several weeks at room or transit temperatures.

For inactivated poliovirus vaccine, at least four inoculations over a period of 1-2 years have been recommended in the primary series, followed by periodic booster immunizations to maintain immunity. Fewer injections seem necessary with the enhanced potency (eIPV) vaccine; killed vaccine induces humoral antibodies, but upon exposure, the virus is still able to multiply in the nonimmunized gut.

The live poliovirus vaccine multiplies, infects, and thus immunizes. In the process, infectious progeny of the vaccine virus are disseminated. Although the viruses, particularly types 2 and 3, mutate in the course of their multiplication in vaccinated children, only extremely rare cases of paralytic poliomyelitis have occurred in recipients of oral polio vaccine or their close contacts. Repeat vaccinations seem to be important to establish permanent immunity. The vaccine induces not only IgM and IgG antibodies in the blood but also secretory IgA antibodies in the intestine, which then becomes resistant to reinfection.

A potential limiting factor for oral vaccine is interference. If the alimentary tract of a child is infected with another enterovirus at the time the vaccine is given, the establishment of polio infection and immunity may be blocked. This may be an important problem in areas (particularly in tropical regions) where enterovirus infections are common.

Trivalent oral poliovirus vaccine is generally used in the United States. Primary immunization of infants begins at 2 months of age. The second and third doses should be given at 2-month intervals thereafter, and a fourth dose at $1^{1}/_{2}$ years of age. A trivalent vaccine booster is recommended for all children entering elementary school. Adults residing in the continental United States have only a small risk of exposure. However, those who are at increased risk because of contact with a person with the disease or who are anticipating travel to an endemic or epidemic area should be immunized.

Although in 1993 more than 150,000 paralytic cases occurred annually around the world, chiefly in developing countries, polio has become a rare disease in the United States because of vaccine effectiveness. In 1987, only five cases were reported. Wild virus, which is found in a naturally infected host, has not been isolated in the United States since 1979, and the disease has almost vanished in all industralized countries. The last case caused by wild poliovirus in the Western Hemisphere occurred in 1990. However, there is a continuing need for adequate vaccination programs in all population groups in order to limit the spread of wild viruses. This is particularly important when wild viruses are introduced from developing countries.

Both inactivated and oral poliovirus vaccines induce antibodies and protect the central nervous system from subsequent invasion by wild virus. However, upon subsequent exposure to wild virus, antibody resulting from the inactivated virus vaccine has little effect on intestinal carriage of virus. The gut develops a far greater degree of resistance after administration of the oral vaccine—resistance that seems to be dependent on the extent of initial vaccine virus multiplication in the alimentary tract rather than on serum antibody level.

Live vaccine should not be administered to immunodeficient or immunosuppressed individuals. On very rare occasions, the live vaccine induces paralytic disease in persons not known to be immunodeficient. Such cases, which are carefully studied by public health agencies, are estimated to number about one vaccine-associated case for every million nonimmune persons vaccinated. *See* VACCI-NATION.

Postpolio syndrome. Paralysis and muscle wasting returns in many individuals decades after their experience with paralytic poliomyelitis. Many reported cases have been in the United States, where in 1993 there were approximately 300,000 persons with a history of polio disease. Although postpolio disease is rare, it is a specific syndrome. It does not appear to be a consequence of persistent virus infection, but seems a result of physiological and aging changes in paralytic individuals already burdened by loss of neuromuscular functions. The normal decline in anterior horn cells and consequently in muscle strength that occurs with aging is much more pronounced in persons who lost many anterior horn cells and perhaps 50% of their muscle strength during their acute bout with polio. However, loss of anterior horn cells as a result of aging seldom occurs before age 60, and most of the postpolio cases reported have occurred before this age. Also, some findings suggest that, rather than from loss of anterior horn cells, the syndrome may result from peripheral disintegration of individual nerve terminals in motor units that were reinnervated during recovery following the acute disease. See AGING. Joseph L. Melnick

Bibliography. E. Kcrstak, *Measles and Poliomyelitis: Vaccines, Immunization and Control*, 1993; S. M. Lemon and S. E. Robertson, Global eradication of poliomyelitis: Recent progress, future prospects, and new research priorities, *Prog. Med. Virol.*, 38:42–55, 1991; W. Melchers et al., The postpolio syndrome: No evidence for poliovirus persistence, *Ann. Neurol.*, 32:728–732, 1992; S. A Plotkin and E. A. Mortimer (eds.), *Vaccines*, 3d ed., 1999.

Pollen

The small male reproductive bodies produced in the pollen sacs of seed plants (gymnosperms and angiosperms). The term comes unchanged from Latin, meaning fine dust or flour, and was popularized by Carl Linnaeus in 1751 (but used since 1523). Pollen is usually carried by wind or insects to the sticky stigmas of angiosperms or to the naked ovules of gymnosperms. Fertilization, which follows pollination within a wide time range—minutes, hours, a year in *Agathis*, 15 months or more in juniper, and 2 years in stone pine (*Pinus pinea*)—is usually effected only by pollen of the same species. *See* FLOWER; PLANT RE-PRODUCTION.

Although the function of pollen has been intensively studied only in very recent times, Mesopotamians 5000 years ago understood the sexuality of date palms sufficiently to ensure their crops. Assyrian sculptures of about 1000 years later show ritual dusting by priests of pollen from male trees onto inflorescences of the female trees. Hand pollination is now a major factor in food production; for example,

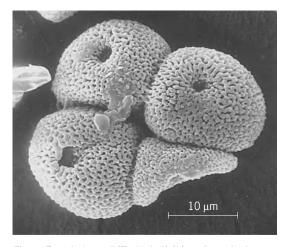


Fig. 1. Tetrad of cattail (*Typha latifolia*), grains cohering, one-pored, exine reticulate. (*Scanning electron micrograph* by C. M. Drew, U.S. Naval Weapons Center, China Lake, California)

in the culture of hybrid maize (*Zea*). Where pesticides have eradicated pollinating insects, hand pollination has become essential. *See* AGRICULTURAL SCI-ENCE (PLANT); BREEDING (PLANT).

Cattail (*Typha*) grains (**Fig. 1**) were made into cakes by natives of India and New Zealand, while cattail and maize pollen has long served in fertility rites of the Apache and Pueblo peoples, and in the Navaho search for the "Path of Peace." Unwittingly, humans have always ingested much pollen in honey and from the air. People may always have been allergic to the massive amounts shed by wind-pollinated trees such as olive and mulberry and by weed herbs in the amaranth, chenopod, daisy, and grass families, all of which should be combated as health hazards through selective planting or control of waste urban areas. *See* ALLERGY.

A palynologist's studies may involve many disciplines, from genetics and classification to aerobiology and the treatment of hay fever; to the study of pollen in pure, toxic, and adulterated honey; to the identification of remains in peats, coals, mineral sediments, feces, and in seasonal layers of ice. Most of these matrices yield datable stratigraphic records. Pollen analysis has become an indispensable tool in tracing vegetation history and climate, especially since the last glaciation. Macerated coals also yield to this technique since they are so rich in pollen and spores. Even the pollen and spore assemblages of Antarctic sediments are studied for correlation with those of other Gondwana segments.

Striking use of palynology has also been made by petroleum researchers in locating ancient shorelines and thus basins of deposition of oil and gas through recognition of land and shore algae, microspores, and other life forms, often present in high percentages in marine beds. *See* MICROPALEONTOLOGY; PA-LYNOLOGY; PETROLEUM MICROBIOLOGY.

The individual grain. On maturation in the pollen sac, a grain may reach 0.00007 mg as in spruce, or less than 1/20 of this weight. A grain usually has two waxy, durable outer walls—the ectexine and

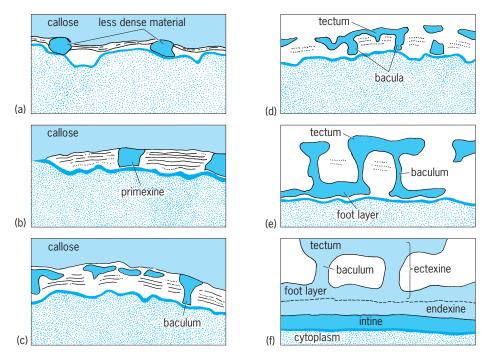


Fig. 2. Growth of hellebore (buttercup family) pollen walls: (*a*,*b*) matrix of sculpture material forms against callose wall; (*c*–e) ectexine (tectum, baculum, and foot layer) develops; and (*f*) endexine and intine appear. (*After P. Echlin, Pollen, Sci. Amer., 218:80–90, 1968*)

the endexine—composed of sporopollenin able to resist most acids and temperatures up to about 570° F (300° C). An inner fragile wall, the intine, surrounds the contents with its nuclei and reserves of starch and oil. Much of this detail can be seen in simple water or glycerin mounts at magnifications of $400 \times$ or less.

Only since the scanning electron microscope became commercially available have great advances been made with high resolution at very high magnifications, the range between 2000 and 12,000× being invaluable for pollen and spore work. However, because the scanning electron microscope registers only surface detail, the transmission electron microscope is used for interpretation of subsurface features far beyond the range of the light microscope. Used optimally in sequence on the same specimen, these two instruments have opened a new era of understanding of both living and fossil material.

The detail secured in studies of developing hellebore pollen grains is exemplified in **Fig. 2**, which should be compared with **Fig. 3** for a wider view of the callosic wall in which grains are embedded. In Fig. 2*a* and *b* is shown the initiation, after meiosis, of the waxy sporopollenin (biopolymer) walls with the deposition of the primexine matrix of fibrillar material in which the rod-layer elements develop. Growth of these little rods (known as columellae or bacula) extend below to form a "foot" layer and above (the limit of the ectexine) to form a "roof" (tectum), which may be perforate, as in Fig. 2*c*-*f*. **Figure** 4 shows various polarity and surface patterns, as revealed by the scanning electron microscope. *See* SCANNING ELECTRON MICROSCOPE. **Diagnostic characters.** Identification depends on interpretation of morphological features. Exine and aperture patterns are especially varied in the more highly evolved dicots, so that recognition at family, genus, or even species level may be possible despite the small surface area available on a grain. Since the morphological characters are conservative in the extreme, usually changing very slowly through geologic time, studies of fine detail serve to establish the lineal descent of many plants living today (**Fig. 5**). Ephemeral parts of the grain may also be useful in

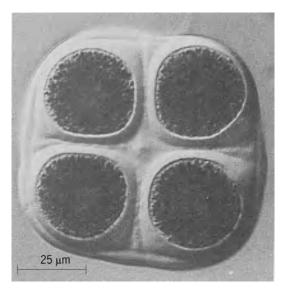


Fig. 3. Young tetrad of *Lavatera* (mallow family), with grains free in callose of pollen mother cell. (*Photomicrograph by Luc Waterkeyn*)

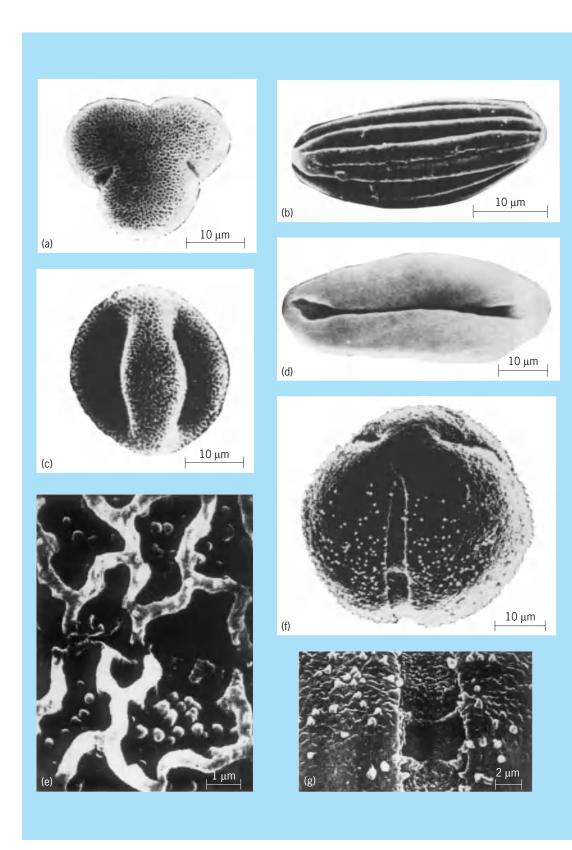


Fig. 4. Scanning electron micrographs of pollen grains, showing polarity and surface patterns. (*a*,*c*) Polar and side (equatorial) views of a reticulate, tricolpate *Homalanthus polyandrus* (spurge family) grain. (*b*,*d*) One-furrowed (monosulcate) grains with short polar axis: (*b*) ridged (plicate) grain of *Ephedra trifurca;* (*d*) smooth grain of nikau palm (*Rhopalostylis sapida*) with strongly defined sulcus. (*e*) Portion of exine of *Pelea barbigera* (orange family) showing an open network of walls (muri) supported by rods; free rods stand on the "floor" enclosed by the muri. (*f*) Tricolporate grain of *Scaevola kauaiensis* (Goodenia family), slightly oblique view showing the furrow, equatorial germ pore, smooth polar area, and scattering of spinules on perforate tectum. (*g*) Detail of sulpture and pore shape of *S. kauaiensis* pollen. (*Scanning electron micrographs a, c, and d by R. S. Thompson and L. M. Cranwell; b by C. M. Drew; e, f, and g by J. V. Ward*)

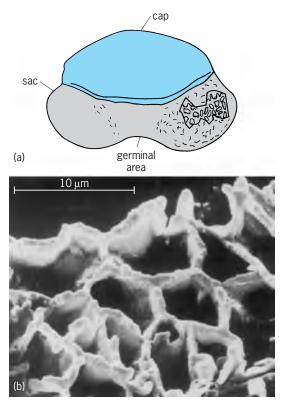


Fig. 5. Subfossil *Pinus* grain, corroded in soil. (a) Sketch showing cap and two sacs in side view, with right sac broken. (b) View into damaged right sac, which lacks its tough outer wall, to show honeycomb (alveolate) ectexine (sexine) structure in the cavity; note absence of rods (columellae). (*Scanning electron micrograph by C. M. Drew*)

diagnosis (**Fig. 6***a*), excessive thickening of the intine under the pores being a distinctive character of certain wind-pollinated types, as in hazel and birch, while hyaline plugs that swell to help split the aperture membranes are found in oak and beech. Color, though mainly yellow, may also be distinctive: the range from whitish to orange, green, or blackish is dramatic in the lily family. The walls of the grain, however, provide the most useful characters.

Ornamentation. Few grains are completely smooth (psilate), although they are so in some palms. Whether with or without a tectum, the sculptural elements occur on the surface. In subtectate grains, where narrow muri represent the tectum, a pattern of short rods may be exposed on the floor between the walls, whose longer rods (columellae) function as supports, like pilings under a wharf. Common patterns include echinate (Hibiscus); reticulate; microechinate (Nothofagus); smooth; foveolate (pitted); baculate (rods untapering, as in Franklandia); clavate or pilate (rods expanded at top, as in *Ilex*); and verrucate (warty). A striking stellate reticulate pattern of low muri topped by wedge-shaped clavae occurs in Croton and Aleurites (spurge family); and an identical pattern distinguishes Stellatopollis, which dates from the Early Cretaceous period.

Size. Extreme variations may occur within a family, but pollen grains range mainly from 15 to 50 micrometers, with the dicot range being from 2 μ m in *Myosotis* to 250 μ m in *Mirabilis*. A unique East African *Crossandra* has three-furrowed grains up to about 600 μ m × 19 μ m. Terrestrial monocots range from about 15 to 300 μ m or more, and a few aquatic types are exceptionally long, with eelgrass (*Zostera*) having pollen measuring 2550 μ m × 3.7 μ m in a class of its own; living gymnosperms range from 15 μ m (*Gnetum*) to about 180 μ m in *Abies* (including sacs), while fossil types range from about 11 μ m to 300 μ m.

Symmetry and shape. In most pollen grains the polar axis runs from the inner (proximal) face to the outer (distal) face, as oriented during tetrad formation. The equator crosses it at right angles. Bilateral grains dominate in the gymnosperms and monocots, the polar

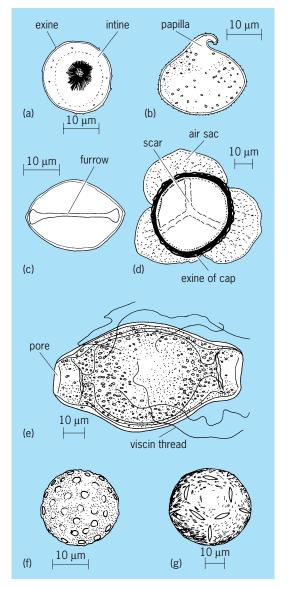


Fig. 6. Some varied pollen types of (*a-d*) gymnosperms and (*e-g*) angiosperms. (*a*) Inaperturate (*Cupressus arizonica*). (*b*) Papillate (*Sequoia sempervirens*). (*c*) Monosulcate (*Zamia floridiana*), a cycad. (*d*) Young three-saccate grain of kahikatea (*Dacrycarpus dacrydioides*) showing rounded germinal area (distal) and vestigial trilete scar on opposite (proximal) side of grain. (*e*) Diporate dicot type (*Fuchsia excorticata*) with viscin threads. (*f*) Many-pored grain of glasswort (*Salicornia australis*, goosefoot family). (*g*) Grain of *Heimerliodendron* (four-o'clock family) with 15 apertures.

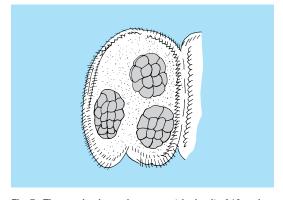


Fig. 7. Three polyads, each a symmetrical unit of 16 grains, lying free in an open anther lobe of *Albizzia julibrissin* (mimosa family).

axis usually being the shorter one, with the single aperture on the distal side. On the other hand, almost all dicot pollen is symmetrical around the polar axis (usually the long axis), with shapes ranging mainly from spheroidal to ellipsoidal, with rounded equatorial outlines, and sometimes "waisted" as in the Umbelliferae. Three-pored grains often have strikingly triangular outlines in polar view. Proximal tetrad scars (three-armed), so common in lower plants, occur freely in ancient conifer grains, but are only rarely traced in those of modern genera (Fig. 6*d*) and much more rarely in dicots, such as in the tropical *Cardiospermum*.

Aggregation. Most grains are free (monads) though often loosely grouped because of the spines, sticky oils, or viscin threads, as in *Fuchsia* (Fig. 6e), which are laid down in the pollen mother cell. It is suggested that the sticky "pollenkitt" may ensure germination of the pollen tube by enabling recognition of a compatible stigmatic surface. Compound grains (polyads) are richly developed in some angiosperm families, commonly occurring in fours (tetrads) as in *Catalpa*, many species of *Epilobium*, and most of the Ericales; or in multiples of four (polyads, **Fig. 7**) up to 64 or more (as in orchid pollinia). Twinned types (dyads) are rare. Permanent association in tetrads is rare in gymnosperms, but may occur freely in *Classopollis*, a fossil type.

Apertures. Rarely lacking, the flexible membranes of apertures are sometimes covered only by endexine. They allow for changes in volume and for the emergence of the pollen tube. They are classified as furrows (colpi), with elongate outlines, and pores usually more or less circular in shape (Figs. 1, *6f*). Short slitlike intermediate forms occur (Fig. 6g). A few families have no apertures (for example, Lauraceae), some as a result of reduction as an adaptation to wind or water pollination involving loss of exine. Gymnosperm (for example, cypresses) may also lack openings, or have one small papilla as in *Sequoia* (Fig. 6*c*). Most have one furrow as in cycads (Fig. 6*c*), or a long weak sulcoid distal area (sulcus) as in some conifers (**Fig. 8g** and *b*).

Angiosperm pollen exhibits a rich variety of furrow and pore types, with the association of pore and furrow (colporate forms) evolving in the dicots in the Albian. The first known tricolpate grains are Aptian in age.

Almost all grasses have one pore and almost all palms have one furrow. Dicot genera are typically tricolpate, triporate, or tricolporate, but may have numerous configurations (for example, 15 short apertures symmetrically arranged, as in Fig. 6g). The number of pores ranges from 1 to about 90. Rarely, paired pores occur, one above and one below the equator, and they are either free, as in *Anacolosa*, or set in a furrow, as in *Ribes* (gooseberry family). A few pore types, such as those of grasses (Fig. 8f) and gourds (**Fig. 9**e), have an operculum (lid).

Dispersal. Grains of small to medium size (15 to $40 \,\mu$ m) travel highest and farthest; theoretically, they could circle in the stratosphere almost forever. In fact, most fall close to the parent plant or are soon washed out of the air by rain, hail, or snow. In studies of the settling velocity of certain Southwest desert types, it has been found that jojoba (Simmondsia) and Ephedra pollen falls, or is trapped by the female flowers, very close to the male plants. For Epbedra, it has been reported that 95% of the grains blow less than 11 ft (3.5 m) from their source. It is all the more remarkable, then, that some South American Ephedra and Nothofagus pollen does reach the subantarctic islands. The known limit of dispersal, 4700 mi (7500 km), was obtained from studies of peat on Marion Island in the Indian Ocean. In these studies, a few types blown 1100 mi (1800 km) from the southern tip of Africa were recovered.

Pollen also travels in rivers and ocean currents. J. G. Kölreuter in 1766 was the first to subdivide pollen according to windborne (anemophilous) and insect-borne (entomophilous) types. The former usually have small volume and little or no sculpture (Figs. 6, 8, and 9), while the latter range from very small to very large and heavy (Figs. 8*e* and 4*d* and *e*) and tend to "clump" because of their oily and sticky surfaces.

The number of grains per pollen sac is usually much lower in entomophilous (about 30 in *Mirabilis*) than in anemophilous types, which have

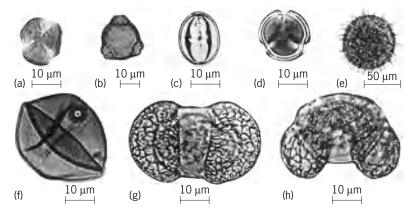


Fig. 8. Dicot aperture types, (a) Tricolpate (*Hebe rapensis*). (b) Triporate (*Casuarina stricta*) (photomicrograph by Isobel C. Cookson). (c) Tricolporate (Griselinia littoralis), side view, unexpanded; (d) polar view, intine present, expanded. (e) Periporate (*Hibiscus trionum*, mallow family). (f) Monocot type, one-pored (*Poa flabellata*, grass family). (g) Conifer type (*Podocarpus spicatus*), polar view, and (h) equatorial view.

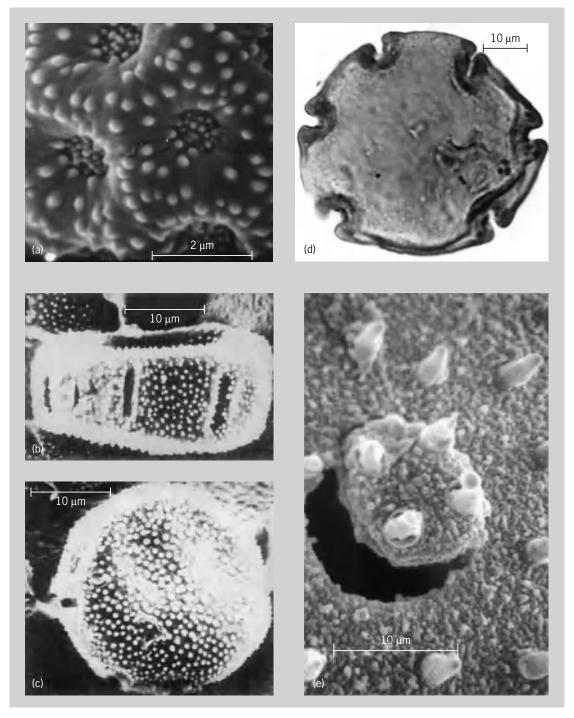


Fig. 9. Some anemophilous dicot types and an entomophilous type of pollen grains. (a) Surface of many-pored Russian thistle (*Salsola kali*), grain with very small spinules. (b, c) Modern and (d) fossil *Nothofagus fusca*-type grains notable for endexinal thickening around apertures; (b, c) *N. truncata*, side and polar views, with spinules dense and bold and apertures slitlike (thickenings hidden under surface); (d) fossil Antarctic grain with very thick rings from the Oligocene. (e) Spiny insect-carried grain of the coyote gourd (*Cucurbita palmata*) showing a portion of exine with pore lid detached. (*Stereoscan electron micrographs a-d and e by C. M. Drew*)

spendthrift yields that form "sulfur clouds" over forests and thick "blooms" on lakes. A single male cone of *Araucaria* may produce 10^9 grains; those of pine 1.5×10^6 . Spruce forests of southern Sweden alone produce something like 75,000 tons (68,000 metric tons) of pollen each year, estimated at about 10^{21} grains.

Grasses yield vastly and are often geared to shed

their pollen for a few minutes at a precise hour of day or in warm or cool weather only. In very cold, wet climates, as in South Georgia Island, the anthers may not open and may be preserved full of pollen in layers of peat. A few monocot species with threadlike pollen shed under water (for example, *Zostera*), while a few others raise their flowers to shed in the air (for example, *Ruppia*). Almost all gymnosperms and most trees, shrubs, and herbs causing hay fever are wind-pollinated.

"Bat flowers" may also yield rich stores of pollen, well exposed on trunks and branches, as in *Kigelia pinnata*, the sausage tree. Such flowers are usually dusky or pallid, and may be odd-smelling; the bats are drawn to them after dark. Hummingbirds, seeking nectar, are also common pollinators, especially in tropical parts of the Americas.

Most flowering plants are solely either wind- or insect-pollinated, but there are odd deviants such as wormwood and sagebrush (Artemisia species) in the daisy family, whose pollen is otherwise almost exclusively carried by insects. Among the maples, box elder (Acer negundo) relies wholly on wind, while Norway maple (A. platanoides) depends on insects; the other species may combine both means of dispersal. The more ancient beech family, with fossil pollen dating back to the Upper Cretaceous, is about half wind- and half insect-pollinated, the oaks (Quercus) and southern beeches (Nothofagus), both wind-pollinated, now being widespread in the Northern and Southern hemispheres, respectively. Pollen of the latter genus is so distinctive (Fig. 9b-d) that it can be recognized with ease in deposits which are up to about 8×10^7 years old. Grasses and sedges, although now predominantly anemophilous, may produce nectar and thus may be reverting, in part, to an earlier condition. See POLLI-NATION.

Pollen record. First, in the Upper Carboniferous, over 3 \times 10⁸ years ago, appeared simple, often one-furrowed pollen, followed closely by types with one encircling sac, and later two separate sacs. Many were produced by gymnosperms, soon to supplant the giant lycopods and horsetails and dominate world vegetation for well over 2×10^8 years. Grains with three sacs had a Lower Cretaceous peak, especially in Gondwana countries. Today, pines and their allies (mainly in the Northern Hemisphere) and podocarps (mainly in the Southern) continue to shed vast crops of pollen, mainly bisaccate. Only a few western Pacific relict genera now have trisaccate grains, some large, as in the New Zealand Dacrycarpus dacrydioides (Fig. 6d), some small, as in the Tasmanian Microcachrys. Cycads, which are native to old land surfaces today, retain the shape and single furrow (Fig. 6c) of the ciliate sperm of their very ancient line of descent.

The question of the nature and rise of the earliest flowering plants (that is, the angiosperms) is being penetrated by modern and fossil studies, but it is still not known where their roots lie in the past.

Angiosperm characters typify *Clavatipollenites hughessi* (**Fig. 10**), a pollen type from Barremian and Aptian beds in England. It is now known that in their history of 1.15×10^8 years the four living genera of the chloranth family share the same diversity of aperture types and the same minute supratectate spinules and granules of their ancient forerunners in the Lower Cretaceous.

With the finding of varied monocot pollen in the eastern United States, the first firm dicot and

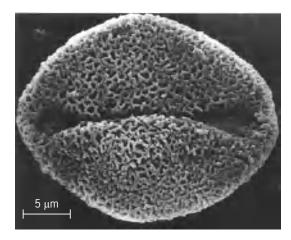


Fig. 10. Polar view of an Aptian *Clavatipollenites* grain, showing the long furrow and surface pattern (as in living *Ascarina*). From the Arundel clay of the Atlantic coastal plane, United States. (*Scanning electron micrograph by J. V. Ward*)

monocot records have been proved to be more or less coeval. These ancient monocots (leaves associated) with larger, also reticulate grains have been described from basal Potomac beds. Their coarse and varied mesh patterns are typical of many modern insect-pollinated monocots, among them the modern lilies.

Also in the Aptian in a few countries appear a very few small dicot types with three furrows, much as in living *Homalanthus* (spurge family; Fig. 4*a* and *b*). It is clear, then, that while the flowering plants came in very gradually, by the Late Cretaceous they had richly diversified their aperture and sculpture types. However, many of the dominants of present landscapes, such as the grasses with their single pore (Fig. 8*f*), the chenopods with their many pores (Figs. 6*f* and 9*a*), and members of the daisy family (usually tricolporate), made slow starts in the early Tertiary before surging to their present importance, along with the more than 250,000 other species of flowering plants now known.

The structure of the exine walls has been found to fall into three basic categories: (1) columellarrestricted to angiosperms (except for Classopollis, an extinct gymnosperm); (2) alveolar (spongy)characteristic of cycads and saccate conifers (Fig. 5a and b); and (3) granular-as in Ephedra and the spinulose Gnetum, but in only a few dicots, such as certain magnoliids. These advances mean that many of the earliest dicots can be identified as such, whereas their pollen might earlier have been confused with that of the smooth-walled gymnosperms. Unfortunately, the taxa of the third group can be separated only through transmission electron microscopy, which reveals a laminated endexine (below the granules) in the gymnosperms and an unlaminated endexine (or none at all) in the few known dicot examples. Lucy M. Cranwell

Bibliography. S. Blackmore and I. K. Ferguson (eds.), *Pollen and Spores: Form and Function*, 1986; K. Faegri and P. Kaland, *Textbook of Pollen* *Analysis*, 4th ed., 1989, reprint, 2000; W. H. Lewis, P. Vinay, and V. E. Zenger, *Airborne and Allergenic Pollen of North America*, 1983; P. D. Moore, J. A. Webb, and M. E. Collinson, *Pollen Analysis*, 2d ed., 1992; J. Muller, Fossil pollen records of extant angiosperms, *Bot. Rev.* (Lancaster), 47:1-142, 1981; C. Newman, Breath of life and sneezes, *Nat. Geog.*, 166:490-520, 1984; K. J. Niklas, Wind pollination: A study in controlled chaos, *Amer. Sci.*, 73:462-470, 1985; A. Traverse, *Paleopalynology*, 1988.

Pollination

The transport of pollen grains from the plant parts that produce them to the ovule-bearing organs, or to the ovules (seed precursors) themselves. In gymnosperms ("plants with naked seeds"), including conifers and cycads, the pollen, usually dispersed by the wind, is simply caught by a drop of fluid excreted by each freely exposed ovule. In angiosperms ("flowering plants"), where the ovules are contained in the pistil, the pollen is deposited on the pistil's receptive end (the stigma), where it germinates. *See* FLOWER.

Ideally, each pollen grain forms a pollen tube which grows down through the slender style (stalk of the pistil) toward one of the ovules enclosed in the basal part or ovary. Here one of the two sperm cells within the pollen tube fuses with the egg cell of the ovule, making possible the development of an embryo, while the other sperm cell combines with two so-called polar nuclei of the ovule, thus starting the formation of endosperm, a tissue rich in reserve food. After this double fertilization, the ovule enlarges and becomes transformed into a seed.

Without pollination, there would be no fertilization; it is thus of crucial importance for the production of fruit crops and seed crops. Pollination also plays an important part in plant breeding experiments aimed at increasing crop production through the creation of genetically superior types. *See* BREED-ING (PLANT); PLANT REPRODUCTION.

Self- and cross-pollination. In most plants, selfpollination is difficult or impossible, and there are various mechanisms which are responsible. For example, in dichogamous flowers, the pistils and stamens reach maturity at different times; in protogyny, the pistils mature first, and in protandry, the stamens mature before the pistils. Selfing is also impossible in dioecious species such as date palms and willows, where some plants bear flowers that have only pistils (pistillate or female flowers), while other individuals have flowers that produce only pollen (staminate or male flowers). In monoecious species such as hazel, where pistillate and staminate flowers are found on the same plant, self-breeding is at least reduced. Heterostyly, as exemplified by certain primroses, is another device that promotes outbreeding. Here some flowers (pins) possess a long pistil and short stamens, while others (thrums) exhibit the reverse condition; each plant individual bears only pins or only thrums.

Pollen transport from stamens to pistils of the same length is easy, while in the other cases it is difficult.

Flower attractants. As immobile organisms, plants normally need external agents for pollen transport. These can be insects, wind, birds, mammals, or water, roughly in that order of importance. In some plants, for example, certain arum lilies, Dutchman's-pipes, and orchids, the pollinators are simply trapped; in the large majority of cases, however, the flowers offer one or more rewards, such as sugary nectar, oil, solid food bodies, perfume, sex, an opportunity to breed (as in figs), a place to sleep (Serapias orchids), or some of the pollen itself. For the attraction of pollinators, flowers provide either visual or olfactory signals. Color includes ultraviolet, which is perceived as a color by most insects and at least some hummingbird species. Fragrance is characteristic of flowers pollinated by bees, butterflies, or hawkmoths, while carrion or dung odors are produced by flowers catering to certain beetles and flies. A few orchids (such as Opbrys and Cryptostylis species), using a combination of olfactory and visual signals, mimic females of certain bees or wasps so successfully that the corresponding male insects will try to mate with them, thus achieving pollination (pseudocopulation).

At close range, pollinating insects are in many cases guided to the nectar by special color patterns, lines, or dots known as nectar or honey guides. Olfactory honey guides and osmophores (particular flower parts specialized for odor production) may be equally important. For better evaporation of odors, certain flowers may also develop considerable heat through a special process known as thermogenic respiration (certain arum lilies and water lilies).

Animal pollinators. While some flowers are "generalists," catering to a whole array of different animals, others, such as *Opbrys*, are highly specialized, being pollinated by a single species of insect only. Extreme pollinator specificity is an important factor in maintaining the purity of plant species in the field, even in those cases where hybridization can easily be achieved artificially in a greenhouse or laboratory, as in most orchids. The almost incredible mutual adaptation between pollinating animal and flower which can frequently be observed exemplifies the idea of coevolution.

Bird pollination. Hummingbirds have good red vision and feed on the wing. Flowers pollinated by these birds are often orange and red and hang down freely or are exposed (*Fuchsia*); they are sturdy, lack a landing platform, are open in the daytime, have little or no odor, and offer a not-too-concentrated nectar. Plants with flowers pollinated by birds other than hummingbirds (such as sunbirds in Africa) usually provide perches.

Bat pollination. Flowers pollinated by bats include those of the sausage tree, saguaro cactus, cup-andsaucer vine, and certain bananas. They are open in the evening, when they emit a rank or musty odor; their petal colors are either murky and nondescript or snow white; they are wide-mouthed and often bell-shaped; and they normally offer both nectar and pollen in abundance. In connection with the poorly developed sonar sense of pollinating bats, these flowers often hang down freely from ropelike twigs (flagelliflory), or are on the main stem or limbs of the tree (cauliflory). These phenomena, however, may also be connected with the dispersal, by bats, of the fruits and seeds, even though the pollinators may have been different animals.

Hawkmoth pollination. Flowers pollinated by hawkmoths have a great deal in common with those pollinated by hummingbirds, but they are white or lightcolored and very fragrant, especially in the evening when they are open.

Butterfly pollination. Flowers pollinated by butterflies open in the daytime, are usually colorful and fragrant (butterfly bush, or *Buddleia*; lilacs), and offer either a flat landing space (*Impatiens*) or, when small, are combined in inflorescences providing flattish surfaces to be walked on (red valerian, *Buddleia*). The narrow flower tubes allow the slender butterfly proboscis to go through, but discriminate against other insects.

Fly and beetle pollination. Both flies and beetles are highly diversified groups, which makes it difficult to describe "typical" fly- or beetle-pollinated flowers. Trap flowers exploit types such as carrion and dung beetles (or flies) which are not at all adapted to flowers. However, many modern flowers cater to animals with highly specialized mouthparts, such as certain longhorn beetles and the bee-fly *Bombylius*.

Bee pollination. By far the most important of all insect pollinators are the social bees, especially honeybees (*Apis*) and bumblebees (*Bombus*). While the former are subtropical in origin, bumblebees are adapted to cold climates. The larvae of both *Apis* and *Bombus* live on pollen and honey (a modified nectar); the adults drink nectar.

Among the physical and behavioral adaptations the adults possess to obtain these foods are long tongues, hairy bodies, special pollen baskets on the hindlegs, and a good color sense. The four broad color regions which bees distinguish, according to Karl von Frisch, are yellow, blue-green, blue, and ultraviolet. They learn very quickly to associate colors or minty odors with food, and remember the link for several weeks. These traits, and the bees' excellent time sense, which enables them to visit flowers when these have the most to offer, are important in the establishment of flower constancy: the bees stay with one, or very few, crop plants only, which is important for seed production.

Honeybees can distinguish various sugars. In contrast to bumblebees, which are "rugged individualists," honeybees have an excellent communication system: a worker bee can indicate both the direction and the distance of a rich food source by means of special dances, so that the resource can be exploited cooperatively. Flowers pollinated by bees, for example, true sage, bleeding heart, and Scotch broom, generally have bright colors, minty odors, and a landing platform; they often display bilateral symmetry. Their pollen grains are sticky, spiny, or highly sculptured for better adherence to the bees.

Wind pollination. Wind pollination is prevalent in the primitive cycads and in conifers, but it is also found in such familiar flowering plants as grasses, rushes, sedges, cattails, plantains, nettles, hemp, sorrel, lamb's-quarters, alder, birch, hazel, poplars, and temperate-region oaks. It offers advantages in regions where animal pollinators are scarce or where these cannot operate due to adverse environmental or seasonal conditions, such as strong winds. Wind-pollinated flowers are usually inconspicuous, being devoid of attractants and rewards for the pollinators, such as showy petals, fragrance, and nectar.

To facilitate exposure to the wind, blooming often takes place in early spring before the leaves have emerged, or the flowers may be placed very high on the plant. In a number of species, the anthers burst open in an explosive fashion, scattering the pollen widely; the stigmas often are long and divided into arms or lobes, or, like those of some willows and cereals, are feathery, which facilitates pollen catching. The latter process is also promoted because, in open areas, wind-pollinated plants of one species often grow together in dense populations. The pollen grains, which are shed singly, are dry, glueless, smooth-surfaced, and in some cases (pine, cedar) provided with air sacs. Wind pollination is a wasteful process because it lacks precision. For example, while one male plant of Mercurialis annua, a common weed, may produce 1.25×10^9 pollen grains, there is only a slim chance that an individual stigma will be hit by more than one or two pollen grains. Undoubtedly in connection with this, the number of ovules per pistil and (later) the number of seeds per fruit is very low (two ovules per flower in beech, hazel, oak, and walnut; one per flower in birch, grasses, elm, stinging nettle, and sweet gale).

Water pollination. Although the pollen grains of some plants can be made to germinate in aqueous sugar solutions, water alone in most cases will destroy them. Accordingly, water is seldom the true medium of pollen dispersal, even in aquatic plants. In ribbon weed (*Vallisneria*) the male and female flowers are released to the surface separately. There, male flowers are driven by the wind until they bump into female flowers; the shock causes the pollen to be catapulted over to the stigma. True water dispersal of pollen (hydrophily), in which the grains are wetted, is found only in the fresh-water hornworts (*Ceratopbyllum*) and marine eelgrasses (*Zostera, Pbyllospadix*). Bastiaan J. D. Meeuse

Ecology. Traditionally pollination biology has considered mechanisms that promote or guarantee the proper transfer of pollen from one individual of a given plant species to another. Often this involves a careful description of the transfer mechanism coupled to a particular pollinating agent (such as hawkmoths, bumblebees, bats, or wind). Pollinator-plant interactions, however, have also been used as a model for understanding broader evolutionary and

ecological issues that pertain to all living systems. Pollination systems have proven to be an excellent model for studying many current issues in population biology.

Since plants have a variety of breeding structures, they provide an opportunity for assessing the reproductive value of different sexual strategies and for modeling the evolution of breeding systems and sexual allocation. A classic problem in animal populations is the number of males and females a mother should produce as offspring (the sex ratio). Hypotheses accounting for the evolution of the sex ratio can be tested by using dioecious plant species, and speculations on sex-ratio evolution can be far more complicated in plants due to their striking variety in methods of sexual allocation (such as gynodioecy, androdioecy, or hermaphrodism).

If pollen movement is considered as a measurable analog of gene movement, then pollination systems are ideal for unraveling patterns of gene flow and the concomitant genetic structuring of plant populations. Pollen transfer can be measured by numerous methods. Gene flow (or the lack of it) has been implicated in retarding (or promoting) local population adaptation to different habitats as well as affecting rates of species formation. The traditional and the ecological approaches to pollination studies are distinguishable by their different views on the important consequences of pollen transfer. The traditional view stresses the mechanism of transfer within individual flowers or plants, while the ecological view stresses the population consequences of transfer among many individuals, even across communities or habitats. See POPULATION ECOLOGY; POP-ULATION GENETICS.

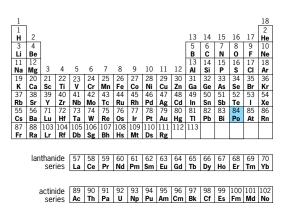
Pollinators are often faced with having to make complicated decisions about when, where, and on what flowers to forage. They therefore have been used as models for testing hypotheses concerning animal behavior such as patterns of learning, the acquisition of information from the environment, or alteration of behavior associated with changes in the environment.

The complex interaction among plants and their pollinators has prompted research on communitylevel ecology as well as population-level approaches. The timing of flowering within given habitats may reflect competitive interactions among plants for pollination, and competition may also affect the kinds of pollinators found in a habitat and their pattern of resource use. The evolutionary ecology of many plants and pollinators may be best understood as a result of complex interactions among populations of many species. *See* ECOLOGICAL COMMUNITIES; ECOL-OGY; POLLEN; POPULATION DISPERSAL. Leslie Real

Bibliography. S. Blackmore and I. K. Ferguson, *Pollen and Spores: Form and Function*, 5th ed., 1990; H. Hartmann, D. Kester, and F. Davies, *Plant Propagation: Principles and Practices*, 6th ed., 1996; R. P. Kapil (ed.), *Pollination Biology: An Analysis*, 1986; E. Ottaviano, M. S. Gorla, and G. Bergamini-Mulcahy (eds.), *Angiosperm Pollen and Ovules*, 1992.

Polonium

A chemical element, Po, atomic number 84. Marie Curie discovered the radioisotope ²¹⁰Po in pitchblende. This isotope is the penultimate member of the radium decay series. All polonium isotopes are radioactive, and all are shortlived except the three α -emitters, artificially produced ²⁰⁸Po (2.9 years) and ²⁰⁹Po (100 years), and natural ²¹⁰Po (138.4 days). *See* PERIODIC TABLE.



Polonium (²¹⁰Po) is used mainly for the production of neutron sources. It can also be used in static eliminators and, when incorporated in the electrode alloy of spark plugs, is said to improve the cold-starting properties of internal combustion engines.

Most of the chemistry of polonium has been determined using ²¹⁰Po, 1 curie of which weighs 222.2 micrograms; work with weighable amounts is hazardous, requiring special techniques. Polonium is more metallic than its lower homolog, tellurium. The metal is chemically similar to tellurium, forming the bright red compounds SPoO₃ and SePoO₃. The metal is soft, and its physical properties resemble those of thallium, lead, and bismuth. Valences of 2 and 4 are well established; there is some evidence of hexavalency. Polonium is positioned between silver and tellurium in the electrochemical series.

Two forms of the dioxide are known: low-temperature, yellow, face-centered cubic (UO_2 type), and high-temperature, red, tetragonal. The halides are covalent, volatile compounds, resembling their tellurium analogs. *See* RADIOACTIVITY; TELLURIUM. Kenneth W. Bagnall

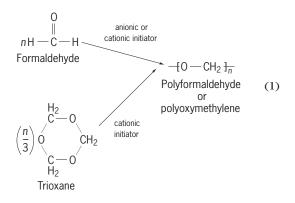
Bibliography. P. W. Atkins et al., *Inorganic Chemistry*, 4th ed., 2006; F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., 1999; D. Sangeeta, *Inorganic Materials Chemistry Desk Reference*, 1997.

Polyacetal

A polyether derived from aldehydes ($R=CH_2$ or CHR') or ketones [R=C(R'R'')] and containing -O-R-O- groups in the main chain. Of the many possible polyacetals, the most common is a polymer or copolymer of formaldehyde, polyoxymethylene

 $(-O-CH_2-)_n$. While the substance paraformaldehyde contains oligomers or low-molecular-weight polyoxymethylenes (*n* very small), high-molecularweight, crystalline polyoxymethylenes constitute an important class of engineering plastics that, in commerce, is often simply referred to as polyacetal. Cellulose and its derivatives also have a polyacetal structure. For discussion of other polyethers and polymers containing acetal side groups. *See* ACETAL; CELLULOSE; FORMALDEHYDE; POLYETHER RESINS; POLYVINYL RESINS.

Chemistry. The carbonyl group is the only unsaturated linkage besides the carbon-carbon double bond that has been successfully polymerized by a chain reaction on a commercial scale. As shown in reaction (1), formaldehyde can be readily polymerized by



using anionic initiators such as triphenylphosphine and, somewhat less readily, by using cationic initiators such as protonic acids. Alternatively, a similar polymer can be obtained by the ring-opening polymerization of trioxane using, for example, a boron trifluoride complex as initiator.

At temperatures above ~ 110 °C (230 °F; the ceiling temperature, above which depolymerization becomes favored over polymerization), the polymers as formed degrade by an unzipping reaction to reform monomer. To prevent this, one of two approaches is commonly used: esterification of the hydroxyl end groups, or copolymerization with a small amount of a monomer such as ethylene oxide or 1,3-dioxolane. Polymerization using gamma radiation has also been reported.

In its linearity, polyoxymethylene resembles polymethylene; the molecules pack well and crystallize readily. However, the polarity associated with the -C-O- linkage results in a higher crystalline melting point than that of polymethylene [$\sim 195^{\circ}C$ $(383^{\circ}F)$ versus ~135°C (275°F)]. Other aldehydes such as acetaldehyde or trichloroacetaldehyde can be polymerized at low temperatures (to avoid aldol condensation) and, preferably, high pressures by generally similar techniques, but the ceiling temperatures of the products are very low $[-35^{\circ}C(-31^{\circ}F)]$ and 13°C (55°F), respectively]. Even acetone has been polymerized on a small scale to yield the polyketal $[-O-C(CH_3)_2-]_n$, by using metallic or organometallic catalysts. The polymerization of sulfur analogs, for example, trithiane (the analog of trioxane), hexafluorothioacetone, and thiocarbonyl fluoride is also possible; these monomers are reactive to free-radical as well as ionic polymerization. *See* FREE RADICAL.

The successful synthesis of high-molecular-weight polyoxethylene on a commercial scale has been achieved after considerable earlier research on the synthesis of fiber-forming polyacetals by the classical reaction of formaldehyde (in the form of paraformaldehyde) with aliphatic diol, as in reaction (2), where R must contain more than five car-

$$nH - C - H + nHO - R - OH \longrightarrow$$

$$-[O - R - O - CH_2]_n + nH_2O \qquad (2)$$

bons to avoid cyclization. An interchange reaction between a diol and an acetal yields a similar structure. The relatively low melting points obtained (60° C or 140°F) can be increased by use of a cycloaliphatic diol, for example, 1,4-cyclohexanediol.

The versatility of polyacetal formation is also evident in the polymerization of diverse monomers such as cyclic formals, as in reaction (3). Various

$$n (CH_2)_m CH_2 \longrightarrow -\{OCH_2 - O - (CH_2)_m\}_n (3)$$

$$Cyclic \text{ formal} \qquad (especially when m is 5,7,8)$$

spiroacetals have been based on reactions of aldehydes or ketones with pentaerythritol, for example, reaction (4). Some of these polymers are unique in

that expansion, rather than contraction, occurs on polymerization. *See* POLYMER; POLYMERIZATION.

Properties and uses. Polyacetals are typically strong and tough, resistant to fatigue, creep, organic chemicals (but not strong acids or bases), and have low coefficients of friction. Electrical properties are also good. Improved properties for particular applications may be attained by reinforcement with fibers of glass or polytetrafluoroethylene, and by incorporation of an elastomeric toughening phase. The combination of properties has led to many uses such as plumbing fittings, pump and valve components, bearings and gears, computer hardware, automobile body parts, and appliance housings. John A. Manson

Bibliography. J. A. Brydson, *Plastics Materials*, 7th ed., 1999; G. W. Ehrenstein and R. P. Theriault,

Polymeric Materials: Structure, Properties, Applications, 2000; G. Odian, *Principles of Polymerization,* 4th ed., 2004; A. B. Strong, *Plastics: Materials and Processing,* 3d ed., 2005.

Polyacrylate resin

Polymers obtained from a variety of acrylic monomers, such as acrylic and methacrylic acids, their salts, esters, and amides, and the corresponding nitriles. The most important monomers are:

 $\begin{array}{ccc} CH_3 & H\\ H_2C = C - COOCH_3 & H_2C = C - COOC_2H_5\\ Methyl methacrylate & Ethyl acrylate\\ H\\ H_2C = C - CN\\ Acrylonitrile \end{array}$

Poly(methyl methacrylate), poly(ethyl acrylate), and a few other derivatives are discussed in this article. *See* POLYACRYLONITRILE RESINS.

Poly(methyl methacrylate) is a hard, transparent polymer with high optical clarity, high refractive index, and good resistance to the effects of light and aging. It and its copolymers are useful for lenses, signs, indirect lighting fixtures, transparent domes and skylights, dentures, and protective coatings.

The monomer is usually made by the dehydration and methanolysis of acetone cyanohydrin as in reaction (1). Polymerization may be initiated

(1)

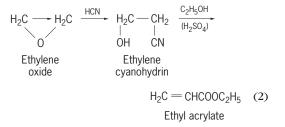
Methyl

methacrylate by free-radical catalysts such as peroxides, or by organometallic compounds such as butyl lithium. The free-radical polymerization, which is used for commercial production, can be carried out in bulk, in solution, and in aqueous emulsion or suspension. By changing the catalyst, temperature, and solvent in a polymerization it is possible to prepare polymers which are predominantly isotactic, syndiotactic, or atactic, or which have stereoregular blocks. Although solution polymerization is not commonly used, bulk polymerization, usually with comonomers, is frequently employed in various casting operations, as in the formation of sheets, rods, and tubes; in the mounting of biological, textile, and metallurgical test specimens; and in dental applications. A 20% reduction in volume accompanies the conversion of monomer to polymer. This makeit difficult to prepare articles to predetermined dimensions by the cast polymerization technique. The problem is minimized by using a syrup of polymer dissolved in monomer, or a dough of finely divided polymer dispersed in a relatively small amount of monomer.

Molding powders suitable for the injection molding of dials, ornamental fixtures, and lenses may be prepared from the granules produced by aqueous suspension polymerization or from the product of bulk polymerization.

Solutions of poly(methyl methacrylate) and its copolymers are useful as lacquers. Aqueous latexes formed by the emulsion polymerization of methyl methacrylate with other monomers are useful as water-based paints and in the treating of textiles and leather. *See* EMULSION POLYMERIZATION.

Poly(ethyl acrylate) is a tough, somewhat rubbery product. The monomer is used mainly as a plasticizing or softening component of copolymers. Ethyl acrylate is usually produced by the dehydration and ethanolysis of ethylene cyanohydrin, which can be obtained from ethylene oxide as shown by reaction (2).



Polymerization may be affected by catalysts of the free-radical type. Copolymerizations with other monomers are frequently carried out in aqueous emulsion or suspension.

Modified acrylic resins with high impact strengths can be prepared. Blends or "alloys" with poly(vinyl chloride) are used for thermoforming impactresistant sheets. Copolymers of ethyl acrylate with 2-chloro-ethyl vinyl ether are useful rubbers which can be vulcanized in the same way as ordinary polydienes. *See* COPOLYMER.

Methyl methacrylate is of interest as a polymerizable binder for sand or other aggregates, and as a polymerizable impregnant for concrete; usually a cross-linking acrylic monomer is also incorporated. The binder systems (polymer concrete) are used as overlays for bridge decks as well as for castings, while impregnation is used to restore concrete structures and protect bridge decks against corrosion by deicing salts.

Other acrylic derivatives are the subject of continuing interest. The butyl and octyl esters of acrylic acid yield rubbery materials, and fluorinated esters have been used as oil-resistant rubbers. Polymers of alkyl-2-cyano-acrylates make excellent adhesives that cure at low temperatures using adventitious moisture as a catalyst. Cross-linkable acrylic copolymers are often used in surface coatings. Polymers of methyl acrylate or acrylamide are water-soluble and useful for sizes and finishes. Addition of poly(lauryl methacrylate) to petroleum lubricating oil improves the flowing properties of the oil at low temperatures and the resistance to thinning at high temperatures. *See* ACRYLONITRILE; PLASTICS PROCESSING; POLYMER-IZATION; RUBBER. John A. Manson

Bibliography. H. Allcock, F. Lampe, and J. Mark, *Con*temporary Polymer Chemistry, 3d ed., 2003; R. G. Gilbert, *Emulsion Polymerization: A Mechanistic Approach*, 1995; G. Odian, *Principles of Polymerization*, 4th ed., 2004; M. P. Stevens, *Polymer Chemistry: An Introduction*, 3d ed., 1998.

Polyacrylonitrile resins

The polyacrylonitrile resins are hard, relatively insoluble, and high-melting materials. Polyacrylonitrile [poly(vinyl cyanide)] is used almost entirely in copolymers. The copolymers fall into three groups: fibers, plastics, and rubbers. The presence of acrylonitrile in a polymeric composition tends to increase its resistance to temperature, chemicals, impact, and flexing. *See* COPOLYMER.

Acrylonitrile is generally prepared by one of the following methods.

Catalyzed addition of hydrogen cyanide to acetylene as in reaction (1):

$$CH \equiv CH + HCN \xrightarrow{catalyst} H_2C = CH \qquad (1)$$

Reaction between hydrogen cyanide and ethylene oxide as in (2):

$$H_2C \xrightarrow{CH_2 + HCN} \xrightarrow{H_2C} = CH_2 + HOH \quad (2)$$

Reaction between ammonia and propylene as in (3):

$$2H_{3}C - CH = CH_{2} + 2NH_{3} + 3O_{2} \longrightarrow$$

$$2H_{2}C = CH + 6H_{2}O \quad (3)$$

$$| CN$$

The polymerization of acrylonitrile can be readily initiated by means of the conventional free-radical catalysts such as peroxides, by irradiation, or by the use of alkali metal catalysts. Although polymerization in bulk proceeds too rapidly to be commercially feasible, satisfactory control of a polymerization or copolymerization may be achieved in suspension and in emulsion, and in aqueous solutions from which the polymer precipitates. Copolymers containing acrylonitrile may be fabricated in the manner of thermoplastic resins.

The major use of acrylonitrile is in the form of fibers. By definition an acrylic fiber must contain at least 85% acrylonitrile; a modacrylic fiber may contain less, from 35 to 85% acrylonitrile. The high strength; high softening temperature; resistance to aging, chemicals, water, and cleaning solvents; and the soft woollike feel or fabrics have made the product popular for many uses such as sails, cordage, blankets, and various types of clothing. Commercial forms of the fiber probably are copolymers containing minor amounts of other vinyl derivatives, such as vinyl pyrrolidone, vinyl acetate, maleic anhydride, or acrylamide. The comonomers are included to produce specific effects, such as improvement of dyeing qualities. *See* MANUFACTURED FIBER.

Copolymers of vinylidene chloride with small proportions of acrylonitrile are useful as tough, impermeable, and heat-sealable packaging films.

Extensive use is made of copolymers of acrylonitrile with butadiene, often called NBR (formerly Buna N) rubbers, which contain 15-40% acrylonitrile. Minor amounts of other unsaturated esters, such as ethyl acrylate, which yield carboxyl groups on hydrolysis, may be incorporated to improve the curing properties. The NBR rubbers resist hydrocarbon solvents such as gasoline, abrasion, and in some cases show high flexibility at low temperatures. *See* RUB-BER.

In the 1960s development of blends and interpolymers of acrylonitrile-containing resins and rubbers represented a significant advance in polymer technology. The products, usually called ABS resins, typically are made by blending acrylonitrile-styrene copolymers with a butadiene-acrylonitrile rubber, or by interpolymerizing polybutadiene with styrene and acrylonitrile. Specific properties depend upon the proportions of the comonomer, on the degree of graftings, and on molecular weight. In general, the ABS resins combine the advantages of hardness and strength of the vinyl resin component with toughness and impact resistance of the rubbery component. Certain grades of ABS resin are used for blending with brittle thermoplastic resins such as poly(vinyl chloride) in order to improve impact strength.

The combination of low cost, good mechanical properties, and ease of fabrication by a variety of methods, including typical metalworking methods such as cold stamping, led to the rapid development of new uses for ABS resins. Applications include products requiring high impact strength, such as pipe, and sheets for structural uses, such as industrial duct work and components of automobile bodies. ABS resins are also used for housewares and appliances and, because of their ability to be electroplated, for decorative items in general. *See* ACRYLONITRILE; PLASTICS PROCESSING; POLYMER; STYRENE. John A. Manson

Bibliography. H. Allcock, F. Lampe, and J. Mark, *Contemporary Polymer Chemistry*, 3d ed., 2003; A. W. Birley, R. H. Heath, and M. J. Scott, *Plastics Materials: Properties and Applications*, 2d ed., 1988; J. Brydson, *Plastics Materials*, 7th ed., 1999; G. Odian, *Principles of Polymerization*, 4th ed., 2004.

Polyamide resins

The product of polymerization of amino acids or the condensation of a diamine with a dicarboxylic acid. Polyamides are a class of polymers in which individual structral units are joined by amide bonds (R—CO—NH₂). Whitish, translucent, and high-melting polyamide resins can be essentially transparent and amorphous when their melts are quenched. On cold drawing and annealing, most become quite crystalline and translucent. However, some polyamides based on bulky repeating units are inherently amorphous. The polymers are used for fibers, bristles, bearings, gears, molded objects, coatings, and adhesives. The term nylon formerly referred specifically to synthetic polyamides as a class. Because of many applications in mechanical engineering, nylons are considered engineering plastics. *See* MANUFACTURED FIBER.

Brief outlines of the preparations of commercial polyamides by the reaction of dicarboxylic acids with diamines and the condensation of amino acids are given in this article. The most common commercial aliphatic polyamides are nylons-6,6; -6; -6,10; -11; and -12. Other types find specialized application.

Nylon-6,6 and nylon-6,10 are products of the condensation reaction of hexamethylenediamine (6 carbon atoms) with adipic acid (6 carbon atoms), and with sebacic acid (10 carbon atoms), respectively. By heating equimolar proportions of the two reactants, a polymeric salt is formed, as in reaction (1),

$$nHOOC(CH_{2})_{4}COOH + nH_{2}N(CH_{2})_{6}NH_{2} \longrightarrow$$
Adipic acid Hexamethylene-
diamine
$$[NH_{3}(CH_{2})_{6}NH_{3}OOC(CH_{2})_{4}COO]_{n}$$
"Nylon-6,6 salt"
$$\downarrow^{heat}$$

$$[-HN - (CH_{2})_{6} - NH - OC - (CH_{2})_{4} - CO -]_{n} + 2nH_{2}O$$
(1)

which, on further heating, yields the polyamide resin shown. Because the end groups on the polymer can react on further heating, as in melt spinning, it is desirable to add a very small amount of a monoacid or monoamine to the polymerizing mixture in order to prevent the formation of material of very high molecular weight.

Nylon-6,6, nylon-6,10, nylon-6,12, and nylon -6 are the most commonly used polyamides for general applications as molded or extruded parts; nylon-6,6 and nylon-6 find general application as fibers.

Nylon-6 and nylon-11 can be obtained by the selfcondensation of ϵ -aminocaproic acid and ω -amino undecanoic acids, respectively, as shown by reactions (2) and (3).

$$\begin{array}{ll} H_2 N(CH_2)_5 COOH & \rightarrow \mbox{ nylon-6} & (2) \\ \hline \epsilon - \mbox{Aminocaproic} & & \\ acid & & \\ NH_2 (CH_2)_{10} COOH & \rightarrow \mbox{ nylon-11} & (3) \\ \hline \omega - \mbox{Aminoundecanoic} & & \\ acid & & \\ \end{array}$$

Each molecule, containing both the amino and carboxylic groups, can condense to yield high polymers by reactions similar to those between the diacids and the diamines.

Nylon-6 is usually made by polymerization of the

lactam of ϵ -aminocaproic acid, as in reaction (4).

$$\begin{array}{c|c} H_2C & \hline C = 0 \\ H_2C & NH \\ n & | \\ H_2C & CH_2 \\ C \\ H_2 \\ \hline H_2 \\ C \\ H_2 \end{array} \xrightarrow[(460-540^\circ F)]{trace} \\ catalyst \\ catalyst \\ c40-280^\circ C \\ (460-540^\circ F) \\ H_2 \\ \hline \end{array}$$

$$\begin{array}{c} H \\ (-N - (CH_2)_5 - C -)_n \\ N \\ \text{Nylon-6} \end{array}$$
 (4)

 \cap

Caprolactam (a common term for ϵ -aminocaprolactam) is produced by rearrangement of cyclohexanone oxime, which is prepared by reaction of cyclohexanone with hydroxylamine sulfate.

The lactam process is generally preferred for commercial operation because it is easier to make and purify the lactam than the ϵ -aminocaproic acid. Nylon-12 is prepared by a similar self-condensation of the lactam of aminododecanoic acid, which is prepared by a series of reactions using butadiene as starting material. Lactams may also be polymerized by anionic catalysis. This process makes possible a onestep casting to form large articles, especially from nylon-6.

As a group, nylons are strong and tough. Mechanical properties depend in detail on the degree and distribution of crystallinity, and may be varied by appropriate thermal treatment or by nucleation techniques. Because of their generally good mechanical properties and adaptability to both molding and extrusion, the nylons described above are often used for gears, bearings, and electrical mountings. Nylon bearings and gears perform quietly and need little or no lubrication. Sintering (powder metallurgy) processes are used to make articles such as bearings, and gears which have controlled porosity, thus permitting retention of oils or inks. Nylon resins are used extensively as filaments, bristles, wire insulation, appliance parts, and film. Properties can be modified by copolymerization.

Reinforcement of nylons with glass fibers results in increased stiffness, lower creep and improved resistance to elevated temperatures. Such formulations, which can be readily injection-molded, can often replace metals in certain applications. The use of molybdenum sulfide and polytetrafluoroethylene as fillers increases wear resistance considerably. For uses requiring impact resistance, nylons can be blended with a second, toughening phase.

Other types of nylon are useful for specialty applications. Solubility may be increased by interference with the regularity and hence intermolecular packing. This may be accomplished by copolymerization, or by the introduction of branches on the amide nitrogen, for example, by treatment with formaldehyde. The latter type of resin may be subsequently cross-linked. Nylons incorporating aromatic structures, for example, based on isophthalic acid, are becoming more common for applications requiring resistance to very high temperatures. Aromatic-nylon fibers based on all-para-substituted monomers can be spun from liquid-crystal suspensions to yield fibers that have exceptional moduli and strengths up to approximately 25,000,000 lb/in.² (160 gigapascals) and 400,000 lb/in.² (2.8 GPa), respectively, in the fiber direction. These values may be compared with theoretical moduli and strengths of approximately 40,000,000 lb/in.² (270 GPa) and approximately 1,000,000-4,000,000 lb/in.² (6.9-27 GPa), respectively.

Heterocyclic amide-imide copolymers are also available commercially, and a wide variety of other amide-containing polymers, such as polyesteramides and polysulfonamides, can be synthesized.

The unsaturated fatty acids in vegetable oil, for example, linoleic acid, may dimerize or polymerize to low polymers through their unsaturated groups. The di- or polycarboxylic acids obtained yield polyamides by condensation with di- or polyamines. These products are employed as coatings and adhesives, especially in epoxy and phenolic resin formulations. *See* HETEROCYCLIC POLYMER; PLASTICS PROCESSING; POLYETHER RESINS; POLYMERIZATION. John A. Manson

Bibliography. H. Allcock, F. Lampe, and J. Mark, *Contemporary Polymer Chemistry*, 3d ed., 2003; J. Brydson, *Plastics Materials*, 7th ed., 1999; R. Puffr and V. Kubanek, *Lactam-based Polyamides*, vols. I and II, 1991; G. Odian, *Principles of Polymerization*, 4th ed., 2004.

Polychaeta

The largest class of the phylum Annelida (multisegmented worms) with about 13,000 described species. Polychaetes (**Fig. 1**) range from less than

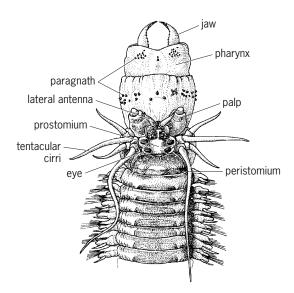


Fig. 2. *Platynereis dumerilii antipoda*, showing the antero-dorsal end with the pharynx everted. (*After P. L. Beesley et al., 2000*)

1 mm (0.04 in.) to up to 3 m (10 ft) in length. Represented worldwide, they are predominantly marine organisms and occur from the intertidal zone to the greatest ocean depths at all latitudes. Some notable exceptions occupy nonmarine habitats such as supralittoral zones (beyond the high-tide mark) and freshwater lakes. The oldest known entire fossilized species dates back to the mid-Cambrian, but records on trace fossils suggest that the origin of the group predates this period. The only synapomorphy (a derived trait shared by two or more taxa that is believed to reflect their shared ancestry) that distinguishes them from other annelids is the pair of nuchal organs, chemosensory structures

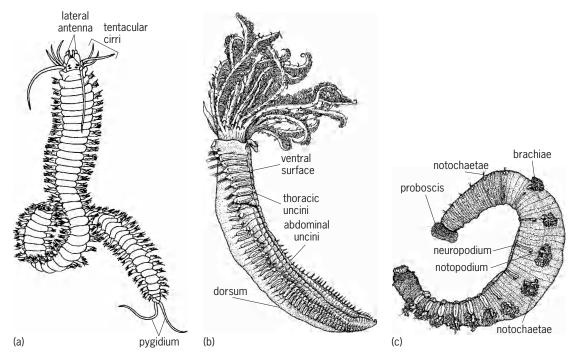


Fig. 1. Examples of polychaetes. (a) Platynereis dumerilii antipoda. (b) Branchiomma sp. (c) Arenicola bombayensi. (After P. L. Beesley et al., 2000)

located on the prostomium (the portion of the head anterior to the mouth). *See* ANNELIDA; DEEP-SEA FAUNA.

External morphology. As in other annelids, the polychaete body consists of two presegmental regions (the prostomium and the peristomium), a segmented trunk, and a postsegmental pygidium (tail region) [**Fig. 2**]. Often the prostomium and the peristomium are fused with the first anterior segments of the trunk, known as the cephalized segments, forming the head of the polychaete. The prostomium varies greatly in shape, from a simple conical structure to a simple T-shape, and may be fused to the peristomium. In many polychaetes, the only peristomial structure visible is the area surrounding the mouth; in a few others, the peristomium is seen as one complete ring separating the prostomium from the first segment.

Head appendages, which are present or absent according to the species considered, include antennae, palps (various sensory, usually fleshy appendages near the oral aperture), peristomial cirri (slender hairlike appendages), and tentacular cirri. Antennae (up to three) are always located on the prostomium. Palps emerge from the prostomium or the peristomium. They are structurally divided into grooved feeding palps and tapering sensory palps. A single pair of grooved palps is present in many polychaetes. In sabellids and serpulids, the grooved palps form a tentacular crown, with each tentacle being made of a long median radiole (spine) and numerous lateral pinnules (branchlets arranged in plumelike rows). Sensory palps are generally a pair of lateral, slender, tapering structures. Peristomial cirri are attached dorsolaterally at the level of the peristomium, while tentacular cirri emerge from the cephalized segments [Figs. 3 and 4].

The trunk of a polychaete comprises serially repeated segments, also called metameres, numbering from few to many. Segments may be similar to one another (homonomous metamerism) or different (heteronomous metamerism), resulting in the latter case in the elaboration of various trunk re-

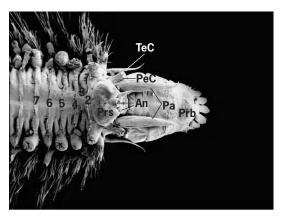


Fig. 3. Hololepidella laingensis; scanning electron microscope view of the anterior end with the pharynx everted. An: antennae; Pa: palps; Prb: proboscis; Prs: prostomium; TeC: tentacular cirrus. Numbers refer to the segments of the trunk. $50 \times .$

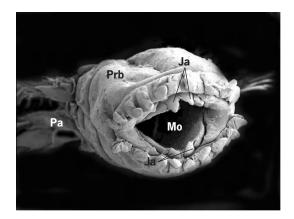


Fig. 4. Gastrolepidia clavigera; scanning electron microscope view of the anterior end. Mo: mouth; Ja: jaw; Pa: palps; Prb: proboscis. $30 \times$.

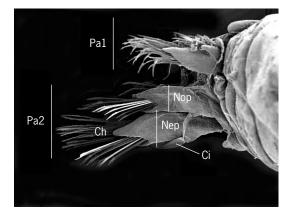


Fig. 5. Gastrolepidia clavigera, scanning electron microscope view of the two first parapodia (Pa1, Pa2). Ch: chaetae; Ci: cirrus; Nop: notopodium; Nep: neuropodium. 40 \times .

gions (thoracic, abdominal, caudal). Each segment basically bears a pair of lateral appendages, the parapodia (**Fig. 5**). A parapodium is equipped with musculature, and bears chaetae (setae) in most taxa. Parapodial structures vary, but they often consist of two rami (branches), a dorsal notopodium and a ventral neuropodium. In many polychaetes, the two rami differ in size, structure, and orientation. Dorsal and ventral cirri are distinctive parapodial features. Cirri are generally slender tapering structures, but in scaleworms some of the dorsal cirri are present as elytra (scales) that are held horizontally above the dorsal surface.

Chaetae are epidermal chitin-sclerotized (hardened) proteinaceous rods (like bristles) sometimes impregnated with calcium carbonate. Chaetae occur in bundles and emerge at the tip of neuropodia and notopodia. They are connected to internal muscles and can be retracted in the parapodium or protruded into the medium. They act in anchoring the parapodium to the substrate during locomotion. Many polychaetes have internal chaetae, the aciculae, which form the internal skeleton or supporting rods for the parapodia. Spines and hooks are thicker chaetae present in some polychaetes. The shape of the chaetae is highly diversified and used in specific diagnosis. **Internal morphology and physiology.** Beneath the epidermis of polychaetes lies a nearly continuous sheath of circular muscle that is internally doubled by thick longitudinal muscles (four bands in most species). In addition, series of oblique muscles join the ventral area of the trunk to the midlateral regions. The innermost body wall component is a peritoneal layer lining the coelom. The coelom is segmented, and in most polychaetes, septa (partitions) separate internally adjacent metameres (**Fig. 6**).

The digestive tract is essentially a simple tube supported by a dorsal mesentery and partial-to-complete septa. The anterior part of the digestive tract is often differentiated in a buccal (mouth) organ that is axial or ventral, well developed or simple. The muscular region of the buccal organ is called the pharynx; the eversible nonmuscular region is often referred to as the proboscis. Feeding methods in polychaetes are diverse. Polychaetaes are raptorial (living on prey) feeders and use their buccal apparatus to seize food items; nonselective or selective deposit feeders; or filter feeders. The proboscis of raptorial feeders is well developed and often bears chitinous structures, the paragnaths (small, sharp jawlike structures) and jaws.

The great majority of polychaetes have a closed circulatory system made of medial dorsal and ventral longitudinal vessels linked by smaller capillaries and gut lacunae (cavities). Some taxa have more contractile hearts. Many species have elaborate branchiae (gills) developed as extensions of the body wall containing a loop of the vascular system and capillaries. Blood generally does not contain cells but respiratory pigments such as hemoglobins and chlorocruorins (green substances that are the probable cause of the color of the blood in some worms). Waste products of metabolism are excreted via nephridia (generally metanephridia). The distribution of nephridia varies, but most commonly nephridia are present as a pair per segment along most of the trunk. See EXCRE-TION.

The nervous system of polychaetes is composed of a brain, or cerebral ganglia, situated in the prostomium. The brain is dorsal and linked to a ventral nerve cord by two circumesophageal connectives. The ventral nerve cord is usually made up of a coupled pair of cords running the length of the trunk. Polychaetes have several sensory structures that include palps, antennae, cirri, eyes, statocysts, nuchal organs, and lateral organs. Statocysts are gravity receptors numbering from one pair to more than 20 pairs. They often appear as simple pits at the surface of the organisms. Nuchal organs are paired structures innervated from the dorsal part of the brain. They are most likely chemoreceptors, probably like the lateral organs that are present segmentally between the notopodia and the neuropodia. Most polychaetes have eyes or photoreceptors located on the prostomium. These structures range from simple pigmented cups with receptor cells to complex compound eyes analogous to those of arthropods. In certain families, eyes may be capable of forming images, but they are more probably structures

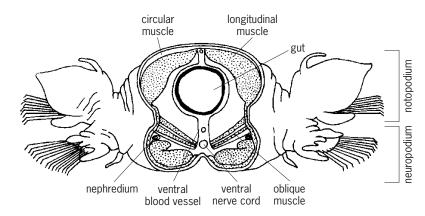


Fig. 6. Section through a polychaete. (After E. E. Ruppert and R. D. Barnes, Invertebrate Zoology, Saunders College Publishing, Orlando, 1994)

integrating light intensity and direction.

Gonochorism (separation of the sexes) is the most common sexual condition, although hermaphrodism (the presence of both male and female reproductive organs) is widespread. Most polychaetes have no permanent gonads. At some period of the year, spermatozoa and ova are differentiated from germ cells located in the peritoneum and are discharged into the coelom before being expulsed to the water medium through specialized gonoducts or nephridia (in a few cases by rupture of the body wall).

Life cycle and reproduction. Life cycle is either semelparous or iteroparous. Semelparity refers to breeding occurring once per lifetime. Iteroparity, where breeding occurs several times in a lifetime, can be divided into annual iteroparity (breeding yearly) and continuous iteroparity (breeding over an extended season). In most polychaetes, fertilization is external and the development gives rise to a pelagic trochophore (free-swimming) larva. Some obvious features of the larvae are the ciliary bands that are used for locomotion and feeding. After the pelagic larval stage, larvae of most species sink to the sea bottom and metamorphose into benthic juveniles that grow into adults. In a few cases, ova may be deposited in gelatinous or mucous balls or strands, sometimes covering the beaches during the spawning season. Ova may be attached to parapodia or laid in tubes, where brood care is displayed. Some onuphids construct special ovicells in which the young develop to a stage resembling adults. Special thoracic pouches or opercular chambers serve as ovaria in some serpulids.

Asexual reproduction is widespread and involves paratomy or architomy. Paratomy refers to the formation of a recognizable complete individual that separates from the parent piece; architomy (far more widespread) is simple fission or fragmentation of the body without prior cephalization (anterior specialization resulting in the concentration of sensory and neural organs in the head).

Synchronous spawning is widespread in polychaetes. In addition to temporal synchrony, many benthic species swarm to the surface of the water column to spawn. These species often undergo drastic morphological changes during swarming that are called epitoky. One case of epitoky involves the Palolo worm, *Eunice viridis*, from the Samoa Islands, in which the worm separates from its posterior part during swarming.

Ecology. In marine habitats, polychaetes may burrow into the muddy or sandy substrates in both intertidal and subtidal areas. Some live naked in the sediments, burrowing freely into them; others live in tubes that range from semipermanent mucous types to well-constructed firm tubes of fine sediments. Tubicolous species are fairly immobile within sediments. Many others live under rocks and crevices, either free-living or with cemented tubes. Polychaetes such as the Christmas tree worm, *Spirobranchus giganteus*, are also found in live corals and can exhibit a wide range of colors even on a single coral.

Members of a few families are parasites. Some species actively bore into calcareous substrata: Boccardia and Polydora species are parasites of oysters and other bivalves, causing blisters to develop in the shell. All histriobdellids occur in the branchial chambers of marine and freshwater crustaceans. A species of ichthyomids are parasites found in the fins of eels. Several polychaetes are commensals (commensalism is a symbiotic relationship in which two different species are associated, wherein one is benefited and the other neither benefited nor harmed), especially polynoids that live in the tubes of terebellids or are found in association with holothuroids, asteroids, ophiuroids, antipatharian corals, or sponges. Some hesionids and flabelligerids are also found in echinoids.

Some nonmarine polychaetes, especially nereids, live in the supratidal zone. Others live in freshwater lakes and streams. Some ampharetids inhabit the inland waterways of central Europe, especially the Danube River and its tributaries. *Manayunkia speciosa* (sabellid) is frequent in the Great Lakes of the northeastern United States and their tributaries. Another species of *Manayunkia* has been observed in hypersaline ephemeral lakes (short-lived lakes which have no inlets/outlets and in which the water evaporates, is absorbed, or is consumed) in the wheat belt of Western Australia.

Geographic distribution. Polychaetes are distributed in all marine habitats and show remarkable specificity according to latitudes, depths, and kinds of substrata. Until recently, many polychaetes were regarded as being cosmopolitan (having worldwide distribution), and many species originally described from Arctic regions were also reported from Antarctica. Similarly, many species in Europe and in Australia were initially thought to be identical, but deeper analyses showed that many Australian species were new, suggesting that cosmopolitanism is less extended in polychaetes than previously thought.

In temperate intertidal zones, syllids, nereids, and polynoids predominate, whereas tropical intertidal zones have large numbers of amphinomids, eunicids, serpulids, and terebellids. Preponderant numbers of maldanids, spionids, and cirratulids are found in soft sediments in moderate depths, whereas abyssal depths of more than 6600 ft (2000 m) yield large numbers of ampharetids, flabelligerids, and other tubicolous kins. The diversity continues equally great from shallowest to greatest depths, but unit members diminish from littoral and upper slope, to deep slope and abyssal depths.

Any major zoogeographic area may expect to support from 350 to 800 species of polychaetes, coming from shallowest to greatest depth and in a variety of sediments. The number of endemic species may range from 25 to 40% of the total number; another 30% may be more widely distributed in adjacent areas; and 20 to 30% may be cosmopolitan. These percentages differ with latitude and place. The number of cosmopolitan species is lowest in high or polar latitudes and highest in warm to tropical latitudes, but such figures reflect the greater knowledge of tropical latitudes.

Economic significance. Some polychaetes are economically important to shipping. The calcareous tube-building *Hydroides norvegica* covers the bottoms of ships with thick layers of its tubes. Polychaetes are also used as bait by amateur anglers all over the world. In Britain, *Neanthes virens* is farmed for air-freighting around northern Europe.

With a few exceptions, polychaetes are harmless for humans, but the chaetae of amphinomids can injury people handling them. Some people develop local allergic reaction to these worms, called fireworms. Some other polychaetes can bite.

Classification. In the first half of the twentieth century, the taxon Polychaeta consisting of Archiannelida, Errantia, and Sedentaria was widely accepted. This system was used until recently, but various phylogenetic analyses made during the 1980s and 1990s provided evidence that the three groups are not reliable natural entities. Today, Polychaeta should be divided into two monophyletic groups, Scolecida and Palpata, with 81 accepted families. Palpata, which includes 72 families, are named for the presence of palps on the head. Palpata comprise two major clades, Canalipalpata and Aciculata. Aciculata are named for the presence of aciculae supporting welldeveloped parapodia. They include the Phyllodocida (27 families), Amphinomida (2 families), and Eunicida (8 families). Canalipalpata have grooved palps, usually peristomial in origin. They include the Sabellida (4 families), Spionida (8 families), and Terebellida (12 families). The Scolecida do not have palps but do have parapodia with similar-sized rami and two or more pairs of pygidial cirri; they include 9 families. Eleven other families are incertae sedis (of uncertain taxonomic position). Igor Eeckhaut; Olga Hartman

Bibliography. T. Bartolomaeus and G. Pürschke, Morphology, Molecules, Evolution, and Phylogeny in Polychaeta and Related Taxa, Springer, Dordrecht, 2005; P. L. Beesley, G. J. B. Ross, and C. J. Glasby, Polychaetes and Allies: The Southern Synthesis—Fauna of Australia, Polychaeta, Myzostomida, Pogonophora, Echiura, Sipuncula, CSIRO Publishing, Melbourne, 2000; G. W. Rouse and K. Fauchald, Cladistics and polychaetes, Zool. Scripta, 26:139-204, 1997; G. W. Rouse and F. Pleijel, Polychaetes, Oxford University Press, New York, 2002.

Polychlorinated biphenyls

A generic term for a family of 209 chlorinated isomers of biphenyl. The biphenyl molecule is composed of two six-sided carbon rings connected at one carbon site on each ring. Ten sites remain for chlorine atoms to join the biphenyl molecule. While the rules of nomenclature would indicate that the term polychlorinated biphenyl can refer only to molecules containing more than one chlorine atom, the term has been used to refer to the biphenyl molecule with one to ten chlorine substitutions.

Polychlorinated biphenyls (PCBs) were discovered in the late nineteenth century, were introduced into United States industry on a large scale in 1929, and have been in use since that time in most industrial nations. The qualities that made PCBs attractive to industry were chemical stability, resistance to heat, low flammability, and high dielectric constant. The PCB mixture is a colorless, viscous fluid, is relatively insoluble in water, and can withstand high temperatures without degradation. However, these characteristics are precisely the qualities that make polychlorinated biphenyl persistent in the environment; mono- and dichlorobiphenyl have halflives in sunlight of about 6 days, but the higherchlorinated isomers are not readily degraded in the environment.

Uses. The major use of PCBs has been as dielectric fluid in electrical equipment, particularly transformers (including transformers used on railway locomotives and self-propelled cars), capacitors, electromagnets, circuit breakers, voltage regulators, and switches. Electrical equipment containing PCBs is no longer produced in the United States, but in 1981 40% of all electrical equipment in use in the United States contained PCBs. PCBs have also been used in heat transfer systems and hydraulic systems, and as plasticizers and additives in lubricating and cutting oils. *See* DIELECTRIC MATERIALS.

Environmental distribution. PCBs have been reported in animals, plants, soil, and water all over the world. The fact that PCBs have been detected in the polar ice caps is an indication of the power of atmospheric transport as a major pathway of global contamination. PCBs are also transported through the biosphere in water and attached to sediments. It is not known what quantities of PCBs have been released to the environment, but major sources are industrial and municipal waste disposal, spills and leaks from PCB-containing equipment, and manufacture and handling of PCB mixtures. *See* ATMOSPHERIC GENERAL CIRCULATION; BIOSPHERE.

PCBs have been reported in fishes, eels, and benthic organisms in the waters of the North Sea near the Netherlands. Fish products from United States waters and the Baltic Sea have been found to contain several parts per million of PCB and have been declared unfit for human consumption. PCBs have been found in all organisms analyzed from the North and South Atlantic, even in animals living under 11,000 ft (3400 m) of water. These phenomena are the result of bioaccumulation and biomagnification in the food chain. Uptake of PCBs by plants has also been reported, but it is generally small and does not directly contribute to substantial human contamination. In a few instances, poultry products, cattle, and hogs have been found to contain high concentrations of PCBs after the animals have eaten feed contaminated with PCBs. *See* FOOD WEB.

Following the initial restriction of PCB production in the 1970s, contamination levels dropped rapidly in the Great Lakes. Since 1979, levels in the Great Lakes have ceased to decline, probably because of atmospheric transport, contaminated ground water, spills and leaks from authorized equipment, continued improper disposal, and the fact that the lowerchlorinated species degraded first, leaving the recalcitrant higher-chlorinated biphenyls. *See* EUTROPHI-CATION.

PCBs have been detected in human adipose tissues and in the milk of cows and humans. The estimated percentage of the United States population with detectable levels of PCBs was nearly 100% in 1981. The estimated percentage of the United States population with greater than 3 ppm PCBs in their tissue was 2.7% in 1972, and less than 1% in 1981. *See* HUMAN ECOLOGY.

PCBs have also been found as low-level, inadvertently generated contaminants in a wide range of chemical products such as certain paints, inks, plastics, and paper coatings. It has been estimated that fewer than 100,000 lb (45,000 kg) of PCBs are inadvertently generated in chemical processes each year in the United States.

Toxicity. PCBs can enter the body through the lungs, gastrointestinal tract, and skin, circulate throughout the body, and be stored in adipose tissue. Except for occupational contact, human exposure is mainly through food. PCBs can cause chloracne, skin discoloration, liver dysfunction, reproductive effects, development toxicity, and oncogenicity in exposed humans. Some PCBs have the ability to alter reproductive processes in mammals. Prenatal exposure in animals can result in various degrees of developmentally toxic effects. There is concern, based on extrapolation from animal studies, that PCBs may be carcinogenic in humans. The problem is complicated by the presence, in most PCB mixtures, of toxic impurities.

Disposal. Incineration of solid and liquid PCB materials is a common and highly efficient method of destruction. While laboratory tests have demonstrated PCB destruction at about 1470° F (800° C), the commercial incinerators operate at temperatures in excess of 1800° F (1000° C). High-efficiency destruction has been reported in large stationary incinerators and boilers, cement kilns, incinerator ships, and smaller mobile thermal destruction units.

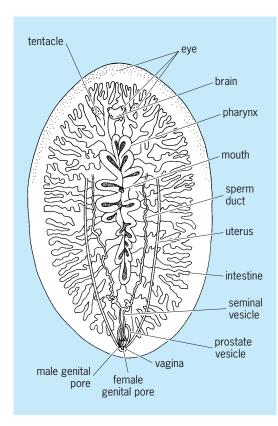
PCB-contaminated fluids have been decontaminated by using chemical reagents to attack the chlorinated molecule. Generally, a metal, such as sodium, is used to remove the chlorine atoms from the biphenyl molecule. Some research has been conducted toward chemical treatment of PCB-contaminated soil by spraying the reagent on soil; the results and environmental consequences are unclear. Microbial degradation of PCB-contaminated soils has been pursued by several investigators, but there is little evidence of commercial-scale success. *See* ECOLOGY; HAZARDOUS WASTE. Glenn Kuntz

Bibliography. M. Barros et al. (ed.), *PCB Seminar Proceedings*, Organization for Economic Cooperation and Development, 1983; W. Bouchard, PCBs in perspective, *Hazard. Mater. Manag. J.*, 1980; O. Hutzinger et al., *Chemistry of PCBs*, 1974, reprint 1983; A. Leifer et al., *Environmental Transport and Transformation of PCBs*, USEPA 560/5-83-025, 1983; R. Lucas et al., *PCBs in Human Adipose Tissue and Mother's Milk*, USEPA Contract 68-01-5848, 1982; S. Safe, *Polychlorinated Bipbenyls* (*PCBs): Mammalian and Environmental Toxicology*, 1987.

Polycladida

An order of marine Turbellaria which are several millimeters to several centimeters in length and whose leaflike bodies have a central intestine with radiating branches. Most of these platyhelminth species live in the littoral zone on the bottom, on seaweed or on other objects, or as commensals in the shells of mollusks and hermit crabs. None are parasitic. Except in warm waters, they are seldom brightly colored.

Usually they have many eyes, and tentacles are frequently present. Frontal organs and statocysts are absent and adhesive organs are rare. The epidermis



Stylochus ellipticus.

is covered with cilia and contains numerous rhabdite glands. Near the middle of the body is the mouth, followed by the plicate pharynx which opens into the central cavity of the intestine. The brain lies anterior to the pharynx, with a number of nerves radiating from it, but the two nerves which parallel the pharynx are usually the largest. Ovaries and testes are numerous and scattered, but yolk glands are lacking. Sperm ducts connect the testes with the copulatory apparatus which is variable in structure and often multiple rather than single. Insemination occurs through copulation or hypodermic impregnation. The entolecithal eggs usually accumulate in the oviducts or uteri and after fertilization pass to the exterior through the vagina. Lang's vesicle, a bursa with a long stalk, is generally present. Müller's larva, a free larval stage known in the Turbellaria, is found in some polyclads, but is lacking in most. This larva may indicate an evolutionary link between the Turbellaria and the Ctenophora, and between the Turbellaria and the Annelida.

Notoplana and *Stylochus* (see **illus**.) are two of the largest and best-known genera of polyclads with representatives from both coasts of North America. In particular, species of *Stylochus* are often large forms 2 in. (5 cm) or more in length. *See* TURBEL-LARIA. E. Ruffin Jones Bibliography. IL C. Brong. (ad.). Klassen und Ord

Bibliography. H. G. Bronn (ed.), *Klassen und Ord*nungen des Tierreichs, vol. 4, 1930–1933.

Polyester resins

A large class of polymeric materials in which ester groups,

0 ∥ −C−0

are in the main chains. Generally, polyesters are prepared by the condensation reactions of glycols (or dialcohols) with dicarboxylic acids, and they range from relatively soft aliphatics to aromatic derivatives which are considered engineering materials. Their properties may be modified by cross-linking, crystallization, plasticizers, or fillers. *See* ESTER.

Commercial products include alkyds, which are used in paints and enamels and in molding compounds; unsaturated polyesters or unsaturated alkyds, which are used extensively with fiberglass for boat hulls and panels; aliphatic saturated polyesters; aromatic polyesters such as poly(ethylene terephthalate), which is used in the form of fibers and films, in blow-molded bottles, and in injection-moldable resins such as poly(butylene terephthalate); and aromatic polycarbonates.

The poly(diallyl esters) are frequently listed with the polyesters; however, they are not true polyesters as defined above.

Alkyds. The alkyds have been in common use as coatings since World War I. They were first commercialized by the General Electric Company as Glyptal, which became a common name. In the beginning, they consisted almost entirely of the reactions

products of *o*-phthalic anhydride and glycerol, and pigment. Because the functionality of the system is greater than two, a cross-linked insoluble polymer is formed as in reaction (1).

Fully cured alkyds are quite hard and brittle. Flexible and tough materials can be produced by incorporating monobasic acids or monohydroxy alcohols in proportions sufficient to increase flexibility but insufficient to prevent curing. Combinations of conventional vegetable drying oils and alkyd resins represent the basis of most of the oil-soluble paints. For example, by heating a mixture of dehydrated castor oil (the glycerol ester of linoleic acid) with suitable proportions of glycerol and phthalic anhydride, an oil-soluble polyester is formed. A common oil paint is produced by the addition of thinners, such as aromatic hydrocarbon solvents, a paint drier such as cobalt octoate, and pigments. By exposure to air in the presence of the paint dryer, the unsaturated diene groups of the linoleic ester polymerize to yield a tough, weather-resistant coating. See DRYING OIL.

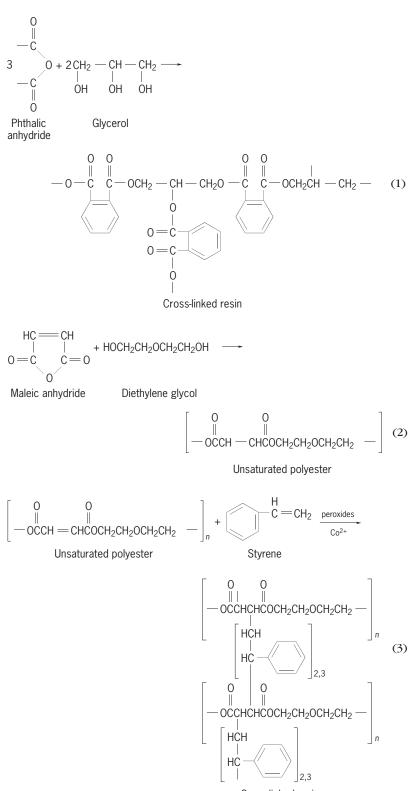
The drying oil-alkyd described above may be further modified by the inclusion of a vinyl monomer, such as styrene, in the original esterification process. Some of the styrene polymerizes, probably as a graft polymer, and the remainder polymerizes and copolymerizes in the final drying or curing of the paint. Low-molecular-weight liquid condensates are useful as polymeric plasticizers for resins such as poly(vinyl chloride). Molding components are also used for such applications as appliance housings and electrical components. *See* COPOLYMER; PAINT AND COATINGS.

Unsaturated polyesters. The unsaturated polyesters were developed during and shortly after World War II. In combination with glass fiber, they found immediate applications as panels, roofing, radar domes, boat hulls, and protective armor for soldiers. Structural applications of this type continued to increase in importance over the years. The compositions are distinguished by ease of fabrication and high impact resistance. *See* POLYMERIC COMPOSITE.

A low-molecular-weight unsaturated polyester intermediate is first produced. The reaction of maleic anhydride with diethylene glycol is typical, as shown in reaction (2). The product is a viscous oil with a molecular weight of 2000–4000. Many other acids (or anhydrides) such as fumaric acid, and glycols such as propylene glycol may be reacted in a similar manner.

The low-molecular-weight unsaturated polyester will cross-link in the presence of a peroxide by copolymerization with styrene or other vinyl monomers. The unsaturated maleic group copolymerizes in essentially a 1:2 or 1:3 ratio with styrene. Therefore, several styrene molecules which react effectively join two ester chains together to yield an insoluble cross-linked structure, as in reaction (3).

The commercially available intermediate unsaturated polyesters usually contain about 30% styrene or other vinyl monomer. On addition of a peroxide or other free-radical catalyst and a paint drier,



Cross-linked resin

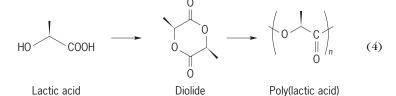
the copolymerization starts. In this stage, the resin may be handled as a viscous fluid for a few minutes to a few hours, depending upon the activity of the catalyst. The viscous liquid may be applied to glass fiber in the form of matt, tow, roving, or cloth, with precautions to eliminate air bubbles and to avoid bubbles that may be caused by overheating as a result of too rapid curing. The surface of the glass fiber must have been given a special finishing treatment in advance for the polyester to adhere strongly. Glass fibers treated with a vinyl silane or an organochrome complex are commercially available.

In the absence of the paint drier, oxygen of the air has an inhibiting effect on the curing process, with the result that the surface of the product remains soft after the inner portions have hardened. In the presence of a paint drier, such as cobalt naphthenate, this skinning effect is eliminated.

A number of modifications of the composition described above have been made. Other acids, other glycols, and various combinations may be used to vary properties of the final product, such as flexibility. The chlorinated derivatives have higher resistance to burning. By varying the free-radical initiator, the optimum temperature required for curing may be varied. There are thermosetting molding compositions which have glass fiber as a filler, and a catalyst which is relatively inactive at ordinary temperatures. When the mixture is cross-linked in the heated mold by the conventional process for thermosetting molding compounds, a fiber-reinforced, high-strength composite material results.

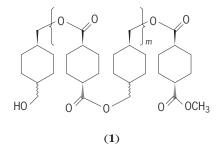
Saturated aliphatic polyesters. Linear polyesters were first studied by W. H. Carothers and coworkers at DuPont in the 1920s and 1930s. Made by the condensation of a diacid such as adipic acid with a diol such as diethylene glycol, the aliphatic polyesters have been frequently used as intermediates in the preparation of prepolymers for making segmented polyurethanes by reaction with diisocyanates. Lactone rings can also be opened to yield linear polyesters; for example, $poly(\epsilon$ -caprolactone) has been used as an intermediate in polyurethane technology, as a polymeric plasticizer for poly(vinyl chloride), and in other specialty applications. *See* RING-OPENING POLYMERIZATION.

A class of aliphatic polyesters that is growing in importance are those derived from bioresources. Poly(lactic acid), made by the polymerization of diolide, in turn derived from corn sugars, has become a commercial material, as in reaction (4). It



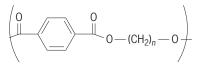
is used in applications demanding biodegradability such as packaging, compost bags, and medical applications. The poly(hydroxyalkanoates), especially poly(hydroxybutyrate), are synthesized within the cells of bacteria in very high yield. Genetic modification of the bacteria has been used to prepare a family of these materials, which have extremely high molecular weights.

One class of aliphatic polyesters with versatile properties is based on hydrogenated terephthalic acids or esters. Poly(cyclohexylidene cyclohexanedicarboxylate) [PCCD; structure 1] is prepared

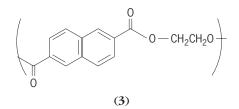


from 1,4-cyclohexanedicarboxylate (DMCD) and 1,4-cyclohexanedimethanol (CHDM). During reduction, a mixture of cis and trans isomers forms, and the properties of the resultant polymers depend strongly on the ratio of those isomers. For example, crystalline polymers with melting points of 230°C (450°F) or greater can be obtained from DMCD with >98% trans content, while transparent amorphous polymers are obtained when the trans content is less than 90%. CHDM can also be combined with terephthalic acid to make poly(cyclohexylidene terepthalate), a fast-crystallizing polymer, with a melting point of about 290°C (550°F). Similarly, combination of CHDM with PET provides polymers known as PCTG or PETG, depending on the ratio of CHDM to ethylene glycol units. These amorphous polymers are known for their transparency and toughness.

Aromatic polyesters. Several aromatic polyesters have achieved general importance as engineering materials, including poly(ethylene terephthalate) [PET; structure **2**a], poly(butylene terephthalate) [PBT; structure **2**b], poly(ethylene naphthoate) [PEN; structure **3**], and poly(trimethylene terephthalate) [PTT; structure **2**c]. PET is the most widely

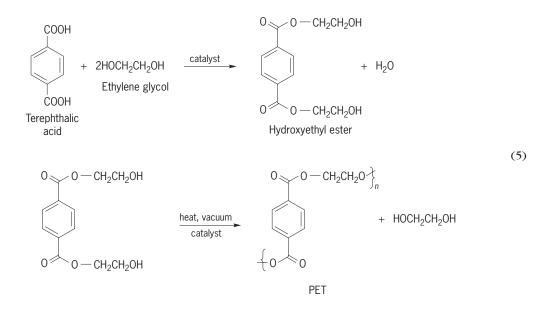


2a (n = 2), **2**b (n = 4), **2**c (n = 3)



used material, in polyester fibers that are available in the United States and in Europe, in films, and in blow-molded bottles.

Typically, PET is prepared by a two-step process from ethylene glycol and terephthalic acid, as in reaction (5). The diacid is first reacted with excess ethylene glycol in the presence of a catalyst to form a hydroxyethyl terephthalate or an oligomer. In the



t

second step, polycondensation to high molecular weight is done at higher temperature under vacuum, eliminating excess ethylene glycol, and leading to an amorphous polymer. As produced in the melt, the polymer is usually amorphous, but it readily crystallizes on reheating, or on extension of the spun filaments as cast or extruded sheets. Molecular weight and crystallinity of PET can be increased by postpolmerization in the solid state as chips, at a temperature somewhat lower than the polymer's melting point of 250°C (482°F). Amorphous PET does crystallize readily, although the crystallization rate can be increased with the use of additives. Conversely, PET can be made completely amorphous with the addition of comonomers, such as diethylene glycol, which retard crystallization.

PET fiber is resistant to mildew and moths. It is used frequently in combination with cotton for clothing. Its chemical and heat resistance placed it in demand for sails and cordage. Thermoplastic molding compounds are used extensively to replace glass in beverage bottles, the recycle of which has become a major business.

PET film, which is stretched and oriented biaxially, is tough, strong, and insensitive to moisture. It is used for special packaging, as photographic film, in electrical transformers and capacitors, and in highstrength laminates. Useful films and fibers are also prepared using 1,4-cyclohexylene glycol as the glycol.

PBT is a thermoplastic, semicrystalline polyester with a melting point of about 225-230°C (437-446°F). Replacement of the ethylene glycol unit with the longer, more flexible 1,4-butanediol imparts an ability to crystallize rapidly, thus making injectionmolding applications possible. PBT is often reinforced with glass fibers or other minerals, and its applications include electronic and communication equipment, automotive uses, appliances, and tools. The crystalline nature of the polymer offers good chemical and solvent resistance and good dimensional stability. PBT has low water uptake, providing consistent dimensional, impact, and electrical properties in a broad range of environments. Preparation of PBT is similar to that of PET, except that a diester, dimethyl terephthalate, is often used as the aromatic monomer, releasing methanol during the polymerization, as in reaction (6).

COOCH₃
+ HOCH₂CH₂CH₂CH₂OH
$$\xrightarrow{\text{Ti(OR)}_4}$$

1,4-Butanediol
COOCH₃
Dimethyl
erephthalate
 $0 \rightarrow 0 - \text{CH}_2\text{CH}_2\text{CH}_2\text{O}$
+ CH₃OH (6)
PBT

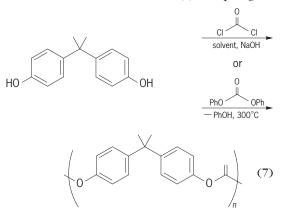
Other polyesters with commercial importance include PEN and PTT. PEN, based on 2,6naphthalene dicarboxylates, is a high-temperature, melt-processible material with excellent barrier, strength, and thermal properties. Although it crystallizes more slowly than PET, its melting point is about 15°C (27° F) higher. PTT has been produced commercially only since 1998 due to the lack of availability of the monomer, 1,3-propanediol. With development of a new monomer process, PTT has been introduced specifically for fiber applications. The fibers have lower modulus and better bending and recovery than PET fibers, opening applications for carpets and other materials. PTT has a melting point similar to PBT (228°C; 442°F) but crystallizes somewhat slower, although faster than PET does. *See* MANUFACTURED FIBER.

All aromatic polyesters include those commonly referred to as polyarylates, made from bisphenol A and mixtures of iso- and terephthalic acids (via their diacid chlorides). Although they may be crystalline, the commercial materials are amorphous, transparent polymers with glass transition temperatures in excess of 175° C (350° F). A special class of ultraviolet-resistant, weatherable polyarylates is based on resorcinol instead of bisphenol A as a monomer. Upon exposure to UV light, the polymer undergoes a backbone rearrangement, leading to a highly absorbent UV-resistant layer, which protects the underlying polymer.

An attractive class of aromatic polyesters are the liquid crystalline polyesters (LCPs), largely based on *p*-hydroxybenzoic acids in combination with other monomers such as terephthalic acid, hydroquinone, biphenol, or biphenol dicarboxylic acid. The LCP polyesters have excellent flow properties in the melt, but orient when spun to provide fibers with very high modulus (stiffness).

Aromatic polycarbonates. Polycarbonates (PC) are a strong, optically clear, tough group of thermoplastic polymers formed most frequently from bisphenol A. The products are noted for high glass-transition temperatures, usually greater than 145° C (284° F), high impact resistance, clarity, and resistance to creep.

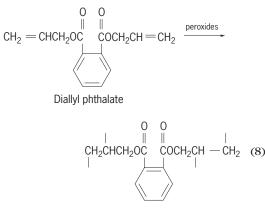
Polycarbonates are prepared commercially by two methods: interfacial reaction of bisphenol A with phosgene, or transesterification with diphenyl carbonate in the melt as in reaction (7). The phosgene



process is milder but requires removal of by-product NaCl, whereas the melt process produces resin pellets directly. Bisphenol A is obtained by the condensation of phenol and acetone.

The polymer is usually available as an injectionmolding resin. Because of its high strength, toughness, and softening point, the resin, both by itself and as a glass-reinforced material, has found many electrical and engineering applications. It is often used to replace glass and metals. Examples include bottles, unbreakable windows, appliance parts, electrical housings, marine propellers, and shotgun shells. Flame-retardant grades are of interest because of low toxicity and smoke emission on burning. Films have excellent clarity and electrical characteristics, and have been used on capacitors and in solar collectors, as well as in electronic packaging. With excellent optical properties, polycarbonate has been the material of choice for optical media such as compact discs and DVDs.

Polydiallyl esters. These are polymers of diallyl esters, such as diallyl phthalate, diallyl carbonate, diallyl phenyl phosphonate, and diallyl succinate, in which cross-linked products are made by polymerization of the allyl groups, as in the case of diallyl phthalate, reaction (8).





Thermosetting molding compounds may be produced by careful limitation of the initial polymerization to yield a product which is fusible. Then the polymerization and curing are completed in the final molding operation. Partially polymerized resins and the monomer are useful in making chemically resistant coatings. The monomer is often used as a crosslinking agent for polyester resins.

Major applications are in electronic components, sealants, coatings, and glass-fiber composites. *See* PLASTICS PROCESSING; POLYMERIZATION.

Daniel J. Brunelle Bibliography. A.-C. Albertsson, *Degradable Alipathic Polyesters*, 2001; Y. Doi and A. Steinbüchel (eds.), *Polyesters III: Applications and Commercial Products*, vol. 4: *Biopolymers*, 2002; G. Odian, *Principles of Polymerization*, 4th ed., 2004; J. Scheirs and T. E. Long (eds.), *Modern Polyesters*, Wiley, West Sussex, England, 2003.

Polyether resins

Thermoplastic or thermosetting materials which contain ether-oxygen linkages, -C-O-C-, in the polymer chain. Depending upon the nature of the reactants and reaction conditions, a large number of polyethers with a wide range of properties may be prepared.

The main groups of polyethers in use are epoxy resins, prepared by the polymerization and cross-linking of aromatic diepoxy compounds; phenoxy resins, high-molecular-weight epoxy resins; poly(ethylene oxide) and poly(propylene oxide) resins; polyoxymethylene, a high polymer of formaldehyde; and poly(phenylene oxides), polymers of xylenols.

Epoxy resins. The epoxy resins form an important and versatile class of cross-linked polyethers characterized by excellent chemical resistance, adhesion to glass and metals, electrical insulating properties, and ease and precision of fabrication.

In the preparation of a typical resin, a lowmolecular-weight diepoxy compound is first mixed with cross-linking agents, fillers, and plasticizers and then allowed to cure either at room temperature or with the application of heat.

The intermediate diepoxy compounds are condensation products of epichlorohydrin and aliphatic or aromatic diols and are available commercially. An example is the product of the reaction of epichlorohydrin with bisphenol A [reaction (1)].

A wide variety of diols may be used in order to obtain a desired balance of properties. The liquid or solid intermediates obtained are quite stable, but do not have useful physical properties until they are polymerized further, by the addition of a curing or cross-linking agent, in a manner similar to addition polymerization. Polymerization may be initiated by catalysts such as boron trifluoride or tertiary amines. Alternatively, compounds containing a reactive hydrogen atom such as organic acids, alcohols, mercaptans, primary and secondary amines, and polyamides can also serve as catalysts. Reaction (2) represents this curing process with an amine compound. For special uses, polyols of greater functionalities are often used. These include epoxidized novalacs, epoxidized polyphenols, and glycidyl derivatives of amino-substituted aromatic compounds.

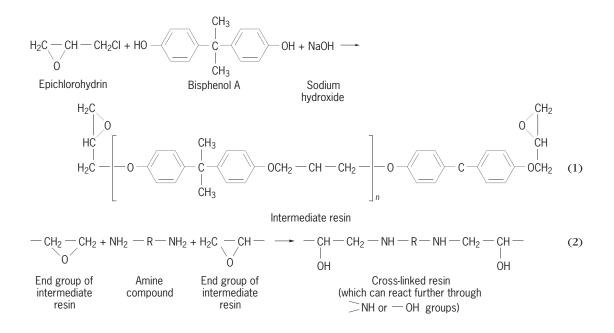
The type of curing agent employed has a marked effect on the optimum temperature of curing and

has some influence on the final physical properties of the product. By judicious selection of the curing system, the curing operation can be carried out at almost any temperature from 0 to 200° C (32 to 390° F).

Various fillers such as calcium carbonate, metal fibers and powders, and glass fibers are commonly used in epoxy formulations in order to improve such properties as the strength and resistance to abrasion and high temperatures. Some reactive plasticizers, such as mercaptans and polyamides, act as curing agents, become permanently bound to the epoxy groups, and are usually called flexibilizers. Rubbery polymers such as carboxy-terminated butadienestyrene copolymers are added to improve toughness and impact strength. *See* COPOLYMER.

Because of their good adhesion to substrates and good physical properties, epoxies are commonly used in protective coatings. Because of the small density change on curing and because of their excellent electrical properties, the epoxy resins are used as potting or encapsulating compositions for the protection of delicate electronic assemblies from the thermal and mechanical shock of rocket flight. Because of their dimensional stability and toughness, the epoxies are used extensively as dies for stamping metal forms, such as automobile gasoline tanks, from metal sheeting. Foams are also made.

By combining epoxies, especially the higherperformance types based on polyfunctional monomers, with fibers such as glass or carbon, exceedingly high moduli and strengths may be obtained. Thus a typical carbon-fiber-reinforced epoxy may have a tensile modulus and strength of 75,000,000 lb/in.² (500 gigapascals) and 300,000 lb/in.² (2 GPa), respectively. Such composites are of importance in aerospace applications where high ratios of a property to density are desired, as well as in such domestic



applications as sports equipment. *See* COMPOSITE MATERIAL; POLYMERIC COMPOSITE.

The adhesive properties of the resins for metals and other substrates and their relatively high resistance to heat and to chemicals have made the epoxy resins useful for protective coatings and for metal-tometal bonding.

Poly(olefin oxide) resins. Poly(ethylene oxide) and poly(propylene oxide) are thermoplastic products whose properties are greatly influenced by molecular weight. The general method of preparation is represented by reactions (3) and (4). Other cyclic

$$nH_2C \xrightarrow{CH_2} CH_2 \xrightarrow{catalyst} [-CH_2 - CH_2 - 0 -]_n \quad (3)$$

Ethylene Poly(ethylene oxide)
$$nH_2C \xrightarrow{CH_2} CH_2 \xrightarrow{catalyst}$$

O Propylene oxide

$$\begin{bmatrix} -CH_2 - CH - 0 - \end{bmatrix}_n \quad (4)$$

oxides, such as tetrahydrofuran, can also be polymerized to give polyethers.

Low-to-moderate molecular-weight poly(ethylene oxides) vary in form from oils to waxlike solids. They are relatively nonvolatile, are soluble in a variety of solvents, and have found many uses as thickening agents, plasticizers, lubricants for textile fibers, and components of various sizing, coating, and cosmetic preparations. The poly(propylene oxides) of similar molecular weight have somewhat similar properties, but tend to be more oil-soluble (hydrophobic) and less water-soluble (hydrophilic).

Nonionic surface-active agents can be prepared from C_{10} - C_{20} fatty alcohols and acids by the condensation of some 5–40 ethylene oxide groups, for example, reaction (5). An interesting commercial example

$$C_{11}H_{23}COOH + 20 H_2C \xrightarrow{O} CH_2 \xrightarrow{NaOH}$$
Lauric acid Ethylene oxide
$$C_{11}H_{23}COO(CH_2CHO)_{20}H \qquad (5)$$
A nonionic detergent

of a block copolymer has been produced by polymerization of propylene oxide onto poly(ethylene oxide) to yield a linear chain with sequences of the two compounds, as in reaction (6), in which x and

$$\begin{bmatrix} -CH_2 - CHO - \end{bmatrix}_x + y H_2C - CHCH_3 \longrightarrow$$
Polyethylene oxide
Propylene oxide

$$\begin{bmatrix} -CH_2 - CHO - \end{bmatrix}_x - \begin{bmatrix} -CH_2 - CH - O - \end{bmatrix}_y \quad (6)$$

y may be in the range 10-100. The hydrophobichydrophilic balance necessary for surface activity is achieved because the polyethylene sequence is relatively hydrophilic and the polypropylene sequence is relatively hydrophobic.

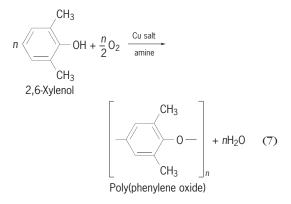
Oil-water emulsions prepared by use of the product have remarkable stability to both hydrophilic and hydrophobic precipitating agents.

While poly(alkylene oxides) are not of interest as such in structural materials, poly(propylene oxides) are used extensively in the preparation of polyurethane forms.

Crystalline, high-molecular-weight poly(propylene oxides) with melting points up to about 74° C (165°F) have been prepared with three groups of catalysts: (1) solid potassium, (2) complexes of ferric or stannic chloride with propylene oxide, and (3) certain metallic alkyls such as aluminum triethyl. By starting with optically active propylene oxide, an optically active polymer is produced.

Phenoxy resins. Phenoxy resins differ from the structurally similar epoxy resins based on the reaction of epichlorohydrin with bisphenol A mainly by possessing a much higher molecular weight, about 25,000. The polymers are transparent, strong, ductile, and resistant to creep, and, in general, resemble polycarbonates in their behavior. Cross-linking may be effected by the use of curing agents which can react with the OH groups. Molding and extrusion may be used for fabrication. The major application is as a component in protective coatings, especially in metal primers.

Poly(phenylene oxide) resins. Poly(phenylene oxide) [PPO] is the basis for an engineering plastic characterized by chemical, thermal, and dimensional stability. The polymers are prepared by the oxidative coupling of 2,6-xylenol, using a cuprous salt and an amine as catalyst, as shown in reaction (7).

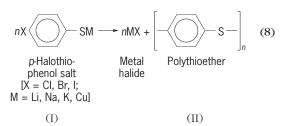


Similar reactions may be used to produce many varied structures. The poly(phenylene oxide) resin is normally blended with another compatible but cheaper resin such as high-impact polystyrene. The blends are cheaper and more processable than PPO and still retain many of the advantages of PPO by itself.

Poly(phenylene oxide is outstanding in its resistance to water, and in its maximum useful temperature range (about $170-300^{\circ}$ C or $340-570^{\circ}$ F). In spite of the high softening point, the resins, including glass-reinforced compositions, can be molded and extruded in conventional equipment. Uses include

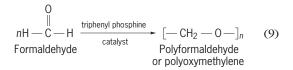
medical instruments, pump parts, and insulation. Structurally modified resins are also of interest and the general family of resins may be expected to replace other materials in many applications.

An analogous polythioether, poly(phenylene sulfide), may be prepared by the reaction of *p*-dichlorobenzene with sulfur or sodium sulfide in a polar solvent yielding a branched polythioether, or by polymerization of metal salts of *p*-halothiophenols (I), yielding a linear polythioether (II) [reaction (8)].



The combination of regular structure with the chemically and thermally stable ring-sulfur bond leads to a combination of a high melting point ($\sim 290^{\circ}$ C or 550° F), thermal stability, inherent resistance to burning, and resistance to most chemicals. Commercial products can be processed by injection or compression molding, as well as by slurry or electrostatic coating; reinforcement with glass fibers is also commonly practiced. The combination of properties is useful in such applications as electronic pumps, and automotive components and coatings, especially where environmental stability is required.

Polyoxymethylene. Polyoxymethylene, or polyacetal, resins are polymers of formaldehyde. Having high molecular weights and high degrees of crystallinity, they are strong and tough and are established in the general class of engineering thermoplastics. Polymerization is accomplished by the use of a catalyst such as triphenyl phosphine, shown in reaction (9). Since the untreated polymers tend to de-



grade by an unzipping or depolymerization reaction, the end groups are frequently esterified to block the initiation of degradation. Alternatively, small amounts of a comonomer such as ethylene oxide or 1,3-dioxolane may be introduced during polymerization.

Though somewhat similar to polyethylene in general molecular structure, polyacetal molecules pack more closely, and attract each other to a much greater extent, so that the polymer is harder and highermelting than polyethylene. Polyacetals are typically strong and tough, resistant to fatigue, creep, organic chemicals (but no strong acids or bases), and have low coefficients of friction. Electrical properties are also good. Improved properties for particular application may be attained by reinforcement with fibers of glass or polytetrafluoroethylene.

The combination of properties has led to many

uses such as plumbing fittings, pump and valve components, bearings and gears, computer hardware, automobile body parts, and appliance housings. Other aldehydes may be polymerized in a similar way.

A chlorinated polyether based on the cationic polymerization of 3.3-bis(chloromethyl-1-oxacyclobutane) was developed as an engineering plastic. While it has excellent chemical resistance and chemical stability, other plastics can give as good performance at lower cost. *See* EPOXIDE; PLASTICS PROCESS-ING; POLYMERIZATION. John A. Manson

Bibliography. J. A. Brydson, *Plastics Materials*, 7th ed., 1999; G. W. Ehrenstein and R. P. Theriault, *Polymeric Materials: Structure, Properties, Applications*, 2000; G. Odian, *Principles of Polymerization*, 4th ed., 2004; A. B. Strong, *Plastics: Materials and Processing*, 3d ed., 2005.

Poly(ethylene glycol)

Any of a series of water-soluble polymers with the general formula $HO(CH_2CH_2O)_nH$. These colorless, odorless compounds range in appearance from viscous liquids for n = 2 to 13 (freezing point range of -15 to 0°C or 5 to 32°F); waxy solids and pastes for n = 22 to 90 (melting point range of 37 to 58°C or 98 to 136°F); and free flowing powder for n = 113 and above (melting point range of 60 to 67°C or 140 to 153°F).

Uses. The low-molecular-weight members, di(ethylene glycol) (n = 2) through tetra(ethylene glycol) (n = 4), are produced as pure compounds and find use as humectants, dehydrating solvents for natural gas, textile lubricants, heat-transfer fluids, solvents for aromatic hydrocarbon extractions, and intermediates for polyester resins and plasticizers. These lower poly(ethylene glycol)s are distilled by-products of ethylene glycol manufacture.

The intermediate members of the series with average molecular weights of 200 to 20,000 are produced as residue products by the sodium or potassium hydroxide-catalyzed batch polymerization of ethylene oxide onto water or mono- or di(ethylene glycol). These polymers, also known as poly(ethylene oxide) or polyoxyethylene, are formed by stepwise anionic addition polymerization and, therefore, possess a distribution of molecular weights. Examples of commercial uses for products in this range are in ceramic, metal-forming, and rubber-processing operations; as drug suppository bases and in cosmetic creams, lotions, and deodorants; as lubricants; as dispersants for casein, gelatins, and inks; and as antistatic agents. These poly(ethylene glycol)s generally have low human toxicity.

The highest members of the series have molecular weights from 100,000 to 10,000,000. They are produced by special anionic polymerization catalysts that incorporate metals such as aluminum, calcium, zinc, and iron, and coordinated ligands such as amides, nitriles, or ethers. These members of the poly(ethylene glycol) series are of interest because of their ability at very low concentrations to reduce

friction of flowing water. *See* ETHYLENE OXIDE; POLY-MERIZATION; RING-OPENING POLYMERIZATION.

Robert K. Barnes; Tacey X. Viegas **Biotechnical applications.** The properties of poly(ethylene glycol) make it suitable for medical and biotechnical applications. These polymers of ethylene oxide are frequently described as amphiphilic, meaning that they are soluble both in water and in most organic solvents. The medical and biotechnical applications derive from this amphiphilicity, from a lack of toxicity and immunogenicity, and from a tendency to avoid other polymers and particles also present in aqueous solution.

Poly(ethylene glycol)s are widely used in a variety of pharmaceutical formulations such as dermal ointments and creams, oral solvents in liquid-filled capsules, and in injectable solution concentrates where the low-molecular-weight polymers (n < 9) have been used in concentrations up to 30% by volume.

Attaching poly(ethylene glycol) to another molecule provides the latter with enhanced solubility in solvents in which poly(ethylene glycol) is soluble, and thus poly(ethylene glycol) is attached to drugs to enhance water and blood solubility. Similarly, poly(ethylene glycol) is attached to insoluble enzymes to impart solubility in organic solvents. These poly(ethylene glycol) enzymes are used as catalysts for industrial reactions in organic solvents. *See* CATALYSIS.

The tendency to avoid other polymers results in formation of two immiscible aqueous layers when a solution of a poly(ethylene glycol) is mixed with certain other polymer solutions. These aqueous twophase systems are used to purify biological materials such as proteins, nucleic acids, and cells by partitioning of desired and undesired materials between the two phases.

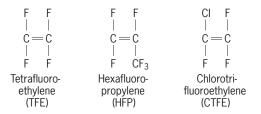
The tendency of poly(ethylene glycol)s to avoid interaction with cellular and molecular components of the immune system results in the material being nonimmunogenic. Covalent attachment of a poly(ethylene glycol) to proteins results in greatly reduced immunogenicity of the poly(ethylene glycol) proteins. This property, along with reduced rate of kidney clearance because of larger size, leads to a greatly enhanced blood circulation lifetime of poly(ethylene glycol) proteins and to their application as pharmaceutical delivery systems. Similarly, adsorption of proteins and cells to surfaces is greatly reduced by attaching poly(ethylene glycol) to the surface, and such coated materials find wide application as biomaterials. See BIOMEDICAL CHEMICAL ENGI-NEERING. J. Milton Harris; Tacey X. Viegas

Bibliography. J. Brydson, *Plastics Materials*, 7th ed., 1999; C. Delgado, G. E. Francis, and D. Fisher, The uses and properties of PEG-linked proteins, *Crit. Rev. Therap. Drug Carrier Sys.*, 9:249–304, 1992; J. Fried, *Polymer Science and Technology*, 2d ed., 2003; J. M. Harris (ed.), *Polyethylene Glycol Chemistry: Biotechnical and Biomedical Applications*, 1992; J. M. Harris and F. M. Veronese (eds.), *Adv. Drug Deliv. Rev.*, 55(10):1259–1350, September 26,

2003; F. M. Veronese and J. M. Harris (eds.), *Adv. Drug Deliv. Rev.*, 54(4):453-610, June 17, 2002.

Polyfluoroolefin resins

Resins distinguished by their resistance to heat and chemicals and by the ability to crystallize to a high degree. Several main products are based on tetrafluoroethylene, hexafluoropropylene, and monochlorotrifluoroethylene. Structural formulas of the monomers are shown below.



Copolymers of TFE and HFP with each other and of TFE and HEP with ethylene are available commercially. Homopolymers of vinylidene fluoride and copolymers with HFP are also in common use. For a description of poly(vinyl fluoride). *See* POLYVINYL RESINS.

Polytetrafluoroethylene. Tetrafluoroethylene can be obtained by the prolysis of monochlorodifluoromethane, which in turn is obtained from a rather complex reaction between anhydrous hydrogen fluoride and chloroform. This reaction is represented by (1), (2), and (3). Although poly-

$$\begin{array}{rcl} \mbox{CHCl}_3 & + & \mbox{HF} & \xrightarrow{\mbox{Sucl}_5} \\ \mbox{Chloro-} & & \mbox{Anhydrous} \\ \mbox{hydrofluoric} \\ \mbox{acid} \\ & & \mbox{FCHCl}_2 + \mbox{F}_2 \mbox{CHCl} + \mbox{CCl}_2 \mbox{F}_2 + \mbox{HCl} & (1) \end{array}$$

ChO

$$2FCHCI_2 + AICI_3 \longrightarrow F_2CHCI + CHCI_3$$
(2)

 $\begin{array}{c} H \\ 2F_2CCI \\ Monochlorotri- \\ fluoromethane \end{array} \quad \begin{array}{c} 700-800^{\circ}C \\ \hline (1300-1500^{\circ}F) \end{array} \quad F_2C = CF_2 + 2HCI \quad (3)$

merization in bulk can proceed with violence, the monomer can be polymerized readily and conveniently in emulsion under pressure, using free-radical catalysts such as peroxides or persulfates. The polymer is insoluble, resistant to heat (up to 275° C or 527° F) and chemical attack, and in addition, has the lowest coefficient of friction of any solid. Because of its resistance to heat, the fabrication of polytetrafluoroethylene requires modification of conventional methods. After molding the powdered polymer using a cold press, the moldings are sintered at 360-400°C (680-750°F) by procedures similar to those used in powder metallurgy. The sintered product can be machined or punched. Extrusion is possible if the powder is compounded with a lubricating material. Aqueous suspensions of the polymer can also be used for

coating various articles. However, special surface treatments are required to ensure adhesion because polytetrafluoroethylene does not adhere well to anything.

Polytetrafluoroethylene (TFE resin) is useful for applications under extreme conditions of heat and chemical activity. Polytetrafluoroethylene bearings, valve seats, packings, gaskets, coatings and tubing can withstand relatively severe conditions. Fillers such as carbon, inorganic fibers, and metal powders may be incorporated to modify the mechanical and thermal properties.

Because of its excellent electrical properties, polytetrafluoroethylene is useful when a dielectric material is required for service at a high temperature. The nonadhesive quality is often turned to advantage in the use of polytetrafluoroethylene to coat articles such as rolls and cookware to which materials might otherwise adhere. A structural derivative containing perfluoroalkoxy side groups (PFA resin) is available. PFA resins behave generally like TFE resins, but can be processed by conventional molding, extrusion, and powder-coating methods.

Poly(chlorotrifluoroethylene). The monomer is prepared by the dechlorination of 1,1,2-trichloro-1,2,2trifluoroethane as in reaction (4).

$$CCI_2FCCIF_2 + Zn \text{ dust} \xrightarrow{C_2H_5OH} CIFC = CF_2 + ZnCI_2 \quad (4)$$

Polymerization can be carried out in aqueous suspension by the free-radical process in which a combination of a persulfate and bisulfite is used as the initiator.

The properties of poly(chlorotrifluoroethylene) [CTFE resin] are generally similar to those of polytetrafluoroethylene; however, the presence of the chlorine atoms in the former causes the polymer to be a little less resistant to heat and to chemicals. The poly(chlorotrifluoroethylene) can be shaped by use of conventional molding and extrusion equipment, and it is obtained in a transparent, noncrystalline condition by quenching. Dispersions of the polymer in organic media may be used for coating.

The applications of poly(chlorotrifluoroethylene) are in general similar to those for polytetrafluoroethylene. Because of its stability and inertness, the polymer is useful in the manufacture of gaskets, linings, and valve seats that must withstand hot and corrosive conditions. It is also used as a dielectric material, as a vapor and liquid barrier, and for microporous filters.

Copolymers. Copolymers of TFE and HFP propylene (fluorinated ethylenepropylene, or FEP resins) and copolymers of TFE with ethylene (ETFE) are often used in cases where ease of fabrication is desirable. The copolymers can be processed by conventional thermoplastic techniques, and except for some diminution in the level of some properties, properties generally resemble those of the TFE homopolymer.

Copolymers of ethylene with CTFE can also be processed by conventional methods, and have better mechanical properties than PTFE, FEP, and PFA resins. **Poly(vinylidene fluoride).** The monomer is prepared by the dehydrohalogenation of 1,1,1chlorodifluoroethane or by the dechlorination of 1,2-dichloro-1,1-difluoroethane as in reactions (5) and (6). Polymerization is effected using free radical catalysis.

$$CH_3CCIF_2 \xrightarrow{-HCI} H_2C = CF_2$$
(5)

$$CH_2 CICCIF_2 \xrightarrow{-Cl_2} H_2 C = CF_2$$
(6)

The properties are generally similar to those of the other fluorinated resins: relative inertness, low dielectric constant, and thermal stability (up to about 150° C or 300° F). The resins (PVF₂ resins) are, however, stronger and less susceptible to creep and abrasion than TFE and CTFE resins.

Applications of poly(vinylidene fluoride) are mainly as electrical insulation, piping, process equipment, and as a protective coating in the form of a liquid dispersion.

Fluorinated elastomers. Several types of fluorinated, noncrystallizing elastomers were developed in order to meet needs (usually military) for rubbers which possess good low-temperature behavior with a high degree of resistance to oils and to heat, radiation, and weathering. Fluorinated acrylic rubbers such as poly(heptafluorobutyl acrylate) and fluorine-containing silicone rubbers have been used to some extent for this purpose. Copolymers of hexafluoropropylene with vinylidene fluoride make up an important class with such applications as gaskets and seals. Copolymers of nitrosomethane with tetrafluoroethylene are showing considerable promise for similar applications. See HALOGENATED HYDROCARBON; PLASTICS PROCESSING; POLYMERIZA-TION. John A. Manson

Bibliography. J. A. Brydson, *Plastics Materials*, 7th ed., 1999; G. G. Hougham et al. (eds.), *Fluoropolymers 1: Synthesis* and *Fluoropolymers 2: Properties*, 1999; G. Odian, *Principles of Polymerization*, 4th ed, 2004.

Polygalales

An order of flowering plants, division Magnoliophyta (Angiospermae), in the subclass Rosidae of the class Magnoliopsida (dicotyledons). The order consists of 7 families and nearly 2300 species, mostly of tropical and subtropical regions. The vast majority of the species belong to only 3 families, the Malpighiaceae (about 1200 species), Polygalaceae (about 750 species), and Vochysiaceae (about 200 species). Within its subclass the order is distinguished by its simple leaves and usually irregular flowers, which have the perianth and stamens attached directly to the receptacle (hypogynous), and often have the anthers opening by terminal pores instead of longitudinal slits. The Barbados cherry (Malpigbia glabra), noted for the high vitamin-C content of its fruits, is a well-known member of the Polygalales. See MAGNOLIOPSIDA; PLANT KINGDOM; ROSIDAE. Arthur Cronquist; T. M. Barkley

Polygon

A geometric figure consisting of an ordered set of three or more (but a finite number of) points called vertices, each vertex connected by line segments called sides or edges to two other vertices. These two sides are said to be adjacent, and so are any two vertices that are end points of a side. The perimeter of the polygon is the sum of the lengths of the sides. The line segments that join two nonadjacent vertices of a polygon are called diagonals. A polygon is said to be directed, or oriented, if a preferred direction is assigned to each side so that at each vertex one of the adjacent sides is directed toward the vertex and the other away from it.

The angle between the two sides of a polygon at a vertex is called an angle of the polygon. Thus an *n*-sided polygon, called an *n*-gon, has *n* vertices and *n* angles. In particular, if *n* is 3, 4, 5, 6, 7, 8, 10, or 12, the polygon is called a triangle (3), quadrangle (4) [or quadrilateral, meaning four sides], pentagon (5), hexagon (6), heptagon (7), octagon (8), decagon (10), or dodecagon (12).

Plane and skew polygons. A plane polygon is one whose vertices all lie in the same plane. Other polygons are called skew polygons, except for the spherical polygons described below. Skew polygons can be constructed in any number of dimensions.

A plane polygon is called ordinary if no point belongs to more than two edges; it is proper if no two adjacent sides are collinear; it is simple if no two edges intersect each other except at vertices. A simple polygon divides the plane into two regions: an unbounded outside region and an inside region whose area is called the area of the polygon.

A simple polygon is called convex if its interior lies entirely on one side of each (infinite) line through any two adjacent vertices. A nonconvex polygon has at least one interior angle that exceeds 180° . The sum of the interior angles in a convex *n*-gon is $(n - 2)180^{\circ}$. A plane polygon is called regular if all its sides are equal and all its angles are equal; it is semiregular if all its angles are equal. *See* REGULAR POLYTOPES.

Triangles. A triangle is equilateral if it has three equal sides, isosceles if two sides are equal, and scalene otherwise. It is a right triangle if one angle measures 90°, obtuse if one angle exceeds 90°, or acute otherwise. Its altitudes are the line segments (or their lengths) drawn from each vertex perpendicular to the opposite side (extended if necessary). These meet in a point *H* called the orthocenter (**Fig. 1**).

The circle passing through the three vertices A, B, and C of the triangle ABC is the circumscribed circle of radius R, having its center at the circumcenter P, which is the point of intersection of the perpendicular bisectors of the sides. Segments that join the midpoint of a side of the triangle to the opposite vertex are the triangle medians; they intersect at the

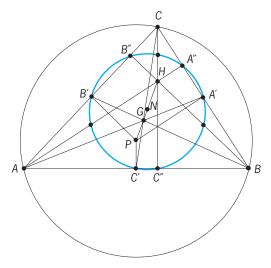


Fig. 1. Triangle *ABC* with nine-point circle. Symbols are explained in text.

center of gravity *G* that trisects both the medians and the segment *PH*, so that HG = 2GP (Fig. 1).

The incenter of the triangle is the point of intersection of the interior angle bisectors; it is the center of the inscribed circle of radius *r*. Three escribed circles lying outside the triangle are each tangent to one side and to the other two sides extended; their radii are rs/(s - a), rs/(s - b), and rs/(s - c), where s =(a + b + c)/2 is the semiperimeter of the triangle.

The area of the triangle *ABC* with sides of lengths *a*, *b*, and *c* is given by Eq. (1).

Area =
$$rs = \sqrt{s(s-a)(s-b)(s-c)} = \frac{abc}{4R}$$
 (1)

Nine-point circle. An important theorem states that the so-called nine-point circle of radius R/2 with center N at the midpoint of HP (Fig. 1) passes through the midpoints of the sides (A', B', and C') and of the segments HA, HB, HC, as well as through the feet of the altitudes (A', B', and C') of the triangle ABC. Feuerbach's theorem states that this circle is also tangent to the inscribed and escribed circles. *See* ANGLE MODULATION; CIRCLE.

Right triangles. In a right triangle *ABC*, the side *c* opposite the right angle *C* is called the hypotenuse. The pythagorean theorem states that its square is equal to the sum of the squares of the other two sides: $a^2 + b^2 = c^2$. If the sides *a*, *b*, and *c* are integers with highest common factor 1 or 2, they form a pythagorean triple (*a*,*b*,*c*), expressible in the form of Eqs. (2),

$$a = m2 - n2 \qquad b = 2mn \qquad c = m2 + n2 \qquad (2)$$
$$m > n$$

where *m* and *n* are integers without a common factor. If n < m < 2n, the smallest angle *A* can be approximated in degrees to within 1 minute by Eq. (3). For example, for the pairs (m,n) = (2,1),

$$A = \frac{86(m-n)(m^2 - n^2)}{m^3 - n^3}$$
(3)

(8,5), and (10,7), this formula approximates by

 $36^{\circ}51.4'$, $26^{\circ}0'$, and $20^{\circ}1.64'$ the angles $A = 36^{\circ}52.2'$, $25^{\circ}59.4'$, and $20^{\circ}0.95'$.

Quadrangles. Quadrangles or quadrilaterals are four-gons. They include (1) the trapezoid that has two parallel sides, whose area is equal to half the sum of these parallel sides times the distance between them; (2) the parallelogram, whose opposite sides are parallel and whose diagonals bisect each other; (3) the rhombus, whose four sides are equal; (4) the rectangle, whose four angles are right angles; (5) the inscribed quadrilateral, whose vertices lie on a circle so that opposite angles are supplementary; and (6) the square, which has equal sides and equal angles, and is included in all the types listed above.

If s = (a + b + c + d)/2 denotes the semiperimeter of an inscribed quadrilateral with sides *a*, *b*, *c*, and *d*, its area is given by Eq. (4).

Area =
$$\sqrt{(s-a)(s-b)(s-c)(s-d)}$$
 (4)

Pentagons. The regular pentagon has five diagonals that form a polygon known as the star pentagon, or pentagram, a nonconvex polygon familiar in Christmas star decorations. The ratio of lengths of the diagonal and side of a regular pentagon, known as the golden ratio, is equal to $(1 + \sqrt{5})/2$, which is close to 144/89. If *A*, *B*, and *C* denote three consecutive vertices of a regular pentagon, and *M* is the midpoint of the diagonal *AC*, then the triangles *ABM* and *CBM* each have angles 36° , 54° , and 90° (**Fig. 2**). A triangle with sides 274, 377, and 466 has angles $36^{\circ}0.84'$, $53^{\circ}59.98'$, and $89^{\circ}59.18'$, close to 36° , 54° , and 90° , and is thus a building block for constructing a nearly regular pentagon without a protractor.

Hexagons. The regular hexagon is commonly found in floor tilings and is noticeable in the cross section of a beehive. When circular objects are closely packed, they tend to assume a regular hexagonal pattern. Crystalline solids having such a form are known as hexagonal close-packed (hcp) structures. *See* CRYSTAL STRUCTURE.

Spherical polygons. A spherical polygon consists of points on a spherical surface called vertices, each connected to two adjacent vertices by great-circle arcs, called sides, that are measured by their central angles. The *n* angles of a convex spherical *n*-gon are the angles at each vertex between the tangent lines to the two sides that meet there, and their sum exceeds $(n - 2)180^{\circ}$ by an amount *E*, called the spherical area enclosed by the polygon. If the spherical radius is *R*, this area is given by Eq. (5).

Area =
$$\frac{E}{180^{\circ}}\pi R^2$$
 (5)

Spherical triangles and polar triangles. Any three distinct points, *A*, *B*, and *C*, on a spherical surface are the vertices of a spherical triangle, whose sides are the (shorter) great-circle arcs a = BC, b = CA, and c = AB, joining the vertices. The relations between the sides and angles of a spherical triangle are investigated in the subject of spherical trigonometry. *See* TRIGONOMETRY.

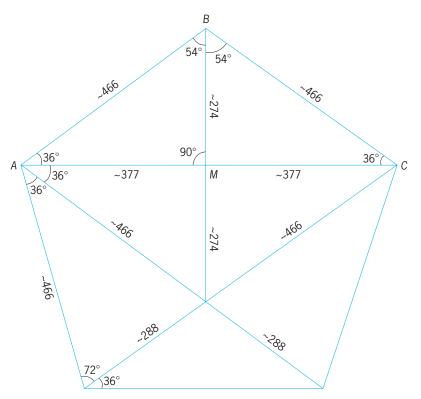


Fig. 2. Regular pentagon, showing approximate construction without a protractor. Symbols are explained in text.

The axis and poles of a great-circle arc are respectively the spherical diameter perpendicular to its plane and the points where this axis meets the spherical surface. Thus, on the Earth's surface (assumed to be spherical) the poles of the Equator are the North and South poles, lying on the Earth's axis of rotation. Given a spherical triangle *ABC*, the polar triangle *A'B'C'* is the triangle whose vertices, *A'*, *B'*, and *C'*, are those poles of the sides *a*, *b*, and *c* nearest to *A*, *B*, and *C*, respectively. The polar triangle of *A'B'C'* is the original triangle *ABC* (**Fig. 3**). If the

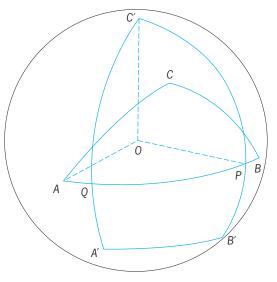


Fig. 3. Polar triangles *A*, *B*, *C* and *A'*, *B'*, *C'*. The center of the sphere is at *O*. *AQ* and *PB* are 90° arcs, so angle C' =arc *PQ* is the supplement of *AB*.

sides A' = C'B' and b' = C'A' of the polar triangle (extended if necessary) meet side *c* of *ABC* in points *P* and *Q* respectively, then arcs *AP* and *QB* measure 90°, so the angular measures of arc *AB* = side *c* and arc *PQ* = angle *C'* are supplementary. Thus, the angular measures of the sides of a spherical triangle are the supplements of the angles of the polar triangle.

Curve approximation. By taking a large number of consecutive points along a smooth closed curve as vertices of a polygon whose side lengths are sufficiently small, the length of the curve can be approximated by the perimeter of the polygon. The curve length is defined if these perimeter lengths converge when the largest side length approaches zero. If the curve forms a knot in space, the knot can be analyzed by projecting an oriented approximating polygon onto a suitable plane and indicating, at each point where two sides cross each other, both of their directions and which side is above the other in projection. J. Sutherland Frame

Bibliography. F. Bachmann and E. Schmidt, *n-Gons*, 1975; H. S. M. Coxeter, *Introduction to Geometry*, 2d ed., rev. printing, 1989; B. Grünbaum and G. C. Shephard, *Tilings and Patterns*, 1986; L. M. Kelly (ed.), *The Geometry of Metric and Linear Spaces*, 1975; K. Miyazaki, *An Adventure in Multidimensional Space: The Art and Geometry of Polygons*, *Polybedra, and Polytopes*, 1986.

Polygonales

An order of flowering plants, division Magnoliophyta (Angiospermae), in the subclass Caryophyllidae of the class Magnoliopsida (dicotyledons). The



Polygonum hydropiper, eastern American smartweed, characteristic of order Polygonales. Sheathing stipules and many tiny flowers are typical of group. (A. W. Ambler, National Audubon Society)

order consists only of the family Polygonaceae, with about 800 species, most abundant in north temperate regions (see illus.). Within its subclass the order is characterized by its well-developed endosperm; an ovary with but one chamber (unilocular), mostly tricarpellate with a single basal ovule which is usually straight and has the micropyle at the opposite end from the stalk (orthotropous); the flowers are often trimerous, that is, with floral parts in sets of three, and usually two or three sets of stamens. Rhubarb (Rheum rhaponticum) and buckwheat (Fagopyrum esculentum) are familiar members of the Polygonales. See BUCK-WHEAT; CARYOPHYLLIDAE; MAGNOLIOPSIDA; PLANT Arthur Cronquist; T. M. Barkley KINGDOM; RHUBARB.

Polyhedron

A solid whose boundary consists of a finite number of polygonal faces, that is, planar regions that are bounded by polygons. The sides of the faces are edges of the polyhedron; the vertices of the faces also are vertices of the polyhedron. *See* PLANE GEOM-ETRY; POLYGON.

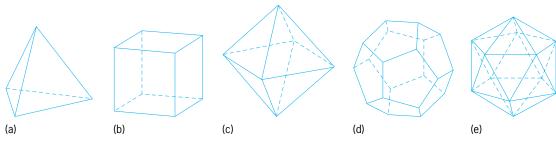
Most polyhedra met in applied geometry are convex and simply connected. A polyhedron is convex if it passes this test: if any face is placed coincident with a plane, then all other points of the polyhedron lie on the same side of that plane. A more informal test is to imagine enclosing the polyhedron within a stretched elastic membrane; the polyhedron is convex if all points on the boundary are in contact with the membrane. A simply connected polyhedron has a boundary that is topologically equivalent to a sphere: if the boundary were made of some perfectly elastic material, then the boundary could be distorted into a sphere without tearing or piercing the surface. A simply connected polyhedron is said to be eulerian, because the number of faces F, the number of edges E, and the number of vertices Vsatisfy Euler's formula (1). The remainder of this ar-

$$F + V = E + 2 \tag{1}$$

ticle will deal only with simply connected convex polyhedra.

A polyhedron has a volume, a measure of the amount of space enclosed by the boundary surface. As noted in cases below, often the volume can be determined if certain other measurements associated with the polyhedron are known.

The fewest number of faces that a polyhedron might have is four. "Poly" signifies "many," but since an *n*-hedron (a polyhedron having *n* faces) can be transformed into an n + 1-hedron by truncating the polyhedron near a vertex, polyhedra exist having any number of faces greater than three. Some polyhedra have names that convey the number of faces (but not the shape) of the polyhedron: tetrahedron, 4 faces; pentahedron, 5 faces; hexahedron, 6 faces; octahedron, 8 faces; dodecahedron, 12 faces; and icosahedron, 20 faces.



The five regular polyhedra: (a) Tetrahedron. (b) Cube. (c) Octahedron. (d) Dodecahedron. (e) Icosahedron.

Many polyhedra have properties that allow them to be placed in one or more of the categories below.

Prismatoid. A prismatoid is a polyhedron in which all vertices lie on exactly two parallel base planes. A base plane might contain only one or two vertices, but if it contains three or more, then that polygonal face is a base of the prismatoid. The height (or altitude) of a prismatoid is the perpendicular distance between the base planes.

Any prismatoid satisfies the necessary requirements for its volume to be given by the prismoidal volume formula: if b denotes the height of the prismatoid, L and U denote the areas of the bases (where one of these areas might equal zero), and M is the area of a plane section parallel to a base and midway between the base planes, then the volume, V, is given by Eq. (2). (The formula is not especially useful in

$$V = \frac{1}{6}b(L + 4M + U)$$
(2)

practice because seldom is the area of the midsection known or easily determined.) Some solids other than prismatoids also have volumes given by this formula. *See* SOLID (GEOMETRY).

A prismoid is a prismatoid that satisfies some additional requirement. However, there is no agreement on precisely what that requirement is, so any use of the word should be accompanied by the user's choice from the many definitions appearing in dictionaries.

Prism. A prism is a prismatoid that has two congruent bases and whose other faces (the lateral faces) are all parallelograms. If the lateral faces are rectangles, then the prism is right; otherwise it is oblique. If *B* represents the area of either base and *b* represents the height, then the volume, *V*, of the prism is given by Eq. (3). A prism can be classified according

$$V = Bb \tag{3}$$

to the type of polygonal region forming the bases: a triangular prism, a square prism, a hexagonal prism, and so forth. *See* PRISM.

Parallelepiped. A parallelepiped is a prism in which the bases are parallelograms. If the bases and all lateral faces are rectangles, then the prism is a rectangular parallelepiped or, more informally, a rectangular box. If l and w denote the length and width of a base, and b is the height of a rectangular parallelepiped, then the volume, V, is given by Eq. (4).

V

$$I = lwb$$
 (4)

Cube. A cube is a rectangular parallelepiped in which all six faces are squares. If *s* denotes the length of an edge of the cube, then the volume, *V*, is given by Eq. (5).

$$V = s^3 \tag{5}$$

See CUBE.

Pyramid. A pyramid is a prismatoid in which one base plane contains a single vertex: the vertex of the pyramid. The polygonal region defined by the vertices on the other base plane is the base of the pyramid. The base might contain any number of sides greater than two, but the other faces (the lateral faces) will all be triangular regions. A pyramid may be classified according to the shape of the base; thus a tetrahedron may also be called a triangular pyramid. A regular pyramid has a base that is a regular polygon, and its lateral faces are all congruent isosceles triangles. If *B* denotes the area of the base, and *h* is the height of the pyramid, then the volume, *V*, is given by Eq. (6).

$$V = \frac{1}{3}Bb \tag{6}$$

Frustrum. A frustrum (alternative name, frustum) of a pyramid is the solid remaining after a pyramid is sliced by a plane parallel to the base and the part containing the vertex is discarded. (The discarded part is another pyramid, similar to the original.) The bases of the frustrum are the base of the original pyramid and the new face created by the slicing. If *B* and *b* denote the areas of the bases, and *b* is the height of the frustrum, then the volume, *V*, is given by Eq. (7).

V

$$V = \frac{1}{3}b(B + \sqrt{Bb} + b)$$
(7)

Regular polyhedra. A polyhedron is regular (or platonic) if all faces are congruent and all dihedral

Regular	Number of elements*		
polyhedron	F	V	E
Tetrahedron	4	4	6
Cube (hexahedron)	6	8	12
Octahedron	8	6	12
Dodecahedron	12	20	30
Icosahedron	20	12	30

angles (the angles between adjacent faces) are equal. There are only five regular polyhedra: the tetrahedron, cube, octahedron, dodecahedron, and icosahedron (see **illus.** and **table**).

The center of each face of a cube is the vertex of a regular octahedron; the center of each face of a regular octahedron is the vertex of a cube. The cube and the regular octahedron, therefore, are said to be reciprocal polyhedra. The regular dodecahedron and the regular icosahedron also form a pair of reciprocal polyhedra. The reciprocal of a regular tetrahedron is another regular tetrahedron. *See* EUCLIDEAN GEOME-TRY; REGULAR POLYTOPES. Harry L. Baldwin, Jr.

Bibliography. H. S. M. Coxeter, Introduction to Geometry, 2d ed., 1989.

Polymer

Polymers, macromolecules, high polymers, and giant molecules are high-molecular-weight materials composed of repeating subunits. These materials may be organic, inorganic, or organometallic, and synthetic or natural in origin. Polymers are essential materials for almost every industry as adhesives, building materials, paper, cloths, fibers, coatings, plastics, ceramics, concretes, liquid crystals, photoresists, and coatings. They are also major components in soils and plant and animal life. They are important in nutrition, engineering, biology, medicine, computers, space exploration, health, and the environment. *See* CE-RAMICS; MANUFACTURED FIBER; NATURAL FIBER; SUR-FACE COATING.

Natural inorganic polymers include diamonds, graphite, sand, asbestos, agates, chert, feldspars, mica, quartz, and talc. Natural organic polymers include polysaccharides (or polycarbohydrates) such as starch and cellulose, nucleic acids, and proteins. Synthetic inorganic polymers include boron nitride, concrete, many high-temperature superconductors, and a number of glasses. Siloxanes or polysiloxanes represent synthetic organometallic polymers. *See* SIL-ICONE RESINS.

Synthetic polymers offer a savings in energy requirements over other materials such as metals. Their lightness reduces the consumption of fuel in vehicles and aircraft, and they outperform most metals when measured on a strength-per-weight basis. As properties have improved, polymers have been developed which can be readily and economically fabricated, and which can be used for engineering purposes such as gears, bearings, and structural members.

History. Most of the early discoveries in polymer science were empirical, with polymer chemists often converting naturally occurring materials into materials with enhanced properties. Charles Goodyear (1839) transformed natural rubber (from the rubber tree *Hevea brasiliensis*) to a less heatsensitive product using sulfur as a crosslinking agent, thus producing an elastic material with greater thermal stability. Alexander Parkes and John W. Hyatt (about

1868) added agents to natural materials such as cellulose and cellulose-derived materials to make them more flexible. Leo H. Baekeland (1907) produced crosslinked polymeric materials from formaldehyde and phenol, called Bakelite. *See* CELLULOSE; PHENO-LIC RESIN; RUBBER.

In the 1920s, Hermann Staudinger demonstrated that natural and synthetic polymers are not aggregates such as colloids, or cyclic compounds such as cyclohexane, but are long chainlike molecules with specific terminal groups. His work, for which he was awarded the Nobel Prize in Chemistry (1953), is considered the basis for modern polymer science.

In 1928, Herman Mark and Kurt Meyer used x-ray techniques to determine the crystallographic structures of cellulose and natural rubber. The following year, Wallace H. Carothers synthesized and characterized linear (aliphatic) polymers. Shortly thereafter, his group at Dupont synthesized nylon, a fiberforming polyamide. Carothers also coined many of the terms used in polymer science. *See* POLYAMIDE RESINS.

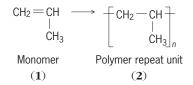
Nomenclature. Since much of polymer science was developed in industry and within a wide range of disciplines, many polymers have both a common name and a structure-based name specified by the International Union of Pure and Applied Chemistry (IUPAC). Many polymers are commonly known by their acronyms. *See* ORGANIC NOMENCLATURE.

For example, polystyrene or PS is named poly(1-phenylethylene) by the IUPAC, while poly(methyl methacrylate) or PMMA is named poly[(1-metho-xycarbonyl)-1-methylethylene]. *See* POLYACRYLATE RESIN; POLYSTYRENE RESIN.

Many companies use trade names to identify the specific polymeric products they manufacture. For example, Fortrel[®] polyester is a poly(ethylene terephthalate) or PET fiber. Many polymers are generically named, such as rayon, polyester, and nylon. *See* POLYESTER RESINS.

Composition

Generally, polymers have structures that can be represented by similar or identical repeat units. The repeat units are derived from smaller molecules, called monomers, which react to form the polymer. For example, propylene monomer (structure 1) and the repeat unit (2) it forms in polypropylene are shown below.



With the exception of its end groups, polypropylene is composed entirely of this repeat unit. The number of units (n) in a polymer chain is called the degree of polymerization or DP. *See* POLYVINYL RESINS. Other polymers, such as proteins, can be described in terms of the approximate repeat unit (3)

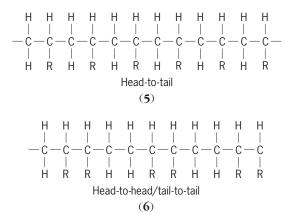
$$(3) \overset{O}{\parallel} \\ (0) \overset{O}{\amalg} (0) \overset{O}{\sqcup} \\ (0) \overset{O}{\sqcup} \\ (0) \overset{O}{\sqcup} \\ (0) \overset{O}{\amalg} (0) \overset{O}{\amalg} (0) \overset{O}{\amalg} (0) \overset{O}{\sqcup} \\ (0) \overset{O}{\amalg} (0) \overset{O}{\amalg} (0) \overset{O}{\sqcup} (0) \overset{O}{\sqcup}$$

where the nature of R (a substituted atom or group of atoms) varies. *See* PROTEIN.

Primary structure. The sequence of repeat units within a polymer is called its primary structure. Unsymmetrical reactants, such as substituted vinyl monomers (4), react almost exclusively to give

$$\begin{array}{c}
H & H \\
C = C \\
H & R
\end{array}$$
Vinyl monomer
$$(4)$$

a "head-to-tail" product (5), in which the R substituents occur on alternate carbon atoms. A variety of head-to-head structures (6) are also possible.



Each R-substituted carbon atom is a chiral center (one atom in a molecule attached to four different groups) with different geometries possible. Arrangements where the substituents on the chiral carbon are random are referred to as atactic structures. Arrangements where the geometry about the chiral carbon alternates are said to be syndiotactic. Structures where the geometry about the chiral atom have the same geometry are said to be isotactic. Both syndiotactic and isotactic polymers are stereoregular. *See* POLYMER STEREOCHEMISTRY AND PROPERTIES.

Stereoregular polymers are produced using special stereoregulating catalyst systems. A series of soluble catalysts have been developed that yield products with high stereoregularity and low chain size disparity. As expected, polymers with regular structures that is, isotactic and syndiotactic structures—tend to be more crystalline and stronger. *See* STEREOSPECIFIC CATALYST.

Polymers can be linear or branched with varying amounts and lengths of branching (**Fig. 1**). Most polymers contain some branching.

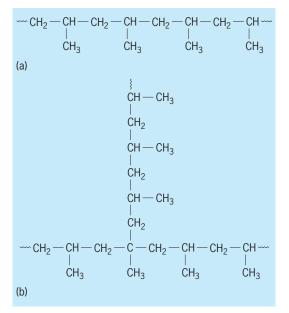


Fig. 1. Polymer structures: (a) linear, (b) branched.

Copolymers are derived from two different monomers, which may be represented as A and B. There exists a large variety of possible structures and, with each structure, specific properties. These varieties include alternating, random, block, and graft (**Fig. 2**). *See* COPOLYMER.

Secondary structure. Secondary structure refers to the localized shape of the polymer, which is often the consequence of hydrogen bonding and other hydrophilic and hydrophobic interactions. (Nonpolar regions tend to clump with other nonpolar regions, while polar regions tend to clump with other polar regions.) Most flexible to semiflexible linear polymer chains tend toward two structures—helical and pleated sheet/skirtlike. The pleated skirt arrangement is most prevalent for polar materials where hydrogen bonding can occur. In nature, protein tissue is often of a pleated skirt arrangement. For both polar and nonpolar polymer chains, there is a tendency

$$-A-B-A-B-A-B-A-B-$$
(a)

$$-A-B-B-A-B-A-A-A-B-B-B-B-A-A-B-B-$$
(b)

$$-A-A-A-A-B-B-B-B-B-A-A-A$$
(c)

$$-A-A-A-A-A-A-A-A-A$$

$$\begin{vmatrix} & & & \\ B & & B \\ & & & \\ B & & B \\ & & & \\ B & & B \\ & & & \\ B & & B \\ & & & \\ B & & B \\ & & & \\ B & & B \\ & & & \\ B & & B \\ & & & \\ B & & B \\ & & & \\ B & & B \\ & & & \\ B & & B \\ & & & \\ B & & B \\ & & & \\ B & & B \\ & & & \\ B & & B \\ & & & \\ B & & B \\ & & & \\ B & & B \\ & & & \\ B & & B \\ & & & \\ B & & B \\ & & \\ (d)$$

Fig. 2. Copolymer structures: (a) alternating, (b) random, (c) block, (d) graft.

toward helical formation with the inner core having "like" secondary bonding forces. *See* HYDROGEN BOND.

Tertiary structure. Tertiary structure refers to the overall shape of a polymer, such as in polypeptide folding. Globular proteins approximate rough spheres because of a complex combination of environmental and molecular constraints, and bonding opportunities. Many natural and synthetic polymers have "superstructures," such as the globular proteins and aggregates of polymer chains, forming bundles and groupings.

Quaternary structure. Quaternary structure refers to the arrangement in space of two or more polymer subunits, often a grouping together of tertiary structures. For example, hemoglobin (quaternary structure) is essentially the combination of four myoglobin (tertiary structure) units. Many crystalline synthetic polymers form spherulites. *See* HEMOGLOBIN.

Synthesis

For polymerization to occur, monomers must have at least two reaction points or functional groups. There are two main reaction routes to synthetic polymer formation—addition and condensation. In chaintype kinetics, initiation starts a series of monomer additions that result in the reaction mixture consisting mostly of unreacted monomer and polymer. Vinyl polymers, derived from vinyl monomers and containing only carbon in their backbone, are formed in this way. Examples of vinyl polymers include polystyrene, polyethylene, polybutadiene, polypropylene (7), and poly(vinyl chloride).

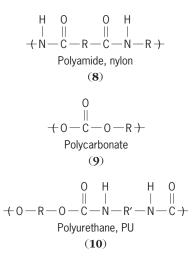


See POLYMERIZATION; POLYOLEFIN RESINS.

The second main route is a step-wise polymerization. Polymerization occurs in a step-wise fashion so that the average chain size within the reaction mixture may have an overall degree of polymerization of 2, then 5, then 10, and so on, until the en-

TABLE 1. Polymers and amounts synthesized annually in the United States			
Polymer	Type of synthesis	Amount, 10 ⁶ lb	
Polyesters	Condensation	5,500	
Ureas	Condensation	1,900	
Phenolics	Condensation	3,300	
Polyethylene	Addition	25,000	
Polypropylene	Addition	10,000	
Polystyrene	Addition	6,000	
Poly(vinyl chloride)	Addition	11,000	
Polyacrylonitrile	Addition	500	
Polyamides, nylons	Condensation	30,000	
Olefins	Addition	2,500	
Polybutadiene	Addition	1,100	
Styrene-butadiene	Addition	2,200	

tire mixture contains largely polymer with little or no monomer left. Polymers typically produced using the step-wise process are called condensation polymers, and include polyamides (8), polycarbonates (9), polyesters, and polyurethanes (10). Condensa-



tion polymer chains are characterized as having a noncarbon atom in their backbone. For polyamides the noncarbon is nitrogen (N), while for polycarbonates it is oxygen (O). Condensation polymers are synthesized using melt (the reactants are heated causing them to melt), solution (the reactants are dissolved), and interfacial (the reactants are dissolved in immiscible solvents) techniques. **Table 1** lists some addition and condensation polymers. *See* POLYURETHANE RESINS.

Properties

The properties of polymeric materials are determined by their molecular properties and morphology. Polymer morphology in turn depends on the polymerization conditions.

Molecular properties. These include molecular size and weight, molecular structure or architecture, molecular weight distribution, polarity, and flexibility of the polymeric chains (or chain segments between crosslinks in cured or vulcanized polymers). Molecular properties taken together determine the attractive forces between the molecules, the morphology or arrangement of masses of molecules, and the general behavior of the polymer.

Molecular weight. The molecular weight of a particular polymer chain is the product of the number of units times the molecular weight of the repeating unit. Polymerization reactions, both synthetic and natural, generally produce products with varying chain lengths. [The exception is the biochemical reactions that produce enzymes (proteins) and nucleic acids (such as DNA), which have precise molecular weights and chain orientations critical to their function.]

A polymer's varying chain lengths result in a product with a distribution of molecular weights. This molecular weight distribution may be narrow or broad, monomodal or polymodal, depending on the polymerization conditions.

Two statistical averages describe polymers, the number-average molecular weight Mn and the weight-average molecular weight $\overline{M}w$. The $\overline{M}n$ is particularly dependent on the number of polymer chains with a specific chain length, while the $\overline{M}w$ is dependent on the size distribution of the polymer chains. The $\overline{M}n$ is determined using procedures that are related to the number of chains of a particular length. Colligative properties are dependent on the number of particles present, and can be used to determine Mn. These techniques include membrane osmometry, vapor-phase osmometry, ebulliometry (boiling-point elevation), cryoscopy (melting-point depression), end-group analysis, and osmodialysis. Light-scattering photometry is the most often used technique for measuring $\overline{M}w$.

Since polymers generally greatly increase the viscosity (resistance to flow) of solutions containing them, viscometry is often used to determine the relative molecular weight of polymers. Today, combinations of gel permeation chromatography and light-scattering photometry can be used to determine both average molecular weight and the molecular weight of narrow molecular weight groupings. *See* LIGHT-SCATTERING TECHNIQUES; MOLEC-ULAR WEIGHT; SCATTERING OF ELECTROMAGNETIC RADIATION; VISCOSITY.

Size. Size is the most important property of polymers. Size allows storage of information (nucleic acids and proteins) and allows polymeric materials to "remember" when reversible movement occurs; in cases where movement is not reversible, polymer structures are a consequence of past treatment such as bending, stretching, melting-any action that distorts or moves polymer chains or segments). For large chains, entanglement occurs, introducing physical crosslinking and enhancing the ability of the polymers to remember and return to original molecular structures after applied stress/strain. An aspect of this remembering involves the need for segments of a polymer chain to accompany adjacent polymer segments. Thus, for space travel applications, even when the vacuum is enough to allow segmental "evaporation" of external polymer chains, some of the chain remains connected, trapped within the interior of the polymer matrix, preventing total removal. Polymeric materials are used extensively on the space shuttles.

Attractive forces. Size also accounts for an accumulation of the interchain (between molecules) and intrachain (within the same molecule) secondary attractive forces. These secondary forces, called van der Waals forces, are composed of London or dispersion forces, dipole-dipole forces, and hydrogen bonding. For a nonpolar polymer, such as polyethylene (**11**), the attractive forces for each

$$+ CH_2CH_2 +_n$$

Polyethylene, PE
(11)

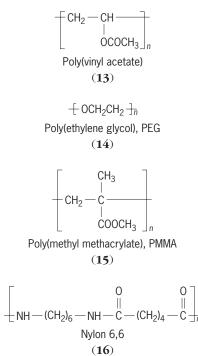
unit are about 2 kcal/mole per repeating unit. For a polyethylene chain 1000 units long, the cumulative dispersion attractive forces would be about 2000 kcal/mole, which is well above the carboncarbon bond strengths so that even low-molecularweight polymer chains are nonvolatile. As the temperature is increased, polymers will decompose prior to volatilization. *See* CHEMICAL BONDING; IN-TERMOLECULAR FORCES.

For polar polymers, such as poly(vinyl chloride) (**12**), attractive forces include both dispersion

$$\begin{array}{c} \left[\begin{array}{c} \mathsf{CH}_2\mathsf{CH} \\ | \\ \mathsf{CI} \end{array} \right]_n \\ \text{Poly(vinyl chloride), PVC} \\ (\mathbf{12}) \end{array}$$

Ρ

and dipole-dipole forces so that the total attractive forces are proportionally larger than those for polyethylene. Examples of polar polymers include poly(vinyl acetate) (13), poly(ethylene glycol) (14), and poly(methyl methacrylate) (15). For polymers that can take advantage of hydrogen bonding, the attractive forces are even greater. Such polymers include poly(vinyl alcohol), proteins, polysaccharides, nucleic acids, and nylons (16).



Hydrogen bonding is so strong in cellulose that cellulose is not soluble in water until the inter- and intrachain hydrogen bonds are broken. Thus, "wood shaping" is often accomplished by breaking the hydrogen bonds by addition of sodium hydroxide. After the wood is shaped, the sodium hydroxide is washed away and the hydrogen bonds form again, allowing the wood to retain a new shape. This principle is similar to that of permanent press clothing, where crosslinks are introduced to retain fabric shape.

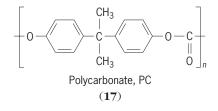
Morphology. As bonding forces and symmetry increase, the tendency to form regular structures increases, and since the inter-and intrachain forces are

inversely proportional to the square of the distance, the chains are able to reside closer together. Polymers often have a combination of ordered regions, called crystalline regions, and disordered or amorphous regions. The amorphous regions contribute to polymers' flexibility and diffusivity of gases and other materials. At low temperatures these regions are rigid, but as the temperature increases there is a temperature range where local or segmental chain mobility occurs. This temperature range is called the glass-phase transition temperature, T_g . Polymers are flexible only above their glass-phase transition temperature.

Regions where polymer chains are aligned in an orderly manner are called crystalline regions. Crystalline regions contribute to strength and resistance to external forces, including chemical attack. Crystalline regions can be melted. At the melting point or range, whole-chain movement can occur.

While crystalline regions add strength to a material, such regions are both rigid and brittle. The amorphous regions, where the chains are in a more random orientation, increase flexibility but decrease the strength of a material. Most commercial polymers have a balance between amorphous and crystalline regions, allowing a balance between flexibility and strength.

Crystallization is favored by the presence of regularly spaced symmetrical units. This is reinforced by the presence of regularly spaced polar groups that can form secondary dipole-dipole interactions. Polymers such as polyamides (nylons), polycarbonates (17), and polyesters tend to form crystalline



polymers because of the presence of additional secondary forces. Polymers such as linear polyethylene (with little or no branching) tend to be crystalline because of their symmetry and the ability of their chains to closely approach one another.

Viscoelastic materials. Polymers are viscoelastic materials, meaning they can behave as liquids and solids. When they behave as liquids, their actions can be described by Newton's law, where the applied stress is proportional to the rate of strain. For materials behaving as solids, the mechanical properties can often be described in terms of a spring using Hooke's law, where the applied stress is proportional to the resultant strain. Stress is force per unit area, and strain or elongation is the extension per unit length. *See* STRESS AND STRAIN.

A reinforced or crystallizable elastomer exhibits a relatively low breaking stress but a high elongation. Ductile polymers, such as polyethylene and polypropylene, "give" or "yield," and at high elongations some strengthening and orientation occur.

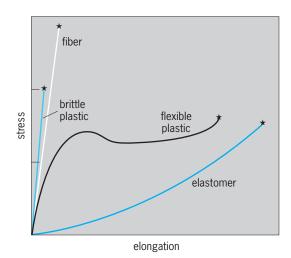


Fig. 3. Stress-elongation behavior of typical classes of polymer (not drawn to scale). The star represents the point at which failure occurs.

A brittle polymer, such as polystyrene, does not give much and breaks at a low elongation. A fiber exhibits high strength, high stiffness, and little elongation. In each case, the particular property is dependent on the temperature and the rate in which the stress is applied. Thus, at low temperatures where segmental and whole-chain mobility does not occur, the material can act as a solid. In addition, when a stress is rapidly applied so that chain or segmental mobility is not able to engage, the material will also act as a solid. Thus, a polymer that is brittle or "hookean" at low temperature may become ductile with newtonian behavior at higher temperatures where segmental or chain mobility can occur (**Fig. 3**). *See* ELASTICITY.

When an applied stress is removed, some or all of the deformation may be recovered, with the material returning to its original shape. The term "creep" is used to describe irreversible deformation. The term "elastic memory" is used to describe situations where a polymer is heated and then deformed and then cooled rapidly enough so that the polymer chains are "frozen" in place, with the material retaining the "deformed" shape. *See* CREEP (MATERIALS).

Industrial Polymer Processes

There are about 10,000 American companies active in the general area of synthetic polymers. These companies can be divided into manufacturers, processors, and fabricators. Over 200 companies produce the bulk polymers that are used by the other two groupings of companies. These polymers are generally sent to processor companies as solid pellets in a variety of shapes, including spheres and rice-grainlike. While most manufacturers produce bulk polymers in large quantities, more companies are producing specialty polymers for specific uses.

While some processors make their own polymers, most purchase the raw polymer material from manufacturers. Processors may specialize in a particular product or group of products (such as textiles or tires) or with certain polymers, such as nylons or polyethylenes, or the production of certain forms of material, such as pipes, adhesives, coatings, composites, or sheets. They may also specialize in a particular type of processing, such as blow molding.

The majority of industrial processors are involved in the fabrication and finishing of polymercontaining products. Fabrication includes machining, forming, and fashioning.

The properties of processed polymers are dependent on the conditions of processing such as cooling rate, temperature, duration of application of heat, and pressure. The drawing or pulling of materials as they exit the processing machine can have a profound effect on the properties of a material. Materials with properties that are the same in all directions are called isotropic. Fibers and films that are pulled in one or two directions are stronger in the direction of the applied stress because of the preferential aligning of the polymer chains in the direction of the pull. Thus, fibers are much stronger in the "length" direction because this is the direction of the applied stress.

There are a number of fabrication techniques. In blow molding, air is blown into a hot tube held in a mold for the production of bottles and other hollow shapes.

In extrusion, a hot melt is forced through a hole, or die, with an opening shaped to produce the desired shape. Extrusion and molding techniques are combined to produce sheet foams. Films can also be made using extrusion when air is forced through a tube. Through control of the air pressure, films of varying thicknesses can be made. Films and sheets can also be produced using calendering, where the hot polymer is forced between tightly fitted rollers. These sheets can be formed into various shapes using thermoforming, where the heated sheet is forced against the contours of a mold by a positive pressure or vacuum.

Injection molding can be used to shape thermoplastics. The material is softened and then forced into a relatively cold mold. Pipes, films, rods, and fibers can be made using injection molding. In casting, fluid materials are poured into molds of the desired shape and then allowed to cool or cure (crosslink). In some sense, painting can be considered to be a casting operation. *See* PLASTICS PRO-CESSING.

Applications

The usefulness of polymers depends on their properties, abundance, availability, and cost, including shipping or transportation, manufacture, fabrication, and environmentally related factors. Polymer properties are related to a variety of factors, including chain size, distribution of chain sizes, resistance to chemical attack, degradation (both biological and physical), flammability, dyeability, comfort, ease of fabrication, weatherability, and strength.

Fibers. Fibers are polymer materials that are strong in one direction, and they are much longer (>100 times) than they are wide. Fibers are usually formed by pulling the fiber in one direction. This orients the polymer chains, giving the fiber preferential strength in the direction of the pull. Fibers formed from symmetrical polymers allow the close approach of the polymer chains, resulting in more effective secondary bonding. The attractive forces between chains need to be relatively high, so polymers that engage in dipole-dipole seconding bonding are preferred. Fibers are to have little or no local mobility so that their use temperature is below the T_g and melting point (T_m) . A small amount of crosslinking is generally introduced to lock in a preferred shape. Table 2 contains selected synthetic fibers.

Elastomers. Elastomers (or rubbers) are polymeric materials that can be distorted through the application of force, and when the force is removed, the material returns to its original shape; this characteristic is called memory. Memory is a result of physical or, more normally, chemical crosslinks. The driving force for the return to the original shape is entropy. In the rest state, the polymer chains in the elastomer have a high degree of disorder. As the material is

TABLE 2. Industrially important synthetic fibers			
Fiber	Definition	Uses	
Acrylic	85% or more acrylonitrile units	Carpeting, skirts, socks, slacks, sweaters, blankets, draperies	
Modacrylic	35-85% acrylonitrile units	Simulated fur, scatter rugs, stuffed toys, paint rollers, carpets, hairpieces	
Polyester	85% or greater ester units	Permanent press wear, skirts, shirts, slacks, underwear, blouses, rope, fish nets, tire cord, sails, thread	
Nylon	Polyamide	Carpeting, upholstery, tents, blouses, sails, suits, stretch fabrics, tire cord, curtains, rope, nets, parachutes	
Polyurethane	85% or more urethane units	Girdles, bras, slacks, bathing suits, pillows	
Rayon	Regenerated cellulose, with substituents replacing not more than 15% of the hydrogen's of the hydroxyl groups	Dresses, suits, slacks, tire cord, ties, curtains, blankets, blouses	
(Rayon) acetate	More than 92% OH substituted	Dresses, shirts, slacks, cigarette filters, upholstery, draperies	
Triacetate	Reaction of cellulose with acetic acid or acetic anhydride	Skirts, dresses, sportswear	
Fibrous glass		Composites, insulation	
Olefins	Polypropylene, polyethylene	Synthetic grass, rugs	

TABLE 3. Synthetic elastomers	
Elastomer	Use
Acrylonitrile-butadiene-styrene, ABS Butadiene rubber, BR Butyl rubber, IIR Chloroprene rubber, polychloroprene, CR Epichlorohydrine (epoxy copolymers) Ethylene-propylene rubbers, EP, EPDM Fluoroelastormers Ionomers, mostly copolymers of ethylene and acid-containing monomers reacted with metal ions	Oil hoses, fuel tanks, pipe, appliance and automotive housings Tire tread, hose, belts Inner tubes, cable sheathing, roofing, seals, tank liners, coated fabrics Tire and cable insulation, hoses, footwear, mechanical automotive products Seals, gaskets, wire and cable insulation Cable insulation, window strips Wire and cable insulation, aerospace applications Golf ball covers, shoe soles, weather stripping
Natural rubber, polyisoprene, NR Nitrile rubber (random copolymer of butadiene and acrylonitrile), NBR Polysulfide Polyurethanes Silicons, mostly polydimethylsiloxanes Styrene-butadiene rubber, SBR	 Tires, bushings, couplings, seals, footwear, belting Seals, automotive parts that are in contact with oils and gas, footwear, hoses Adhesives, sealants, hose binders Sealing and joints, printing rollers, tires, footwear, wire and cable covering Medical-application body parts, outer space applications, flexible molds, gaskets, seals Tire tread, footwear, wire and cable covering

distorted, the polymer chains become more organized. When the distorting force is released, the elastomer returns to the original disordered state, taking advantage of the increase in disorder or entropy that results when it moves from a situation of greater order to a situation of lesser order. Products with low amounts of crosslinking, such as rubber bands, are more easily deformed or stretched. As the amount of crosslinking increases, the force needed to distort the elastomer increases and the material becomes harder. In order for elastomers to easily stretch, the attractive forces between chains must be low and the material must be above the temperature where local segmental mobility occurs-that is, it must be above the T_g . These properties are found in many hydrocarbon-intense polymers (Table 3).

Plastics. Plastics are materials that have properties between fibers and elastomers-they are hard and flexible. Polymers that may be fibers or elastomers may also be plastics (Table 4). Crystalline nylon is a good fiber, whereas less crystalline nylon is a plastic.

Coatings and adhesives. Coatings and adhesives are generally derived from polymers that are members of other groupings. For instance, polysiloxanes are elastomers, but also are used as adhesives. Poly(methyl methacrylate) is a plastic and coatings. Coatings or coverings are generally highly viscous (low-flowing) materials. The major use of coatings is for protection. Coatings resist normal wear and tear and degradation due to rain, sunlight, snow, ionizing radiation, and temperature. They protect housing, documents, chairs, stoves, cars, tables, bridges, and so on. Coatings must adhere to the surface to which they are applied. Coatings are generally a mixture of a liquid (vehicle), binder (adhesive), and one or more colorants (pigments). They generally contain a number of additives, which provide added protection against ionizing radiation and mildew, and decreased drying time.

In comparison to coatings that are required to adhere to only one surface, adhesives are used to join two surfaces together. Many good adhesives are also good coating materials. Adhesion for both coatings and adhesives can occur for a number of reasons, including physical interlocking, chemical adhesion where primary bonding occurs between the adhesive and the surfaces being joined, secondary

Plastic	Use
Epoxies	Coatings, laminates, composites, molding, flooring
Urea-formaldehyde resins	Molding compounds, dinnerware
Phenol-formaldehyde resins	Molding compounds
Melamine-formaldehyde resins	Dinnerware, table tops
Polytetrafluoroethylene	Electrical components, nonsticking surfaces, gaskets
Polypropylene	Automotive parts, toys, housewares, appliance parts
Polystyrene	Containers, recreational equipment, housewares, appliance parts
Poly(vinyl chloride) and copolymers	Pipes, fittings, sheets, flooring materials, automotive parts
Polycarbonates	Tail light lens, bullet-resistant vests, appliance housings, signs, bottles
Polysulfones	Mechanical parts, small appliances, electrical connectors
Poly(phenylene sulfide)	Electrical and mechanical parts
Polyesters	Apparel, home furnishings
Polyethylenes	Containers, bottles, housewares, pipe and fittings
Styrene-acrylonitriles	Appliance housings, housewares
Poly(methyl methacrylate)	Signs, glazing, lighting, fixtures, automotive lenses, solar panels
Poly(phenylene oxide)	Business machine housings, electrical parts, automotive parts

bonding where hydrogen bonding or polar bonding occurs between the adhesive and bonded materials, and resistance to flow (viscosity) where movement is restricted because of the viscous nature of the adhesive material. *See* ADHESIVE.

Some industrially important adhesives and coatings are the following:

Acrylic acid and acrylic ester polymers Acrylonitrile-butadiene	Polysiloxanes Polyurethane resins Poly(vinyl acetate)
copolymers	Poly(vinyl alcohol)
Aromatic polyamides	Poly(vinyl butyral)
Butyl rubber	Poly(vinyl chloride)
Cellulose and starch	Resorcinol-formal-
derivatives	dehyde
Epoxy resins	Styrene-butadiene
Phenol-formaldehyde	copolymers
Poly(alkyl cyanacrylates) Polyamides	Unsaturated poly- ester resins
Polychloroprene	Urea-formaldehyde
Polyisobutylene	Vinyl acetate-ethyl-
Poly(methyl meth- acrylate)	ene copolymers

Laminates. The combination of an adhesive and adherent is a laminate. Laminates include products that are produced with wood as the adherent and phenolic-formaldehyde, urea-formaldehyde, epoxy, resorcinol-formaldehyde, and polyester resins as the adhesive. The major wood laminate is plywood. Laminates of paper or textile include Formica and Micarta. Laminates are also used as mechanical, electrical, and general-purpose structural materials. *See* COMPOSITE LAMINATES.

Sealants and caulks. Sealants and caulks provide barriers to gases, liquids, and solids; maintain pressure differences; and moderate thermal and mechanical shock. While adhesives are used for "load transfer" requiring high shear and tensile strengths, sealants act as insulators and shock attenuators and do not require high shear and tensile strengths.

Composites. Composites are materials that contain strong fibers (reinforcement) embedded in a matrix or resin. Composites are found in automobiles, fighter and bomber planes, the "reusable" space shuttle, shafts of golf clubs (graphite), synthetic body parts, and marine craft (fibrous glass). *See* POLYMERIC COMPOSITE.

Films. Films are two-dimensional forms of plastic that are thick enough to be coherent, but thin enough to be easily folded and creased without cracking. Sheeting is a two-dimensional form of plastic that is generally thicker than a film but generally does not easily fold or crease without cracking.

Polyblends. Polyblends are made by mixing polymeric components together. Most are heterogeneous materials in which one polymeric material is embedded in another polymer. The components of polyblends adhere using secondary forces.

Liquid crystals. Liquid crystals (LCs) are materials that undergo physical reorganization where at least one of the rearranged structures involve molecular alignment along a preferred direction, causing the material to exhibit behavior that is directionally dependent. LCs serve as one of the major building blocks of communications industries, including computers and active signs. *See* LIQUID CRYSTALS.

Ceramics. Most ceramics contain large amounts of inorganic polymeric materials. While traditional ceramics were based on natural clays, today's ceramics generally are based on synthetic materials. Ceramics are generally brittle, strong, resistant to chemicals, and high-melting. The new high-temperature superconductors are ceramic material.

Cements. The term "cement" is used to describe a wide variety of materials, such as the adhesive materials used by dentists and the concrete (portland cement) used in constructing roads and buildings. Most cements are based on polymeric materials. *See* CEMENT.

Smart materials. Most smart materials are polymeric. They are materials that react to an externally applied force (electrical, bending, pressure, temperature, light, magnetic), and then the interaction is used to achieve a defined engineering or scientific goal. Windows that regulate the amount of light that is transmitted through them as the amount of sunlight varies are an example of smart materials because the windows are part of a more complete energy-control system.

Additives

Processed polymeric materials are generally a combination of the polymer and the materials that are added to modify its properties, assist in processing, and introduce new properties. The latter materials are called additives.

Additives can be solids, liquids, or gases. Thus, for many latex paints, the polymer is poly(methyl methacrylate), titanium IV oxide is added as the pigment, and other materials are added to assist the paint to "cover" the surface, to protect the material from ultraviolet radiation degradation, and to help it resist mildew. *See* INHIBITOR (CHEMISTRY).

Some typical additives are:

Antiblocking agents Antifoaming agents Antifogging agents Antimicrobial agents Antioxidants Antistatic agents Antizonates Blowing agents Coloring agents Coupling agents Curing agents Extenders Fillers Finishing aids Flame retardants Foaming agents Impact modifiers Low-profile materials Lubricants Mold release agents Odorants/fragrances Plasticizers Preservatives Reinforcements Slip agents Softeners Stabilizers—radiation and heat Viscosity modifiers

Plasticizers. Plasticizers are added to increase the flexibility and to reduce the melt temperature range and viscosity. Most plasticizers can be considered

to be near-solvents that allow segmental mobility to occur. In natural systems, water often acts as a plasticizer, allowing proteins, nucleic acids, and polysaccharides to be flexible. For synthetic polymers, most plasticizers are designed to be nonmobile and nonvolatile.

Antioxidants. Antioxidants are added to retard oxidative degradation generally brought about by the formation of free radicals through application of heat, ionizing radiation, or mechanical shear. *See* AN-TIOXIDANT.

Colorants. Many materials contain colorants or coloring agents. Colorants are generally added to increase the esthetic value of a coating.

Fillers and reinforcements. Fillers are relatively inert materials, while reinforcements improve the properties of the material to which they have been added. For many composites, addition of reinforcing fibers produces dramatic improvements in the strength of the material. While most fillers are added to simply increase the bulk of the material, some provide needed increases in the mechanical properties of the material. Fillers are generally inexpensive, lightweight, and readily available in large quantity. Fillers include natural organic materials such as wood product discard materials, corncobs, and peanut shells. Glass of many forms as well as a number of minerals (including sand and diatomaceous earth, mica, and talc) are also used as fillers.

Recycling and Degradation

Polymers offer a wide variety of techniques for recycling and degradation. Recycling of simple polymeric materials such as bottles and containers is fairly straightforward. Many polymers are thermoplastics, that is, they can be reshaped through application of heat and pressure and used in the production of other thermoplastic materials. The recycling of thermosets, polymers that do not melt but degrade prior to softening, is more difficult. These materials are often ground into a fine powder, are blended with additives (often adhesives or binders), and then are reformed.

Most glass contains some reused glass, and much paper contains recycled paper.

The recycling of more complex objects such as tires is more difficult since they are composed of a number of different materials that must be separated before real recovery of the material is possible. *See* RECYCLING TECHNOLOGY.

The mechanism for the degradation of polymers can be divided into two general catagories biological and physical. Many microorganisms contain enzymes that are "sent out" by the organism to locate food and render it usable to the organism. Some of these organisms have enzymes that degrade the polar linkages that are present in polyesters, polyamides, and other condensation polymers. Thus, these polymers can be degraded by microorganisms. *See* BIODEGRADATION.

Physical degradation includes "simple" hydrolysis (with catalysis by the acid found in acid rain), heat, and light. Again, condensation polymers are particularly sensitive to hydrolysis. Many vinyl polymers are degraded through exposure to light as well. *See* HYDROLYSIS.

The Future

Materials in today's society must be varied to perform many tasks. Often they must perform them repeatedly and in a special manner. Plastic hinges must be able to work thousands of times. The human heart, a complex muscle largely made up of protein polymers, provides about 2.5 billion beats within a lifetime, moving oxygen throughout the approximately 144,000 km of a circulatory system and delivering about 8000 liters of blood daily with little deterioration of the cell walls. Nerve impulses travel within the body about 300 m per minute, and macromolecules are the materials that allow this rapid and precise transfer of nerve impulses. Bones are made of polymeric materials, and have a strength about five times that of steel on a weight basis. In this century, synthetic polymers should approach the properties of these biological polymers.

Polymers are part of the so-called nanorevolution where single polymer strands, one atom thick, will be the connectors; polymer sheets one molecule thick will act as adhesives and insulators for communication (including computers); small motors the breadth of 1/20th of a human hair will go from zero to 500,000 revolutions per second in a real "blink of an eye." *See* NANOTECHNOLOGY.

Catalysis for production of polymers and the use of polymers as catalysts continue. The presence of soluble stereoregulating catalysts used in the production of polymers will allow better control and enhanced properties for these materials. The use of polymers for multisite synthesis will continue to be developed. *See* CATALYSIS.

The use of superstructures will continue, including in the mimicry of nature. For instance, the mucus of snails contains protein "ball-bearing-like" spheres: efforts are under way to mimic these proteins using synthetic polymeric materials. The synthesis of biomedical materials, including skin and tendons, will continue along with the creation of materials for replacement of structural body parts.

Manufacturing companies are dividing into those that produce bulk polymers on a large scale and others that produce specialty polymers on a smaller scale. The development, use, and production of specialty polymers for specific applications will increase. The use of, and improvement in properties of, the older "standby" polymers such as polyethylene will also continue. Charles E. Carraher, Jr.

Bibliography. H. Allcock, F. Lampe, and J. Mark, *Contemporary Polymer Chemistry*, 3d ed., 2003; E. A. Grulke et al. (eds.), *Polymer Handbook*, 2 vols., 4th ed., 2003; G. Odian, *Principles of Polymerization*, 4th ed., 2004; M. P. Stevens, *Polymer Chemistry: An Introduction*, 3d ed., 1998; A. N. Wilkinson and A. J. Ryan (eds.), *Polymer Processing and Structure Development*, 1999.

Polymer composites

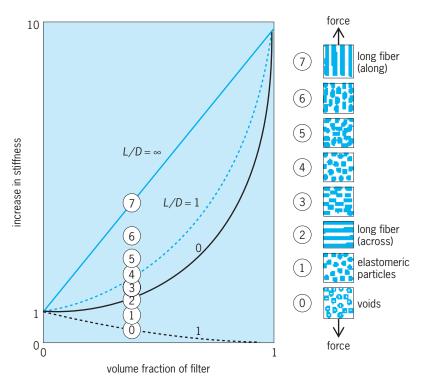
Polymer materials or matrices that are combined with fillers or fibers. Fillers are intentionally placed in polymers to make them stronger, lighter, electrically conductive, or cheaper. Any filler will affect the mechanical behavior of a polymeric material. For example, long fibers will make the material stiffer but usually denser, whereas foaming will make it more compliant but much lighter. A filler such as calcium carbonate will decrease the polymer's toughness, while making it considerably cheaper (**Fig. 1**). Polymer composites are also found in nature, such as wood (cellulose fibers bonded with lignin) and bone (minerals bonded with collagen).

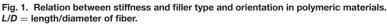
Polymer composites have become prevalent in automotive and aircraft design and manufacture. For example, newer commercial aircraft can have up to 80% of their structure made of polymer composites, including the whole fuselage. The automotive industry introduced composites in their body panel design in the early 1950s. Today, many automobiles not only have body panels that are compression- or injectionmolded out of fiber-reinforced polymers, but also polymer-composite structural components as well as many under-the-hood applications such as valve covers and oil pans. Other applications of polymer composites include filament-wound pressure vessels and pipes for the chemical industry, braided fiberreinforced epoxy tubes used in the construction of composite bicycles, and resin transfer-molded boats, skies, and snowboards, to name a few. See COMPOS-ITE MATERIAL.

Polymer matrix. In a composite structure, the task of the polymer matrix is to hold the fibers in place or bond them together into one integral structure. Polymer matrices can be either thermoplastic or cross-linked polymers. Thermoplastic polymers are divided into amorphous and semicrystalline thermoplastics, and cross-linked polymers can be either elastomers or thermosets. *See* POLYMER.

Amorphous thermoplastics. Amorphous thermoplastics, such as polystyrene, polycarbonate, and acrylics, exhibit "useful" behavior below a temperature called the glass transition temperature. An amorphous polymer will flow when subjected to temperatures above its glass transition. Amorphous polymers are brittle below their glass transition temperature and therefore fail at low strains. As the temperature increases, the strength of the amorphous thermoplastic decreases, since the thermoplastic becomes leathery in texture and is able to withstand larger deformations. Above the glass transition temperature, the strength decreases significantly, as the maximum strain continues to increase, until the flow properties have been reached, at which point the mechanical strength is negligible.

Some amorphous thermoplastics can be made highly impact resistant (less brittle) through copolymerization. The most common example is acrylonitrile-butadiene-styrene (ABS). Since butadiene chains vitrify (become glasslike) at temperatures below -50° C (-58° F), ABS is very tough at room temperature in contrast to polystyrene and acrylics





by themselves. *See* BRITTLENESS; COPOLYMER; PLAS-TICITY; STRESS AND STRAIN.

Semicrystalline thermoplastics. Semicrystalline thermoplastics, such as polyethylene, polypropylene, and polyamide, are leathery and tough at room temperature, since their amorphous regions usually vitrify at very low temperatures; for example, polypropylene has a typical glass transition temperature of -10° C (14°F). The temperature at which a semicrystalline polymer becomes useless is the melting temperature. The melting temperature and the strength and stiffness of semicrystalline plastics increase with the degree of crystallinity. The brittle behavior that polypropylene would exhibit below its glass transition temperature can sometimes pose a problem in design. However, this can be mitigated through copolymerization, as polypropylene is often copolymerized with ethylene. With semicrystalline polymers, an increase in toughness between the glass transition temperature and the melting temperature is observed. Room temperature usually falls in a range between the glass transition temperature and the melting point for most semicrystalline thermoplastics.

Cross-linked polymers. Cross-linked polymers, such as thermosets and elastomers, behave completely differently from thermoplastic polymers. A coarse cross-linked system, typical of elastomers such as natural rubber and styrene-butadiene rubber, renders a material that will no longer flow and that exhibits a low modulus (stiffness) above the glass transition temperature. The glass transition temperature of these materials is usually below $-50^{\circ}C$ ($-58^{\circ}F$), so they are soft and flexible at room temperature. On the other hand, highly cross-linked

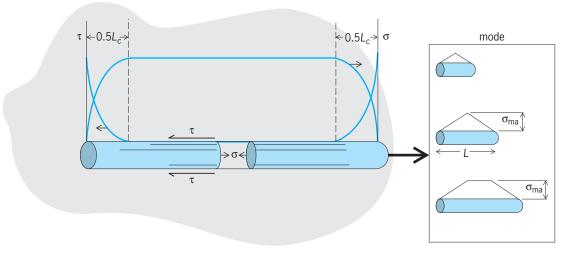


Fig. 2. Load transfer from matrix to fiber in a composite. $L_{C} =$ critical length. $\sigma_{ma} =$ matrix stress.

systems, typical in thermosets such as epoxy and unsaturated polyester, are very stiff and show only a small decrease in stiffness as the material is raised above the glass transition temperature; that is, the decrease in properties becomes smaller as the degree of cross-linking increases. The strength of a thermoset remains fairly constant up to the thermal degradation temperature of the material. *See* RUBBER.

Fiber reinforcement. In a well-designed composite material, the fiber is the load-bearing structure of a part. The purpose of introducing a fiber into a matrix is to transfer the load from the weaker material to the stronger one. This load transfer occurs over the length of the fiber (Fig. 2). The length necessary to complete the load transfer from the matrix to the fiber, without fiber or matrix fracture, usually is referred to as the critical length. The critical length is a function of fiber diameter, shear strength of the matrix, and tensile strength of the fiber. Experimental evidence suggests that fiber aspect ratios (fiber length to diameter) of 100 or higher are required to achieve maximum strength in a fiber-reinforced composite structure. If composites have fibers that are shorter than their critical length, they are referred to

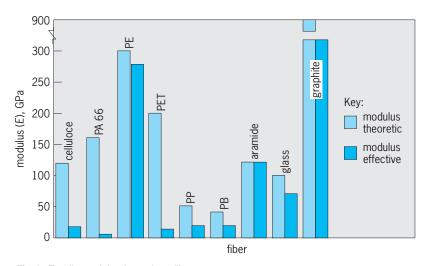


Fig. 3. Tensile modulus for various fibers.

as short-fiber composites; and if the fibers are longer, they are referred to as long-fiber composites.

Traditionally, fibers were glass or carbon. However, fibers also can be made out of polymers by orienting or aligning the long molecular chains. The properties of polymeric fibers (polyethylene and aramid) are comparable to the properties of graphite and glass fibers, differing mostly in that polymers are less dense than inorganic fibers (see **table**). *See* CAR-BON; GLASS; GRAPHITE; MANUFACTURED FIBER.

Molecular orientation and polymeric fibers. If a thermoplastic is stretched at a temperature high enough that the polymer chains can slide past each other but low enough that stresses can build up, the orientation generated during stretching is retained within the polymer component. Any degree of molecular orientation results in property variations within thermoplastic polymers. The stretching will lead to decreased strength and stiffness properties perpendicular to the orientation, and increased properties parallel to the direction of deformation. In addition, highly oriented materials tend to split along the orientation direction under small loads. *See* STRENGTH OF MATERIALS.

If the material is stretched such that a needlelike, or fibrilic, morphological structure results, the stiffness of the material will be very high. The high strength and stiffness of fibrilic structures is taken advantage of when making synthetic fibers. **Figure 3** shows the theoretical and achievable elastic moduli of various synthetic fiber materials. *See* ELASTICITY; YOUNG'S MODULUS.

High-stiffness and high-strength synthetic fibers are increasingly important for lightweight highstrength applications. Extended-chain ultrahighmolecular-weight polyethylene fibers have been available commercially only since the mid-1980s. They are manufactured by drawing or extending fibers of small diameters at temperatures below the melting point. The modulus and strength of the fibers increase with the drawing ratio or stretch. The tensile modulus of superdrawn ultrahigh-molecular-weight high-density polyethylene fibers reaches a maximum modulus of 200 GPa (the modulus of steel) at draw ratios of 250. A draw ratio is the ratio of the final length to the initial length of the fiber. In addition to amorphous and semicrystalline thermoplastics, there is a whole family of thermoplastic materials whose molecules do not relax and thus retain their orientation, even in the melt state. This class of thermoplastics is the liquid crystalline polymers. One such material is aramid fiber, commonly known as Kevlar[®], which has been available in the market for several years.

Mechanical properties. Fiber-reinforced polymers are often anisotropic; that is, they show different properties in different directions. As a result, the usual relations that are used to represent and analyze homogeneous materials are not valid. The most common model to analyze fiber-reinforced composites is the two-dimensional laminate structure (**Fig. 4**). In this simplified case, *L* and *T* define the longitudinal and transverse directions, respectively.

The longitudinal and transverse properties can be calculated using (among other models) the Halpin-Tsai model. The actual behavior of the composite is related not only to the individual properties of the matrix and the fibers but also to the volume fraction of fibers or fillers present in the composite, as well as the orientation of the loads relative to the fiber direction. In addition to the Halpin-Tsai model, several other models are in use today to predict the elastic properties of aligned fiber-reinforced laminates. Most models predict the longitudinal modulus quite accurately, but not the transverse. The longitudinal modulus is directly proportional to the fiber content for composites with unidirectional reinforcement, increasing linearly with respect to the fiber volume fraction. The increase of transverse modulus as a function of fiber content is not as dramatic, and in simple models often is approximated to be equal to the stiffness of the matrix.

The loading in a laminated structure is not always aligned with the transverse and longitudinal directions of the reinforcement. In other words, the laminate and its properties are often rotated by an angle θ . The stiffness of a laminate decreases as one rotates away from the longitudinal axis for an aligned fiberreinforced composite with different volume-fraction fiber content (**Fig. 5**). For high volume-fraction fiber contents, only a slight misalignment of the fibers from the loading direction results in a drastic reduction of the properties. This is critical in composite structures that are constructed with aligned fibers, such as aircraft components.

With help of commercially available computeraided technology (CAD), one can design and pre-

Mechanical properties of selected fibers				
Fiber	Tensile strength, MPA	Tensile modulus, GPa	Elongation to break, %	Specific gravity
Polyethylene Aramid Graphite S-glass	3000 2760 2410 4585	172 124 379 90	2.7 2.5 0.6 2.75	0.97 1.44 1.81 2.50

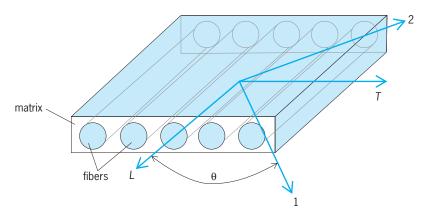


Fig. 4. Unidirectional continuous fiber–reinforced laminated structure. L = longitudinal direction. T = transverse direction.

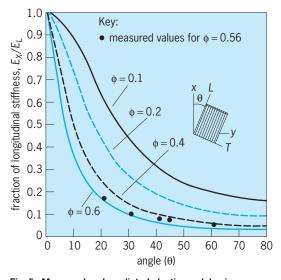


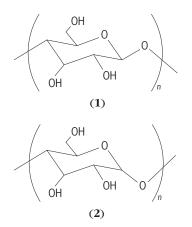
Fig. 5. Measured and predicted elastic modulus in a unidirectional fiber-reinforced laminate as a function of angle between the loading and fiber direction.

dict properties of composites with reinforcing fibers that have orientation distributions that result from flow and deformation during manufacturing. Such systems are typical of compression-molded automotive body panels, as well as injection-molded fiberreinforced components. Tim A. Osswald

Bibliography. G. W. Ehrenstein, Faserverbund-Kunststoffe, 2d ed., Hanser, Munich, 2006; S. I. Krishnamachari, Applied Stress Analysis of Plastics, Van Nostrand Reinhold, 1993; L. E. Nielsen and R. F. Landel, Mechanical Properties of Polymers and Composites, 2d ed., Marcel Dekker, 1994; T. A. Osswald, E. Baur, and E. Schmachtenberg, International Plastics Handbook, Hanser, Munich, 2006; T. A. Osswald and G. Menges, Materials Science of Polymer for Engineers, 2d ed., Hanser, Munich, 2003; S. W. Tsai, J. C. Halpin, and N. J. Pagano, Composite Materials Workshop, Technomic, 1968.

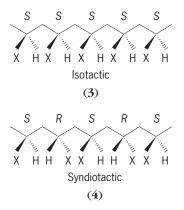
Polymer stereochemistry and properties

The properties of polymers are significantly influenced by their stereochemistry. Cellulose (structure 1) and amylose (2), the main components of



wood and starch, respectively, are among the most abundant polymers on Earth. Both are produced from D-glucose and are stereoisomers of each other. These two have different linkages between 1- and 4-carbons; the former has a $1,4-\beta$ -structure and the latter a 1,4- α -structure. This difference in the linkage of these polysaccharides leads to different higherorder structures of the polymer chains; cellulose has a stretchable polymer chain, which facilitates the regular arrangement of the polymer chains through the interchain hydrogen bond, whereas amylose has a helical chain, in which about seven glucose residues construct one turn of a helix. Because of their structural differences, their properties are also very different; for instance, cellulose is not soluble in water because of the strong interchain hydrogen bonds, but amylose is. Human beings cannot digest cellulose to D-glucose to use as an energy source, but can digest amylose or starch to p-glucose.

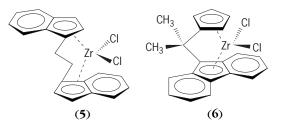
Vinyl polymers. When vinyl monomers (CH₂= CH—X) like propylene (X = CH₃), styrene (X = C₆H₅), vinyl acetate (X = OCOCH₃), and *N*-isopropylacrylamide (X = CONHCH(CH₃)₂) are polymerized, each step of the monomer addition to a propagating chain end yields an asymmetric *R* or *S* center. The continuing formation of the same center produces an isotactic polymer (**3**), and the alternating formation of the *R* and *S* centers produces a syndiotactic polymer (**4**). This kind of arrangement of side chains



has been called stereoregularity or tacticity. The control of tacticity was first realized by G. Natta in 1956 for the stereospecific coordination polymerization of propylene using the Ziegler-Natta catalyst consisting of TiCl₃—Al(C₂H₅)₃. The isotactic polymer had properties superior to those of the atactic polypro-

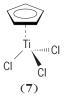
pylene, which has randomly arranged side-chain methyl groups. Isotactic polypropylene is a hard, durable material with a high melting point ($\sim 165^{\circ}$ C or 330°F) and high mechanical strength, whereas the atactic polypropylene is a greasy material. Isotactic polypropylene is one of the most widely used polymers.

Nearly 100 million tons per year of the polyolefins polypropylene and polyethylene are produced worldwide. The production of these polyolefins is steadily increasing due to their low cost and high performance. They have been used for a number of products, such as grocery bags, containers, gasoline tanks, and automobile bumpers. The Ziegler-Natta catalyst is an insoluble and heterogeneous catalyst, and therefore control of the polymerization reaction with the catalyst is difficult. This defect has been improved by the discovery of new homogeneous catalyst systems composed of metallocenes (5, 6).



By changing the ligand structures on zirconium, the tacticity of polypropylene can be varied from highly isotactic to syndiotactic.

Polystyrene is also a very popular polymer and has been industrially prepared by radical polymerization, which usually produces an atactic polymer. The polymer is used as kitchen utensils, video cassettes, and machine enclosures. Styrene is also polymerized by coordination polymerization using a catalyst containing a half metallocene (7), which gives a highly



syndiotactic polymer. The properties of this stereoregular polymer significantly differ from those of the atactic polymer. Polystyrene obtained by radical polymerization becomes soft at around 100° C (212° F), while the syndiotactic polymer has high crystallinity and a melting point as high as 270° C (520° F). Syndiotactic polystyrene has both high thermal and chemical stability and therefore can be used as an engineering plastic.

The properties of poly(vinyl alcohol) (8) are also

$$(-CH_2C -)_n \\ 0H \\ (8)$$

significantly influenced by its stereoregularity, and both the isotactic and syndiotactic polymers are known to have higher melting points than the atactic polymer. Poly(vinyl alcohol)s with different tacticities have been prepared by cutting the ether linkage of poly(vinyl ether)s obtained by cationic polymerization under various conditions. However, this method is not a practical process. The polymer is manufactured by the saponification (hydrolysis) of poly(vinyl acetate), which is obtainable only by the radical polymerization of vinyl acetate. For this polymerization, there exist almost no temperature and solvent effects on tacticity, making it is difficult to change the properties of poly(vinyl alcohol) by these processes.

N-isopropylacrylamide is also polymerized only by a radical process. The stereochemistry of this polymerization is difficult to change by the reaction conditions, as is the polymerization of vinyl acetate. Because the polymer exhibits a unique phase transition in an aqueous solution and becomes a crosslinked gel at around 32°C (90°F), it has been extensively studied without paying attention to the polymer structure, particularly the tacticity. However, recent progress in the stereocontrol during the radical polymerization of N-isopropylacrylamide using a catalytic amount of Lewis acids, such as the trifluoromethanesulfonates of yttrium (Y) and ytterbium (Yb), enables the production of isotactic-rich polymers. The phasetransition temperature of the isotactic-rich polymer decreases with an increase in the isotacticity.

Polyester and polyamide. A polyester, poly(L-lactic acid) (9), is prepared from L-lactic acid, which is man-

$$\begin{pmatrix} 0 & | & | \\ 0 & | & | \\ C & | & C \\ 0 & | & C \\ C & | & C \\ C & H_3 \\ (9) \end{pmatrix}$$

 \sim

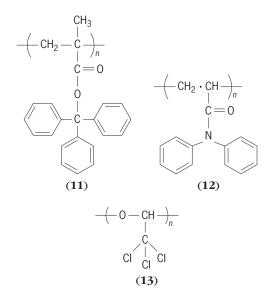
ufactured by the fermentation of glucose from starch. The polymer has the continuous L-configuration, and is a totally stereoregular and optically active polymer. This polymer has drawn much attention as being an environmentally friendly, biodegradable polymer. Its stereoisomer, poly(D-lactic acid), is prepared from D-lactic acid. Most properties of these two polymers are identical, but they differ in biodegradability. The degradation is catalyzed by enzymes, which often discriminate enantiomers like poly(L-lactic acid) and poly(D-lactic acid). To change the polymer properties, the preparation of the atactic polymer containing both L- and D-isomers is effective.

Protein analogue homopolymers, that is, poly(Lamino acids) (10), have been synthesized from



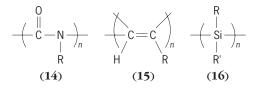
L-amino acids and are totally stereoregular polyamides. The incorporation of D-amino acids into the L-amino acid sequences yields an atactic polymer, the properties of which significantly differ from those of the corresponding homopolymers.

Higher-order structure: helicity. Most stereoregular polymers have a helical conformation at least in the solid state, and some of them can maintain the helicity in solution. It is also well known that most proteins have a partial helical conformation to exert their sophisticated functions in the living systems, and therefore some poly(t-amino acid)s also have a helical conformation in solution. Polymethacrylate (11), poly(N,N-disubstituted acrylamide) (12), and polychloral (13) are examples that can be synthe-



sized as one-handed helical polymers using optically active initiators. Due to the steric hindrance of the bulky side chains, these polymers can maintain the helical structure that is constructed via the polymerization processes. Some of the helical polymers exhibit high chiral recognition and have been used as the chiral stationary phases in high-performance liquid chromatography (HPLC) to separate enantiomers.

Polyisocyanates (14), stereoregular polyacetylenes (15) with the cis-transoid structure, and polysilanes (16) have a dynamic conformation,



which can be changed through the incorporation of a tiny chiral component or through the interaction with chiral stimuli. For instance, poly(*N*-hexyl isocyanate) (**12**, $R = n-C_6H_{13}-)$ takes a prevailing one-handed helical structure by introducing a 1% optically active R group. Poly(4carboxyphenylacetylene) (**13**, $R = -C_6H_4-COOH)$ can change its helicity through the interaction with

chiral amines. This change is readily detected by circular dichroism (CD) spectroscopy, which shows intense CD peaks in the ultraviolet (UV) and visible regions above 300 nanometers depending on the chirality of the amines. Because the amines with analogous structures induce the same CD pattern, the CD measurement can identify the absolute configuration of the chiral amines. The behavior of polysilanes (16) is similar to that of polyisocyanates (14), and the introduction of a very small amount of a chiral side chain forces the entire polymer chain to have the same helicity. See METALLOCENE CATALYST; POLYACRYLATE RESIN; POLYESTER RESINS; POLYMER; POLYMERIZATION; POLYOLEFIN RESINS; POLYSTYRENE RESIN; POLYURETHANE RESINS; POLYVINYL RESINS; STEREOCHEMISTRY. Yoshio Okamoto

Bibliography. M. Farina, Stereochemistry of linear macromolecules, *Top. Stereochem.*, vol. 17, no. 1, 1987; T. Nakano and Y. Okamoto, Synthetic helical polymers: Conformation and function, *Chem. Rev.*, 101:4013, 2001; Y. Okamoto et al., *Lewis Acidcatalyzed Tacticity Control during Radical Polymerization of (Metb)acrylamides*, ACS Symp. Ser. 854, ed. by K. Matyjaszewski, vol. 59, 2003.

Polymer-supported reaction

An organic chemical reaction where one of the species, for example, the substrate, the reagent, or a catalyst, is bound to a cross-linked, and therefore insoluble, polymer support. A major attraction of polymer-supported reactions is that at the end of the reaction period the polymer-supported species can be separated cleanly and easily, usually by filtration, from the soluble species. This easy separation can greatly simplify product isolation procedures, and it may even allow the polymer-supported reactions to be automated. Often it is possible to reuse or recycle polymer-supported reactants. Partly for this reason and partly because they are insoluble, involatile, easily handled, and easily recovered, polymer-supported reactants are also attractive from an environmental point of view.

Features. Polymer-supported reactants are usually prepared in the form of beads of about 50–100 micrometers' diameter. Such beads can have a practically useful loading only when their interiors are functionalized, that is, carry functional groups. With a typical polymer-supported reactant, more than 99% of the reactive groups are inside the beads. An important consequence is that for soluble species to react with polymer-supported species the former must be able to diffuse freely into the polymer beads.

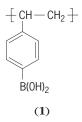
Two major morphological types of polymer beads have been used: microporous and macroporous. The former are lightly cross-linked gels, and they are significantly porous only when swollen by solvent. This property makes the choice of reaction solvent for use with a microporous polymer-supported reactant crucial, especially as the swelling properties of the beads may change substantially as the supported functional groups undergo chemical change. Macroporous beads have a highly open structure that is maintained by a high degree of cross-linking. While functional groups within the pores are accessible in many solvents, functional groups in the highly cross-linked and relatively rigid structure surrounding the pores are much less accessible. Thus, macroporous polymer beads effectively have a lower maximum capacity than microporous polymer beads. *See* GEL; POLYMER.

The high concentration of functional groups in a polymer bead can result in a microenvironment for reaction that is significantly different from that in solution, and this result can lead to distribution effects and to increases or decreases in the rates of polymer-supported reactions compared with their nonsupported analogs. A rigid, very lightly loaded polymer-supported reactant may display a degree of site isolation; that is, reactions between supported groups may be reduced or, to a limited extent, prevented. The use of a spacer arm between the functional groups and the polymer backbone can often enhance reaction rates.

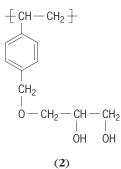
The most common polymer supports are crosslinked polystyrenes. They may be appropriately functionalized by electrophilic aromatic substitution. For less reactive electrophiles the polystyrene beads can be lithiated first. Chloromethylated polystyrene beads are easily functionalized by reaction with a wide range of nucleophiles. *See* ELECTROPHILIC AND NUCLEOPHILIC REAGENTS.

Substrates. In a typical application of reactions involving polymer-supported substrates, a substrate is first attached to an appropriately functionalized polymer support. The synthetic reactions of interest are then carried out on the supported species. Finally, the product is detached from the polymer and recovered. Often the polymer-supported species has served as a protecting group that is also a physical "handle" to facilitate separation. The fact that it is not possible to separate one polymer-supported species from another means, however, that in most cases it is usually necessary to carry out all the polymersupported reactions cleanly and in high yields. As a result of much research effort, this goal has proved possible in the extremely important cases of "solidphase" peptide and nucleotide syntheses, but such is not always the case.

Two other types of application of polymersupported substrates have proved to be of interest. The first exploits easy attachment reactions, for example, the reactions of polymer-supported boronic acid (1) with diols to give cyclic boronate

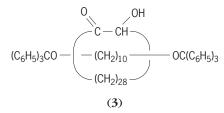


esters, and the reactions of polymer-supported diols such as polymer (2) with ketones to give ketals. Such

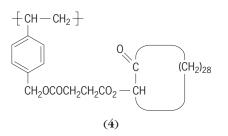


reactions can form the basis of a type of affinity chromatography. Thus, if the polymer-supported diol (2) is reacted with a mixture containing ketones, the latter will bind to the polymer through ketal formation. After washing away unreacted species, ketal formation is reversed and the ketones released, free of other types of species. *See* LIQUID CHROMATOG-RAPHY.

The second application involves isolating (fishing out) interesting species formed in small amounts. The classical example is concerned with the synthesis of rotaxane (3). The capping of decane-1,10-diol



with trityl groups was carried out in the presence of the polymer-supported cyclic compound (4). At



the time of the capping reaction, a small amount of the decane derivative was threaded through the 30-membered ring of compound (4); and because of the large terminal trityl groups, it was trapped, that is, a rotaxane was formed. The yield of rotaxane was less than 1%. However, by repeatedly reusing the polymer-supported species the amount of polymersupported rotaxane was built up to a significant level. The ring was then detached from the support by hydrolysis of the ester linkages, and rotaxane (3) was isolated.

Polymer-supported reagents. Reactions involving polymer-supported reagents are generally more useful than those involving polymer-supported substrates, because no attachment or detachment reactions are needed, and it is not necessary for all the polymer-supported species to react in high yield. Indeed, polymer-supported reagents are often used in excess to drive reactions to high conversions.

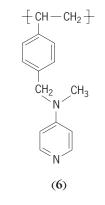
The most accessible polymer-supported reagents are various anionic forms of commercial macroporous anion-exchange resins. An example is a compound with structure (5), where X^- represents



one of the following anions: hydrogen chromate (HCrO_4^-) , borohydride (BH_4^-) , fluoride (F^-) , cyanide (CN^-) , or periodate (IO_4^-) . These can be considered as polymer-supported quaternary ammonium salts; since the support is macroporous, they can be used in a wide variety of solvents. *See* QUATERNARY AM-MONIUM SALTS.

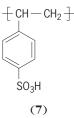
One of the most interesting applications of polymer-supported reagents involves the use of a mixture of two of them. For example, cleavage of certain 1,2-diols by IO_4^- gives rise to very sensitive dialdehydes. By using polymer-supported IO_4^- and BH₄ anions together, the sensitive products of diol cleavage were immediately reduced to stable alcohols. Since the polymer-supported oxidizing and reducing agents are on separate beads, they do not come into contact and so cannot destroy each other. However, the species in solution can pass freely between the two reagents, and the overall reaction proceeds smoothly.

Polymer-supported reagents can also be covalently bound species, and approximately a hundred polymer-supported reagents of this type have been described. These include polymer-supported carbodiimides, peroxy acids, phosphorus and sulfur ylides, acylating agents, and bases such as polymersupported dimethylaminopyridine (6).



Polymer-supported catalysts. These are the most attractive type of polymer-supported reactants; in such reactions, the loadings of catalytic sites need not be high, not all sites need be active and, in most cases, the polymer-supported catalyst is recovered in a suitable form for reuse. Accordingly a large number and variety of polymer-supported catalysts have been studied.

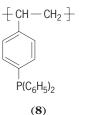
The earliest examples were polymer-supported sulfonic acids (7), that is, the acid forms of com-



mercial strong acid cation-exchange resins. These have been used to catalyze, for example, esterifications, ester hydrolyses, ketalizations and deketalizations. Many variations have been prepared that are even stronger acids, for example, fluorinated acids and complexes formed from polymer-supported proton acids with Lewis acids. Polymer-supported acids are used in a flow system for the synthesis of methyl *t*-butyl ether, the favored antiknock additive to gasoline. Polymer-supported bases have also been widely studied. *See* GASOLINE.

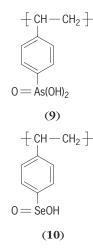
An area of major interest is polymer-supported phase-transfer catalysts, because of the relative ease of studying some of the many factors involved. Polymer-supported quaternary ammonium and phosphonium salts, polymer-supported crown ethers, and polymer-supported poly(ethylene glycols) have been used to catalyze a wide range of nucleophilic displacements and other reactions such as dichlorocarbene generation. *See* MACROCYCLIC COM-POUND; POLY(ETHYLENE GLYCOL).

An area of great commercial interest involves polymer-supported transition-metal complex catalysts for hydrogenation, hydroformylation, olefin isomerization, or olefin oligomerization, and analogs of homogeneous transition-metal complex catalysts have been prepared by replacing some of the ligands with a polymer-supported ligand. The most common ligands are polymer-supported phosphines such as the polymer-supported analog of triphenylphosphine (8). The commercial interest arises mainly



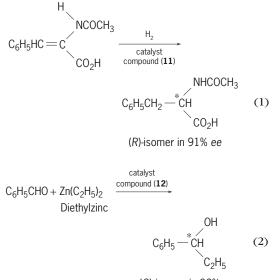
from the possibility of recovering the precious metals. The problem is that the metal tends to leach slowly from the support.

Other types of polymer-supported catalysts include polymer-supported arsonic acid (9), and polymer-supported selenic acid (10). These have



been used to catalyze epoxidations of olefins and the conversion of ketones into esters. Unlike the nonpolymeric analogs, these polymer-supported catalysts are nontoxic, because the polymers are insoluble and involatile. Polymer-supported porphyrinmanganese(III) complexes have also been used to catalyze epoxidations.

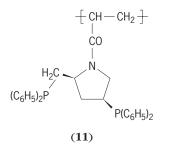
Some molecules, including many important compounds found in nature, exhibit a property known as handedness, or chirality. The structural feature commonly responsible for chirality is a tetrahedral carbon atom bearing four different substituents [indicated by an asterisk (*) in the products of reactions (1) and (2)]. Single molecules of this type can exist in two



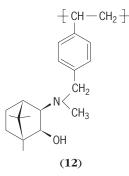
(S)-isomer in 92% ee

forms, left-handed and right-handed, known as enantiomers. In the absence of a chiral agent, enantiomers are formed in equal amounts.

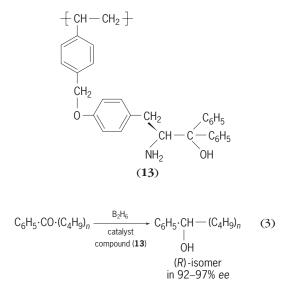
Asymmetric synthesis, that is, the synthesis of enantiomers in other than 50:50 ratio, is important, especially as a source of homochiral intermediates in the synthesis of pharmaceuticals. The most efficient type of asymmetric syntheses are those achieved by chiral catalysts. One of the more successful ones was a polymer containing the bisphosphine moieties [compound (11)], which had been treated with tris-



triphenylphosphinerhodium chloride $([(C_6H_5)_3P]_3$ -RhCl). This polymer was used to achieve the asymmetric hydrogenation in reaction (1). Other examples are the use of polymer-supported catalyst (**12**)



for reaction (2) and polymer-supported catalyst (**13**) for reaction (3).



Reactions (2) and (3) have been carried out by using the polymer-supported catalysts in flow systems. In flow systems, the functional polymer beads are placed in a tube. Solutions of other reactants are caused to flow through the tube and to react under the influence of the polymer. Under optimum conditions, the solution leaving the tube contains just the chemical product, which is thus produced continually. *See* ASYMMETRIC SYNTHESIS; CATALYSIS; OR-GANIC SYNTHESIS; STEREOCHEMISTRY. **Combinational synthesis.** In combinatorial synthesis a short linear synthesis is carried out on polymersupported substrates. This approach is being used in the pharmaceutical industry to identify lead compounds, which can thus be identified in weeks rather than years. Philip Hodge

Bibliography. A. Akelah and A. Moet, *Functionalised Polymers and Their Applications*, 1990; A. Kirsching (ed.), *Immobilized Catalysts: Solid Phases, Immobilization and Applications*, 2004; D. C. Sherrington and P. Hodge, *Syntheses and Separations Using Functional Polymers*, 1988.

Polymer thermodynamics

The relationships between properties of polymers, especially those affected by changes in temperature and the conversion of energy from one form to another. Like other materials, the bulk properties of polymers are determined by their chemical structure, but unlike many other classes of engineering materials, the range of properties that can be obtained is extremely wide. This is a consequence of the chemical and structural diversity of polymeric molecules.

Structure. Polymers are formed by linking together monomers (small molecules) to form chains of various "architectures," of which the linear chain is one example. Polymers formed from a single monomer species are called homopolymers, while those formed by the polymerization of more that one monomer species are called copolymers. In the copolymerization of two types of monomer species into a linear polymer, there are three key molecular features that may be varied independently: (1) the sequence in which the monomer units are joined, (2) the molecular weight distribution of the polymer, and (3) the tacticity (stereoregularity) of the polymer.

Possible sequences include block copolymers such as AAABBB, alternating copolymers such as ABAB, and random copolymers, such as ABBAABA, where A and B denote the two different monomers. The manner in which polymers are synthesized invariably leads to a distribution of molecular weights which are typically characterized by an average value and a quantity called the polydispersity (the width of the distribution). The tacticity of a polymer describes the orientation of functional groups attached to the backbone of the molecule. For example, the groups may be attached to one side only (isotactic); or they may alternate (syndiotactic); or they may be orientated randomly (atactic).

Phase behavior. Polymeric molecules typically have negligible vapor pressure and exist only in solid or liquid forms. However, the distinction between these phases is less clear than it is in typical substances composed of small molecules. Some polymers form a regular crystalline solid with a well-defined melting temperature T_m . However, irregularities of molecular size and structure typically lead to solid phases that are amorphous or, at most,

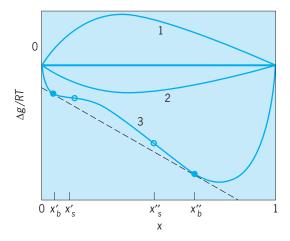


Fig. 1. Three possible forms of Δg for a binary-liquid mixture shown as a function of the mole fraction *x* of one component at constant temperature *T* and pressure *p*. (*R* = gas constant.) Solid curves show Δg for (1) a totally immiscible system, (2) a totally miscible system, and (3) a partially miscible system. On curve 3, filled circles identify the binodal points at mole fractions x'_b and x''_b , while open circles identify the spinodal points with mole fractions x'_s and x''_s . The broken line is the common tangent which connects the binodal points on curve 3.

semi-crystalline. Thermodynamic studies on such materials reveal a gradual transition from an amorphous glassy phase to a viscoelastic liquid as the temperature is raised above the glass-transition temperature T_g (the temperature above which the polymers becomes flexible). Above either T_m or T_g , the viscosity drops very rapidly with increasing temperature.

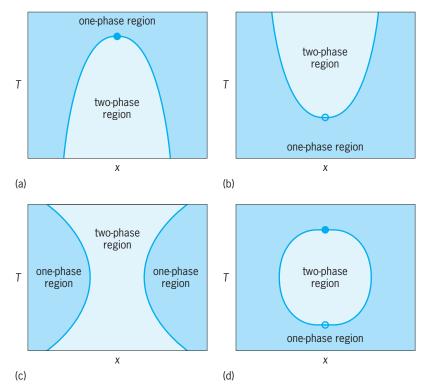


Fig. 2. Typical liquid-liquid miscibility diagrams for a binary mixture. (a) System with upper critical solution temperature (UCST). (b) System with lower critical solution temperature (LCST). (c) System without a critical solution point. (d) System with both a UCST and an LCST.

Miscibility of blends and solutions. Polymers are frequently blended in order to obtain materials with desired properties. In addition, polymers may be synthesized or processed in solution. In all these situations, the thermodynamics of mixing are of considerable importance. The most important issue is whether or not the components are miscible, and the key to understanding this is the behavior of the Gibbs free energy G of the system. For a system containing fixed amounts of the components under constant temperature T and pressure p, the fundamental thermodynamic criterion governing the number of phases present at equilibrium is that G should be minimized. Thus, a mixture held at constant T and p will split into different phases only if this leads to reduction in G. It follows that the phase behavior of a two-component (binary) mixture may be determined by examining the change of the molar Gibbs free energy upon mixing, Δg , as a function of the mole fraction x of one component. The fundamental criterion requires that, for stability as a single phase, $\Delta g(x)$ must be negative and have positive curvature with respect to the mole fraction x. Three possible scenarios are illustrated in Fig. 1. Curve 1 represents a system (for example, polymer A + polymer B or polymer + solvent) which is totally immiscible in all proportions: $\Delta g(x) \ge 0$ for all x. Curve 2 represents a system which is totally miscible in all proportions: $\Delta g(x) \leq 0$ and $(\partial^2 \Delta g / \partial x^2)_{T,p} \geq 0$ for all x. Curve 3 represents a system which is partially miscible, such that the two conditions for phase stability are satisfied only for $x \leq x'_s$ and $x \geq x''_s$, where x'_s and x_s'' are known as the spinodal compositions. Whenever this type of behavior is encountered, a miscibility gap exists within which a single homogeneous phase is not possible. Application of the fundamental criterion (that is, minimization of G) to this case shows that the miscibility gap spans the wider region $x_b'' \le x \le x_b'$ between the two binodal points. These are the points joined by a common tangent bounded from above by $\Delta g(x)$. A mixture prepared with overall composition such that $x_b'' \le x \le x_b'$ will, at equilibrium, exist as two separate phases with compositions x'_b and x''_b .

The molar Gibbs free energy change upon mixing is related to the corresponding changes Δb in molar enthalpy and Δs in molar entropy by the equation $\Delta g = \Delta b - T \Delta s$, where T is thermodynamic temperature. The entropy change is generally positive, and it does not vary greatly with molecular size. Thus the term containing Δs favors mixing, especially as the temperature is raised. The enthalpic term may be positive or negative, and its magnitude typically increases with molecular size so that, in systems containing polymers, Δb is important in determining the phase behavior. When considering how miscibility changes with temperature, a number of interesting possibilities arise (Fig. 2). As with mixtures of many ordinary liquids, the miscibility gap (that is, $x_h'' - x_h'$) may narrow with increasing temperature and eventually vanish at an upper critical solution temperature (UCST) above which the components are fully miscible (Fig. 2a). This is typical of systems in which

both Δb and Δs are positive and only weakly dependent upon *T*. Alternatively, the gap may narrow with decreasing temperature and vanish at a lower critical solution temperature (LCST; Fig. 2*b*). This situation is usually a consequence of specific interactions between the molecules (such as hydrogen bonding) which lead to negative values of Δb at low temperatures. In some cases, these two kinds of behavior combine, leading to a miscibility gap which first narrows with increasing temperature and then widens again without reaching a critical solution point (Fig. 2*c*). Occasionally, the binodal compositions define a closed loop within which the system is immiscible and one then has both a UCST and an LCST (Fig. 2*d*).

The phenomenon of partial miscibility is of course not unique to polymer blends and solutions. Many binary liquid mixtures exhibit a UCST if the system does not first freeze. The interesting features for polymer systems are that (1) partial miscibility is so common, and (2) both UCST and LCST behavior are frequently observed.

The actual process of dissolution in a polymersolvent mixture is itself interesting. A lump of polymer immersed in the solvent usually first swells as the solvent is absorbed. Sometimes the mutual solubility is slight and the process stops there with a gellike polymer-rich phase, containing some solvent, coexisting with a solvent-rich phase which may contain only a vanishingly small concentration of polymer. In a miscible system, the polymer typically proceeds to dissolve after an initial period of swelling. Often the process is very slow, even with vigorous agitation.

Vapor-liquid equilibrium. Although polymers have negligible vapor pressure, situations of vapor-liquid equilibrium arise when polymers are dissolved in low-molecular-weight solvents. The question to be addressed in this case is the value of the saturated vapor pressure p_{σ} of the solvent above a polymer solution of specified concentration. This quantity varies smoothly from the vapor pressure of the pure solvent through to zero as the weight fraction of the polymer is increased. However, the relation is found to deviate greatly from the ideal-solution theory of elementary thermodynamics.

Thermodynamic models. Because of the huge number of possible blends and solutions of polymers, much of modern polymer thermodynamics is concerned with modeling the liquid-liquid and vaporliquid equilibrium and other properties of such systems. The goal is to develop models that can predict the properties of mixtures containing polymers based on a minimum set of experimental data. Thermodynamic models fall into two broad categories: excess Gibbs free energy models and equation of state theories. The former seek to relate Δg to temperature, pressure, and composition and may be used to predict phase equilibrium, whereas the latter express the system pressure in terms of temperature, density, and composition. Equation-of-state theories also lead to Δg (and hence to predictions of phase behavior). They also permit other thermodynamic properties, such as compressibility and thermal expansivity, to be predicted. *See* COPOLY-MER; ENTHALPY; FREE ENERGY; GLASS TRANSITION; POLYMER; POLYMERIZATION; THERMODYNAMIC PRIN-CIPLES. J. P. Martin Trusler

Bibliography. F. W. Billmeyer, Jr., *Textbook of Polymer Science*, 3d ed., Wiley, Chichester, 1984; H.-G. Elias, *An Introduction to Polymer Science*, VCH, New York, 1997; J. I. Kroschwitz, *Concise Encyclopedia of Polymer Science and Engineering*, Wiley-Interscience, New York, 1990; F. Rodriguez, *Principles of Polymer Systems*, 4th ed., Taylor & Francis, Washington DC, 1996; S. I. Sandler, Models for thermodynamic and phase equilibria calculations, Dekker, New York, 1994.

Polymerization

The repetitive linking of small molecules (monomers) to make larger molecules (polymers). In general, polymers are named after the small molecules from which they are made. For example, the gaseous monomer ethylene can be polymerized to make solid polyethylene. A polymer molecule may contain hundreds of thousands repeat units, or monomers. Polymerization requires that each small molecule have at least two reaction points or functional groups. The polymerization of a monomer usually involves many steps. In each step, new bonds are formed, resulting in the extension and growth of the respective chain. Synthetic polymers are usually polydisperse; that is, in any given sample, the many polymer chains present will have varying lengths, while being of the same chemical composition. Consequently, the value describing a molecular weight or size of a polymer sample always refers to an average, such as a weight average or number average. Size and architecture of polymer molecules are of greatest importance to the polymer chemist, as they determine the ultimate material properties, and hence the necessary polymerization process.

Polymer Structures/Architectures

Depending on the particular chemistry of the polymerization process, the monomers can be combined in a variety of ways, resulting in polymers of different shapes and sizes. Possible shapes include linear, branched, ladderlike, star-shaped, arborescent, and cross-linked (**Fig. 1**).

Two samples of the same polymer, but with different architectures, will have distinctly different material properties, such as crystallinity, solubility, and mechanical strength, as well as the ability to swell, melt, or deform.

Homopolymers and copolymers. If the polymer synthesis involves only one monomer (except polycondensates from combinations of a-a/b-b type monomers; see below), the resulting polymer chains will feature only one structural repeat unit, and the polymer is called a homopolymer (**Fig. 2**).

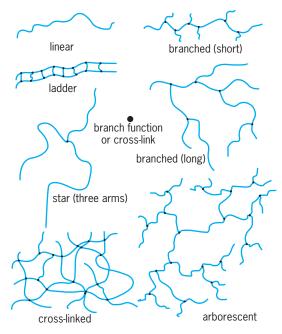


Fig. 1. Various polymer shapes/architectures.

In contrast, copolymers are derived from two or more monomers, resulting in two or more structural repeat units within the copolymer molecule (Fig. 2). Several types of copolymers are distinguished, depending on the sequential placement of the monomers, as well as the molecular architecture, for example, linear or branched.

In the case of copolymers with two different mers (repeat structures), A and B, the exact relative placement of A and B can be controlled by choosing the appropriate polymerization process, resulting in very different material properties. In the case of the random copolymer, A and B are distributed statistically along the chain. In alternating copolymers, A and B alternate along the chain, thus making it possible to view this copolymer as a homopolymer made of AB mers.

Block copolymers can be regarded as two block sequences of homopolymer, poly A and poly B, which are attached to each other. The properties of the resulting materials are a combination of the properties of the individual homopolymers, which offers a vast parameter space for macromolecular design of target materials. The number of block sequences in one chain is usually two (diblock) or three (triblock); the individual sequence length, however, can vary widely from several to several thousand mers. The synthesis of block copolymers has become increasingly important in recent years. Graft copolymers feature a homopolymer block sequences of poly B are grafted. *See* COPOLYMER.

Orientation of repeat units. In a homopolymer or a homopolymer block sequence, the monomer molecules usually have two different sides in regard to their orientation within the polymer chain: head and tail. Depending on which side of the monomer adds to another, there are three general coupling modes, head-to-tail, head-to-head, or tail-to-tail (**Fig. 3**).

Usually, the placement is head-to-tail. However, depending on the monomer structure and reactivity, other possibilities may have to be taken into consideration during the polymer synthesis. In some instances, process conditions may be chosen that favor head-to-head or tail-to-tail coupling, should this be desirable with respect to the target material.

Another variable is that every R-substituted carbon in a vinyl monomer is a potential center of

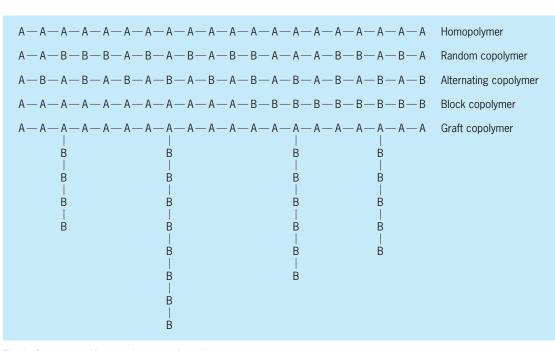


Fig. 2. Structures of homopolymers and copolymers.

asymmetry, with two possible lateral orientations of R in respect to the polymers chain, resulting in a wide range of possible material characteristics (**Fig. 4**).

If the polymerization process offers no control during the monomer addition, the placement will be statistical, resulting in an atactic polymer. The reaction conditions, for example, using an appropriate catalyst system, will allow for precise control over the placement of the R substituents. If all remain on the same side of the chain, the polymer is called isotactic. If the placement alternates the sides of the chain, the polymer is called syndiotactic.

Polymer size. The size of a polymer molecule is commonly referred to as molecular weight and given in terms of g/mol. Unlike organic molecules with a defined structure and molecular weight, polymers of the identical chemical composition can vary greatly in size, depending on how many mers constitute one polymer molecule. The general influence of molecular weight on the mechanical strength is shown in **Fig. 5**.

Even a sample of a chemically pure polymer is composed of chains of various sizes. Molecular-weight characterization is an integral part of any polymerization process. Possible analytical approaches include using the colligative properties, viscosity, and light-scattering characteristics of polymer solutions. One has to distinguish between methods relative to known calibration standards and direct measurements. Methods based on the colligative properties enable the determination of the number-average molecular weight \overline{M}_n , essentially determining the number of polymer molecules in a given sample. Viscosity measurements yield the viscosity-average molecular weight \overline{M}_{ν} , depending on the hydrodynamic volume of the polymer molecule. Modern developments in light-scattering instruments have enabled great advances in the direct determination of the weight-average molecular weight \overline{M}_w on a routine basis, which are becoming more independent of measurements relative to known calibration standards. The molecular-weight uniformity of a sample, or polydispersity, is usually characterized by the polydispersity index,

$$PDI = \frac{\overline{M}_w}{\overline{M}_n}$$

For comparison, the polydispersity of a small organic molecule, such as styrene, would have a value of PDI equal to unity, meaning the sample is monodisperse. Polymer PDI values may approach 1 in special cases of controlled/living polymerization, but usually are significantly larger than 1. *See* LIGHT-SCATTERING TECHNIQUES; MOLECULAR WEIGHT; VISCOSITY.

Polymer Synthesis

There are two distinct major types of polymerization processes: condensation polymerization, in which the chain growth is accompanied by elimination of small molecules such as water (H_2O) or methanol (CH₃OH), and addition polymerization, in which the polymer is formed without the loss of other materials.

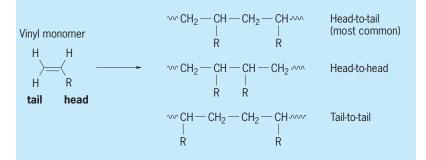


Fig. 3. Three general coupling modes of monomers.

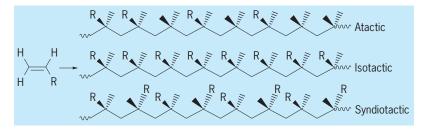
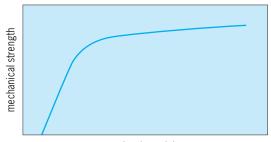


Fig. 4. Placement of the R substituents.

Condensation (step-growth) polymerization. The formation of macromolecules by the condensation process, as in the production of polyamides, polyesters, and polysulfides, requires the elimination of small molecules; at the same time, strongly polar and strongly attracting groups are produced. Most condensations proceed in a stepwise fashion. Common types of condensation polymers are polyamide, polyester, polyurethane, polysiloxane, phenolformaldehyde, urea-formaldehyde, melamine-formaldehyde, polysulfide, polyacetal, protein, wool, and silk. *See* PHENOLIC RESIN; POLYACETAL; POLYAMIDE RESINS; POLYESTER RESINS; POLYSULFIDE RESINS; POLYURETHANE RESINS; PROTEIN; SILICONE RESINS; SILK; WOOL.

An example of the condensation process is reaction (1) of ε -aminocaproic acid in the presence of a catalyst to form the polyamide, nylon-6. The repeating structural unit is equivalent to the starting material minus H and OH, the elements of water. A similar product would be obtained by reaction (2) of a diamine and a dicarboxylic acid. In both cases, the



molecular weight

Fig. 5. Dependency of mechanical strength on molecular weight.

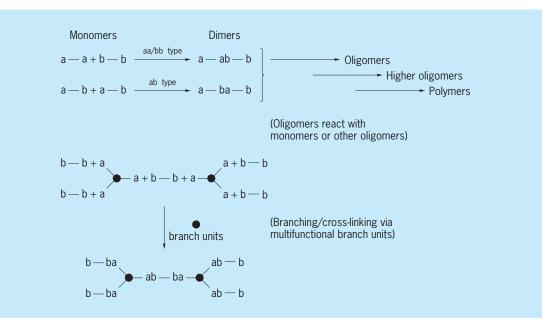


Fig. 6. General condensation behavior.

molecules formed are linear because the total functionality of the reaction system (functional groups per molecule) is always two (Fig. 1).

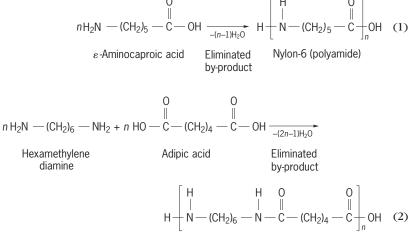
If a trifunctional material, such as a tricarboxylic acid, were added to the nylon-6,6 polymerizing mixture, a branched polymeric structure would result, because two of the carboxylic groups would participate in one polymer chain, and the third carboxylic group would start the growth of another (Fig. 1). Under appropriate conditions, these chains can become bridges between linear chains, and the polymer is cross-linked (Fig. 1).

This general condensation behavior is sketched in **Fig. 6**, where "a" could stand for a carboxylic acid function and "b" could correspond to the amine function discussed above.

The scheme also makes clear that all monomers, oligomers, and polymers in the reaction mixture pos-

sess equivalent chain end functionality, and as a consequence react with each other in a statistical fashion. This is in stark contrast to the mechanism of chain-growth polymerizations discussed below.

In the equimolar, stepwise reaction of a dicarboxylic acid a-a with a diamine b-b as an example, after 50% of the groups have reacted, the average degree of polymerization (degree of polymerization) of the polymer formed is 2; after 90% reaction, the degree of polymerization is 10; and after 99.5% reaction, the degree of polymerization is 200. If the molecular weight of the repeating unit is 100, then the average molecular weight of the polymer is 20,000. Polymers of considerable molecular weight thus are produced only at very high monomer conversions (**Fig.** 7). Furthermore, in order for the reaction to approach 100% completion and thus to



Nylon-6,6 (polyamide)

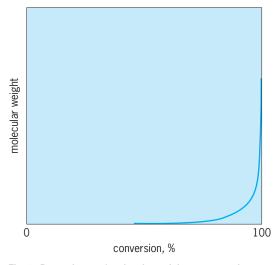


Fig. 7. Dependency of molecular weight on conversion (polycondensation).

obtain a high-molecular-weight product, the small molecules formed during the condensation must be removed from the reaction mixture.

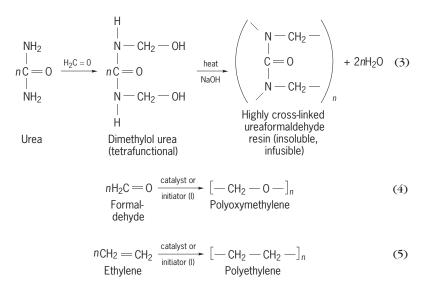
In practice, condensations are usually started under moderate conditions of temperature and pressure and completed at high temperature (removal of by-product) and low pressure to yield linear products in the molecular-weight range of about 5000–30,000. In certain cases, interfacial polymerization is useful. If reactants can be dissolved in immiscible solvents, polymerization may take place at low temperatures and resources. The linear products, thermoplastic condensation resins, are used in fibers, films, coatings, molding compounds, and adhesives.

Useful condensation polymers that are highly cross-linked are prepared from low-molecularweight polyfunctional reaction systems. The condensations of formaldehyde with phenol and with urea, and other condensations, are discussed in the Encyclopedia articles on specific products. The reaction of formaldehyde with urea is shown in (3).

In the intermediate compound, dimethylol urea is tetrafunctional. The H- and HO- groups on one molecule react with the HO- and H- groups on other molecules to form water. Each urea unit finally becomes bound to other urea units via four methylene (-CH₂-) bridges. Because the final product is cross-linked and infusible, the final shaping operation usually must coincide with the final curing or cross-linking. In practice, the soluble, low-molecularweight intermediate condensate can be isolated and mixed with strengthening fillers, coloring pigments, and curing catalysts to yield a molding powder. By subjecting the urea-formaldehyde molding powder to heat and pressure in a mold, the curing reaction takes place, some of the water is driven off as steam, and some is adsorbed by the filler. The molded object, such as a piece of dinnerware, is now insoluble and infusible. Because all the molecules are joined together, the molecular weight of a highly cross-linked polymer is a meaningless term since its value is effectively infinite.

The end groups of the polymer molecules are the functional groups that have not reacted at any stage. It is apparent that at exactly 100% conversion in a difunctional condensation, the reaction system would consist of only one molecule. The fact that the end groups of a condensation polymer can always undergo further reaction creates a difficulty in the high-temperature-melt spinning of polyamides and polyesters. To prevent subsequent changes in molecular weight, such as might occur in the melt during spinning at elevated temperatures, a monoacid or monoalcohol (molecular-weight modifier) is added to the original polymerization mixture. The excess of hydroxyl groups, for example, places a limit on the chain growth. That limit is reached when all the acid groups have reacted and all the end groups are hydroxyl.

Some examples of addition polymerization are reactions (4) and (5). The structure of the repeating unit is the difunctional monomeric unit, or mer. In the presence of catalysts or initiators, the monomer



yields a polymer by the joining together of n mers. If n is a small number, 2–10, the products are dimers, trimers, tetramers, or oligomers, and the materials are usually gases, liquids, oils, or brittle solids. In most solid polymers, n has values ranging from a few score to several hundred thousand, and the corresponding molecular weights range from a few thousand to several million. The end groups of these two examples of addition polymers are shown to be fragments of the initiator.

Addition (chain-growth) polymerization. Unsaturated compounds such as olefins and dienes polymerize without the elimination of other products. Certain ring structures, for example, lactams or alkylene oxides, may polymerize by an opening of the ring. The molecular weight and structure of the polymer are determined by the reaction conditions, that is, the nature of the catalyst or initiator, the temperature, and the concentration of reactants, monomer, initiator, and modifying agents. In general, a ceiling temperature exists above which polymerization cannot occur. Most addition polymerizations proceed by a chain reaction in which the average chain length of the polymer formed initially is high and may increase further through secondary branching reactions as the polymerization approaches completion.

The molecular-weight range for many useful addition polymers is relatively high, typically from 20,000 to several million, as compared with the molecular weight range of 5000–30,000 for typical condensation polymers.

The types of catalysis or initiation which are effective for addition (chain-growth) polymerization fall into four groups: (1) free-radical catalysis by peroxides, persulfates, azo compounds, oxygen, and ultraviolet and other radiation; (2) acid (cationic) catalysis by the Lewis acids, such as boron trifluoride, sulfuric acid, aluminum chloride, and other Friedel-Crafts agents; (3) basic (anionic) catalysis, by chromic oxide on silica-alumina, nickel, or cobalt on carbon black, molybdenum on alumina, and complexes of aluminum alkyls with titanium chloride. (4) The precise mechanisms of the fourth group are difficult to elucidate. However, evidence suggests that, at least in some cases, the mechanism involves an anionic process modified by coordination of the monomer and polymer with a surface.

It is convenient to discuss the mechanism and experimental methods of free-radical initiation as one subject, and to treat the remaining three types under the heading complex or ionic catalysts. *See* CATALY-SIS.

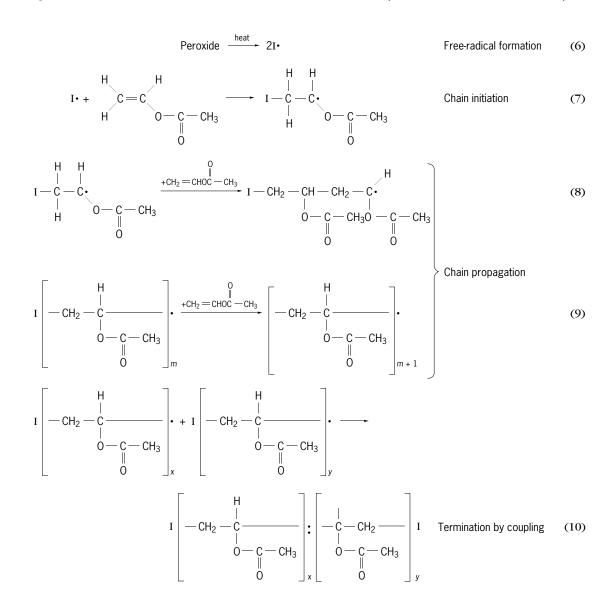
Free-radical catalysis. Among the several kinds of polymerization catalysis, free-radical initiation has been most studied and is most widely employed. Atactic polymers are readily formed by free-radical polymerization, at moderate temperatures, of vinyl and diene monomers and some of their derivatives.

At an appropriate temperature, a peroxide decomposes to yield free radicals. In the presence of a monomer, the greater proportion of these radicals adds to the monomer and thereby initiates chain growth. The growing chains may terminate by coupling, by disproportionation, or by transfer to monomer, polymer, or added materials (transfer agents, retarders, and inhibitors). Reactions (6)-(10) illustrate the initiation, propagation, and termination by coupling in vinyl acetate.

If transfer occurs to the unreacted monomer or polymer already formed, higher-molecular-weight branched structures will be produced; and if branching is excessive, insoluble products may be formed. If the radical produced in the transfer process is not sufficiently active to initiate a new chain, the transfer agent is called an inhibitor or a retarder. Mercaptans (RSH), carbon tetrachloride, and various organic solvents are examples of transfer agents, whereas amines and phenols are often used as inhibitors or retarders.

The rate of chemically initiated free-radical polymerization is increased by raising the temperature or increasing the concentration of monomer and initiator, whereas the molecular weight of the polymer is increased by increasing the monomer concentration, by lowering the temperature, and by lowering the concentration of initiator and transfer agents. *See* FREE RADICAL.

Complex or ionic catalysis. Some polymerizations can be initiated by materials, often called ionic catalysts,



that contain highly polar reactive sites or complexes. The term "heterogeneous catalyst" is often applicable to these materials because many of the catalyst systems are insoluble in monomers and other solvents. These polymerizations are usually carried out in solution from which the polymer can be obtained by evaporation of the solvent or by precipitation on the addition of a nonsolvent.

A general mechanism is shown in reactions (11),

Initiative complex:

Complex catalyst + $M \rightarrow CM^*$

Initiation:

 $CM^* + M \rightarrow M^*CM$ (11) Termination by decomposition of complex:

 $M_x^*CM \rightarrow M_x + CM$

Termination by transfer to monomer

$$M_x^*CM + M \rightarrow M_x + M^*CM$$

in which the growing chain is represented as an activated complex with the complex catalyst, without attempting to specify whether separate ions or free radicals are involved.

A distinguishing feature of complex catalysts is the ability of some representatives of each type to initiate stereoregular polymerization at ordinary temperatures or to cause the formation of polymers that can be crystallized. The polymerization process may often be visualized as the formation of an activated complex of the monomer with the complex catalyst. For stereoregular growth to take place, the entering monomer must collide with the complex and form a new, transient complex in which the new monomer molecule is held in a particular orientation. As reaction takes place, the new monomer assumes an activated condition within the complex catalyst and, at the same time, pushes the old monomer unit out. Chain growth is therefore similar to the growth of a hair from the skin. If conditions favor certain orientations of the old and new monomer units, a stereoregular polymer results. See SINGLE-SITE CATALYSTS (POLYMER).

The effect of conditions on rates of polymerization and on molecular size and structure is not yet fully understood. In general, the rate of polymerization is proportional to the concentrations of complex catalyst and monomer. The effect of temperature on the rate depends upon the stability and activity of the complex catalyst at the temperature under consideration. If the complex catalyst decomposes on increasing the temperature, the rate of polymerization will be reduced. The effect of temperature upon molecular weight also depends upon the stability of the complex catalyst and upon the relative rates of propagation and termination. In some cases, at an optimum temperature of polymerization, the molecular weight depends upon the product of the ratio of the rate of the propagation to termination and the monomer concentration, and in other cases, only upon that ratio of rates.

Examples of polymerization with the different

types of complex catalysts are briefly described below.

1. *Lewis acids*. Carbonium-ion or cationic catalysts such as BF_3 , $AlCl_3$, or H_2SO_4 usually require the presence of a promoter such as H_2O or HCl, as in reaction (12). The point of chain growth is a carbo-

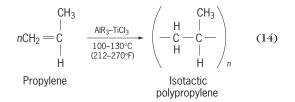
nium (positive) ion, which is associated with a negative counterion. Polymerization in the presence of Lewis acids takes place very rapidly at low temperatures, -100 to 0° C (-150° to $+32^{\circ}$ F). The order of reactivity of some olefins in Lewis acid catalysis is vinyl ethers > isobutylene > α -methyl styrene > isoprene > styrene > butadiene.

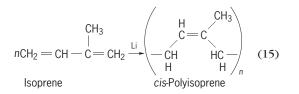
2. *Lewis bases*. Carbanions of anionic catalysts such as sodium, lithium, and lithium butyl function at moderate to low temperatures, -70 to $+150^{\circ}$ C (-94 to $+300^{\circ}$ F). Inert hydrocarbon or ether solvents are generally used as reaction media. The reaction is shown as (13).

$$\operatorname{LiC}_{4}H_{9} + \operatorname{CH}_{2} = \operatorname{CH}_{2} \longrightarrow \begin{pmatrix} H & H \\ | & | \\ C_{4}H_{9} - C - C^{-} \\ | & | \\ H & H \end{pmatrix}^{+} \operatorname{Li} (13)$$
Lithium Ethylene Initiating complex

It has been suggested that the point of chain growth is a negative or anionic center, although not usually dissociated from its positive counterion. The order of reactivity of some monomers in carbanion or anionic polymerization is acrylonitrile > methacrylonitrile > methyl methacrylate > styrene > butadiene. *See* HOMOGENEOUS CATALYSIS; REACTIVE INTERMEDIATES.

3. *Heterogeneous catalysts*. Certain heavy metals or metal oxides on supports and complexes of aluminum alkyls with titanium chloride function as catalysts at moderate to high temperatures, $50-220^{\circ}$ C ($120-430^{\circ}$ F). Inert hydrocarbon or ether solvents are generally used as reaction media. The catalysts may be used in a fixed bed or as a slurry. Two examples are shown as reactions (14) and (15).





See HETEROGENEOUS CATALYSIS.

Polymerization processes. There are a number of processes for achieving polymerization, designed for various conditions and end products.

Bulk process. This consists of polymerization of the pure monomer in liquid form. On initiation by heat or light or very small amounts of azobisisobuty-ronitrile, a very pure polymer can be formed. The monomer and polymer are poor heat conductors; therefore the temperature of bulk polymerization is difficult to control. A further disadvantage is that small quantities of unreacted monomer are difficult to remove from the polymer. Polymerization in solution offers a means of carrying out the polymerization at lower monomer concentrations. Because solvents frequently act as transfer agents, polymerization in solution generally leads to the formation of lower-molecular-weight products.

Aqueous emulsion. This type of polymerization has the advantages of giving a high rate of polymerization, a high molecular weight, and ease of temperature control. A liquid monomer is emulsified in water by use of a surface-active agent, such as soap. The soap micelles provide the polymerization centers. The free radicals (from a water-soluble initiator or growing chains of low molecular weight) diffuse into the soap micelle, in which they react to form a relatively linear polymer of high molecular weight. The polymer particles of small diameter, 50-150 nanometers, are in stable suspension because the soap of the original micelle remains adsorbed in the outer layer of the polymer particle. The rate of emulsion polymerization and the molecular weight of the polymer increase with increasing numbers of micelle particles per unit volume. The product, a stable colloidal suspension of the polymer in water, usually is called a polymer latex or polymer emulsion. Polymer latexes are used directly for water-based paints, for adhesives, and for treating textiles. When polar solvents or electrolytes are added to the colloidal suspension, the polymer coagulates, and it can be separated and dried. To produce polymer emulsions which have the desired mechanical and thermal stability, it is frequently necessary to use moderately high concentrations of surface-active agents and protective colloids. Therefore, emulsion polymers are generally less pure than bulk ones. See EMULSION POLYMERI-ZATION

Redox initiation. This system was developed for polymerization in aqueous emulsions. In the presence of a water-soluble reducing agent such as sodium bisulfite or ferrous sulfate, the peroxide decomposes more rapidly at a given temperature, and consequently polymerization at useful rates can take place at lower temperatures. By the use of the redox system, the temperature for the commercial emulsion

polymerization of styrene and butadiene is lowered from 50 to 5° C (120 to 40° F) to form "cold" rubber, which is higher in molecular weight than the "50°" or "hot" synthetic rubber.

Suspension polymerization. This system offers several advantages. By the use of a very small amount of surface-active materials and mechanical aggregation, the monomer can be dispersed as droplets in the water. The monomer is not colloidally dispersed, but is temporarily broken up into droplets which would coagulate if the stirring were discontinued. In the presence of a peroxide which is soluble in the monomer, bulk polymerization takes place in the droplets. When the polymerization reaches some 15-40%, the droplets containing the dissolved polymer become sticky and may coalesce. Various agents, such as talc and metallic oxides, have been recommended for use in very small amounts to prevent coagulation in the sticky stage. At higher degrees of conversion, the droplets will be transformed to hard balls of polymer containing dissolved monomer. At the completion of polymerization, the balls or beads settle out. These may be dried easily and are ready for use. In the suspension system, the possibility of producing a pure polymer in bulk polymerization is combined with the ease of temperature control in the aqueous emulsion polymerization. There is the additional advantage that the product of suspension polymerization can be easily isolated for use.

Ring opening. In this type, some cyclic monomers can undergo ring opening to yield high polymers; the reaction may proceed by a chain- or stepgrowth mechanism. Commercially important products are the polymerization of caprolactam, caprolactone, and ethylene oxide. *See* ACID AND BASE; ALKENE; CHAIN REACTION (CHEMISTRY); CHEMICAL DYNAMICS; INHIBITOR (CHEMISTRY); MANUFACTURED FIBER; ORGANIC REACTION MECHANISM; OXIDATION-REDUCTION; POLYMER; RING-OPENING POLYMERIZA-TION. Ralf M. Peetz; John A. Manson

Polymerizations in Supercritical Carbon Dioxide

The use of supercritical carbon dioxide as a solvent or dispersing medium for the production of polymers has the potential for offsetting the environmental concerns associated with conventional solvents. A supercritical fluid is typically a gas that above a certain temperature (referred to as the critical temperature; 88.3°F or 31.3°C for carbon dioxide) cannot be liquefied by an increase in pressure. The supercritical state exhibits properties intermediate between those of a gas and a liquid, and density-dependent properties such as solvent dielectric constant and viscosity can be tailored by simple changes in pressure or temperature. The unique properties of supercritical fluids (solvating ability similar to conventional liquids coupled with high diffusion rates and gaslike viscosities) offer many potential advantages in terms of reaction control. In addition, there are obvious environmental advantages of any technology with the potential to dramatically reduce waste streams. Supercritical carbon dioxide is compatible with a variety of chemistries, and is an attractive solvent

alternative in free-radical solution, precipitation, and dispersion processes; in cationic polymerizations; and for some metal-catalyzed polymerizations. A particularly attractive aspect of polymerizations in supercritical carbon dioxide involves the productisolation step. Since carbon dioxide is a gas at standard pressure and temperature, simply venting carbon dioxide from the reaction vessel at the end of the polymerization leaves behind the solid or liquid polymer product. In addition, the employment of supercritical fluid extraction processes provides the potential for removal of residual reactants and other contaminants from the polymer when the presence of such impurities is undesirable.

Carbon dioxide can be employed in the solution polymerization of amorphous fluoropolymers and siloxanes. For solution processes, the monomer and initiator employed, as well as the resulting polymer, are soluble in carbon dioxide. After the polymerization is complete, carbon dioxide is vented, and the polymer remains in the reactor. Two key examples are the solution polymerization of 1,1'dihydroperfluorooctyl acrylate (FOA), initiated by a free-radical initiator [reaction (16)], and the freeradical telomerization of tetrafluoroethylene [reaction (17)].

$$H_{2}C = CH \qquad \xrightarrow[dioxide]{initiator} (16)$$

$$\downarrow = 0 \qquad \xrightarrow[dioxide]{dioxide} (CH_{2} - CH -)_{n} (16)$$

$$\downarrow = 0 \\ OCH_{2}C_{7}F_{15} \qquad OCH_{2}C_{7}F_{15}$$

$$FOA \qquad Poly(FOA)$$

 $F_{2}C = CF_{2} \xrightarrow[dioxide]{group} R_{f} \leftarrow CF_{2} - CF_{2} \rightarrow I \quad (17)$ Tetrafluoroethylene

Amorphous fluoropolymers and siloxanes are the exception rather than the rule when considering polymer solubilities in supercritical carbon dioxide. Most polymers are insoluble, precluding the use of solution processes; thus application of heterogeneous methodologies such as precipitation and dispersion polymerization processes is appropriate.

Precipitation polymerization processes in supercritical carbon dioxide are useful for the preparation of many polymers. They have been employed most successfully in the preparation of water-soluble polymers such as poly(acrylic acid), which is widely used in preparation of superabsorbent materials. In this type of process, monomer and initiator are soluble in carbon dioxide, and the resulting polymer is not. Consequently, the polymer precipitates as it is formed. Venting the reaction vessel after polymerization is complete enables the isolation of dry free-flowing powders. For applications where even low levels of impurity are unacceptable, supercritical fluid extraction techniques may be applied to remove residual monomer, initiator, and other impurities. Carbon dioxide shows particular promise as a replacement for the toxic, environmentally detrimental solvents, such as benzene and methylene chloride, that have been used in the precipitation polymerization of acrylic acid.

An important development in this area involves dispersion polymerizations in supercritical carbon dioxide. In a dispersion polymerization, the monomer and initiator are soluble in carbon dioxide and the resulting polymer is insoluble, as in precipitation polymerization. However, in dispersion polymerization, specially designed surface-active agents (surfactants) stabilize the precipitated polymer as a colloid. This results in uniform spherical polymer particles similar to those achieved in conventional emulsion polymerizations. However, dispersion polymerizations conducted in supercritical carbon dioxide offer many unique opportunities resulting from carbon dioxide's ability to plasticize the growing polymer particles, thereby allowing for the formation of interpenetrating networks, core-shell materials, drug delivery systems, and so forth. The ease with which carbon dioxide can be vented provides a simple workup and isolation of the resulting polymer. In addition, carbon dioxide may find application as a propellant for aerosol coatings; thus a dispersion of polymer in carbon dioxide could be used directly in spray coating and similar applications. For applications where high purity is necessary or recovery of stabilizer is important, supercritical fluid extraction techniques may also prove useful. See CAR-BON DIOXIDE; COLLOID; CRITICAL PHENOMENA; SUR-FACTANT. T. J. Romack; J. M. DeSimone

Living/Controlled Polymerizations (Block Copolymers)

As described, chain-growth polymerizations originate from an initiator fragment and involve the sequential addition of new monomer molecules to the growing chain end. The growth of any particular chain is stopped by transfer or termination reactions. In recent years, the toolbox of macromolecular engineering has grown to include a variety of polymerization chemistries that address the transfer and termination problem by either suppression or prevention, or by rendering them reversible. The result can be a polymer chain that stops growing when the monomer is depleted in the reaction mixture, but starts growing again when new monomer is added. Hence, the molecular weight \overline{M}_n scales linearly with conversion (**Fig. 8**).

In such a living/controlled polymerization, a defined number of initiator fragments start a defined number of polymer chains, all of which ideally grow at the same rate, while no new chains are started. This results in a very uniform molecular weight distribution in the sample and a low polydispersity, often with PDI < 1.1. This enables the synthesis of polymers with defined molecular weights and high uniformity. Because any impurity in the reaction mixture has the potential to act as a transfer or termination agent, or an initiator site, particular attention has to be paid to the purity of the monomers and solvents.

Even after isolation of the polymer, the chain ends often remain functional for either a next

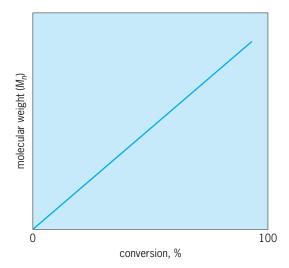


Fig. 8. Living/controlled polymerization, showing molecular weight as a function of monomer conversion.

polymerization or appropriate functionalization reactions. In the case of the living cationic polymerization of isobutylene, the initiator fragment may have two, three, or more equivalent functional groups that are able to start a growing chain. In the given example, starting from tricumyl chloride, the living cationic polymerization of isobutylene progresses in the presence of the Lewis acid TiCl₄ at a temperature of -80 to -60° C. Termination and transfer reaction are prevented by additives, such as *N*,*N*-dimethylacetamide and 2,6-di-*tert*-butylpyridine (**Fig. 9**).

After reaction completion, the result is a starshaped polyisobutylene polymer with three functional ends, also called a tritelechelic polymer. As the functional groups at the chain end remain unchanged, they can be converted into a wide variety of alternative functional groups by conventional organic chemistry, depending on the end use of the telechelic polymer. This general principle has enabled the synthesis of polymers with precisely controlled molecular weights and various chain end functionalities. The polymers can be prepared and used as macromolecular building blocks in various applications later. In (Fig. 9), the resulting end-functional polyisobutylene polymer is subjected to another polymerization, in which the functional chain ends serve as initiating sites for the living polymerization of styrene at temperatures below -50° C. The overall result is a star-shaped block copolymer, in which three block copolymer arms emanate from the trifunctional core, with polyisobutylene on the inside of the star and polystyrene on the outside of the star, the corona. Such a combination of the rubbery polyisobutylene and the glassy polystyrene demonstrates the scope of this synthetic approach to generate a variety of high-tech materials, such as for biotechnological applications. The example shows the combination of a rubbery with a glassy polymer, the two of which are incompatible and phase separate in bulk, resulting in a soluble and fusible material called a thermoplastic elastomer (TPE). Thermoplastic elastomers are physically cross-linked rubbers, in which the cross-link can be reversed by dissolving the polymer in an appropriate solvent for both blocks, or heating above the glass transition temperature (T_{q}) of the glassy polymer. See RUBBER.

Living polymerization techniques exist for monomers that are polymerized radically or anionically. Historically, this field started with the living anionic polymerization of styrene, leading to the anionic block copolymers of the KRATON[®] type, a linear triblock copolymer of polystyrene(glassy)*co*-polydiene(rubber)-*co*-polystyrene(glassy). Newer and commercially important developments include the industrially attractive radical techniques, which have developed into a wide array of synthetic polymerization methodologies, including atom-transfer radical polymerization (ATRP),

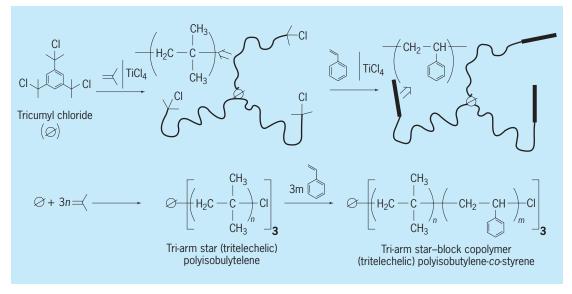


Fig. 9. Complex architectures via prevention of termination and transfer reaction.

stable free-radical polymerization (SFRP), and reversible addition-fragmentation transfer (RAFT). *See* CONTROLLED/LIVING RADICAL POLYMERIZATION; MACROMOLECULAR ENGINEERING.

Enzyme Catalysis

A new field has been receiving considerable attention both academically and industrially. Enzymes can by "recruited" for polymerization processes in different ways. One possibility is to use enzymatic catalysis to synthesize the monomer feedstock. An example is the synthesis of 1,3-propanediol from the renewable feedstock corn sugar in a fermentation reaction for the Sorona® process. The other possibility is to catalyze a wide variety of polymerization reactions under in-vitro conditions. The potential of enzymes spans from condensation polymerizations over various step-growth polymerizations. It includes the controlled synthesis of numerous examples of polysaccharides, polyesters, polyphenols, polyurethanes, and even the semiconducting polyaniline, as well as different vinyl polymers. In addition to novel polymer syntheses, enzymes are used for reactions involving the modification of natural polymers. It is even possible to synthesize block copolymer architectures.

Under in-vivo conditions, enzymes polymerize monomers with high precision in regard to the polymer structure, including molecular weight. This can be potentially exploited under in-vitro conditions, and polymer products can be obtained that would be very difficult or impossible to make by conventional polymerization methods. However, the development of such polymerization processes involve enzyme engineering—the targeted mutation of the respective enzyme by means of accelerated evolution strategies—in order to achieve the desired polymer structure. Combinatorial methods play an important role in the selection process. *See* ENZYME.

Ralf M. Peetz

Bibliography. R. F. Brady, *Comprehensive Desk Reference of Polymer Characterization and Analysis*, 2003; J. Brandrup and E. H. Immergut (eds.), *Polymer Handbook*, vol. 1: *Polymerization and Depolymerization*, 4th ed., 1999; J. A. Brydson, *Plastics Materials*, 7th ed., 1999; N. P. Cheremisinoff, *Handbook of Polymer Science and Technology*, vol. 1: *Synthesis and Properties*, 1989; R. T. Fenner, *Principles of Polymer Processing*, 1980; H. R. Kricheldorf, O. Nuyken, and G. Swift, *Handbook of Polymer Synthesis*, 2d ed., 2005; H. F. Mark et al. (eds.), *Encyclopedia of Polymer Science and Engineering*, 3d ed., 12 vols., 2003-2004; G. Odian, *Principles of Polymer Synthesis*, 2d ed., 1991.

Polymicrobial diseases

Contraction of multiple microbial infections. Such infections, which can occur in humans and animals, may be sequential or concurrent. For example, the initial infection may predispose the host to subsequent microbial infections, or the initial infecting microorganism may induce a state of immunosuppression. Concurrent infections are caused by various combinations of viruses, bacteria, fungi, and parasites. *See* BACTERIA; DISEASE; FUNGI; PARASITOLOGY; VIRUS.

Polymicrobial diseases are not new. Infections involving numerous pathogens were reported early in the twentieth century as complex infections, complicated infections, dual infections, mixed infections, secondary infections, co-infections, synergistic infections, or concurrent infections. In some instances, they were referred to as complexes or syndromes. In humans, acute necrotizing ulcerative gingivitis and respiratory diseases were recognized very early as having a polymicrobial etiology. In the 1920s, Haemophilus influenzae or Streptococcus pneumoniae were routinely found in individuals with viral respiratory disease. A similar relationship was seen during the influenza pandemics in 1918, 1957, and 1968-1969. In animals, foot rot and chronic nonprogressive pneumonia in sheep were among the first diseases found to have multiple etiologies. Respiratory disease in cattle was later shown to involve multiple infections by microorganisms. See INFECTION; INFECTIOUS DISEASE; PATHOGEN.

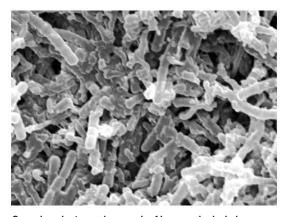
Interactions leading to diseases. Microorganisms can interact in a variety of ways, causing both acute and chronic diseases in humans and animals (Tables 1 and 2). One microorganism generates a niche for other pathogenic microorganisms to colonize; the presence of one microorganism predisposes the host to colonization by other microorganisms; two or more nonpathogenic microorganisms together cause pathology; or one microorganism interferes with infection or colonization of other microorganisms. Many of these polymicrobial interactions occur within biofilms that form on natural or artificial surfaces within the human host. Biofilms are sessile aggregate communities of bacteria attached to surfaces in hydrated polymeric matrices of exopolysaccharides. Organisms in biofilms are very hardy and resistant to antimicrobial agents as well as to host innate and acquired immune defenses. See BIOFILM.

Synergistic infections. Synergism results when one microorganism creates an environment that favors the infection and colonization of other, often pathogenic microorganisms. Recent examples are the concurrent infection of human metapneumovirus and coronavirus in severe acute respiratory syndrome (SARS) and concurrent infection of human metapneumovirus and respiratory syncytial virus in bronchiolitis and other respiratory infections. Other examples include measles virus, human T-lymphotropic virus type I (HTLV-I), HTLV-II, and human immunodeficiency virus (HIV), which induce immunosuppression favoring secondary infections with other viruses, bacteria, fungi, and protozoa. Concurrent Epstein-Barr virus and retrovirus infections may result in multiple sclerosis, but this is still to be confirmed.

Infections predisposing to disease. The presence of one microorganism may predispose the host to

TABLE 1. Combinations of microorganisms in polymicrobial diseases of humans	
Disease	Etiologic agents
Synergistic polymicrobial infections	
Severe acute respiratory syndrome	Metapneumovirus and coronavirus
Bronchiolitis Measles	Metapneumovirus and respiratory syncytial virus Measles and Mycobacterium tuberculosis and Staphylococcus aureus
AIDS	HTLV-I, HTLV-II, and/or HIV-1, HIV-2; HIV and Mycobacterium tuberculosis; HBV or HCV and HIV-1; HIV and enteric viruses, Acinetobacter radioresistens, Mycobacterium tuberculosis, Ehrlichia chaffeensis, Candida albicans, Histoplasma capsulatum, Cryptosporidium parvum, Trichomonas vaginalis, and others
Multiple sclerosis	Epstein-Barr virus and retrovirus (unconfirmed)
Infections predisposing to polymicrobial disease	
Respiratory disease	Influenza viruses, parainfluenza viruses, respiratory syncytial viruses, adenoviruses, measles viruses, rhinoviruses and coronaviruses with Streptococcus pneumoniae, Streptococcus pyogenes, Haemophilus influenzae, Staphylococcus aureus, Neisseria meningitidis, Mycobacterium tuberculosis, or Bordetella pertussis
Otitis media	Respiratory tract viruses and bacterial infections
Invasive streptococcal disease	Varicella-zoster virus and Streptococcus pyogenes
Additive polymicrobial infections	
Periodontal infections Pertussis	Over 500 bacterial species and fungi; some viruses are suspect Nasopharynx can be colonized with species of pathogenic bacteria in addition to <i>Bordetella pertussis</i>
Lung abscesses	Nocardia asteroides and Cryptococcus neoformans
Tuberculosis	Herpes zoster and Mycobacterium tuberculosis
Cystic fibrosis	Pseudomonas aeruginosa, Stenotrophomonas maltophilia, Prevotella oris, Fusobacterium gonidoformans, Bacteroides fragilis, Leptotrichia-like spp., Abiotrophia defecta, Citrobacter murliniae, Lautropia mirabilis, and Sarcina ventriculi
Peritonitis	Aerobic and anaerobic gram-positive and gram-negative bacteria
Hepatitis	Any of hepatitis B, C, and D viruses; also hepatitis C virus and HIV
Gastroenteritis	Norwalk-like virus and Aeromonas sobria or Escherichia coli
Urinary tract infection	Combinations of Corynebacterium urealyticum, Gardnerella vaginalis, Anaerococcus lactolyticus, Bacteroides vulgatus, Dialister invisus, Fusobacterium nucleatum, Lactobacillus iners, Leptotrichia amnionii, Prevotella buccalis, Prevotella ruminicola, Rahnella aquatilis, and Streptococcus intermedius
Endocarditis	Staphylococcus, Streptococcus, and the HACEK group
Microbial interference	
AIDS	Flavivirus interferes with HIV infection
Staphylococcus aureus-related disease	Steptococcus pneumoniae suppresses the carriage of S. aureus

Disease	Etiologic agents
Synergistic polymicrobial infections	
Immunosuppression in poultry	Infectious bursal disease virus, hemorrhagic enteritis virus, chicken anemia virus, avian leukosis virus, reticuloendotheliosis virus, and avian reovirus
Immunosuppression in cattle	Bovine respiratory syncytial virus and bovine diarrhea virus
Immunosuppression in pigs	Porcine reproductive and respiratory syndrome virus
Infections predisposing to polymicrobial disease	
Bovine respiratory disease complex	Respiratory infections in cattle involving 13 different viruses followed by secondary infection with one or more of 12 different bacteria
Porcine respiratory disease complex	Respiratory infections in swine involving 5 different viruses followed by secondary infection with one or more of 8 different bacteria
Additive polymicrobial infections	
Porcine atrophic rhinitis	Nasal infections in swine involving Pasteurella multocida and Bordetella bronchiseptica
Ovine conjunctivitis	Eye infections in sheep with Branhamella ovis, Mycoplasma conjunctivae, and Chlamydia psittaci
Ovine foot rot	Hoof infections in sheep with <i>Dichelobacter nodosus</i> and mixtures of other bacteria
Ovine pneumonia	Lung infections in sheep involving Mycoplasma ovipneumoniae often with mixtures of other viruses and bacteria
Bovine gastroenteritis	Enteric in cattle infections involving Salmonella enterica serovar Dublin, S. enterica serovar Typhimurium, coronavirus, and bovine diarrhea virus
Porcine gastroenteritis	Enteric infections involving bovine diarrhea virus and transmissible gastroenteritis virus
Poultry enteritis mortality syndrome	Enteric infections in turkeys involving coronavirus and Escherichia coli
Infectious coryza	Eye infections in poultry with Haemophilus paragallinarum and Ornithobacterium rhinotracheale



Scanning electron micrograph of human gingival plaque showing a mixed biofilm of rods and cocci. (Courtesy of John E. Laffoon and Janet M. Guthmiller, Dows Institute for Dental Research, College of Dentistry, University of Iowa, Iowa City)

colonization or infection by a second microorganism. Respiratory viral infections are notorious for predisposing humans and animals to secondary bacterial infections. In humans, influenza virus increases the incidence of bacterial pneumonia by *Staphylococcus aureus*, *Haemophilus influenzae*, and *Streptococcus pneumoniae*. In addition, influenza virus and respiratory syncytial virus increase the incidence of bacterial otitis media. In cattle, 13 different viruses can predispose the bovine respiratory tract to coinfection by any one of 12 different bacteria. In pigs, 5 different viruses can predispose the porcine respiratory tract to co-infection by any one of 8 different bacteria.

Additive infections. Multiple pathogens are involved in the etiology of bacteremia, abdominal abscess/secondary peritonitis, lung abscess, odontogenic infections, brain abscess/subdural empyema, chronic otitis media/mastoiditis, liver infections, and soft tissue infection/fasciitis. Human gingival plaque mixed biofilms (see illustration) contain interdependent microorganisms such as Streptococcus gordonii, S. mitis, S. oralis, S. sanguis, Haemophilus parainfluenzae, Propionibacterium acnes, Veillonella atypica, Actinomyces naeslundii, Fusobacterium nucleatum, Actinobacillus actinomycetemcomitans, Porphyromonas gingivalis, Treponema denticola, and Prevotella intermedia. Likewise, in many human abscesses there are extensive numbers of aerobic, anaerobic, gram-positive, and gramnegative bacteria.

Microbial interference. Some microorganisms can actually suppress the colonization of other microorganisms. GB virus C (GBV-C) is a flavivirus which is not known to be pathogenic in humans. However, infection with GBV-C is associated with prolonged survival among HIV-positive people. The reason for this association is not known, but likely is due to a concept called viral interference.

Microbial interference also occurs among colonized bacteria. For example, the presence or carriage of *Steptococcus pneumoniae* also suppresses the carriage of *Staphylococcus aureus* in children. Attempts to reduce the carriage of *S. pneumoniae* by vaccinating children with pneumococcal conjugate vaccines not only reduces the nasopharyngeal carriage of vaccine-type *S. pneumoniae* but results in a shift toward the carriage of nonvaccine *S. pneumoniae* serotypes or toward the presence of higher carriage rates of *Staphylococcus aureus*, including the methicillin-resistant form.

Future research areas. Polymicrobial diseases are serious diseases whose etiologic agents are sometimes difficult to diagnose and treat. Future work will undoubtedly focus on improving the identification of multiple microorganisms in polymicrobial diseases, defining microbe-to-microbe and microbe-to-host interactions, and improving prophylaxis and treatment of these diseases.

The diverse microbiota infecting and colonizing humans is currently being defined by metagenomic analyses, quantitative 16S ribosomal DNA hybridization assays, real-time polymerase chain reaction, terminal restriction fragment length polymorphism, and length heterogeneity PCR. Flow-cells are being used to study microbial interactions in biofilms. These techniques, however, cannot presently be utilized in clinical laboratories to assist clinicians in daily practice. Clinical microbiology laboratories presently identify polymicrobial intra-abdominal or soft-tissue infections as fecal flora or mixed grampositive flora and struggle to identify viral agents in mixed respiratory tract infections. As a result, critical-care physicians have trouble justifying invasive sampling techniques such as broncho-alveolar lavage to identify the true etiology of mixed pulmonary infections. See MICROBIOTA (HUMANS).

Traditional therapies of many of these diseases are generally targeted at individual etiologic agents without consideration for effect on the polymicrobial communities. For example, treatment with broad-spectrum antibiotic agents may kill anaerobic flora in the intestinal tract, favoring infection and colonization of multiresistant bacteria such as the vancomycin-resistant enterococcus. In some cases, however, there are beneficial consequences in that treatment of the primary infectious agent may reduce or eliminate infection by a concurrent or sequential agent. In Africa, tuberculosis is prominent among patients with HIV infection. The use of antiretroviral drugs reduces the incidence of tuberculosis to the level observed immediately after HIV seroconversion. See ACQUIRED IMMUNE DEFICIENCY SYNDROME (AIDS); TUBERCULOSIS.

Probiotics are novel prophylactic or therapeutic strategies for managing polymicrobial diseases. *Saccharomyces boulardii*, lactobacilli, or enterococci prevent or reduce the duration of gastroenteritis caused by bacteria and rotaviruses, and treatment with lactobacilli such as *Lactobacillus fermentum*, *L. rhamnosus*, or *L. crispatus* have the potential to reduce vaginal infections. Future studies are needed to determine the probiotic mechanisms of microbial inhibition as well as the optimal conditions for use.

Summary. Polymicrobial diseases have a multifactorial etiology and develop as a result of complex

interactions among environmental factors, host factors (genetic predisposition, stress, nutritional state, etc.), and exposure to pathogens. Understanding the role of predisposing factors, interactions among etiologic agents, and host responses to infection may ultimately assist in better strategies for prophylaxis and/or treatment. Kim A. Brogden; Janet M. Guthmiller

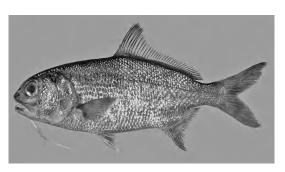
Bibliography. L. O. Bakaletz, Developing animal models for polymicrobial diseases, *Nat. Rev. Microbiol.*, 2:552-568, 2004; K. A. Brogden and J. M. Guthmiller (eds.), *Polymicrobial Diseases*, ASM Press, Washington, DC, 2002; K. A. Brogden and J. M. Guthmiller, Polymicrobial diseases, a concept whose time has come, *ASM News*, 2003; K. A. Brogden, J. M. Guthmiller, and C. E. Taylor, Human polymicrobial infections, *Lancet*, 365:253-255, 2005; H. Smith, The role of microbial interactions in infectious disease, *Pbil. Trans. Roy. Soc. Lond. B Biol. Sci.*, 297:551-561, 1982.

Polymixiiformes

The only order of the superorder Polymixiomorpha, comprising one family (Polymixidae) with one genus (*Polymixia*) and ten extant species. Several fossil genera are known from the Upper Cretaceous. Polymixiidae was originally placed with the Beryciformes; however, recent evidence suggests that the family may be the sister group to all other acanthomorphs (spiny-rayed fishes). The family is known commonly as beardfishes. *See* BERYCIFORMES; OSTE-ICHTHYES; TELEOSTEI.

The beardfishes are so called for the pair of long hyoid (between the lower jaws) barbels (see **illustration**). They are further distinguished by a moderately elongate and compressed body; subocular shelf, orbitosphenoid, and basisphenoid present; two supramaxillae; four branchiostegal rays; a continuous dorsal fin, with 4–6 spines and 26–38 soft rays; an anal fin with 4 short spines and 13–17 soft rays; subabdominal pelvic fins with one spinelike ray and six soft rays; 16 branched caudal rays; about 33–38 lateralline scales; and two sets of intermuscular bones, which is a condition otherwise unknown among the acanthomorphs.

Polymixiids are bathydemersal [living and feeding



Pacific beardfish (Polymixia berndti). (Photo by Jack E. Randall)

on the bottom below 200 m (660 ft)] fishes to depths of 700 m (2300 ft), occurring in warm temperate to tropical waters of the Atlantic, Indian, and Western Pacific oceans. They feed on other fishes, worms, and crustaceans and attain a maximum length of 38 cm (15 in.). Herbert Boschung

Bibliography. R. Froese and D. Pauly (eds.), FishBase, World Wide Web electronic publication, version (05/2005); G. D. Johnson and C. Patterson, Percomorph phylogeny: A survey of acanthomorphs and a new proposal, *Bull. Mar. Sci.*, 52(1):554-626, 1993; J. A. Moore, Polymixiidae, pp. 960-962 in K. E. Carpenter (ed.), *The Living Marine Resources of the Western Central Atlantic: FAO Species Identification Guide for Fishery Purposes*, vol. 2, FAO, Rome, 2003 (dated 2002); J. S. Nelson, *Fishes of the World*, 4th ed., Wiley, New York, 2006; M. L. J. Stiassny, The limits and relationships of the acanthomorph teleosts, *J. Zool. Lond.* (B), 1:411-460, 1986.

Polymorphism (crystallography)

The existence of different crystal structures with the same chemical composition. If only one chemical element is present, the forms are called allotropes. Graphite and diamond are allotropes of carbon, whereas quartz and cristobalite are polymorphs of silica (silicon dioxide, SiO₂). Although properties are different in these forms, reversible transformations, which involve small shifts in atom positions and no bulk transport of material, are common. The quartz transformation at $1063^{\circ}F$ (573°C) is a reversible, atom-displacement transformation. *See* SILICA MINERALS.

In metals and ceramics, similar transformations are called martensitic. Advantage is taken of the localized nature of reversible transformation in steel by controlling the melting atmosphere, temperature, composition, mechanical working (alloying), and tempering and quenching operations. *See* HEAT TREATMENT (METALLURGY).

Transformations. Transformation from one polymorphic form to another is a response to changed conditions, most commonly temperature. The best example is carbon. Diamond is made only under enormous pressures and, when heated to very high temperatures, is transformed irreversibly from a three-dimensional, tetrahedrally coordinated structure to the two-dimensional layered structure of graphite. The large spacing between layers accounts for the lubricity of graphite. The adjustability of this spacing reflects weak interlayer forces. This adjustability is exploited in the production of intercalation compounds. *See* DIAMOND; GRAPHITE; INTERCA-LATION COMPOUNDS.

With temperature as the only variable, tin undergoes a dramatic reversible transformation, resulting in spontaneous brittle fracture. Normally metallic, tin expands by 21 vol % and becomes nonmetallic below 64° F (18°C). *See* TIN.

Control of mechanical properties. Control over transformations to achieve desirable properties as either devices or structural materials in extreme environments is a frequent objective. In the case of tin, reversibility on the atomic scale can have devastating consequences for bulk properties. Similar transformations may be beneficial in the right place and in the desired degree. Such transformation is attempted with metals and ceramics.

As with tin, but at much higher temperatures, zirconium dioxide (ZrO2) contracts on heating and expands on cooling. Addition of other oxides gives a new structure that does not transform. This structure can be toughened by small internal regions of zirconium oxide that do transform. In transforming during cooling, these regions expand, occupying more volume. As they push against the surrounding material, internal cracks are sealed off and do not propagate, which could result in fracture. This procedure is called transformation toughening. It makes it possible to modify ceramic materials so that they exhibit the desired mechanical properties at higher temperatures than metal alloys. See CERAMICS; CRYS-TAL STRUCTURE. Doris Evans

Bibliography. W. L. Bragg and G. F. Claringbull, *Atomic Structures of Minerals*, 1964; J. Donahue, *The Structures of the Elements*, 1974, reprint 1982; H. Kammimura, Graphite intercalation compounds, *Phys. Today*, 40(12):64–71, 1987.

Polymorphism (genetics)

A form of genetic variation, specifically a discontinuous variation, occurring within plant and animal species in which distinct forms exist together in the same population, even the rarest of them being too common to be maintained solely by mutation. Thus the human blood groups are examples of polymorphism, while geographical races are not; nor is the diversity of height among humans, because height is "continuous" and does not fall into distinct tall, medium, and short types. *See* MUTATION.

Control mechanisms. Distinct forms must be controlled by some switch which can produce one form or the other without intermediates such as those arising from environmental differences. This clearcut control is provided by the recombination of the genes. Each gene may have numerous effects and, in consequence, all genes are nearly always of importance to the organism by possessing an overall advantage or disadvantage. They are very seldom of neutral survival value, as minor individual variations in appearance often are. Thus a minute extra spot on the hindwings of a tiger moth is in itself unlikely to be of importance to the survival of the insect, but the gene controlling this spot is far from negligible since it also affects fertility. See RECOMBINATION (GE-NETICS).

Genes having considerable and discontinuous effects tend to be eliminated if harmful, and each gene of this kind is therefore rare. On the other hand, those that are advantageous and retain their advantage spread through the population so that the population becomes uniform with respect to these genes. Evidently, neither of these types of genes can provide the switch mechanism necessary to maintain a polymorphism. That can be achieved only by a gene which has an advantage when rare, yet loses that advantage as it becomes commoner.

Natural selection. Occasionally there is an environmental need for diversity within a species, as in butterfly mimicry. Mimicry is the resemblance of different species to one another for protective purposes, chiefly to avoid predation by birds. It takes two forms. (1) In Batesian mimicry, palatable species resemble those that are distasteful or poisonous. Such mimics generally copy several protected models and are therefore polymorphic. (2) In Mullerian mimicry, protected species resemble one another. These attempt to reduce the risk of predation by producing a uniform type with conspicuous warning patterns, such as those of bees and wasps, and are not therefore affected by polymorphism.

Sexual dimorphism falls within the definition of genetic polymorphism. In any species, males and females are balanced at optimum proportions which are generally near equality. Any tendency for one sex to increase relative to the other would be opposed by selection.

Batesian mimicry and sexual dimorphism, however, are rather special instances. In general, a gene having both advantageous and disadvantageous effects may gain some overall advantage and begin to spread because one of the features it controls becomes useful in a new environment. A balance is then struck between the advantages and disadvantages of such a gene, ensuring that a proportion of the species carry it, thus giving rise to permanent discontinuous variation, that is, to polymorphism. *See* PROTECTIVE COLORATION.

Occurrence in a population. Polymorphism is increasingly known to be a very common situation. Its existence is apparent whenever a single gene having a distinct recognizable effect occurs in a population too frequently to be due merely to mutation, certainly if present in over 1% of the individuals. Even if recognized by some trivial effect on the phenotype, it must in addition have important other effects. About 30% of the people in western Europe cannot taste as bitter the substance phenylthiourea. This is truly an insignificant matter; indeed, no one even had the opportunity of tasting it until the twentieth century. Yet this variation is important since it is already known that it can affect disease of the thyroid gland. See GENETICS; POPULATION GENETICS. E. B. Ford

Methods of study. Genetic polymorphisms can be of several types, depending upon the level of study: visible, chromosomal, protein (enzymes), and deoxyribonucleic acid (DNA).

Visible polymorphisms. Visible polymorphisms are easily studied by observing and counting the frequencies of the different morphs in populations. Examples of such polymorphisms are the color and banding patterns of the land snail (*Cepaea*), and eye color and hair color in human populations.

Chromosomal polymorphisms. Various cytological techniques are required for the study of chromosomal polymorphisms. Tissues with chromosomes particularly favorable for study are stained with a dye which makes the chromosomes easily visible through a microscope. The most favorable tissues for such studies are the salivary glands of Drosophila (and some other Diptera), which have "giant" polytene chromosomes. Such chromosomes have a characteristically complex banding pattern which allows identification of each part of each chromosome. For example, if a population is polymorphic for inversions, the order of genes in some sections of some chromosomes will be inverted with respect to the gene order of other chromosomes. Inversions will be evident by an inversion of the banding pattern in the inverted segment. Such inversion polymorphisms are common in many species of Drosophila and have been the subject of considerable research. See CHROMOSOME; CHROMOSOME ABERRATION.

Most organisms, however, do not have giant polytene chromosomes. To study chromosomal polymorphisms, it is necessary to study cells undergoing division, either mitosis or meiosis. It is possible to obtain particularly favorable tissue at the right stage of mitosis by tissue culture techniques. However, meiotic cell divisions of the gonadal tissues are often more informative to study than mitotic divisions. When homologous chromosomes pair at meiosis, it is possible to detect structural differences between the chromosomes. For example, an individual heterozygous for an inversion polymorphism will display a loop in the section of the chromosome involved in the inversion. *See* MEIOSIS.

In addition to inversion polymorphisms, other chromosomal polymorphisms are known. Translocation polymorphisms (involving an exchange of chromosomal material between nonhomologous chromosomes) are known in many plants. Aneuploid polymorphisms (addition or deletion of whole chromosomes) have been studied in several mammal species.

Techniques of studying chromosomes have made it easier to detect chromosomal polymorphisms. Chromosomes which are not polytene can be induced to show characteristic banding patterns. Thus it is becoming possible to detect chromosomal polymorphisms in almost any organism by looking at any tissue. It is not necessary to use cells in meiosis, nor is it necessary that the individual be a heterozygote in order to detect the variation. Such banding techniques have been a great advance in studying chromosome variation and have proved especially successful in studying human chromosomes.

Protein polymorphisms. There are three basic types of protein polymorphisms: deficiency, antigenic, and electrophoretic. Deficiency polymorphisms occur when certain genotypes in a population do not produce a certain enzyme or the enzyme produced is nonfunctional. By assaying individuals for the presence or absence of a certain enzyme activity, it is possible to detect such polymorphisms. Examples of this type of biochemical polymorphism in human

populations are glucose-6-phosphate dehydrogenase deficiency and lactose deficiency in adults.

Antigenic polymorphisms are detected by preparing antibodies specific for different antigens. Thus if a population is polymorphic for an antigen, tissues from different genotypes will interact with different antibody preparations. For example, blood from humans with type A blood reacts with anti-A serum; blood from type B individuals reacts with anti-B serum; blood from type AB individuals reacts with both anti-A and anti-B sera; blood from type O individuals reacts with neither antiserum. A widespread polymorphism in vertebrates is that of transplantation antigens, mainly responsible for separation of grafts between individuals of a species. *See* ANTI-BODY; ANTIGEN; BLOOD GROUPS; TRANSPLANTATION BIOLOGY.

Electrophoretic polymorphisms are very common and occur in high frequencies in plants, animals, and humans. Proteins from different individuals are extracted and applied to an electrophoresis gel; such gels can be made from starch or polyacrylamide and have a consistency somewhat like gelatin. An electrical field is applied across the gel. Since proteins usually have a net electrical charge, they migrate in the gel toward the pole opposite to the charge on the protein. The rate of migration is determined by the size of the electrical charge as well as the size and shape of the protein. Thus if a mutation which results in a change in the amino acid composition of a certain protein has occurred in the gene coding for that protein, the change is often detected by the change in electrophoretic mobility of the protein. Thus different genotypes in a population produce proteins which migrate differently in an electrophoretic gel. The easiest proteins to study by this method are enzymes because simple techniques are available for visualizing many enzymes in the gel. See ELECTROPHORESIS; PROTEIN.

Electrophoretic studies have been made on a variety of proteins in many species of organisms. Much genetic variation has been revealed. On the average, about one-half of all proteins studied are electrophoretically polymorphic in any population; of course, different populations and different species may vary in their degree of electrophoretic variation. About 30% of the more than 100 proteins studied in humans are polymorphic for electrophoretic variation. Such studies have had very important implications for population genetics and evolutionary studies.

DNA polymorphisms. Advances in deoxyribonucleic acid (DNA) technology have made it possible to study polymorphisms at the most basic level. Two methodologies have been developed. One involves the use of enzymes called restriction endonucleases, and the other determines actual DNA sequences. Restriction endonucleases recognize specific sets of DNA sequences, usually a contiguous set of four to eight bases. After DNA has been digested with such enzymes, the sizes of the DNA fragments produced can be determined by electrophoresis. Individuals polymorphic for the recognition sites of restriction

enzymes will produce differently sized fragments. Hence, such polymorphisms are called restriction fragment length polymorphisms (RFLP). They are ubiquitous and are extremely useful genetic markers for a large number of purposes, including identifying the genetic location of disease-causing genes. In addition, highly repetitive DNA sequences have been studied by using restriction enzymes. The resulting size patterns produced appear to be virtually unique for unrelated individuals. This technology, called DNA fingerprinting, has played a role in some legal cases. *See* FORENSIC MEDICINE; RESTRIC-TION ENZYME.

A direct way of studying DNA polymorphism is to determine actual DNA sequences of the same genetic region in several individuals. Results from DNA sequence data indicate that different copies of the same gene may differ by 0.5-2.0% of their base positions. The number of alleles (defined on the nucleotide level) is thus tremendous. However, most differences among alleles are silent; that is, they do not change the amino acid sequence of the protein being coded.

In addition to base-pair differences revealed by RFLP analysis and DNA sequencing, insertions or deletions of various lengths of DNA bases are also common. These may be a few base pairs in length or extend up to several thousand bases. These latter insertions or deletions may be caused by transposable elements, segments of DNA that have the ability to move around the genome. *See* HUMAN GENETICS; TRANSPOSONS. Jeffrey R. Powell

Bibliography. J. A. Bryant and V. L. Dunham, DNA Replication in Plants, 1988; S. Heinemann and J. Patrick (eds.), Molecular Neurobiology: Recombinant DNA Approaches, 1987; J. B. Jenkins, Human Genetics, 1990; K. Kammermeyer and V. L. Clark, Genetic Engineering Fundamentals: An Introduction to Principles, 1989; M. Levitan, Textbook of Human Genetics, 1988.

Polynomial systems of equations

A collection of equations in several unknowns, expressed in terms of addition, subtraction, and multiplication. System (1) is a system of polynomial equa-

$$x^{2} + xz + y^{2} - 1 = 0$$

$$x^{3} - 11x^{2} + 23x + y - 1 = 0$$

$$x + z + 1 = 0$$
(1)

tions in the three unknowns (or variables) x, y, and z. Typically, the objective is to solve such a system, which means, for Eqs. (1), to find all choices of numbers to substitute for x, y, and z so that all three equations are simultaneously satisfied (in other words, all equal zero). For example, x = 0, y = 1, and z = -1 satisfies Eqs. (1). There are exactly six solutions to Eqs. (1). They are $(x,y,z) = (0,\pm 1,-1)$, $(3,\pm 2,-4)$, $(8,\pm 3,-9)$.

A system of polynomial equations may have any number of equations and any number of variables.

Polynomial systems arise commonly in science and engineering, for example, as chemical equilibrium systems, as equations of kinematics (in robotics), as design criteria systems (in the design of mechanisms), and as systems expressing the intersection of curves and surfaces (in computer graphics and geometric modeling). Solving such systems is challenging, although certain special cases, for example, systems of linear equations and polynomials in only one variable, are more routine.

General system. Since a general polynomial system may have more than three variables, it is customary to denote the variables by using subscripts, as x_1 , x_2, \ldots, x_n , where *n* denotes the number of variables. Similarly, the number of equations may be other than three; in this article, this number is denoted by *m*. (The most common case is m = n.) Then, a precise mathematical description of a general polynomial system is: a collection of mathematical equations of the form of system (2). Each $f_i(x_1, x_2, \ldots, x_n)$, $i = 1, 2, \ldots, m$, is a sum of terms of the form shown as expression (3), where the coefficient *a*

 $f_m(x_1, x_2, \ldots, x_n) = 0$

$$a x_1^{i_1} x_2^{i_2} \dots x_n^{i_n}$$
 (3)

is a constant, and the exponent i_j of the variable x_j is a nonnegative whole number. (A zero exponent means the variable is omitted.) The values of the coefficients and exponents may change from term to term. In this notation, the second equation of system (1) is $1 x_1^3 x_2^0 x_3^0 - 11 x_1^2 x_2^0 x_3^0 + 23 x_1^1 x_2^0 x_3^0 + 1 x_1^0 x_2 x_3^0 - 1 x_1^0 x_2^0 x_3^0 = 0$.

The expressions $f_i(x_1, x_2, \ldots, x_n)$ are called polynomials in several variables (or multivariable polynomials). The problem posed by system (2) is to characterize the set of solutions, where a solution is a choice of values for the variables $x_1 = a_1, x_2 = a_2, \ldots, x_n = a_n$, which simultaneously satisfies each equation of the system. For example, the objective might be to count the solutions or to compute some (or all) of them explicitly. Solution sets of polynomial systems can be geometrically complicated, and most solution methods are limited to finding subsets of solutions (for example, all nonsingular solutions). There is no practical reliable method for computing the full solution set to general polynomial systems. Characterizing the solution sets to polynomial systems is the subject matter of algebraic geometry, while techniques for solving polynomial systems are usually classified as part of numerical analysis. See ALGEBRAIC GEOMETRY; NU-MERICAL ANALYSIS.

Polynomial equations are examples of general nonlinear equations, where transcendental as well as polynomial terms may appear. (For example, terms in general nonlinear equations may include exponential or trigonometric functions.) Thus, methods for solving systems of general nonlinear equations are applicable to systems of polynomials.

Special cases. The most classical special case of polynomial systems is polynomials in one variable. [Thus, in terms of system (2), m = 1 and n = 1.] The degree of a polynomial in one variable is the largest degree occurring in any term. [In terms of expression (3), it is the largest i_{1} .] The fundamental theorem of algebra states that a polynomial has the same number of solutions as its degree. For example, a degree-10 polynomial has 10 solutions. (Solutions that are singular have to be counted by their multiplicity.) For systems with more than one variable, the fundamental theorem of algebra generalizes to results usually known as Bezout's theorem, which give the generic number of solutions in terms of the degrees of the equations. In one variable, equations of degree 2, 3, and 4 have solutions that can be expressed by elementary formulas (the quadratic, cubic, and quartic formulas, respectively). A famous consequence of Galois theory is that no such formulas exist for polynomials of degree greater than four. In computations, these formulas are tricky to use, and numerical techniques (rather than formulas) are often more reliable. No analog of these formulas for multivariable polynomial systems exists. See EQUATIONS, THEORY OF; GROUP THEORY.

The other special case of note occurs when the equations are linear. [In terms of expression (3), $i_1 + i_2 + \cdots = i_n \leq 1$.] Cramer's rule provides an elementary formula for the solutions of a linear system, which is computationally useful when n = 2 or n = 3. For larger linear systems, numerical methods, such as gaussian elimination, are faster and more reliable. Numerical methods for solving linear systems and polynomials in one variable are highly developed and reliable, compared to the general case, discussed below. *See* LINEAR SYSTEMS OF EQUATIONS.

Methods of solution. Methods for solving polynomial systems of equations include single-solution methods, polynomial continuation, the method of resultants, and generalized bisection.

Single-solution methods. These methods aim to find only a single solution. They are usually for systems of general nonlinear equations, not specifically for polynomial equations. Examples include Newton's method (and its variants), steepest-descent leastsquares methods, and single-path continuation methods. These approaches require a start point, an estimate of the solution being sought. In some cases, the method can be shown to be globally convergent, in which case a solution will be found no matter where the start point is chosen. On the other hand (and more typically), these are local methods, in that convergence can be guaranteed only if the start point is sufficiently close to the target solution.

Polynomial continuation. This technique is a multiplestart-point method, guaranteed to find all geometrically isolated solutions but not the positivedimensional solution components. A start system is generated that has certain generic properties of the target system (that is, the system to be solved). Then the start system is transformed by a continuous parameter (the continuation parameter) into the target system, and at the same time the solutions to the start system are transformed into the solutions to the target system.

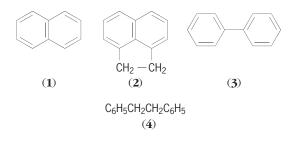
Method of resultants. The classical method of resultants (also called elimination) has been revived and made more computationally feasible via the technique of Gröbner bases. The given polynomial system is transformed into triangular form (that is, the *i*th equation involves only the first *i* variables). Then the triangular system can be solved by using a one-variable method.

Generalized bisection. By using ideas from interval analysis (in particular, the interval Newton's method), a multivariable numerical method has been developed that is analogous to (classical) one-variable bisection techniques. Alexander P. Morgan

Bibliography. O. Balci, R. Sharda, and S. A. Zenios (eds.), Computer Science and Operations Research: New Developments in Their Interfaces, 1992; D. Bini and V. Pan, Polynomial and Matrix Computations: Fundamental Algorithms, vol. 1, 1994; I. N. Herstein, Abstract Algebra, 3d ed., 1996; D. Kahaner et al., Numerical Methods and Software, 1989; A. Morgan, Solving Polynomial Systems Using Continuation for Scientific and Engineering Problems, 1987; W. Rheinboldt, Methods for Solving Systems of Nonlinear Equations, 2d ed., 1998.

Polynuclear hydrocarbon

One of a class of hydrocarbons possessing more than one ring. The aromatic polynuclear hydrocarbons may be divided into two groups. In the first, the rings are fused, which means that at least two carbon atoms are shared between adjacent rings. Examples are naphthalene (1), which has two six-membered rings, and acenaphthene (2), which has two sixmembered rings and one five-membered ring.



In the second group of polynuclear hydrocarbons, the aromatic rings are joined either directly, as in the case of biphenyl (3), or through a chain of one or more carbon atoms, as in 1,2-diphenylethane (4).

The higher-boiling polynuclear hydrocarbons found in coal tar or in tars produced by the pyrolysis or incomplete combustion of carbon compounds are frequently fused ring hydrocarbons, some of which may be carcinogenic. *See* AROMATIC HYDROCARBON; STEROID. C. K. Bradsher

Polyol

Polyols or, as they are sometimes called, polyhydric alcohols, are compounds containing more than one hydroxyl group (—OH). Each hydroxyl is attached to separate carbon atoms of an aliphatic skeleton. This group includes glycols, glycerol, and pentaery-thritol and also such products as trimethylolethane, trimethylolpropane, 1,2,6-hexanetriol, sorbitol, inositol, and poly(vinyl alcohol).

Polyols are obtained from many plant and animal sources and are synthesized by a variety of methods. 1,2-Glycols and glycerol are produced by the hydrolysis of epoxides or chlorohydrins. Formaldehyde reacts with acetaldehyde, propionaldehyde, and butyraldehyde to form, respectively, pentaerythritol, trimethylolethane, and trimethylolpropane. Catalytic hydrogenation of sugars produces sorbitol; 1,2,6-hexanetriol is obtained by the reduction of 2hydroxyadipaldehyde. Saponification of poly(vinyl acetate) is employed in the industrial manufacture of poly(vinyl alcohol), which is shown below.

Polyols such as glycerol, pentaerythritol, trimethylolethane, and trimethylolpropane are used in making alkyd resins for decorative and protective coatings. Glycols, glycerol, 1,2,6-hexanetriol, and sorbitol find application as humectants and plasticizers for gelatin, glue, and cork. Explosives are made by the nitration of glycols, glycerol, and pentaerythritol.

The polymeric polyols used in manufacture of the urethane foams represent a series of synthetic polyols which came into prominence in the 1960s. Generally polyoxyethylene or polyoxypropylene adducts of di- to octahydric alcohols, these cover a molecular weight range of 400 to 6000. *See* GLYCEROL. Philip C. Johnson

Polyolefin resins

Polymers derived from hydrocarbon molecules that possess one or more alkenyl (or olefinic) groups. The term polyolefin typically is applied to polymers derived from ethylene, propylene, and other alphaolefins, and from isobutylene, cyclic olefins, and butadiene and other diolefins. Polymers produced from other olefinic monomers, such as styrene, vinyl chloride, and tetrafluoroethylene, generally are considered separately. *See* ALKENE.

Polyethylene. The term polyethylene generally refers to any homopolymer or copolymer in which ethylene is the major component monomer. Polyethylene is a semicrystalline polymer of low to moderate strength and high toughness. The melting point varies from approximately 105 to 125° C (220 to 257° F), depending on the degree of crystallinity. The density of polyethylene reflects the degree of crystallinity, and is generally

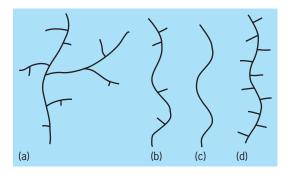


Fig. 1. Polyethylene structures. (a) Conventional lowdensity polyethylene. (b) Linear low-density polyethylene and high-density polyethylene with alpha-olefin comonomer. (c) Linear high-density polyethylene with no comonomer [a polymethylene chain]. (d) Metallocene low-density polyethylene.

used in classifying commercial grades. Low-density polyethylene, having densities ranging from 0.905 to 0.936 g/cm³, can be subdivided into high-pressure low-density polyethylene (LDPE), and low-pressure low-density polyethylene copolymers, referred to as linear low-density polyethylene (LLDPE). This last term denotes a copolymer of polyethylene that is produced by the low-pressure polymerization process, as opposed to low-density polyethylene (or conventional polyethylene), which is produced by an older high-pressure polymerization process. In a typical linear low-density polyethylene, the percentage of comonomer would be on the order of 2-3 mol %, typically butene. High-density polyethylene (HDPE) covers the density range from 0.941 to 0.967 g/cm³. See COPOLYMER.

The various types of polyethylene are characterized by differences in structure (Fig. 1). High-density polyethylene generally consists of a polymethylene $(CH_2)_n$ chain with no, or very few, side chains to disrupt crystallization, while linear low-density polyethylene contains side chains whose length depends on the comonomer used. The spacing and uniformity of the incorporation of the comonomer in linear low-density polyethylene depends on both the process and the type of comonomer, but the incorporation of comonomer is generally nonuniform. To some extent, the uniformity tends to decrease with the size of the comonomer. Polyethylene with structures different from those produced with previousgeneration catalysts are being produced with metallocene catalysts. Metallocene materials can cover the density range of both the high-density polyethylene and the linear low-density polyethylene, with the important distinction that the structures are very uniform in both molecular weight and comonomer incorporation. See METALLOCENES.

Polymerization. The monomer for polyethylene is produced by cracking aliphatic hydrocarbons separated in the refining process. Low-density polyethylene is formed by the polymerization of ethylene at high pressures (15,000-50,000 lb/in.² or 100-350 megapascals) and at temperatures of 150-300°C (300-570°F) in the presence of a small amount of organic peroxide or oxygen. The density is controlled by regulating the temperature, where lower temperatures lead to higher densities. Linear low-density polyethylene is produced commercially by the copolymerization of ethylene with alphaolefins ranging from C_3 to C_8 at temperatures up to 250°C (480°F) at low pressure (300–1100 lb/in.² or 2–7.6 MPa) in the presence of a chromium catalyst or titanium-based Ziegler-Natta catalyst. *See* CATALYSIS; STEREOSPECIFIC CATALYST.

High-density polyethylene is formed at temperatures similar to those for linear low-density polyethylene and at pressures of generally less than 1000 lb/in.² (7 MPa) by using either chromium or Ziegler catalysts. The molecular-weight distribution of high-density polyethylene is largely controlled by the type of catalyst, while the molecular weight is most often controlled with hydrogen level.

Metallocene, or more generally single-site, catalysts have a transition metal sandwiched between organic ring compounds. They allow molecular weight and comonomer distributions to be closely controlled, such that each polymer molecule closely resembles the others. Additionally, these catalyst systems permit the incorporation of comonomers once thought incompatible with previous catalyst generations. *See* POLYMERIZATION; SINGLE-SITE CATALYSTS (POLYMER SCIENCE).

Properties. The physical properties of polyethylene, regardless of process, are primarily dependent on the degree of crystallinity, which is related to the density and the molecular weight. Uniformity in both molecular weight and comonomer can also affect the observed properties. The stiffness, yield strength, and thermal and mechanical properties increase with the crystallinity, which can be systematically reduced by the addition of comonomer. Failure properties, such as toughness and ultimate tensile strength, increase with molecular weight, while to a lesser extent failure properties tend to decrease with broadening of the molecular-weight distribution for a given peak molecular weight. Polyethylene generally shows excellent toughness at low temperatures, because of the low glass transition temperature (generally believed to be between -80 and $-120^{\circ}C$ or -112 and -184°F). Environmental stress cracking, in the presence of detergents, is reduced for polyethylene copolymers as the size of the alphaolefin comonomer is increased and as the density is decreased. Clarity is generally increased with decreasing crystallinity, although the morphology of the crystalline phase can also affect clarity.

The melt flow properties are affected by molecular weight and long-chain branching. Choice of fabrication technique to produce items of commerce from polyethylene generally determines the melt flow properties needed and therefore the type of polyethylene used. Among commodity polymers, polyethylene is unique with respect to the ease of formation of long-chain branching, the presence of which can dramatically affect melt fabrication options. Generally, the ease of flow for both extrusion and injection molding increases with molecular-weight distribution and long-chain branching, but decreases with molecular weight. Consequently, low-molecular-weight resins are favored for injection-molding applications. Processes requiring melt strength and strain hardening of the molten polymer, such as blown film and foam, favor high molecular weight, the presence of long-chain branching, and to a lesser extent broadening of the molecular-weight distribution. For a given molecular weight, long-chain branching can either enhance or diminish the melt viscosity at low shear rates (melt strength), with the high levels in most lowdensity polyethylenes yielding lower viscosity at a given molecular weight. *See* PLASTICS PROCESSING.

Use and fabrication. Polyethylene is used in greater volume worldwide than any other plastic, because it is relatively inexpensive, extremely versatile, and adaptable to a large array of fabrication techniques. Its advantages are chemical inertness (resistance to solvents, acids, and alkalis) and good dielectric and barrier properties. Major uses include packaging films, plastic containers, molded articles, foam, pipe, cable sheathing, and coatings.

Because of its broad melting range, low-density polyethylene can be used in heat-seal operations. Film applications include food packaging bags, shrink wrap, and garment bags. Blow-molded containers are used in milk and other household containers, while injection-molded items include housewares, pails, and toys. Because of its low degree of crystallinity, low-density polyethylene is limited for applications requiring high stiffness, temperature resistance, tensile strength, and good barrier properties.

Linear low-density polyethylene is a versatile material that finds wide application in plastic films. Films produced by extrusion, blown, and cast film processes are widely used as garbage bags, stretch cling films, and other items that require the properties of flexibility, tear and puncture resistance, and toughness. The toughness of linear low-density polyethylene is generally superior to low-density polyethylene and increases with the size of the comonomer. The very low density formulations of linear low-density polyethylene are also used for heat-seal applications, and are stronger than those using low-density polyethylene. While the relatively narrow molecularweight distribution makes extrusion-coating applications somewhat difficult, specific resin grades are commercially available for this purpose. The molded items made from linear low-density polyethylene are similar to those made from low-density polyethylene, such as trash cans, food containers, and closures. Sheathing and flexible pipe are applications that take advantage of the flexibility and low-temperature toughness of linear low-density polyethylene.

High-density polyethylene is used in food packaging applications, because it gives very good moisture barrier properties. However, like all polyethylenes, it is not used in applications requiring an oxygen barrier. While high-density polyethylene has good chemical resistance, it is prone to environmental stress cracking in the presence of detergents and other surfactants. Longer-chain comonomers and higher-molecular-weight resins improve the environmental stress-crack resistance of high-density polyethylene. This is an important consideration in applications where polyethylene contacts chemicals, such as dishwashing detergent and laundry bleach. Because of the higher crystallinity, high-density polyethylene is used in applications where higher stiffness is required, such as injection-molded beverage cases, thinwalled blow-molded milk bottles, grocery bags, and thermoformed sheet for pickup truck bedliners. Large blow-molded applications, which take advantage of the higher rigidity high-density polyethylene, include large industrial containers. Fibers, having very high tensile strengths and tensile moduli, have been produced by a process known as gel spinning from very high molecular weight high-density polyethylene. In gel spinning, very high molecular weight polyethylene that has been swollen with solvent is spun into an extremely high tenacity fiber. This fiber approaches the strength of spider silk, and is used in medical and other applications where a very high strength fiber is desirable, for example, in fishing line. See SPIDER SILK.

Ethylene copolymers. Copolymers of polyethylene are generally limited in comonomer content to preserve some load-bearing properties in the form of enough stiffness to retain their shape. Higher levels of comonomer form rubbery polymers. Nearly all of these rubbery ethylene copolymers are used in compounded mixtures, which include the impact modification of rigid polymers such as polyethylene and polypropylene. These materials generally range in comonomer content from 25 to 60% by weight, with propylene being the most widely used comonomer to form ethylene-propylene rubber. In addition to propylene, small amounts of a diene are sometimes included, forming a terpolymer, that is, a copolymer in which three monomers are used. The distinguishing attribute of this terpolymer is that it can be vulcanized by means of peroxides. Compounded products containing ethylene-propylene rubber and terpolymer have many automotive uses such as in bumpers, dashboard panels, steering wheels, and assorted interior trim.

Another major use of ethylene copolymers is to produce poly(ethylene-co-vinyl acetate). The properties of this polymer are governed by the percentages of ethylene and vinyl acetate. Poly(ethylene-co-vinyl acetate) is polymerized by using a high-pressure process similar to that used in producing low-density polyethylene. As with all polyethylenes, increasing the amount of vinyl acetate comonomer decreases crystallinity, reduces stiffness, increases clarity, and reduces the melting point. Chemical resistance is similar to low-density polyethylene, while the clarity, low-temperature flexibility, and impact strength are superior. These materials are more permeable to oxygen, carbon dioxide, and water vapor than lowdensity polyethylene. Applications include specialty film for heat sealing, adhesives, flexible hose and tubing, footwear components, bumper components, and gaskets. Foamed and cross-linked poly(ethyleneco-vinyl acetate) is used in energy-absorbing applications.

An interesting category of ethylene copolymers is ionomers. These copolymers are produced from the copolymerization of ethylene with a comonomer containing a carboxylic group (COOH). Most often, the comonomer is methyl acrylic acid. When this copolymer is reacted with a salt of sodium or zinc, a polymer with very high tensile strength is obtained having outstanding toughness and abrasion resistance. Ionomers exhibit good adhesion to metal, glass, and many fibers. They are melt-processible, with the ionic bonds acting like cross-links at lower temperatures, yet allowing flow at higher temperatures. Because of toughness, ionomers are widely used in the covers for golf balls. Bruce Bersted

Polypropylene. Commercial polypropylene homopolymers are high-molecular-weight, semicrystalline solids having melting points around 160-165°C (320-329°F), low density (0.90-0.91 g/cm³), excellent stiffness and tensile strength, and moderate impact strength (toughness). Incorporating small amounts of ethylene or higher alpha-olefins in the isotactic polypropylene structure produces random copolymer resins with lower melting points and densities, reduced stiffness and tensile strengths, but higher impact strengths and clarity than homopolymer polypropylene resins. Intimate blending of homopolymer polypropylene with ethylene-propylene rubber produces impact copolymers that possess both high strength and excellent toughness.

Polypropylene was discovered in the 1950s and is now the second most important thermoplastic material, after polyethylene. The attractive combination of low density, excellent mechanical properties over a wide temperature range, low electrical conductivity, and low cost leads to the widespread use of polypropylene homopolymers and random and impact copolymers in products such as automotive parts, appliances, battery cases, carpeting, electrical insulation, fiber and fabrics, film wraps, food packaging and storage containers, clothing, geotextiles, luggage, medical equipment, rope, and tamper-resistant closures.

Production. Propylene, $CH_3CH=CH_2$, is produced with ethylene in large quantities at low cost from the cracking of oil and other hydrocarbon feedstocks. It can also be made by propane dehydrogenation. *See* PETROLEUM PRODUCTS.

The polymerization of propylene can yield the three basic types of polypropylene materials—isotactic, syndiotactic, and atactic—which differ by the orientation of the methyl groups (CH₃) along the polymer chain (**Fig. 2**). Idealized isotactic polypropylene has the same methyl group configuration at each tertiary carbon atom along the polymer backbone, while in perfectly syndiotactic polypropylene the methyl group configuration at the tertiary carbon positions alternate along the chain. Atactic polypropylene possesses a random orientation of methyl groups. These variations in polymer structure produce polypropylenes with widely

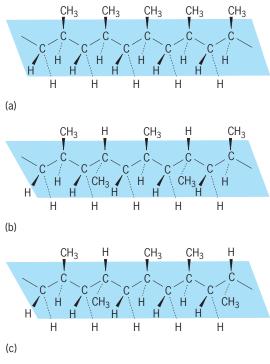


Fig. 2. Idealized structure of polypropylene materials. (a) Isotactic; methyl group always on same side of chain. (b) Syndiotactic; methyl group position alternating. (c) Atactic; methyl groups randomly distributed.

different properties. The stereoregular structures of the isotactic and syndiotactic polypropylene chains result in high-melting, crystalline solids, whereas atactic polypropylene is an amorphous, frequently sticky material. In isotactic and syndiotactic polypropylenes, occasional errors in the orientation of the methyl groups disrupt the crystallization of these polymers, lowering the melting point and decreasing stiffness. The choice of catalyst and the polymerization conditions determine the type of polypropylene produced (syndiotactic, isotactic, or atactic) and the number of defects contained in the polymer. Nearly all of the polypropylene resins that are commercially produced possess the isotactic chain structure, although a small amount of atactic polypropylene is manufactured for use in adhesives and roofing materials.

Sterospecific, titanium-based catalysts are used to produce most isotactic polypropylenes. Titanium trichloride catalysts, used with alkylaluminum cocatalysts, were the earliest and the simplest catalysts, yielding the desired isotactic polypropylene as well as a significant amount of the atactic form as a by-product. Organic compounds containing oxygen, sulfur, or nitrogen were later found to act as catalyst modifiers, increasing the stereoregularity of the isotactic polymer chains while decreasing the amount of atactic polypropylene produced. The most advanced titanium-based catalysts contain titanium tetrachloride and an organic diester modifier supported on magnesium chloride. These titanium/magnesium/diester catalysts, used with an additional silane-ether modifier and trialkylaluminum cocatalyst, produce highly isotactic polypropylene in

very high yields with virtually no atactic polypropylene by-product.

Metallocene and other single-site catalysts have been developed for the stereospecific polymerization of propylene. These single-site catalysts are complexes of zirconium and hafnium that have very precise molecular structures. The ability to tailor the structures of these catalysts has allowed the structure and properties of the resulting polymers to be precisely engineered. Single-site catalysts are now used commercially to produce isotactic polypropylene resins having very narrow molecular-weight distributions, very high melt flow rates, and elastomeric properties.

Polypropylene is produced in a variety of slurryphase and gas-phase processes. Slurry-phase polymerizations are done in hydrocarbon diluents, such as hexane or heptane, or in liquid propylene (also called a bulk slurry process). Gas-phase polymerizations are done in stirred-bed and fluid-bed processes. Hydrogen is used as a chain transfer agent to control the molecular weight of the polymer. More than one reactor in series may be required to produce polymer blends such as impact copolymers. The high-activity titanium/magnesium/diester catalysts are used in both slurry and gas-phase processes, and their development has allowed many processes to be substantially simplified. Single-site catalysts also may be used in the simplified slurry and gas-phase processes. In addition, the advent of the single-site catalyst has led to the commercialization of a high-temperature solution process for producing polypropylene.

Applications. Polypropylene homopolymer, random copolymer, and impact copolymer resins are tailored to fit specific polymer applications and fabrication methods and to achieve the desired product performance. Low-molecular-weight resins, used for melt spun and melt blown fibers and for injection-molding applications, are produced by single-site catalysts or by oxidative degradation of higher-molecular-weight polymers at elevated temperatures. The latter materials, often called controlled rheology resins, have narrower molecularweight distributions and lower viscoelasticity. The brittleness of polypropylene homopolymer, particularly at temperatures below $0^{\circ}C$ (32°F), is greatly reduced by blending it with ethylene-propylene rubber. Compounding with mineral fillers and glass fibers improves product stiffness and other properties. Higher-stiffness resins also are produced by increasing the stereoregularity of the polymer or by adding nucleating agents. Polypropylene resins are used in extrusion and blow-molding processes and to make cast, slit, and oriented films. Stabilizers are added to polypropylene to protect it from attack by oxygen, ultraviolet light, and thermal degradation; other additives improve resin clarity, flame retardancy, or radiation resistance. See RHEOLOGY; STA-BILIZER (CHEMISTRY).

Other poly(alpha-olefins). Isotactic polymers of 1butene and other higher alpha-olefins are produced by using the same stereospecific catalyst systems used for polypropylene. Poly(1-butene) is a tough and flexible resin that crystallizes from the melt in a metastable form (melting point 130° C or 266° F) before transforming to another, more stable crystalline form (melting point 138° C or 280° F). This polymer has been used in the manufacture of film and pipe. Poly(4-methyl-1-pentene) is a transparent polymer with a high melting point (240° C or 464° F) and exceedingly low density (0.83 g/cm³). This material is clear and can be autoclaved, and is used in manufacture of chemical and medical equipment. Other poly(alpha-olefins) have been synthesized, including poly(1-pentene), poly(3-methyl-1-butene), and poly(vinylcyclohexane), but they have attracted little or no commercial interest.

Polyisobutylene. High-molecular-weight rubbery solids are produced when isobutylene is polymerized at low temperature via carbocationic intermediates using a Lewis acid catalyst such as aluminum trichloride. Copolymerization of isobutylene with around 3% isoprene produces butyl rubber. Chlorination or bromination of butyl rubbers is used to produce halogenated butyl rubbers. All three materials can be vulcanized and have low permeability to gases, making them useful as sealants, inner tubes, and tubeless tire liners. Low-molecular-weight polyisobutylenes are prepared near 0°C (32°F) and have molecular weights of 200–3000. They are used in formulations for caulking, sealants, and lubricants. *See* RUBBER.

Polybutadiene and polyisoprene. Butadiene and isoprene can be polymerized by anionic, free-radical, and transition-metal catalysts to give a number of polymer structures. The commercially important forms of polybutadiene and polyisoprene are similar to the *cis*-1,4-poly(diene) structures possessed by natural rubber, as shown in the structure below.

$$\begin{pmatrix} H \\ C = C \\ C H_2 \\ C H_2 \\ C H_2 \\ H$$

Synthetic *cis*-1,4-polyisoprene is produced by alkyllithium catalysts, and high *cis*-1,4-polybutadiene (that is, mostly *cis* double bonds) is obtained by using a catalyst containing titanium, iodine, and trialkyl-aluminum. The 1,4-polybutadiene produced with lithium catalysts contains roughly equal amounts of *cis* and trans structures, but it is suitable for a variety of compounding applications. Elastomeric copolymers of these dienes with styrene and acrylonitrile can also be produced. *See* POLYACRYLONI-TRILE RESINS; POLYMER; POLYSTYRENE RESIN.

Steven A. Cohen Bibliography. B. H. Bersted, On the effects of very low levels of long chain branching on rheological behavior in polyethylene, *J. Appl. Polym. Sci.*, 30:3751– 3765, 1985; J. Boor, Jr., *Ziegler-Natta Catalysts and Polymerizations*, 1979; J. Brandrup and E. H. Immergut (eds.), *Polymer Handbook: Polymerization and Depolymerization*, vol. 1, 4th ed., 1999; J. D. Chadwick, Ziegler-Natta catalysts, in Encyclopedia of Polymer Science and Technology, vol. 8, 3d ed., pp. 517-536, 2003; N. P. Cheremisinoff (ed.), Handbook of Polymer Science and Technology, vol. 1: Synthesis and Properties, 1989; Y. V. Kissin and R. B. Lieberman, Olefin polymers, in Kirk-Othmer Encyclopedia of Chemical Technology, vol. 17, 4th ed., pp. 702-839, 1996; R. Leino, Singlesite catalysts, in Encyclopedia of Polymer Science and Technology, vol. 4, 3d ed., pp. 136-179, 2003; R. Lieberman, Propylene polymers, in Encyclopedia of Polymer Science and Technology, vol. 11, 3d ed., pp. 287-358, 2004; I. I. Rubin, Handbook of Plastic Materials and Technology, 1990; J. Scheirs and W. Kaminsky (eds.), Metallocene Based Polyolefins, 2000; J. L. Throne, Plastics Process Engineering, 1979; R. Wehrenberg, New organic fibers, Plast. World, pp. 60-62, October 1985.

Polyoma virus

A papovavirus (member of a DNA-containing group of animal viruses) that infects rodents. This virus can induce a wide variety of tumors when inoculated into newborn animals. The simple structure has made the virus a popular model for the study of lytic infection (a viral infection that results in the lysis of the host cell and release of viral progeny) by animal viruses, and viral-induced tumorigenesis (tumor formation). As a result, much is known of the biology of this virus. Discoveries made through studying its action have influenced the work on most other cancers. *See* TUMOR VIRUSES.

Polyoma virus is endemic in most wild populations of mice but causes few harmful effects. Tumors produced by this virus are unknown in the wild. Inoculation of large quantities of virus into newborn rodents, however, induces a variety of tumor types, particularly sarcomas and carcinomas. These tumors contain the viral genome but produce few infectious viral particles. Infected animals produce neutralizing antibodies directed against structural components of the virus; tumor-bearing animals produce antibodies to antigens (T antigens) which are present in tumor cells but not in the virus particle. *See* ANTIBODY; ANTIGEN.

The icosahedral viral particle consists of DNA and protein only. The genome is a small, double-stranded, closed circular DNA molecule approximately 5300 base pairs in length. It encodes the T antigens expressed early in the productive cycle and in transformed cells, and the viral capsid proteins, expressed late in the productive cycle. *See* DEOXYRIBONUCLEIC ACID (DNA).

The virus is easily propagated to high titer in mouse embryo tissue culture, resulting in cell lysis and the production of a hemagglutinin (a red blood cell-clumping antibody). This "productive infection" occurs only in mouse cells. Polyoma virus will infect cells of other rodents, but little viral proliferation occurs. Small numbers of "abortively infected" cells retain the viral genome and become transformed. Transformed cells are tumorigenic when injected into syngeneic (having the same genotype and thus being genetically compatible for transplant purposes) animals, and contain the T antigens but produce no virus. The expression of the T antigens has been shown to be necessary and sufficient to induce cell transformation.

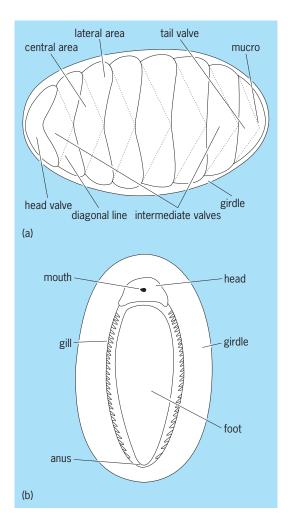
During lytic infection, the three T antigens (large, middle, and small T) are required for initiating replication of the viral genome through an interaction between the large T antigen (LT) and viral DNA, and for creating a suitable environment for high levels of DNA synthesis to occur (all three T antigens). Only one enzymatic activity has been shown to be a property of the T antigens, a DNA unwinding action of the LT molecule involved in initiating DNA synthesis. All the other effects of the T antigens on cells are caused by their ability to bind to, and change the activities of, cellular proteins. Major binding proteins to LT include members of the retinoblastoma associated protein (p105Rb) family, and a number of proteins involved in control of RNA transcription. Middle and small T antigens interact with the core heterodimer of protein phosphatase 2A. Middle T antigen (MT) alone also binds members of the SRC [the sarcomainducing (transforming) gene of Rous sarcoma virus] family of tyrosine kinases, type I phosphatidyl inositol 3' hydroxyl kinase, and the adaptor proteins ShcA and Grb2. These MT binding proteins are normally involved in stimulating cell growth in response to extracellular signals. See IMMUNOLOGY; ONCOLOGY; S. M. Dilworth VIRUS

Bibliography. T. L. Benjamin, Polyoma virus: Old findings and new challenges, *Virology*, 289:167-173, 2001; S. M. Dilworth, Polyoma virus middle T antigen and its role in identifying cancer-related molecules, *Nat. Rev. Cancer*, 2:951-956, 2002; J. L. Sever and D. L. Madden (eds.), *Polyomaviruses and Human Neurological Disease*, 1985; J. Tooze (ed.), *DNA Tumor Viruses*, 2d ed., 1981.

Polyplacophora

A class within the phylum Mollusca whose members are popularly called chitons, coat-of-mail shells, or sea cradles. Polyplacophora, or Loricata, was formerly combined with Aplacophora in the class Amphineura. All chitons are marine, and they typically live in the intertidal zone, although some live in deeper waters. They are found from subarctic to tropical latitudes but are most abundant in warmer waters. There are roughly 750 chiton species living today.

Polyplacophorans exhibit bilateral symmetry and are oval or elongate in outline. Chitons vary in length from a few millimeters ($^{1}/_{16}$ in.) to over 30 cm (12 in.), though most are a few centimeters (about 1 in.) long. They are flattened dorsoventrally and bear eight shell plates on their back. The plates, termed valves, are formed from crystals of the mineral aragonite (a type of calcium carbonate), and the upper layer of the valves contains canals filled with sensory



Polyplacophoran mollusk. (a) Dorsal view. (b) Ventral view.

organs called esthetes. The larger of these structures serve as photoreceptors and hence act as eyes.

Chitons possess three valve types-head, intermediate, and tail (see illus.). The valves are laterally embedded in the fleshy but tough girdle, which typically bears small aragonite spicules or scales of varying shapes. The width of the girdle differs from species to species, in some cases completely covering the valves (such as in the large Pacific gumboot chiton, Cryptochiton), but in most species covering only a small lateral region of the valves. Ventrally, polyplacophorans possess a large foot which they use for moving about and for attaching to rocks in a manner similar to a limpet gastropod. This allows the chitons to cling tenaciously to the substrate. The gills hang along both sides of the animal, in a deep groove between the foot and the shell. The anus is located near the posterior end of the foot. At the front of the foot is the head which, although without tentacles or eyes, bears the mouth with its associated rasping structure (the radula, which is characteristic of most mollusks). A unique feature of the chiton radula is that the largest teeth are hardened with magnetite, an iron mineral.

Most chitons feed by scraping algae and diatoms from the rock surface with their mineral-hardened radula. Others feed on small crustaceans which they trap with their foot (*Placipborella* and *Craspedochiton*), or on wood (*Xylochiton* and *Ferreiraella*). Chitons are usually mobile, although some, such as the California troglodyte chiton *Nuttalina fluxa*, remain within small depressions they have carved out of the rock. Chitons typically spend most of the daytime hidden in crevices or under rocks. They forage when submerged under water at night, returning to inactivity, and often to the same home location, by the time the tide recedes or dawn comes.

Most chiton species have separate sexes and spawn eggs and sperm. The eggs develop initially into a free-swimming form which looks like the trochophore larva characteristic of most annelids. Unlike clams and snails, chiton larvae do not pass through a veliger stage. More than 30 chiton species brood embryos until the young are sufficiently developed to remain attached to the substrate.

Although the fossil record of chitons is sparse, isolated valves are known from sediments as old as the Late Cambrian (about 500 million years ago). The radula is known from only one fossil species (*Glaphurochiton*), and girdle spicules are known from only a few fossils. Otherwise, the chiton fossil record consists of isolated or articulated shell plates. Chitons were affected by both the Permian-Triassic and the Cretaceous-Tertiary extinctions, when the diversity of this group dropped considerably. Modern chitons (the crown group) first appeared in the late Paleozoic (around 320 million years ago).

Chitons are usually allied with the aplacophorans on the basis of similarities in the nervous system, in the possession of spicules, and in a similar overall shape. One main difference between these two classes is that chitons bear a shell whereas aplacophorans do not. Determining the relationship between the shell of the chiton and that of other mollusks has not been straightforward. Although the general structure of the chiton shell is similar to that of other shell-bearing mollusks, there are some features unique to the chiton shell. These include the presence of eight valves and the perforation of the upper shell layer with esthete-containing canals. Characters of the valves, the gills, and the radula typically have been used to differentiate major groups within the Polyplacophora. See APLACOPHORA; LIMPET; MOL-LUSCA; MONOPLACOPHORA. Michael Vendrasco

Bibliography. R. C. Brusca and G. J. Brusca, *Invertebrates*, 2d ed., pp. 701–770, Sinauer Associates, Sunderland, MA, 2002; G. Burghardt and L. Burghardt, *A Collector's Guide to West Coast Chitons*, San Francisco Aquarium Soc. Spec. Publ., no. 4, 1969; A. Solem, *The Shell Makers: Introducing Mollusks*, pp. 48–55, Wiley, New York, 1974.

Polyploidy

The occurrence of related forms possessing chromosome numbers which are multiples of a basic number (n), the haploid number. Forms having 3n chromosomes are triploids; 4n, tetraploids; 5n, pentaploids, and so on. Autopolyploids are forms derived by the multiplication of chromosomes from a single diploid organism. As a result the homologous chromosomes come from the same source. These are distinguished from allopolyploids, which are forms derived from a hybrid between two diploid organisms. As a result, the homologous chromosomes come from different sources. About one-third of the species of vascular plants have originated at least partly by polyploidy, and as many more appear to have ancestries which involve ancient occurrences of polyploidy. The condition can be induced artificially with the drug colchicine and the production of polyploid individuals has become a valuable tool for plant breeding.

Polyploid series are irregularly distributed through the plant kingdom. They are particularly common in some families, notably the Gramineae and Rosaceae, and are rare in others, such as the Fagaceae and Umbelliferae. Not infrequently there exist related genera of the same family, one with and one without polyploidy, such as *Thalictrum* and *Aquilegia* of the family Ranunculaceae, or *Salix* and *Populus* of the family Salicaceae. Polyploidy has a significantly higher frequency in perennial herbs than in annual herbs and woody plants. A polyploid series often cited is that of the wheats, in which the basic chromosome number is 7 and somatic chromosome numbers of 14, 28, and 42 occur.

In animals, most examples of polyploidy occur in groups which are parthenogenetic, such as crustaceans of the genus Artemia, certain earthworms, weevils of the family Curculionidae, moths of the genus Solenobia, sawflies of the genus Diprion, lizards of the genus Cnemidophorus, and fishes of the genera Carassius and Poecilopsis; or in species which reproduce asexually by fission, such as the flatworm Dendrocoelum infernale. Sexual polyploids have been found in the frog genera Odontophrynus, Ceratophrys, and Phyllomedusa, as indicated both by chromosome numbers and nuclear DNA content. In many animals the sex chromosome mechanism is so upset by polyploidy that sterile intersexes are produced. Because hybrid sterility in animals is usually developmental rather than segregational in nature and is not eliminated by chromosome doubling, allopolyploids can occur only rarely. Developmental hybrid sterility, as in the mule, is the result of genes, contributed by the parents, interacting disharmoniously in the hybrid to disturb gonad development, sex-cell differentiation, and the course of meiosis, irrespective of chromosome pairing. Segregational hybrid sterility is the result of nonhomology between parental chromosomes reflected by their inability or reduced ability to pair at meiosis, or of abnormal segregation of genes or chromosomal segments from meiotic pairs formed by chromosomes differing with respect to inversions, translocations, or other rearrangements.

Hybridization. Polyploidy and hybridization are usually associated with each other in evolution. When distantly related species are crossed, the sterile F_1 hybrid often has little or no meiotic chromosome

pairing. However, in some cells the chromosome complement undergoes a doubling, leading to the formation of a few seeds from which second generation hybrids are formed. This fertile polyploid derivative possesses only bivalents, so that it breeds true for the intermediate condition. Such a polyploid is designated allopolyploid. Well-known examples are bread wheat (Triticum aestivum, 2n = 42), cultivated tobacco (Nicotiana tabacum, 2n = 48), and *Raphanobrassica* (2n = 36), which is the hybrid between radish (*Raphanus sativus*, 2n = 18) and cabbage (*Brassica oleracea*, 2n = 18). Polyploids have also originated from hybrids between closely related species (Primula kewensis), or between subspecies of the same species (Dactylis glomerata). These are designated as segmental allopolyploids and autopolyploids of hybrid origin, respectively. Such polyploids may form varying numbers of trivalent and quadrivalent chromosome configurations and so may segregate in the direction of their parental types. Autopolyploids derived from a single diploid race have been produced many times artificially but are rare in nature since they are usually weaker and less fertile than their diploid progenitor. In many plant genera the simultaneous occurrence of several polyploids of these three types, plus others which combine characteristics of auto- and allopolyploidy, along with their diploid ancestors, has produced the polyploid complex. This consists of a series of usually distinct diploid species or subspecies, which represent extremes of morphological and ecological variation, plus a much larger number of polyploids, which form a network of connecting links between the diploids. Examples are Bromus, Vaccinium, Galium, and Antennaria.

Distribution. The geographic distribution of polyploids in relation to their diploid ancestors does not follow any consistent set of rules. Certain northern regions, such as Iceland and Spitsbergen, have floras with particularly high percentages of polyploidy, but so do some regions with subtropical or even tropical climates, such as New Zealand and Sri Lanka. Lowland and adjacent high alpine floras in western Europe do not differ significantly in percentage of polyploids. Diploid representatives of individual polyploid complexes tend to occupy geologically older habitats, and tetraploids are most prevalent in regions newly open to colonization. Consequently, the larger, older, and more stable landmasses have relatively high percentage of diploids, while islands with floras derived through immigration, and areas disturbed by glaciation, volcanic activity, or other causes have high percentages of polyploidy.

Plant breeding. Artificial polyploids have been produced in most of the major species of crop plants by treating seeds or cuttings with colchicine. Generally, they have been less useful than their diploid progenitors because of their slower growth and reduced fertility, but economically valuable autopolyploids have been produced in rye, sugarbeets, rapeseed oil, red clover, snapdragons, marigolds, various orchids, and some other plants. Success has been achieved only when the doubling has been accompanied by intervarietal hybridization and selection. Although several artificial allopolyploids such as wheat-rye and wheat-Agropyron or quack grass have achieved partial success, none has yet been grown on a commercial scale. Artificial autopolyploids have also made possible the transfer of genes for disease resistance from wild species to cultivated species, even in instances in which the species are so distantly related that the F1 hybrids between their normal diploid forms are completely sterile. Examples are the transfer of rust resistance from goat grass (Aegilops umbellulata) to bread wheat and of resistance to both tobacco mosaic and black shank diseases from wild species of Nicotiana (N. glutinosa, N. plumbaginifolia) to cultivated tobacco. Polyploidy is, therefore, a useful tool in plant breeding when combined with hybridization and selection. See BREEDING (PLANT); CHROMOSOME ABERRATION; GENE; GENETICS; PLANT EVOLUTION; G. Ledyard Stebbins SPECIATION.

Polyploid and polytene tissues. In addition to polyploid organisms in which all of the body cells contain multiples of the basic chromosome number, most plants and animals contain particular tissues that are polyploid or polytene. Both polyploid and polytene cells contain extra copies of the deoxyribonucleic acid (DNA), but they differ in the physical appearance of the chromosomes. In polytene cells the replicated copies of the DNA remain physically associated to produce giant chromosomes that are continuously visible and have a banded pattern. The term polyploid has been applied to several types of cells: multinucleate cells; cells in which the chromosomes cyclically condense but do not undergo nuclear or cellular division (this process is termed endomitosis); and cells in which the chromosomes appear to be continually in interphase, yet the replicated chromosomes are not associated in visible polytene chromosomes. See CHROMOSOME.

There are many examples of polyploid and polytene tissues. Binucleate cells are prevalent in vertebrates, for example in liver cells. The megakarocytes of vertebrates can attain ploidy levels of 64n. Vertebrates also contain polytene cells such as the trophoblast cells of the mammalian placenta. These cells most commonly are 256-512n but can be 4096n. The suspensor cells of plant embryos are also polytene. The Diptera exhibit an extreme use of polytenization; almost all of the larval tissues become polytene. The Drosophila polytene salivary gland chromosomes (1024-2048n) have been used extensively for cytologically mapping the position of genes and chromosomal rearrangements. Although the advantages that polyploid and polytene cells provide to the organism have not been tested directly, there is a good correlation between polytene cells and rapidly growing, nutritional organs. The fact that these cells are permanently in interphase may facilitate high metabolic activity, and the large size of the cells may be important in the physical structure of tissues. See CHROMOSOME ABERRATION; GENETICS; MITOSIS Terry L. Orr-Weaver

Bibliography. V. Y. Brodsky and I. V. Uryvaeva, Genome Multiplication in Growth and Development: Biology of Polyploid and Polytene Cells, 1984; V. Grant, *Plant Speciation*, 2d ed., 1981; W. H. Lewis (ed.), *Polyploidy: Biological Relevance*, 1980.

Polypodiales

The largest order of modern ferns, commonly called the true ferns, with approximately 250 genera and 9000 species; also known as Filicales. Although well represented in the temperate regions, they reach their greatest development in the moist tropics. They vary in habit from small filmy structures to large treelike plants (**Fig. 1**). Many are epiphytic (live perched on other plants) and a number are climbing species. A few are aquatic. Perhaps the most striking species are the tropical tree ferns with their upright, unbranched stems and terminal clusters of large graceful leaves. The Polypodiales differ from the other fern orders in being leptosporangiate—that is, their sporangium, or spore sac, arises from a single surface cell—and in having small sporangia with a definite number of spores. The wall of the sporangium

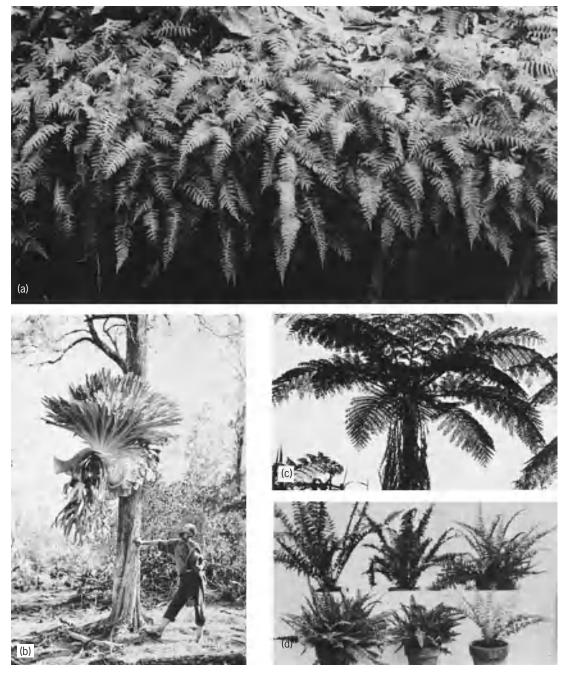


Fig. 1. Representative Polypodiales. (a) Polypody fern on a cliff edge (photograph by C. H. Jones). (b) Largest known specimen of the staghorn fern, an epiphyte on trees in the tropics. Photosynthesis occurs chiefly in the upright leaves. The rounded leaves pressed against the tree cover masses of roots. Reproductive structures develop on the pendant leaves (photograph made on island near Brisbane, Australia, by C. J. Chamerlain, World Book Co.). (c) Tree fern (U.S. Forest Service). (d) Cultivated Boston fern and five of its many mutant varieties, all originated from mutations in vegetative cells (from E. N. Transeau, H. C. Sampson, and L. H. Tiffany, Textbook of Botany, rev. ed., Harper, 1953).

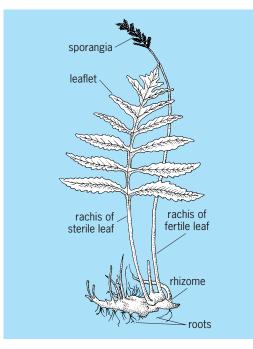


Fig. 2. The sensitive fern (Onoclea sensibilis), a representative of the Polypodiopsida. (After W. W. Robbins, T. E. Weier, and C. R. Stocking, Botany: An Introduction to Plant Science, 3d ed., John Wiley and Sons, 1964)

is almost encircled with a ring of cells having unevenly thickened walls. This ring is called the annulus. When the sporangium is mature, the annulus, acting as a spring, causes the sporangium wall to rupture, thus discharging the spores. These plants are valued for their beauty and for the clues they give to the evolutionary history of the Polypodiales which extends back through the coal measures of the Paleozoic. *See* PALEOBOTANY.

Structure. The sporophyte is the conspicuous phase of the true ferns, and like other vascular plants it has true roots, stems, and leaves (**Fig. 2**). In most ferns, especially those of the temperate regions, the mature stem is usually a creeping rhizome (underground stem) without aerial branches. However, in several species the stems are branched, and in some they are erect. Whereas in the tropics the leaves are usually persistent and evergreen, in temperate regions the leaves of most species die back each year and are replaced by new ones the next growing season. *See* LEAF; STEM.

Characteristic of this order is the apparent uncoiling of the leaves from the base toward the apex, a phenomenon known as circinate vernation (**Fig. 3**). Few ferns have simple leaves, the common type being the compound pinnately divided leaves known as fronds. The leaflets of the fronds are called pinnae, and if they are further divided, the subdivisions are known as pinnules. The branching of the veins in the leaf is characteristically dichotomous and open.

The branching of the adventitious root is always monopodial (having one main branch). This is usually true also of the stem, although branching occasionally may be dichotomous. Nearly all members of this order display a large apical meristematic cell in both stem and root. *See* APICAL MERISTEM; ROOT (BOTANY).

The internal structure of the blade of the leaf and of the root is very similar to that of these organs in the seed plants. The main difference is the presence of large intercellular spaces in the fern leaf and the frequent lack of apparent distinction between the spongy and palisade cells of the mesophyll, possibly because most ferns grow in the shade. The vascular structure of the ferns includes the four types of stele shown in Fig. 4: (1) the protostele, in which the phloem (food-conducting tissue) completely surrounds the xylem (water-conducting tissue), the latter being a solid strand with no central pith; (2) the amphiphloic siphonostele, the xylem of which forms a cylinder enclosing the pitch, with phloem bordering the xylem on both the inside and outside; (3) the ectophloic siphonostele, which has only external phloem surrounding the xylem that is around the pith; and (4) the dictyostele, in which the stele is broken up into a number of vascular bundles either scattered or arranged in a circle. The dictyostele is the common type, with the amphiphloic siphonostele occurring rather frequently and the other two types being comparatively rare. No



Fig. 3. Fern leaves pushing upward through a cement sidewalk. Growth pressure may amount to hundreds of pounds per square inch. (After E. N. Transeau, H. C. Sampson, and L. H. Tiffany, Textbook of Botany, rev. ed., Harper, 1953)

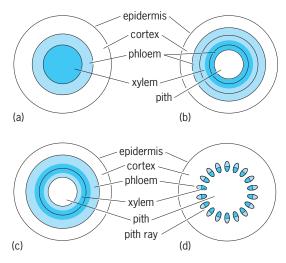


Fig. 4. Types of steles in the Polypodiales. (a) Protostele. (b) Amphiphloic siphonostele. (c) Ectophloic siphonostele. (d) Dictyostele with central pith surrounded by xylem and phloem in the form of discrete vascular bundles. Bands of parenchyma cells known as pith rays are located between the bundles. (*After H. J. Fuller and O. Tippo, College Botany, rev. ed., Holt, 1954*)

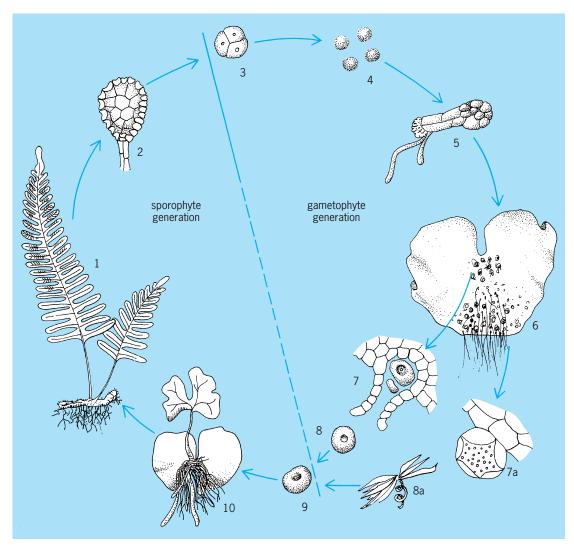


Fig. 5. Life history of a fern: (1) the fern plant, or sporophyte, bearing sori, or clusters of sporangia, on its leaves; (2) a sporangium; (3) a tetrad of young spores; (4) the four mature spores which have come from the tetrad shown in 3; (5) a spore germinating into a young gametophyte; (6) mature gametophyte, bearing sexual organs; (7) archegonium; (7a) antheridium; (8) egg cell, or female gamete; (8a) sperm or male gamete; (9) fertilized egg; (10) young sporophyte growing out of a fertilized egg, the whole still attached to the remains of the gametophyte. (*After E. W. Sinnott and K. S. Wilson, Botany: Principles and Problems, 6th ed., McGraw-Hill,* 1962)

cambium is present in the true ferns so that all the tissue is of primary origin. The development is mesarch (centripetal and centrifugal) since the metaxylem surrounds the protoxylem. The protoxylem consists of spirally thickened tracheids, the thickenings appearing as bands on the cell walls. Thick-walled sclerenchyma cells are numerous in most fern stems. Except in the few species which have a protostele, leaf gaps are present where the leaf (vascular) traces emerge from the stele. *See* PHLOEM; PITH; PLANT ANATOMY; PLANT TISSUE SYSTEMS; SCLERENCHYMA; XYLEM.

Alternation of generations. The life cycle of the fern consists of two independent (self-sustaining) alternating generations (**Fig. 5**). The common leafy fern plant is the sporophytic (spore-producing) generation. When the mature spores are discharged and reach a suitable substrate, they germinate and produce a small, flat, green, heart-shaped structure known as the prothallium or gametophytic (gamete-

producing) generation. The gametophyte produces the sex organs antheridia (male) and archegonia (female). The gametes (sperm and egg) unite in fertilization and the resultant cell, or zygote, develops into the spore-bearing (sporophytic) fern plant.

In all ferns, the spores are produced in special multicellular organs known as sporangia. Except for a few genera, the sporangia are arranged in groups or clusters called sori (**Fig. 6**). These are on the lower (abaxial) surface of the leaves or fertile fronds, either along the midrib of the pinnae, near the leaf margins, or scattered. Usually each sorus is covered by a flaplike structure called the indusium, which may be of various shapes and sizes. However, a few ferns have naked sori, and others have a false indusium formed by the folding or inrolling of the leaf margin. Varying degrees of specialization occur between the vegetative, or sterile, leaf and the reproductive, or fertile, leaf, with most species having leaves able to perform both functions. As a result of meiotic divisions, each

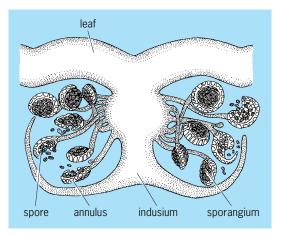


Fig. 6. Diagram of section through a fern leaf, showing details of a sorus. (After W. W. Robbins, T. E. Weier, and C. R. Stocking, Botany: An Introduction to Plant Science, 3d ed., John Wiley and Sons, 1964)

sporangium produces a number of tetrads of spores from the spore mother cells. In most species these spores are all alike (homosporous). While the number of spores produced in each sporangium is small, usually 48–64, though sometimes as low as 16 or as high as 512, the total number produced by a single individual may be surprisingly large. For one species of *Dryopteris* it has been estimated that 50 million spores may be produced in a single season. When the spores are mature and conditions are favorable, the annulus ruptures the sporangium, throwing the spores some distance. *See* MEIOSIS.

The spore is the first cell of the gametophytic generation. Each spore may germinate and produce a short filament of cells which eventually develops a triangular apical meristematic cell. By repeated cell division, the apical meristematic cell produces the heart-shaped or lobed prothallium or gametophyte (Fig. 7). In most species the gametophyte is monoecious; that is, both male and female sex organs are on the same thallus. While the gametophyte is still quite young, antheridia develop on the lower surface. The archegonia are produced later when the prothallium has matured and developed a thickened archegonial cushion just posterior to the growing point (meristem). Since gametophytes usually occur in various stages of development, it is likely that this difference in maturation time of the cells ensures the

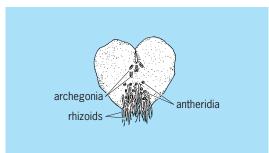


Fig. 7. Diagram of a fern prothallium. (After W. W. Robbins, T. E. Weier, and C. R. Stocking, Botany: An Introduction to Plant Science, 3d ed., John Wiley and Sons, 1964)

crossing of gametes from different prothalli or gametophytes. The form of the gametophyte and the number of rhizoids have been found to be strongly influenced by the quality and intensity of the light reaching them. Light seems to be needed for germination of the spores. With strong white or blue light the normal flat thallus with numerous rhizoids is formed, while with low intensity and red light the gametophyte remains filamentous, with few rhizoids formed.

The antheridium is a rounded structure directly attached to the thallus or sometimes borne on a short stalk. Although each antheridium produces only a few sperms, the presence of several antheridia on each prothallus ensures a large number of sperms. Each sperm is large and spirally coiled, and has many cilia.

The archegonia develop in such manner that when mature, the archegonial venter containing the egg is embedded in the prothallium, whereas the neck cells project above the surface. The mature archegonium liberates a substance into the water which elicits a chemotactic response from the sperms, attracting them to the venter and the egg cell therein where fertilization is accomplished and a zygote is formed.

The ferns are characterized by regularity in the development of the embryo. A vertical division of the zygote followed by a transverse division produces a quadrat of cells. Each of these cells then divides to produce a group of eight cells in four pairs. The anterior pair of inner cells gives rise to the underground stem or rhizome, whereas the posterior pair produces the foot, a short-lived absorbing organ. The anterior pair of outer cells forms the first leaf and the other pair produces the primary root. The first leaves usually do not suggest the species, but these are followed by a series of transitional forms until finally the characteristic type of leaf is produced. The production of spores by the adult plant completes the sporophytic generation of the life cycle.

Because of the fact that the ferns have long been of special interest to botanists, a great deal of morphological and paleobotanical evidence has been accumulated which strongly points to their origin from the Psilotopsida. There is also considerable evidence to show that some ancient group of the Polypodiopsida may have given rise to some of the seed plants (Magnoliophyta), as exemplified by the Cycadicae and their modern descendants. *See* POLYPODIO-PHYTA; PSILOTOPHYTA. Paul A. Vestal

Bibliography. H. C. Bold, *Morphology of Plants and Fungi*, 4th ed., 1980; F. Scagel et al., *An Evolutionary Survey of the Plant Kingdom*, 1965; G. M. Smith, *Cryptogamic Botany*, vol. 2, 2d ed., 1955; K. R. Sporne, *The Morphology of Pteridophytes*, 4th ed., 1975.

Polypodiophyta

A division of the plant kingdom, commonly called the ferns, which is widely distributed throughout the world but is most abundant and varied in moist, tropical regions. The Polypodiophyta are sometimes treated as a class Polypodiopsida of a broadly defined division Tracheophyta (vascular plants). The group consists of five living orders (Ophioglossales, Marattiales, Polypodiales, Marsileales, and Salviniales), plus several orders represented only by Paleozoic fossils. The vast majority of the nearly 10,000 species belong to the single order Polypodiales, sometimes also called Filicales.

Vegetative features. The Polypodiophyta ordinarily have well-developed roots, stems, and leaves that contain xylem and phloem as conducting tissues. The central cylinder of vascular tissue in the stem usually has well-defined parenchymatous leaf gaps where the leaf traces depart from it. The leaves are spirally arranged on the stem and are usually relatively large, with an evidently branching vascular system. In most kinds of ferns the leaves, called fronds, are compound or dissected. *See* LEAF; PHLOEM; ROOT (BOTANY); STEM; XYLEM.

Life cycle. The Polypodiophyta show a welldeveloped alternation of generations, both the sporophyte and the gametophyte generation being detached and physiologically independent of each other at maturity. The sporophyte is much the more conspicuous, and is the generally recognized fern plant. On some or all of its leaves it produces tiny sporangia which in turn contain spores. *See* REPRO-DUCTION (PLANT).

The spores are produced by reduction division and represent the first step in the gametophyte generation. They are usually distributed by the wind. The spore gives rise to a small, thalloid, usually green and photosynthetic gametophyte that is seldom as large as the nail of a person's little finger. Sperms and eggs are produced in characteristic male and female sex organs (antheridia and archegonia) of the gametophyte. The egg in the archegonium is fertilized in place by the sperm, which swims through a film of water to reach it.

The fertilized egg is the first step in the sporophyte generation. The young sporophyte begins its growth as a parasite on the parent gametophyte, but it soon establishes its independence by putting forth a leaf and sending a root down into the substrate. The mature sporophyte with roots, stems, and leaves develops from this tiny plantlet produced by the gametophyte.

Evolutionary history. The Polypodiophyta originated from the Rhyniophyta during the Devonian Period. Some of the fossil connecting forms, such as *Protopteridium*, could be referred about as well to the one division as to the other. The Polypodiophyta became abundant and varied in the Carboniferous Period, but many of the fossils which have caused the Carboniferous to be called the age of ferns actually represent "seed ferns," belonging to a different division (Pinophyta). Although somewhat diminished from their Carboniferous abundance, the Polypodiophyta are still an important part of the flora of the Earth. *See* COENOPTERIDALES; MARAT-TIALES; OPHIOGLOSSALES; POLYPODIALES. **Polypodiopsida**

A class (also known as Filicineae) of the plant division Polypodiophyta containing a large group of plants commonly called ferns. They are widely distributed throughout the world with their greatest development in the moist tropics. *See* POLYPODIO-PHYTA.

The stems, which may be unbranched or branched, frequently bear several large, spirally arranged leaves which are usually compound. In the majority of the species, the green leaves, sometimes called fronds, are the conspicuous organs of the sporophyte. The sporangia (spore sacs) are grouped in special structures called sori on the under margin or abaxial (lower) side of the leaves. Most of the sporangia are borne in groups. The fern sporophyte possesses true stems, leaves, and roots. Leaf and branch gaps both are present in the conducting cylinder or stele. Antherozoids are multiflagellate. *See* LEAF; ROOT (BOTANY); STEM.

Polypodiopsida is an old class with a good representation in the Paleozoic flora. Some of the plant fossils resemble contemporary living species. *See* COENOPTERIDALES; MARATTIALES; OPHIOGLOSSALES; PALEOBOTANY; POLYPODIALES. Paul A. Vestal

Polypteriformes

A distinctive and apparently ancient order of actinopterygian fishes comprising the subclass Cladistia, commonly known as bichirs and reedfishes. The following combination of characters distinguishes the order from all other actinopterygian fishes: elongate body covered with thick, rhombic, ganoid scales; slitlike spiracle behind each eye; wellossified internal skeleton, including vertebrae with ossified centra and neural canals; maxillae firmly united with the skull; basically heterocercal tail, but with a symmetrical caudal fin, and with the upper part continuous with the dorsal fin; dorsal series of free sail-like finlets, each supported by a spine; lobed pectoral fin with its rays supported by ossified radials; pair of enlarged gular plates; very large intercalar bone; no branchiostegals; bilobed swim bladder (the left lobe less developed than the right) located ventral to and attached to the esophagus, serving as lungs; functional gills; young with a pair of feathery external gills, which are assimilated with age; spiral valve-type intestine; and four rather than five pairs of gill arches as is typical for actinopterygians (see illustration). See SCALE (ZO-OLOGY); SWIM BLADDER.



Bichir (Polypterus endlicheri), length to 3 ft (0.9 m). (After G. A. Boulenger, Catalogue of the Fresh Water Fishes of Africa in the British Museum, vol. 1, 1909)

Arthur Cronquist

For many years, Polypteriformes were regarded to be either members of the Sarcopterygii or a taxon more closely related to them than to the Actinopterygii. They are now regarded as the sister group of all other actinopterygians. The order is known from the Eocene and presently comprises a single extant family, Polypteridae, and two genera, *Polypterus*, with about 15 species, and *Erpetoichthys*, with one species. Both genera are confined to freshwaters of tropical Africa. They are common along lake shores and in swamps and are able to withstand anoxic (oxygen-deficient) conditions by breathing atmospheric oxygen. *See* ACTINOPTERYGII; SARCOPTERY-GII.

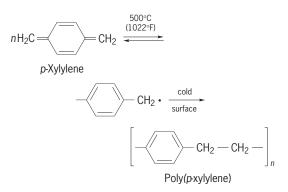
Erpetoichthys. *Erpetoichthys* (synonym *Calamoichthys*) *calabaricus* differs from *Polypterus* in having a snakelike body and no pelvic fins versus an elongate body and pelvic fins. It is endemic to the coastal areas adjacent to the Gulf of Guinea, where it is restricted to reedy habitats in slow-flowing rivers and standing waters. It moves snakewise over the bottom, feeding at night on worms, crustaceans, and insects, but it may also sidewind quite rapidly through the water. The reedfish can breathe air and thus can tolerate low oxygen concentrations. Larvae have external gills and resemble salamander larvae. Maximum total length is 90 cm (36 in.).

Polypterus. Species of *Polypterus* are generally similar in habitat and habits in preferring shallow, vegetated, still or slow-moving waters of streams, flood zones, freshwater lagoons, and lakes. They are primarily carnivorous, with the adults feeding mostly on fishes and the young feeding more on worms, mollusks, insect larvae, and other crustaceans. They may lie quietly or glide snakelike through the water. In spite of their armorlike scale covering, they are very flexible and agile when pursuing prey. The majority of the species occur in the Zaire (Congo) River basin. Maximum total length is 97 cm (38 in.), with most species much smaller. Herbert Boschung; Reeve M. Bailey

Bibliography. W. E. Burgess, Bichirs and rope fish, *Trop. Fish Hobbyist*, 31(5):14-17, 1983; R. Froese and D. Pauly (eds.), FishBase, World Wide Web electronic publication, version (05/2005); J. S. Nelson, *Fisbes of the World*, 4th ed., Wiley, New York, 2006; R. Sacca and W. Burggren, Oxygen uptake in air and water in the air-breathing reedfish *Calamoicbthys calabaricus*: Role of skin, gills and lungs, *J. Exp. Biol.*, 97:179-186, 1982; G. Sterba, *Freshwater Fishes of the World: A Studio Book*, Viking Press, New York, 1963.

Poly(p-xylylene) resins

Linear, crystallizable resins based on an unusual polymerization of *p*-xylene and derivatives. The polymers are tough and chemically resistant, and may be deposited as adherent coatings by a vacuum process. Synthesis begins, as in the reaction below, with



the pyrolysis of xylene to form di-*p*-xylylene. After purification, the di-*p*-xylylene is pyrolyzed to give the monomer *p*-xylylene, a diradical. Condensation of the monomer on cold substrates results in polymerization.

The vapor deposition process makes it possible to coat small microelectronic parts with a thin layer of the polymer. Other potential uses are as coatings for heat-exchange tubes and for encapsulation. Specific properties may be modified by varying the structure of the original xylene, for example, by chlorination. *See* XYLENE. John A. Manson

Bibliography. G. Odian, *Principles of Polymerization*, 4th ed., 2004; M. P. Stevens, *Polymer Chemistry: An Introduction*, 3d ed., 1998.

Polysaccharide

A class of high-molecular-weight carbohydrates, colloidal complexes, which break down on hydrolysis to monosaccharides containing five or six carbon atoms. The polysaccharides are considered to be polymers in which monosaccharides have been glycosidically joined with the elimination of water. A polysaccharide consisting of hexose monosaccharide units may be represented by the reaction below.

$$nC_6H_{12}O_6 \rightarrow (C_6H_{10}O_5)_n + (n-1)H_2O_5$$

The term polysaccharide is limited to those polymers which contain 10 or more monosaccharide residues. Polysaccharides such as starch, glycogen, and dextran consist of several thousand D-glucose units. Polymers of relatively low molecular weight, consisting of two to nine monosaccharide residues, are referred to as oligosaccharides. *See* DEXTRAN; GLUCOSE; GLYCOGEN; MONOSACCHARIDE; STARCH.

Polysaccharides are either insoluble in water or, when soluble, form colloidal solutions. They are mostly amorphous substances. However, x-ray analysis indicates that a few of them, such as cellulose and chitin, possess a definite crystalline structure. As a class, polysaccharides are nonfermentable and are nonreducing, except for a trace of reducing power due, presumably, to the free reducing group at the end of a chain. They are optically active but do not exhibit mutarotation, and are relatively stable in alkali. *See* CELLULOSE; OPTICAL ACTIVITY.

The polysaccharides serve either as reserve nutrients (glycogen and inulin) or as skeletal materials (cellulose and chitin) from which relatively rigid mechanical structures are built. Some polysaccharides, such as certain galactans and mannans, however, serve both functions. Through the action for acids or certain enzymes, the polysaccharides may be degraded to their constituent monosaccharide units. Some polysaccharides yield only simple sugars on hydrolysis; others yield not only sugars but also various sugar derivatives, such as D-glucuronic acid or galacturonic acid (known generally as uronic acids), hexosamines, and even nonsugar compounds such as acetic acid and sulfuric acid.

The constituent units of the polysaccharides molecule are arranged in the form of a long chain, either unbranched as in cellulose and amylose, or branched as in amylopectin and glycogen. The linkage between the monosaccharide units is generally the 1,4- or 1,6-glycosidic bond with either the α or β configuration, as the case may be. The branched glycogen and amylopectin contain both the 1,4 and 1,6 linkages. However, other types of linkage are known. In plant gum and mucilage polysaccharides, 1,2, 1,3, 1,5, and 1,6 linkages occur more commonly than the 1,4 type.

In an attempt to systematize the carbohydrate nomenclature, the generic name glycan was introduced as synonymous with the term polysaccharide. This term is evolved from the generic word glycose, meaning a simple sugar, and the ending, "an," signifying a sugar polymer. Examples of established usage of the "an" ending are xylan for polymers of xylose, mannan for polymers of mannose, and galactomannan for galactose-mannose copolymers. Cellulose and starch are both glucans or glucoglycans, since they are composed of glucose units.

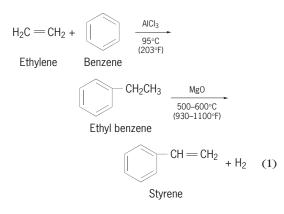
Polysaccharides are often classified on the basis of the number of monosaccharide types present in the molecule. Polysaccharides, such as cellulose or starch, that produce only one monosaccharide type (D-glucose) on complete hydrolysis are termed homopolysaccharides. On the other hand, polysaccharides, such as hyaluronic acid, which produce on hydrolysis more than one monosaccharide type (*N*-acetylglucosamine and D-glucuronic acid) are named heteropolysaccharides. *See* CARBOHYDRATE. William Z. Hassid

Bibliography. L. Stryer, *Biochemistry*, 4th ed., 1995; G. Zubayet et al., *Principles of Biochemistry*, 3 vols., 1994.

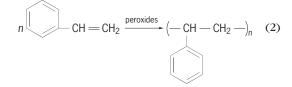
Polystyrene resin

A hard, transparent, glasslike thermoplastic resin. Polystyrene is characterized by excellent electrical insulation properties, relatively high resistance to water, high refractive index, clarity, and low softening temperature.

Styrene produced by the dehydrogenation of ethyl benzene, which in turn is obtained by the alkylation of benzene with ethylene as in reaction (1).



Free-radical catalysts such as peroxides are often used for polymerization and copolymerization in bulk, solution, and in aqueous emulsion and suspension, as in reaction (2).



Ionic and complex catalysts may also be used in bulk and in solution.

The high-molecular-weight homopolymers, copolymers, and polyblends are used as extrusion and molding compounds for packaging, appliance and furniture components, toys, and insulating panels. For a discussion of copolymers of styrene with unsaturated polyesters and with drying oils. *See* POLYESTER RESINS.

The copolymer of styrene and butadiene was the major synthetic rubber of World War II. During the 1940s, the redox system of polymerization was developed in which the presence of a reducing agent caused the peroxide to yield free radicals more rapidly at lower temperatures. At the lower temperature of the redox polymerization, a more linear copolymer called cold rubber is obtained in high conversion with improved physical properties. Styrenebutadiene copolymers are still used for automobile tires and in various rubber articles. *See* POLYMERIZA-TION; RUBBER.

High-styrene-butadiene copolymers (containing more than 50% styrene) are resinous rather than rubbery. The latexes, as produced by emulsion polymerization, achieved wide usage in water-based paints.

Copolymers with 20–30% acrylonitrile may be used when a somewhat higher softening point is desired without sacrifice of transparency. Fabrication is, however, more difficult.

By sulfonation of the copolymer of styrene and

divinyl benzene, an insoluble polyelectrolyte is produced. This product in the form of its sodium salt is employed as a cationic-exchange resin which is used for water softening.

The effects of blending small amounts of a rubbery polymer, such as butadiene-styrene rubber, with a hard, brittle polymer are most dramatic when the latter is polystyrene. The polyblend may have impact strength greater than ten times that of polystyrene. Even better results may be obtained by graft polymerization of styrene in the presence of the rubber.

Other complex polyblends and interpolymers of styrene, butadiene, and acrylonitrile in various combinations (ABS resins) are important as molding resins which have an excellent balance of properties and low cost. *See* POLYACRYLONITRILE RESINS.

ABS resins and their acrylic counterparts (interpolymers of methyl methacylate, butadiene, and styrene, or MBS resins) are used as toughening agents for polymers such as polyvinyl chloride.

Block copolymers of styrene with butadiene, the so-called thermoplastic elastomers, are of interest as resins for the injection molding of such articles as toys, footwear components, and pharmaceutical items. These copolymers typically consist of rigid polystyrene end units connected by elastomeric polybutadiene units. They may be fabricated by conventional hot thermoplastic techniques, but on cooling behave like typical elastomers.

The strength of amorphous, atactic polystyrene may be increased by orientation induced by uniaxial or biaxial stretching even though crystallinity is not produced. Oriented products in the form of filaments, sheet, and film are available.

Crystallizable, isotactic polystyrene can be formed in the presence of (1) triphenylmethyl potassium in hexane solution, (2) the alfin catalyst (sodium allyl-sodium isopropoxide) in hexane and benzene solutions, and (3) a Ziegler catalyst (titanum tetrachloride-aluminum triethyl) in petroleum ether solvent. The softening temperature of the crystalline polymer is substantially greater than that of the amorphous product.

Besides the many applications of styrene in combination with other materials as in rubber and paints, the homopolymer and the polyblends are employed in the injection molding of toys, panels, and novelty items, and in the extrusion of sheets. The sheets are used for panels, or they may be further shaped by vacuum forming for uses such as liners for refrigerator doors.

Polystyrene may also be fabricated in the form of a rigid foam, which is used in packaging, foodservice articles, and insulating panels. *See* ACRYLONI-TRILE; COPOLYMER; PLASTICS PROCESSING; POLYMER; STYRENE. John A. Manson

Bibliography. H. Allcock, F. Lampe, and J. Mark, *Con*temporary Polymer Chemistry, 3d ed., 2003; J. A. Brydson, *Plastics Materials*, 7th ed., 1999; G. Odian, *Principles of Polymerization*, 4th ed., 2004; M. P. Stevens, *Polymer Chemistry: An Introduction*, 3d ed., 1998.

Polysulfide resins

Resins that vary in properties from viscous liquids to rubberlike solids. Organic polysulfide resins are prepared by the condensation of organic dihalides with a polysulfide as in the reaction below. By the

CICH₂CH₂CI +
$$nNa_2S_4$$
 \longrightarrow
1,2,-Dichloro-
ethane tetrasulfide
 $(-CH_2 - CH_2 - S - S -)_n + 2nNaCI$
Polysulfide resin

use of other dichlorides, such as bis(2-chloroethyl) ether, $ClCH_2CH_2OCH_2CH_2CI$, by changing the functionality of the sodium sulfide, or by incorporating a trifunctional halide to yield branching, the properties may be varied. Molecular weights can also be adjusted by treatment with chain-breaking disulfides. The condensation is usually conducted in an aqueous medium from which the product may be separated and dried. Many of the polysulfide resins have an odor which is generally characteristic of monomeric sulfur compounds but is usually milder in nature.

The linear, high-molecular-weight polymers can be cross-linked or cured by reaction with zinc oxide. Compounding and fabrication of the rubbery polymers can be handled on conventional rubber machinery. The polysulfide rubbers are distinguished by their resistance to solvents such as gasoline, and to oxygen and ozone. The polymers are relatively impermeable to gases. The products are used to form coatings which are chemically resistant and special rubber articles, such as gasoline bags.

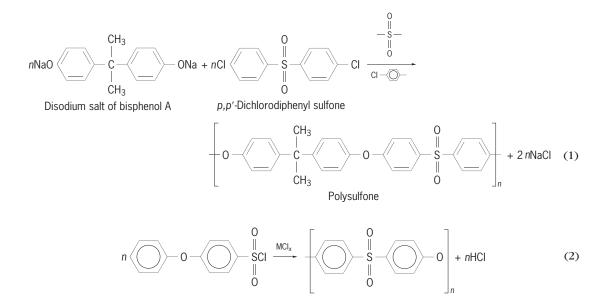
Low-molecular-weight, liquid polymers prepared by cleavage of the —S—S— linkages in the conventional polysulfide are useful as curable ingredients in caulking compounds. In addition, they are used as reactive diluents to reduce the viscosity of epoxy resins and, following reaction with free epoxy groups, to flexibilize and toughen the resin.

The polysulfide rubbers were among the very first commercial synthetic rubbers. Although the products are not as strong as other rubbers, their chemical resistance makes them useful in various applications. The polysulfide rubbers were also among the first polymers to be used in solid-fuel compositions for rockets. *See* ORGANOSULFUR COMPOUND; POLY-MERIZATION; PROPELLANT; RUBBER. John A. Manson

Bibliography. J. A. Brydson, *Plastics Materials*, 7th ed., 1999; J. Mark et al. (eds.), *Physical Properties of Polymers*, 3d ed., 2004; G. Odian, *Principles of Polymerization*, 4th ed., 2004.

Polysulfone resins

Plastics whose molecules contain sulfone groups $(-SO_2-)$ in the main chain, as well as a variety of additional constituents such as ether or isopropylidene groups. Polysulfones based on aromatic backbones



constitute a useful class of engineering plastics, owing to their high strength, stiffness, and toughness together with high thermal and oxidative stability, low creep, transparency, and the ability to be processed by standard techniques for thermoplastics. Aliphatic polysulfones are less stable to thermal degradation or hydrolysis, so they are not of commercial interest. *See* PLASTICS PROCESSING.

Three major types of polysulfones have been of commercial interest: polysulfone, polyethersulfone, and polyphenylsulfone. A typical synthesis for polysulfone involves the nucleophilic reaction of the disodium salt of bisphenol A with p,p'dichlorodiphenylsulfone in dimethyl sulfoxide [reaction (1)]. Friedel-Crafts-type condensations of chlorosulfonyl goups with aromatic groups can also be used to prepare polyethersulfones [reaction (2)], where MCl_r is a metal halide such as FeCl₃ or SbCl₅, although this is not practiced commercially. Polyether sulfones can also be prepared by the self-condensation of 4-chlorophenyl-4'hydroxyphenylsulfone. Many variations of these reactions are possible to yield a wide range of structures, including aromatic polysulfones with even higher thermal performance. See FRIEDEL-CRAFTS RE-ACTION.

The aromatic structural elements and the presence of sulfone groups are responsible for the resistance to heat and oxidation; ether and isopropylidene groups contribute some chain flexibility. Aromatic polysulfones can be used over wide temperature ranges, to ~150°C (300°F) for polysulfone and to 200°C (390°F) for polyethersulfone and polyphenylsulfone.

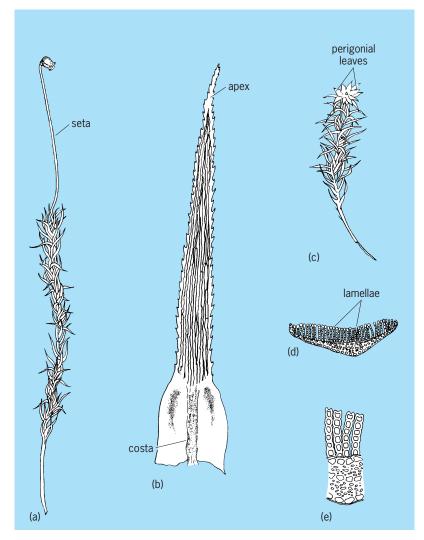
Polyphenylsulfone is also known for its supertough impact strength. Resistance to hydrolysis at high temperatures and to most acids, alkalis, and nonpolar organic solvents is excellent; however, the resins may be attacked or dissolved by polar solvents, especially if the component is under stress. While resistance to ionizing radiation is high, protection against ultraviolet light is necessary for outdoor applications. Dimensional stability and electrical properties such as dielectric loss and strength are retained well during service, and flammability is low, especially for polyphenylsulfone. Polysulfones also withstand sterilization and exposure to sterilants better than most other polymers, and polyphenylsulfone can even replace metals in applications requiring repeated sterilization. *See* PHOTODEGRA-DATION.

Because of the combination of properties discussed, polysulfone resins find many applications in electronic and automotive parts, medical instrumentation subject to sterilization, chemical and food processing equipment, and various plumbing and home appliance items. The inherent low flammability of polyphenylsulfone makes it one of the few materials that can pass the stringent regulations for aircraft interiors. Coating formulations are also available, as well as grades reinforced with glass or carbon fibers. *See* COPOLYMER; HETEROCYCLIC POLY-MER; ORGANOSULFUR COMPOUND.

Theodore T. Moore; John A. Manson Bibliography. J. A. Brydson, Plastics Materials, 7th ed., Butterworth-Heinemann, Woburn, MA, 1999; R. J. Cotter, Engineering Plastics: A Handbook of Polyarylethers, Gordon and Breach, Basel, Switzerland, 1995; M. J. El-Hibri et al., Poly(arylethersulfones), in O. Olabisi (ed.), Handbook of Thermoplastics, Mercel Dekker, 1997; Kirk-Othmer Encyclopedia of Chemical Technology, 4th ed., vol. 19, Wiley, 1996; V. Lakshmana Rao, Polyether sulfones J. Macromol. Sci.-Rev. Macromol. Chem. Phys., C39:655-711, 1999; F. Parodi, Polusulfones, in Comprehensive Polymer Science: The Synthesis, Characterization, Reactions, and Applications of Polymers, vol. 5, ed. by G. C. Eastmond et al., Pergamon Press, Oxford, U.K., 1989.

Polytrichidae

A subclass of the mosses (class Bryopsida), consisting of one family and 19 genera. The plants are acrocarpous and perennial, and the leaf structure



Polytrichum commune, member of the order Polytrichales, an acrocarpous perennial moss. (a) Female plant. (b) Serrate apex of leaf. (c) Male plant. (d) Cross section of leaf showing lamellae, which extend parallel to one another along the midrib of the leaf. (e) Lamellae enlarged; each of the lamellae is several cells high and one cell wide. (After W. H. Welch, Mosses of Indiana, Indiana Department of Conservation, 1957)

(see **illus**.) is characteristic (though similar to that of the Dawsoniidae); the peristome structure is unique.

The plants grow on soil, often in dry habitats, and they are usually unbranched; annual increments often grow up through the male inflorescences, but do not give the appearance of forked stems. The plants are nearly always dioecious, with both male and female inflorescences terminal. The stems generally have well-developed vascular tissues. The long, narrow leaves are sheathed at the base; above the base the midrib often occupies most of the leaf and is covered with few to many longitudinal plates of green cells on the upper surface, sometimes also on the lower. Stomata are usually present at or near the base of the capsule. The peristome consists of a single series of 32-64 (or rarely 16) tongue-shaped teeth attached to a drumlike expansion of the tip of the columella. The tooth cells become long and fiberlike, thickened all around, and U-shaped, extending from the tip of one tooth to the tip of the next. The calyptrae are cucullate and often matted with hairs. The chromosome number is 7, sometimes in polyploid series. *See* BRYOPHYTA; BRYOPSIDA; DAWSONI-IDAE. Howard Crum

Polytropic process

A process which occurs with an interchange of both heat and work between the system and its surroundings. The nonadiabatic expansion or compression of a fluid is an example of a polytropic process. The interrelationship between the pressure (P) and volume (V) and pressure and temperature (T) for a gas undergoing a polytropic process are given by Eqs. (1) and (2), where a and b are the polytropic constants for

$$PV^a = \text{constant}$$
 (1)

$$\frac{P^{\sigma}}{T} = \text{constant}$$
 (2)

the process of interest. These constants, which are usually determined from experiment, depend upon the equation of state of the gas, the amount of heat transferred, and the extent of irreversibility in the process. Once these constants are known, Eqs. (1) and (2) can be used with the initial-state conditions $(P_1 \text{ and } T_1 \text{ or } V_1)$ and one final-state condition (for example, P_2) to determine the temperature or specific volume of the final state. From this information, and other equations of thermodynamics, the entropy change and the enthalpy or internal energy change for the process can be evaluated, so that the power produced by a turbine or necessary to drive a compressor can be computed.

Equations (1) and (2) for a polytropic process may be compared with Eq. (3) for the isothermal expansion or compression of an ideal gas, and Eqs. (4)

$$PV = \text{constant}$$
 (3)

$$PV^{-(C_p/C_v)} = \text{constant}$$

$$\frac{P^{(R/C_p)}}{T} = \text{constant}$$
(4)

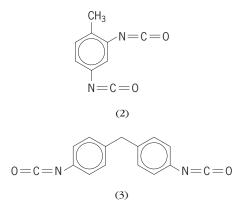
for the reversible adiabatic (isentropic) expansion or compression of an ideal gas. In Eqs. (3) and (4), R is the gas constant, and C_p and C_v the ideal gas heat capacity at constant pressure and volume, respectively. *See* GAS; ISENTROPIC PROCESS; ISOTHERMAL PROCESS; THERMODYNAMIC PROCESSES. Stanley I. Sandler

Polyurethane resins

Polymeric materials that are usually produced by reacting polyfunctional isocyanates $[R(N=C=O)_n,$ where R is alkyl or aryl] with a macrodiol, or polyol, and a glycol extender in the presence of a catalyst. The macrodiols are based on polyethers or polyesters or a combination of both. Some of these have terminal hydroxyl groups, while those reacted with an excess of diisocyanate have terminal isocyanate groups, as in (structure 1), the prepolymer of propylene oxide [with R(NCO)2]. In recent years, diamines have been used as extenders to achieve higher reaction rates in molding or spray applications. The polymerization reactions are done in a mold (casting, injection molding) or continuously in an extruder or on a conveyor (block foam). See GLYCOL; ISO-CYANATE; POLYMER; POLYESTER RESINS; POLYETHER RESINS; POLYMERIZATION; POLYOL.

Polyurethanes are the most versatile class of reaction polymers, providing products ranging from soft thermoplastic elastomers or elastomeric (spandex) fibers to flexible or rigid foams. The elastomers are used mainly in automotive bumper applications produced by the reaction injection-molding process. Spandex fibers are used in sport clothing applications. Flexible foams are used in upholstery materials for furniture and bedding, automotive seating, and carpet underlay. Rigid foams are used as insulation materials in building and construction, refrigerated appliances, and in vehicles. In 2000, the total polyurethane global consumption was about 8000 kilotons. *See* PLASTICS PROCESSING.

The commodity aromatic isocyanates are toluene diisocyanate (a distilled 80:20 mixture of the 2,4- and 2,6-isomers; TDI; structure **2**), 4,4' diphenylmethane diisocyanate (MDI; **3**), and polymeric MDI (PMDI).

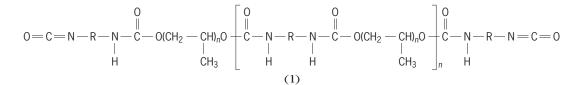


PMDI is obtained in quantitative yield in the phosgenation of aniline/formaldehyde resins. The diisocyanate, MDI, is separated from PMDI by thin-film vacuum distillation, and the distillation residue is sold as PMDI. The manufacture of TDI involves nitration of toluene, hydrogenation of the dinitrotoluenes to give diaminotoluenes, and phosgenation to give the diisocvanate.

When color stability is required in polyurethane resins, the more costly aliphatic diisocyanates are used. Examples are hexamethylene diisocyanate (HDI), isophorone diisocyanate (IPDI), and the reduced MDI (HMDI).

Foams. The diisocyanates used in polyurethanes are often prereacted with a small amount of the diols to form isocyanate-terminated prepolymers. Sometimes the lower-boiling aliphatic diisocyanates are converted into higher-boiling derivatives (allophanates, biurets, or trimers). In the manufacture of foams, water often is added as a comonomer. The water reacts with an isocyanate group to form carbon dioxide and an amine. The latter reacts with excess isocyanate to give a urea. In this manner, poly(urethane ureas) are generated. Poly(urethane ureas) are also obtained when diamines are used as extenders. The generated carbon dioxide is used as a co-blowing (foaming) agent in the production of foams. *See* AMINE; UREA.

Most commercial flexible foam is produced from TDI, a polyether triol, water, a surfactant, and tin/amine catalysts. Flexible foam blocks are produced continuously on a conveyor, while molded flexible foam is produced by reaction molding,



as in automotive seating applications. Semiflexible molded polyurethane foams are used in the construction of automotive instrument panels and dashboards. *See* CATALYSIS; SURFACTANT.

In the production of rigid polyurethane foams, chlorofluorocarbons (CFCs) were used as blowing agents. Because of the mandatory phase-out of CFCs caused by its effects on the ozone layer, new blowing agents, such as hydrogen-containing CFCs or hydro-carbons, are used today. In addition, poly(urethane isocyanurate) [PUIR] rigid foams are produced by using an excess of PMDI and a trimerization catalyst. The excess isocyanates undergo a cyclotrimerization reaction to form isocyanurate bonds, which have better thermal stability than polyurethane bonds. PUIR foams are used to insulate the main fuel tank of the space shuttle as well as liquefied natural-gas (LNG) tankers. *See* FOAM.

Elastomers. Polyurethane elastomers are segmented copolymers. The hard segments are formed in the reaction of the glycol extender with MDI, while the soft segments are obtained from the reaction of the macrodiol with MDI. Linear thermoplastic polyurethane elastomers (TPUs) are produced by reaction extrusion from MDI, polyols, and glycol extenders, while thermoset polyurethane elastomers are produced from TDI prepolymers and the diamine extender MOCA (3,3'-dichloro-4,4'diaminodiphenylmethane), or from MDI prepolymers by reaction injection molding. In the latter reaction, diethyltoluenediamine (DETDA) is used as chain extender. Also, amine-terminated polyols are used in reaction injection molding and spray applications. Sometimes, monomers containing unsaturated groups are used, which can be cured with peroxides. Polyurethanes with the rheology of natural rubber, called millable gums, are made by mixing hydroxylterminated prepolymers and isocyanate cross-linking agents, such as the 2,4-TDI dimer, and other additives on a rubber mill. Soft cast elastomers are also used for puncture-free tires. Other tire applications include solid industrial tires for off-road machines, pneumatic tires for tractors, and skate board and roller skate wheels.

Segmented elastomeric fibers are produced from MDI, using hydrazine or ethylene diamine extenders in a dry spinning process, in which the polymer solution is isocyanate prepolymers extruded through a spinerette into a column of circulated hot air. Wet spinning processes are also used, in which isocyanate terminated prepolymers are extruded into a diamine bath. *See* MANUFACTURED FIBER; RUBBER.

Adhesives. Reconstructed wood products, such as particleboard, increasingly are using PMDI as a liquid adhesive. The use of PMDI as a binder for foundry cores, rubber waste products, or solid rocket fuels is also known. Athletic surfaces are sometimes constructed from ground rubber tire scrap bonded with isocyanate prepolymers. Also, flexible foam scrap is bonded with isocyanate prepolymers to form carpet underlay. *See* ADHESIVE; COMPOSITE MATERIAL; WOOD PRODUCTS.

Polyurethane sealant formulations use TDI or MDI prepolymers made from polyether polyols. The sealants contain 30 to 50% of the prepolymer; the remainder consists of pigments, fillers, plasticizers, adhesion promoters, and other additives.

Coatings. A wide variety of tough and abrasionresistant polyurethane coatings are also available. Many are formed by reacting triols, such as castor oil, with excess diisocyante followed by moisture curing. Polyurethane alkyd coatings are produced by reacting an unsaturated drying oil with glycerol followed by curing with diisocyanates. Often polyurethane coatings are directly sprayed onto the surface of the substrate.

Solutions of fully reacted polyurethanes based on MDI or IPDI are also used in coating applications. Because of concern over volatile organic solvents, aqueous polyurethane dispersions with ionic groups in the polymer backbone are used. *See* PAINTS AND COATINGS. Henri Ulrich

Bibliography. W. F. Gum, W. Riese, and H. Ulrich (eds.), *Reaction Polymers*, Carl Hanser, Munich, 1992; P. Herrington and K. Hock (eds.), *Flexible Polyuretbane Foams*, Dow Chemical, 1991; N. M. K. Lamba, K. A. Woodhouse, and S. L. Cooper, *Polyuretbanes in Biomedical Applications*, CRC Press, 1997; G. Oertel, *Polyuretbane Handbook*, Carl Hanser, Munich, 1993; D. Randall and S. Lee (eds.), *The Polyuretbane Book*, Wiley, 2002; H. Ulrich, *The Chemistry and Technology of Isocyanates*, Wiley, 1993; H. Ulrich, Polyurethanes, in *Encyclopedia of Polymer Science and Technology*, 3d ed., Wiley, 2003.

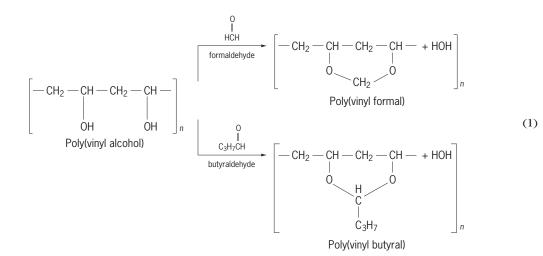
Polyvinyl resins

Polymeric materials generally considered to include polymers derived from monomers having the structure

$$H_2C = C \begin{bmatrix} R_1 \\ R_2 \end{bmatrix}$$

in which R₁ and R₂ represent hydrogen, alkyl, halogen, or other groups. This article refers to polymers whose names include the term vinyl. Of these polymers, several have been used for a number of years, such as poly(vinyl chloride), poly(vinyl acetate), poly(vinylidene chloride), poly(vinyl alcohol), poly(vinyl acetals), and poly(vinyl ethers). Indeed the terms vinyls and vinyl resins are frequently used to refer to the first three polymers of this group. Some polyvinyl resins of more recent origin are poly(vinyl fluoride), poly(vinylpyrrolidone), and poly(vinylcarbazole). For discussions of other vinyltype polymers. *See* POLYACRYLATE RESIN; POLYACRY-LONITRILE RESINS; POLYFLUOROOLEFIN RESINS; POLY-OLEFIN RESINS; POLYSTYRENE RESIN.

Many of the monomers can be prepared by addition of the appropriate compound to acetylene. For example, vinyl chloride, vinyl fluoride, vinyl acetate, and vinyl methyl ether may be formed by the reactions of acetylene with HCl, HF, CH₃OOH, and CH₃OH, respectively. Processes based on ethylene



Ac

as a raw material have also become common for the preparation of vinyl chloride and vinyl acetate.

The polyvinyl resins may be characterized as a group of thermoplastics which, in many cases, are inexpensive and capable of being handled by solution, dispersion, injection molding, and extrusion techniques. The properties vary with chemical structure, crystallinity, and molecular weight.

Poly(vinyl acetals). These are relatively soft, waterinsoluble thermoplastic products obtained by the reaction of poly(vinyl alcohol) with aldehydes [reaction (1)]. Properties depend on the extent to which alcohol groups are reacted. Poly(vinyl butyral) is rubber and tough and is used primarily in plasticized form as the inner layer and binder for safety glass. Poly(vinyl formal) is the hardest of the group; it is used mainly in adhesive, primer, and wire-coating formulations, especially when blended with a phenotic resin.

Poly(vinyl butyral) is usually obtained by the reaction of butyraldehyde with poly(vinyl alcohol). The formal can be produced by the same process, but is more conveniently obtained by the reaction of formaldehyde with poly(vinyl acetate) in acetic acid solution.

Poly(vinyl acetate). Poly(vinyl acetate) is a leathery, colorless thermoplastic material which softens at relatively low temperatures and which is relatively stable to light and oxygen. The polymers are clear and noncrystalline. The chief applications of poly(vinyl acetate) are as adhesives and binders for water-based or emulsion paints.

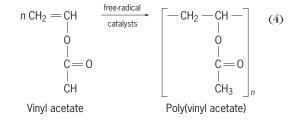
Vinyl acetate is conveniently prepared by the reaction of acetylene with acetic acid, as in reaction (2).

As acetaldehyde has become readily and cheaply available from the oxidation of olefins or hydrocarbons, it has been used to an increasing extent as a starting material. This is shown by reaction (3). Vinyl

$$CH_{3}CHO + O C - CH_{3} \xrightarrow{FeCl_{3}} C - CH_{3} \xrightarrow{FeCl_{3}} C - CH_{3} \xrightarrow{FeCl_{3}} C - CH_{3} \xrightarrow{O} C - CH_{3$$

acetate is also prepared by the vapor-phase reaction of ethylene, oxygen, and acetic acid using a palladium chloride catalyst.

Polymerization and copolymerization may be conveniently effected by free-radical catalysis in aqueous emulsion and suspension systems. Vinyl acetate copolymerizes readily with various other vinyl monomers; however, it does not copolymerize with styrene by the free-radical process. The polymerization process is represented by reaction (4).



See COPOLYMER.

Anhydrous solid polymers and copolymers may be used directly in adhesives.

Aqueous dispersions, produced by the emulsion polymerization process and commonly called polymer emulsions, or latexes, are used for treating textiles and paper, as adhesives, and as water-based paints. The water-based paints, prepared by pigmenting vinyl acetate polymer and copolymer emulsion achieved wide usage because of low cost of materials, ease of application, and resistance to weathering. As water is removed from the latex by evaporation or absorption, the suspended polymer particles coalesce into a tough film. The character of the film may be modified by the use of comonomers in the original polymerization or by the addition of plasticizers to the final emulsions.

Vinyl acetate is also frequently used as a comonomer with ethylene or vinyl chloride. *See* ETHYLENE.

Poly(vinyl alcohol). Poly(vinyl alcohol) is a tough, whitish polymer which can be formed into strong films, tubes, and fibers that are highly resistant to hydrocarbon solvents. Although poly(vinyl alcohol) is one of the few water-soluble polymers, it can be rendered insoluble in water by drawing or by the use of cross-linking agents.

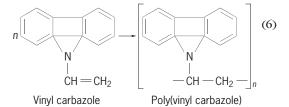
So far, vinyl alcohol itself, H_2C =CHOH, has not been isolated; reactions designed to produce the monomer yield the tautomeric acetaldehyde instead. However, the polymer can be produced on a commercial scale by the hydrolysis of poly(vinyl acetate), reaction (5).

Two groups of products are available, those formed by the essentially complete hydrolysis (97% or greater) of poly(vinyl acetate), and those formed by incomplete hydrolysis (50–90%).

The former, the completely hydrolyzed products, may be plasticized with water or glycols and molded or extruded into films, tubes, and filaments which are resistant to hydrocarbons; these can be rendered insoluble to water by cold-drawing or heat or by the use of chemical cross-linking agents. On cold drawing, the degree of crystallinity is substantially increased. These products are used for liners in gasoline hoses, for grease-resistant coating and paper adhesives, for treating paper and textiles, and as emulsifiers and thickeners. Insolubilized fibers have found uses in Japan for clothing, industrial fabrics, and cordage. Films are strong, tough, and relatively impermeable to oxygen but tend to be sensitive to moisture.

The partially hydrolyzed products are generally more water-soluble and less subject to crystallization by drawing. These materials are used as emulsifying solutions, in adhesive formulations, and textile sizes, and in water-soluble films.

Poly(vinyl carbazole). Poly(vinyl carbazole) is a tough, glassy thermoplastic with excellent electrical properties and the relatively high softening temperature of $120-150^{\circ}$ C (240-300°F). Polymerization, as represented by reaction (6), can be carried out in



bulk by free-radical catalysis. Uses of the product have been limited to small-scale electrical applications requiring resistance to high temperatures.

Poly(vinyl chloride). Poly(vinyl chloride) (PVC) is a tough, strong thermoplastic material which has an excellent combination of physical and electrical properties. The products are usually characterized as plasticized or rigid types. Poly(vinyl chloride) [and copolymers] is the second most commonly used poly(vinyl resin) and one of the most versatile plastics.

The plasticized types, either soft copolymers or plasticized homopolymers, are somewhat elastic materials which are familiar in the form of shower curtains, floor coverings, raincoats, dishpans, dolls, bottle-top sealers, prosthetic forms, wire insulation, and films, among others.

Rigid poly(vinyl chloride) products, which may consist of the homopolymer, copolymer, or polyblends, are used in the manufacture of phonograph records, pipe, chemically resistant liners for chemical-reaction vessels, and siding and window sashes.

The monomer is frequently prepared from chlorine, acetylene, and ethylene by a combination of processes which affords complete utilization of the chlorine, as shown, for example, in reactions (7) and (8).

$$CH_{2} = CH \xrightarrow{Cl_{2}} CH_{2}CI - CH_{2}CI \longrightarrow$$
Ethylene 1,2-Dichloro-
ethane
$$CH_{2} = CHCI + HCI \quad (7)$$
Vinyl Hydrogen
chloride chloride
$$CH = CH \xrightarrow{HCI} CH_{2} = CHCI \quad (7)$$

$$\begin{array}{c} \mathsf{CH} = \mathsf{CH} & \longrightarrow & \mathsf{CH}_2 = \mathsf{CHCl} & (8) \\ \mathsf{Acetylene} & \mathsf{Vinyl\ chloride} \end{array}$$

The polymerization of vinyle chloride, reaction (9), and its copolymerization with other vinyl

$$n \operatorname{CH}_{2} = \operatorname{CH} \xrightarrow{\operatorname{peroxides}} \begin{bmatrix} -\operatorname{CH}_{2} - \operatorname{CH}_{-} \\ | \\ \operatorname{CI} \end{bmatrix}_{n}$$
(9)
Vinvl chloride Poly(vinvl chloride)

monomers may be initiated by peroxide or azo compounds and carried out in bulk or in aqueous emulsion or suspension systems. The structure and properties of the product are quite dependent on polymerization temperature; the lower the temperature, the higher the softening point of the resin. Polymers prepared in bulk or suspension are used for many applications as molded, extruded, or calendered objects. Polymers prepared in emulsion are often molded by dipping or pouring techniques.

Because poly(vinyl chloride) products have a tendency to lose hydrogen chloride at high temperatures, a stabilizer such as a tin or lead compound is included in the final composition. Fillers are also commonly incorporated.

Copolymers with monomers such as vinyl acetate or propylene can be processed at lower temperatures than the homopolymer. The acetate copolymer is especially useful in floor tiles and phonograph records, though homopolymers are used as well.

Blends or "alloys" of poly(vinyl chloride) with small amounts of rubbery materials such as the interpolymer of acrylonitrile, butadiene, and styrene (ABS) have been produced for applications such as panels and pipe in which impact resistance, as well as hardness and strength, is desired.

Chlorination of poly(vinyl chloride) is sometimes effected to obtain a stiffer resin, at the expense of processability.

Poly(vinylidene chloride). Poly(vinylidene chloride) is a tough, horny thermoplastic with properties generally similar to those of poly(vinyl chloride). In comparison with the latter, poly(vinylidene chloride) is softer and less soluble; it softens and decomposes at lower temperatures, crystallizes more readily, and is more resistant to burning.

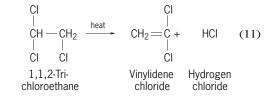
Because of its relatively low solubility and decomposition temperature, the material is most widely used in the form of copolymers with other vinyl monomers, such as vinyl chloride. The copolymers are employed as packaging film, rigid pipe, and as filaments for upholstery and window screens.

Vinylidene chloride is normally prepared by the pyrolysis of 1,1,2-trichloroethane. The latter is obtained by the chlorination of 1,2-dichloroethane which, in turn, is formed by the addition of chlorine to ethylene. This preparation is shown by reactions (10) and (11).

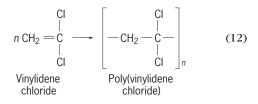
$$\begin{array}{cccc} CH_2 = CH_2 & \stackrel{+ Cl_2}{\longrightarrow} & CH_2 - CH_2 & \stackrel{+ Cl_2}{\longrightarrow} \\ & & & | & & | \\ CI & CI & CI \\ Ethylene & 1,2-Dichloro- \\ & & ethane \\ & & & CI \\ HCI + & CH - CH_2 & (10) \\ & & | & | \\ CI & CI \end{array}$$

1,1,2-Tri-

chloroethane



Polymerization as well as copolymerization may be initiated by peroxides and other free-radical catalysts and is most satisfactorily effected by emulsion and suspension techniques. Reaction (12) represents the polymerization.



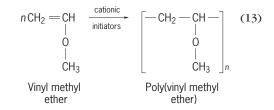
Because of the relatively low decomposition temperature of poly(vinylidene chloride), a stabilizer such as an amine is normally included in the composition.

Films of poly(vinylidene chloride), and especially the copolymer containing about 15% of vinyl chloride, are resistant to moisture and gases. Also, they can be heat sealed and, when oriented, have the property of shrinking on heating. By warming a food product wrapped loosely with a film of the polymer, a skintight, tough, resistant coating is produced.

By cold-drawing, the degree of crystallinity, strength, and chemical resistance of sheets, filaments, and even piping can be greatly increased.

Poly(vinyl ethers). Poly(vinyl ethers) exist in several forms varying from soft, balsamlike semisolids to tough, rubbery masses, all of which are readily soluble in organic solvents. Polymers of the alkyl vinyl ethers are used in adhesive formulations and as softening or flexibilizing agents for other polymers.

The monomers may be prepared by the reaction of alcohols with acetylene in the presence of alkali. Polymerization may be effected in bulk or solution at temperatures of -100 to 25° C (-148 to 77° F) by use of cationic initiators such as boron trifluoride, reaction (13). By careful choice of conditions, it is possi-



ble to achieve stereoregular polymerizations which yield partially crystalline polymers that are harder and tougher than the amorphous products.

Poly(vinyl methyl) ether is soluble in cold water but precipitates when the temperature is raised to about 35° C (95° F). The other alkyl vinyl ether polymers are insoluble in water. Copolymers of vinyl methyl ether with maleic anhydride are useful in textile applications.

Poly(vinyl fluoride). Poly(0vinyl fluoride) is a tough, partially crystalline thermoplastic material which has a higher softening temperature than poly(vinyl chloride). Films and sheets are characterized by high resistance to weathering.

Polymerization can be effected in the presence of oxygen and peroxidic catalysts, reaction (14). Be-

$$n \operatorname{CH}_{2} = \operatorname{CH} \xrightarrow{\operatorname{peroxides}} \begin{bmatrix} -\operatorname{CH}_{2} - \operatorname{CH}_{-} \\ | \\ F \end{bmatrix}_{n}$$
(14)
Vinyl fluoride Poly(vinyl fluoride)

cause of the low boiling point $(-88^{\circ}\text{C or} - 126^{\circ}\text{F})$ and high critical temperature of the monomer, polymerization is accomplished by use of pressure techniques similar to those employed in the highpressure process for polymerizing ethylene. Like other poly(vinyl halides), poly(vinyl fluoride) tends to lose the halogen acid at elevated temperatures.

Films are used in industrial and architectural applications. Coatings, for example, on pipe, are resistant to highly corrosive media.

Poly(vinyl pyrrolidone). Poly(vinyl pyrrolidone) is a water-soluble polymer of basic nature which has film-forming properties, strong absorptive or complexing qualities for various reagents, and the ability to form water-soluble salts which are polyelectrolytes. The polymer can be prepared by free-radical polymerization in bulk or aqueous solution, reaction (15). Iso-

$$\begin{bmatrix} CH_2 - CH_2 & \text{free-radical} \\ | & | & | \\ CH_2 & C = 0 \\ N & | \\ CH = CH_2 \end{bmatrix} \begin{bmatrix} CH_2 - CH_2 \\ | & | \\ CH_2 & C = 0 \\ N & | \\ -CH - CH_2 - \end{bmatrix}_n$$
(15)

tonic solutions were used in Germany in World War II as an extender for blood plasma. The main uses today are as a water-solubilizing agent for medicinal agents such as iodine, and as a semipermanent setting agent in hair sprays. Certain synthetic textile fibers containing small amounts of vinylpyrrolidone as a copolymer have improved affinity for dyes. *See* PLASTICS PROCESSING; POLYMER; POLYMERIZATION. John A. Manson

Bibliography. H. Allcock, F. Lampe, and J. Mark, *Contemporary Polymer Chemistry*, 3d ed., 2003; G. Odian, *Principles of Polymerization*, 4th ed., 2004; M. P. Stevens, *Polymer Chemistry: An Introduction*, 3d ed., 1998.

Pomegranate

A small deciduous tree, *Punica granatum*, belonging to the plant order Myrtales. Pomegranate is grown as an ornamental as well as for its fruit. Propagation is by cuttings and occasionally by layering. The pomegranate is a native of Asia, and was described by writers as early as 300 B.C. It was originally known for its medicinal qualities, and cures for various ills were attributed to the fruit juice, the rind, and the bark of the roots. The fruit is a reddish, pomelike berry, containing numerous seeds embedded in crimson pulp, from which an acidic, reddish juice may be obtained. Limited quantities are grown in California and the Gulf states. *See* FRUIT, TREE; MYR-TALES; PLANT PROPAGATION. J. H. Clarke

Poplar

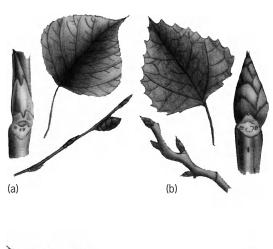
Any tree of the genus *Populus*, family Salicaceae, marked by simple, alternate leaves which are usually broader than those of the willow, the other American representative of this family. Poplars have scaly buds, bitter bark, flowers and fruit in catkins, and a five-angled pith. *See* SALICALES; WILLOW.

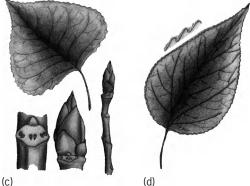
Some species are commonly called cottonwood because of the cottony hairs attached to the seeds. Other species, called aspens, have weak, flattened leaf stalks which cause the leaves to flutter in the slightest breeze. One of the important species in the United States is the quaking, or trembling, aspen (P. tremuloides), which attains a height of 90 ft (27 m) in the Rockies but is smaller in the East. This tree is widely distributed in North America from Labrador to Alaska. In the West it extends south through the Rockies and California to New Mexico and Lower California, and in the eastern United States it grows as far south as West Virginia. It is readily recognized by its comparatively small, finely toothed leaves, and by the shiny pointed winter buds (illus. a). The soft wood of this species is used for paper pulp. The bigtooth aspen (P. grandidentata) attains a height of 60-70 ft (18-21 m), and occasionally a diameter of 2 ft (0.6 m). The bigtooth aspen has a more restricted range in the northeastern quarter of the United States. The leaves are usually larger, $2^{1}/_{2}$ -4 in. (6-10 cm) long, and have larger teeth (illus. b), hence the common name. The buds are plumper and somewhat downy.

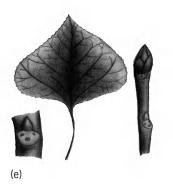
The European aspen (*P. nigra*; illus. *e*), which is similar to the quaking aspen, is sometimes planted, and its variety, *italica*, the Lombardy poplar of erect columnar habit, is used in landscape planting.

The black cottonwood (*P. trichocarpa*) is the largest American poplar and is also the largest broad-leaved tree in the forests of the Pacific Northwest. It attains a height of 175-225 ft (53-69 m) and a diameter of 7-8 ft (2.1-2.4 m). This tree ranges from southern Alaska to California and eastward through Washington and Oregon to Idaho, Montana, and Nebraska. The hairy fruit of the black cottonwood is a three-valved capsule.

The cottonwood or necklace poplar (*P. deltoides*; illus. *c*), native in the eastern half of the United States, is a fast-growing tree which usually attains 80-100 ft (24-30 m) in height and 3-4 ft (0.9-1.2 m) in diameter, but under favorable conditions in the





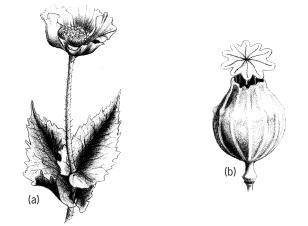


Common poplar species. (a) Quaking aspen (*Populus tremuloides*); leaf scar with an axial bud, leaf, and twig. (b) Bigtooth aspen (*P. grandidentata*); leaf, terminal bud, and twig. (c) Cottonwood poplar (*P. deltoides*); leaf, leaf scar, terminal bud, and twig. (d) Baldwin poplar (*P. balsamifera*); leaf showing detail of margin. (e) Lombardy poplar (*P. nigra* var. *italica*); leaf scar, leaf, and twig.

Mississippi Valley, it may attain a height of 150 ft (45 m) and a diameter of 7-8 ft (2.1–2.4 m). The leaves are broadly triangular (illus. *c*), hence the specific name, and the large terminal buds contain a pleasant-smelling balsamic resin. In the balsam or tacamahac poplar (*P. balsamifera*; illus. *d*), the resin is used in medicine as an expectorant. The wood is used for veneer, boxes, crates, furniture, paper pulp, and excelsior. It is also planted as a shade tree and used in shelter belts. *See* FOREST AND FORESTRY; TREE. Arthur H. Graves; Kenneth P. Davis

Poppy

A plant, *Papaver somniferum* (Papaveraceae), which is probably a native of Asia Minor. It is cultivated extensively in China, India, and elsewhere. This plant is the source of opium, obtained by cutting into the fruits (capsules) soon after the petals have fallen (see **illus.**). The white latex (juice) flows



Opium poppy (*Papaver somniferum*). (a) Stem with flower. (b) Capsule of poppy.

from the cuts and hardens when exposed to the air. This solidified latex is collected, shaped into balls or wafers, and often wrapped in the flower petals. This is the crude opium, which contains at least 20 alkaloids, including morphine and codeine. These drugs are used in medicine to allay pain, induce sleep, and relax spasms. Opium is one of the most useful drugs, but it is habit-forming and consequently should be used with the utmost caution. The opium habit is deleterious physically, mentally, and morally, and misuse of the drug is an extremely serious problem. *See* ALKALOID; OPIATES; PAPAVERALES. Perry D. Strausbaugh; Earl L. Core

Population dispersal

The process by which groups of living organisms expand the space or range within which they live. Dispersal operates when individual organisms leave the space that they have occupied previously, or in which they were born, and settle in new areas. Natal dispersal is the first movement of an organism from its birth site to the site in which it first attempts to breed. Adult dispersal is a subsequent movement when an adult organism changes its location in space. As individuals move across space and settle into new locations, the population to which they belong expands or contracts its overall distribution. Thus, dispersal is the process by which populations change the area they occupy.

Migration. This is the regular movement of organisms during different seasons. Many birds, butterflies, salmon, caribou, and other species migrate

between wintering and breeding ranges. Such migratory movement is marked by a regular return in future seasons to previously occupied regions, and so usually does not involve an expansion of population range. Some migratory species show astounding abilities to return to the exact locations used in previous seasons, even if experimentally transported away from their normal movement paths. Other species show no regular movements, but wander aimlessly without settling permanently into a new space. Wandering (called nomadism) is typical of species in regions where the availibility of food resources are unpredictable from year to year. Many birds in the harsh deserts of Australia are nomads. Neither migration nor nomadism is considered an example of true dispersal. See MIGRATORY BEHAVIOR.

Dispersal unit. Virtually all forms of animals and plants disperse. In most higher vertebrates, the dispersal unit is an entire organism, often a juvenile or a member of another young age class. In other vertebrates and many plants, especially those that are sessile (permanently attached to a surface), the dispersal unit is a specialized structure (disseminule). Seeds, spores, and fruits are disseminules of plants and fungi; trochophores and planula larvae are disseminules of sea worms and corals, respectively. Many disseminules are highly evolved structures specialized for movement by specific dispersal agents such as wind, water, or other animals.

Transportation. Disseminules or dispersing individuals can be transported passively (moved along with a dispersal agent) or actively (under their own efforts). Passive dispersal may involve movement by wind (anemochory) or water (hydrochory). Wind dispersal is common in plants and many small invertebrates such as spiders, and allows dispersers to be scattered over a large spatial area. Species that are associated with highly disturbed habitats (for example, dandelions) are often wind-dispersed, because patches of disturbed habitat may be widely scattered across a landscape, forcing dispersers to travel great distances.

Water dispersal is common in all marine and other aquatic organisms. Marine plankton usually contains many dispersers, such as the larval forms of organisms from the intertidal or benthic zones. Many terrestrial organisms found along shorelines, rivers, and streams also disperse passively by using water currents. Perhaps the best-known example of a disseminule that is evolved for dispersal using water currents is the coconut. Coconut palms are widely distributed on many tropical beaches, mainly because their disseminule, the coconut itself, can float across large expanses of ocean and still be viable when washed ashore.

Dispersal using animal agents (zoochory) involves attachment to skin, feathers, or scales of the animal agent with barbed or sticky surfaces on the disseminule, or ingestion of the disseminule by the dispersal agent. Many disseminules are highly specialized to be carried or eaten by a specific agent, and many agents are highly evolved to make use of specific disseminules. This close evolution between a disperser and its animal vector is an example of the process of coevolution. For example, phainopeplas, birds that inhabit the deserts of the southwestern United States, eat mistletoe berries for the thin layer of sugary pulp that surrounds the seed. The seeds pass through the digestive system of the bird and are excreted onto the branches of desert trees and shrubs, where the seeds stick and grow into new mistletoe plants. Thus, the phainopepla is an effective dispersal agent for the plant, while the plant is a reliable food source for the bird.

A special case of zoochory involves transport by humans. The movement of people and cargo by cart, car, train, plane, and boat has increased the potential dispersal of weedy species worldwide. Many foreign aquatic species have been introduced to coastal areas by accidental dispersal of disseminules in ship ballast water, which is usually dumped overboard when ships reach their destination. The zebra mussel is one exotic species that arrive in this manner and is now a major economic problem throughout the Great Lakes region of North America. Countries such as Australia and New Zealand take precautions to ensure that airplane passengers do not accidentally carry seeds of foreign plants into the countries in their luggage or on their shoes.

Some organisms have been deliberately introduced by humans into new areas. Domestic animals and plants have been released throughout the world by farmers. A few pest species were deliberately released by humans; European starlings, for example, were originally introduced to the United States via New York City. However, they have spread across most of North America and are considered to be one of the worst avian pests ever released on this continent.

Seed dispersal systems and coevolution. Some of the most highly coevolved dispersal systems are those in which the disseminule must be eaten by an animal. Such systems have often evolved a complex series of signals and investments by both the plant and the animal to ensure that the seeds are dispersed at an appropriate time and that the animal is a dependable dispersal agent. Such highly evolved systems are common in fruiting plants and their dispersal agents, which are animals called frugivores. Fruiting plants cover their seeds with an attractive, edible package (the fruit) to get the frugivore to eat the seed. To ensure that fruits are not eaten until the seeds are mature, plants change the color of their fruits as a signal to show that the fruits are ready for eating. Thus, when cherries change from green to dark red, potential frugivores (including humans) are alerted that the fruit is ready to eat.

When eaten by a frugivore, seeds must be able to survive passage through the animal's digestive system without being destroyed. These seeds are often covered with a tough outer coating to protect the seed from the digestion process. In some species, the seed must be scarred by the digestion process itself for the seed to be able to germinate. A loss of the frugivorous dispersal agent in these latter systems means that seeds will be unable to sprout because of lack of scarification. A number of Central American trees produce large numbers of huge fruits, which fall to the ground and rot under the parent tree, without germination. It has been speculated that these trees were coevolved to have their fruits dispersed by the large frugivorous members of the Pleistocene mammal fauna such as the gomphotheres. Extinction of these large mammals left the trees without an effective dispersal agent. The range of many of these tree species has decreased until the trees have become quite rare—an example of how ineffective dispersal can limit populations. However, cattle have been introduced to fill the niche of the extinct frugivores, and these tropical trees have begun to expand their range again as cattle disperse their seeds.

Many plants in the tropical rainforests are coevolved to have their seeds dispersed by specific animal vectors, including birds, mammals, and ants. Many tropical trees, shrubs, and herbaceous plants are specialized to have their seeds dispersed by a single animal species. Temperate forest trees, in contrast, often depend on wind dispersal of both pollen and seeds.

Dispersal barriers. Dispersal barriers are physical structures that prevent organisms from crossing into new space. Oceans, rivers, roads, and mountains are examples of barriers for species whose disseminules cannot cross such features. These types of barriers are often regions of unsuitable temperature, humidity, or light conditions in which disseminules die before reentering suitable conditions. Dispersal barriers do not have to be large. For example, several species of rare forest butterflies refuse to cross open, nonforested space, no matter how small the open gap; some neotropical forest birds do not cross open water. Thus, populations of such species can quickly become isolated due to lack of successful dispersal.

It is believed that the creation of physical barriers such as mountains and rivers is the primary factor responsible for the evolution of new species. A widespread species can be broken into isolated fragments by the creation of a new physical barrier. With no dispersal linking the newly isolated populations, genetic differences that evolve in each population cannot be shared between populations. Eventually, the populations may become so different that no interbreeding occurs even if dispersal pathways are reconnected. The populations are then considered separate species. Therefore, an understanding of the role that dispersal plays in maintaining genetic variation among populations is an important subject in evolutionary biology. *See* SPECIATION.

Conservation importance. Dispersal is of major concern for scientists who work with rare and endangered animals. Extinction is known to be more prevalent in small, isolated populations. Conservation biologists believe that many species exist as a metapopulation, that is, a group of populations interconnected by the dispersal of individuals or disseminules between subpopulations. The interruption of dispersal in this system of isolated populations can increase the possibility of extinction

of the whole metapopulation. Conservation plans sometimes propose the creation of corridors to link isolated patches of habitat as a way of increasing the probability of successful dispersal. Corridors have been proposed for species such as mountain lions in highly fragmented habitats in the western United States, but are controversial because little data usually exist on how the species involved actually disperses. *See* EXTINCTION (BIOLOGY); POPU-LATION DISPERSION. John B. Dunning, Jr.

Bibliography. C. S. Elton, *The Ecology of Invasions* by Animals and Plants, 1958; B. Sleeper, Animal Migrations: Wildlife in Motion, 1994; N. C. Stenseth and W. Z. Lidicker (eds.), Animal Dispersal: Small Mammals as a Model, 1994; I. R. Swingland and P. J. Greenwood, *The Ecology of Animal Movement*, 1983.

Population dispersion

The spatial distribution at any particular moment of the individuals of a species of plant or animal. Under natural conditions organisms are distributed either by active movements, or migrations, or by passive transport by wind, water, or other organisms. The act or process of dissemination is usually termed dispersal, while the resulting pattern of distribution is best referred to as dispersion. Dispersion is a basic characteristic of populations, controlling various features of their structure and organization. It determines population density, that is, the number of individuals per unit of area, or volume, and its reciprocal relationship, mean area, or the average area per individual. It also determines the frequency, or chance of encountering one or more individuals of the population in a particular sample unit of area, or volume. The ecologist therefore studies not only the fluctuations in numbers of individuals in a population but also the changes in their distribution in space. See POPULATION DISPERSAL.

Principal types of dispersion. The dispersion pattern of individuals in a population may conform to any one of several broad types, such as random, uniform, or contagious (clumped). Any pattern is relative to the space being examined; a population may appear clumped when a large area is considered, but may prove to be distributed at random with respect to a much smaller area.

Random or haphazard. This implies that the individuals have been distributed by chance. In such a distribution, the probability of finding an individual at any point in the area is the same for all points (**Fig. 1***a*). Hence a truly random pattern will develop only if each individual has had an equal and independent opportunity to establish itself at any given point. In a randomly dispersed population, the relationship between frequency and density can be expressed by Eq. (1), where *F* is percentage frequency, *D* is

$$F = 100(1 - e^{-D})$$
 (1)

density, and e is the base of natural or napierian logarithms. Thus when a series of randomly selected

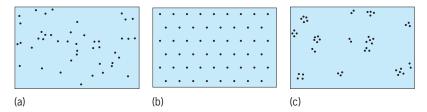


Fig. 1. Basic patterns of the dispersion of individuals in a population. (a) Random. (b) Uniform. (c) Clumped, but groups random. (*After E. P. Odum, Fundamentals of Ecology, Saunders*, 1953)

samples is taken from a population whose individuals are dispersed at random, the numbers of samples containing $0, 1, 2, 3, \ldots, n$ individuals conform to the well-known Poisson distribution described by notation (2).

$$e^{-D}, De^{-D}, \frac{D^2}{2!}e^{-D}, \frac{D^3}{3!}e^{-D}, \dots, \frac{D^n}{n!}e^{-D}$$
 (2)

Randomly dispersed populations have the further characteristic that their density, on a plane surface, is related to the distance between individuals within the population, as shown in Eq. (3), where \overline{r} is the

$$D = \frac{1}{4\bar{r}^2} \tag{3}$$

mean distance between an individual and its nearest neighbor. These mathematical properties of random distributions provide the principal basis for a quantitative study of population dispersion. Examples of approximately random dispersions can be found in the patterns of settlement by free-floating marine larvae and of colonization of bare ground by airborne disseminules of plants. Nevertheless, true randomness appears to be relatively rare in nature, and the majority of populations depart from it either in the direction of uniform spacing of individuals or more often in the direction of aggregation.

Uniform. This type of distribution implies a regularity of distance between and among the individuals of a population (Fig. 1b). Perfect uniformity exists when the distance from one individual to its nearest neighbor is the same for all individuals. This is achieved, on a plane surface, only when the individuals are arranged in a hexagonal pattern. Patterns approaching uniformity are most obvious in the dispersion of orchard trees and in other artificial plantings, but the tendency to a regular distribution is also found in nature, as for example in the relatively even spacing of trees in forest canopies, the arrangement of shrubs in deserts, and the distribution of territorial animals.

Contagious or clumped. The most frequent type of distribution encountered is contagious or clumped (Fig. 1*c*), indicating the existence of aggregations or groups in the population. Clusters and clones of plants, and families, flocks, and herds of animals are common phenomena. The degree of aggregation may range from loosely connected groups of two or three individuals to a large compact swarm composed of all the members of the local population.

Furthermore, the formation of groups introduces a higher order of complexity in the dispersion pattern, since the several aggregations may themselves be distributed at random, evenly, or in clumps. An adequate description of dispersion, therefore, must include not only the determination of the type of distribution, but also an assessment of the extent of aggregation if the latter is present.

Analysis of dispersion. If the type or degree of dispersion is not sufficiently evident upon inspection, it can frequently be ascertained by use of sampling techniques. These are often based on counts of individuals in sample plots or quadrats. Departure from randomness can usually be demonstrated by taking a series of quadrats and testing the numbers of individuals found therein for their conformity to the calculated Poisson distribution which has been described above. The observed values can be compared with the calculated ones by a chi-square test for goodness of fit, and lack of agreement is an indication of nonrandom distribution. If the numbers of quadrats containing zero or few individuals, and of those with many individuals, are greater than expected, the population is clumped; if these values are less than expected, a tendency toward uniformity is indicated. Another measure of departure from randomness is provided by the variance:mean ratio, which is 1.00 in the case of the Poisson (random) distribution. If the ratio of variance to mean is less than 1.00, a regular dispersion is indicated; if the ratio is greater than 1.00, the dispersion is clumped.

In the case of obviously aggregated populations, quadrat data have been tested for their conformity to a number of other dispersion models, such as Neyman's contagious, Thomas' double Poisson, and the negative binomial distributions. However, the results of all procedures based on counts of individuals in quadrats depend upon the size of the quadrat employed. Many nonrandom distributions will seem to be random if sampled with very small or very large quadrats, but will appear clumped if quadrats of medium size are used. Therefore the employment of more than one size of quadrat is recommended.

A measure of aggregation that does not depend on quadrat size of the mean density of individuals per quadrat and that can be applied to patterns consisting of a mosaic of patches with different densities has been developed by Morisita. His index of dispersion is a ratio of the observed probability of drawing two individuals randomly from the same quadrat to the expected probability of the same event for individuals randomly dispersed over the set of quadrats being studied. Index values greater than 1.0 indicate clumping, and values between 0 and 1.0 point to regularity of dispersion.

The fact that plot size may influence the results of quadrat analysis has led to the development of a number of techniques based on plotless sampling. These commonly involve measurement of the distance between a randomly selected individual and its nearest neighbor, or between a randomly selected point and the closest individual. At least four different procedures have been used (**Fig. 2**). The

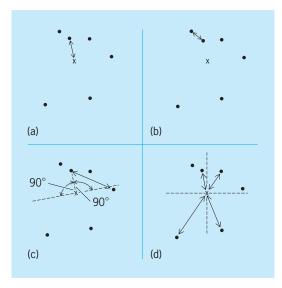


Fig. 2. Distances measured in four methods of plotless sampling. (a) Closest individual. (b) Nearest neighbor. (c) Random pairs, with 180° exclusion angle. (d) Point-centered quarter. The cross represents the sampling point in each case. (After P. Greig-Smith, Quantitative Plant Ecology, Butterworths, 1987)

closest-individual method (Fig. 2*a*) measures the distance from each sampling point to the nearest individual. The nearest-neighbor method (Fig. 2*b*) measures the distance from each individual to its nearest neighbor. The random-pairs method (Fig. 2*c*) establishes a base line from each sampling point to the nearest individual, and erects a 90° exclusion angle to either side of this line. The distance from the nearest individual lying outside the exclusion angle to the individual used in the base line is then measured. The point-centered quarter method (Fig. 2*d*) measures the distance from each sampling point to the nearest individual in each quadrant.

In each of these four methods of plotless sampling, a series of measurements is taken which can be used as a basis for evaluating the pattern of dispersion. In the case of the closest-individual and the nearestneighbor methods, a population whose members are distributed at random will yield a mean distance value that can be calculated by use of the densitydistance equation (3). In an aggregated distribution, the mean observed distance will be less than the one calculated on the assumption of randomness; in a uniform distribution it will be greater. Thus the ratio $\overline{r}_A/\overline{r}_E$, where \overline{r}_A is the actual mean distance obtained from the measured population and \overline{r}_E is the mean distance expected under random conditions, affords a measure of the degree of deviation from randomness.

Students of human geography have used the nearest-neighbor measure as a basis for a highly sophisticated methodology to analyze the dispersion of towns, department stores, and other features of land-use patterns.

Additional information about the spatial relations in a population can be secured by extending these procedures to measurement of the distance to the second and successive nearest neighbors, or by increasing the number of sectors about any chosen sampling point. However, since all of these methods assume that the individuals are small enough to be treated mathematically as points, they become less accurate when the individuals cover considerable space.

Factors affecting dispersion. The principal factors that determine patterns of population dispersion include (1) the action of environmental agencies of transport, (2) the distribution of soil types and other physical features of the habitat, (3) the influence of temporal changes in weather and climate, (4) the behavior pattern of the population in regard to reproductive processes and dispersal of the young, (5) the intensity of intra- and interspecific competition, and (6) the various social and antisocial forces that may develop among the members of the population. Although in certain cases the dispersion pattern may be due to the overriding effects of one factor, in general populations are subject to the collective and simultaneous action of numerous distributional forces and the dispersion pattern reflects their combined influence. When many small factors act together on the population, a more or less random distribution is to be expected, whereas the domination of a few major factors tends to produce departure from randomness.

Environmental agencies of transport. The transporting action of air masses, currents of water, and many kinds of animals produces both random and nonrandom types of dispersion. Airborne seeds, spores, and minute animals are often scattered in apparently haphazard fashion, but aggregation may result if the wind holds steadily from one direction. Wave action is frequently the cause of large concentrations of seeds and organisms along the drift line of lake shores. The habits of fruit-eating birds give rise to the clusters of seedling junipers and cherries found beneath such perching sites as trees and fencerows, as well as to the occurrence of isolated individuals far from the original source. Among plants, it seems to be a general principle that aggregation is inversely related to the capacity of the species for seed dispersal.

Physical features of the habitat. Responses of the individuals of the population to variations in the habitat also tend to give rise to local concentrations. Environments are rarely uniform throughout, some portions generally being more suitable for life than others, with the result that population density tends to be correlated directly with the favorability of the habitat. Oriented reactions, either positive or negative, to light intensities, moisture gradients, or to sources of food or shelter, often bring numbers of individuals into a restricted area. In these cases, aggregation results from a species-characteristic response to the environment and need not involve any social reactions to other members of the population. *See* ENVI-RONMENT.

Influence of temporal changes. In most species of animal, daily and seasonal changes in weather evoke movements which modify existing patterns of dispersion. Many of these are associated with the disbanding of groups as well as with their formation. Certain birds, bats, and even butterflies, for example, form roosting assemblages at one time of day and disperse at another. Some species tend to be uniformly dispersed during the summer, but flock together in winter. Hence temporal variation in the habitat may often be as effective in determining distribution patterns as spatial variation.

Behavior patterns in reproduction. Factors related to reproductive habits likewise influence the dispersion patterns of both plant and animal populations. Many plants reproduce vegetatively, new individuals arising from parent rootstocks and producing distinct clusters; others spread by means of rhizomes and runners and may thereby achieve a somewhat more random distribution. Among animals, congregations for mating purposes are common, as in frogs and toads and the breeding swarms of many insects. In contrast, the breeding territories of various fishes and birds exhibit a comparatively regular dispersion. *See* REPRODUCTIVE BEHAVIOR.

Intensity of competition. Competition for light, water, food, and other resources of the environment tends to produce uniform patterns of distribution. The rather regular spacing of trees in many forests is commonly attributed largely to competition for sunlight, and that of desert plants for soil moisture. Thus a uniform dispersion helps to reduce the intensity of competition, while aggregation increases it. *See* POP-ULATION ECOLOGY.

Social factors. Among many animals the most powerful forces determining the dispersion pattern are social ones. The social habit leads to the formation of groups or societies. Plant ecologists use the term society for various types of minor communities composed of several to many species, but when the word is applied to animals it is best confined to aggregations of individuals of the same species which cooperate in their life activities. Animal societies or social groups range in size from a pair to large bands, herds, or colonies. They can be classified functionally as mating societies (which in turn are monogamous or polygamous, depending on the habits of the species), family societies (one or both parents with their young), feeding societies (such as various flocks of birds or schools of fishes), and as migratory societies, defense societies, and other types. Sociality confers many advantages, including greater efficiency in securing food, conservation of body heat during cold weather, more thorough conditioning of the environment to increase its habitability, increased facilitation of mating, improved detection of, and defense against predators, decreased mortality of the young and a greater life expectancy, and the possibility of division of labor and specialization of activities. Disadvantages include increased competition, more rapid depletion of resources, greater attraction of enemies, and more rapid spread of parasites and disease. Despite these disadvantages, the development and persistence of social groups in a wide variety of animal species is ample evidence of its overall survival value. Some of the advantages of the society are also shared by aggregations that have no social basis. *See* ECOLOGICAL COMMUNITIES; SO-CIAL MAMMALS.

Optimal population density. The degree of aggregation which promotes optimum population growth and survival, however, varies according to the species and the circumstances. Groups or organisms often flourish best if neither too few nor too many individuals are present; they have an optimal population density at some intermediate level. The condition of too few individuals, known as undercrowding, may prevent sufficient breeding contacts for a normal rate of reproduction. On the other hand, overcrowding, or too high a density, may result in severe competition and excessive interaction that will reduce fecundity and lower the growth rate of individuals. The concept of an intermediate optimal population density is sometimes known as Allee's principle. See POPULATION GENETICS.

Francis C. Evans

Bibliography. A. A. Berryman (ed.), Dynamics of Forest Insect Populations, 1988; A. A. Berryman (ed.), Dynamics of Insect Populations: Patterns, Causes, Implications, 1988; P. Greig-Smith, Quantitative Plant Ecology, 3d ed., 1983; S. K. Jain and L. W. Botsford (eds.), Applied Population Biology, 1992; S. A. Levin (ed.), Applied Mathematical Ecology, 1989; A. G. Pakes and R. A. Maller, Mathematical Ecology of Plant Species Competition, 1990; E. C. Pielou, Population and Community Ecology: Principles and Methods, 1974.

Population ecology

The study of spatial and temporal patterns in the abundance and distribution of organisms and of the mechanisms that produce those patterns. Species differ dramatically in their average abundance and geographical distributions, and they display a remarkable range of dynamical patterns of abundance over time, including relative constancy, cycles, irregular fluctuations, violent outbreaks, and extinctions. The aims of population ecology are threefold: (1) to elucidate general principles explaining these dynamic patterns; (2) to integrate these principles with mechanistic models and evolutionary interpretations of individual life-history tactics, physiology, and behavior as well as with theories of community and ecosystem dynamics; and (3) to apply these principles to the management and conservation of natural populations.

Definition of a population. A population is the total number of individuals of a given biological species found in one place at one time. In practice, ecologists often deal with density—numbers per unit area for land organisms and numbers per unit volume in aquatic systems—or even weight rather than raw numbers. What may be described as an "individual" depends on the kind of organism and the aim of the scientific enquiry. In most animals, the life cycle starts with a fertilized egg, passes through a largely irreversible process of coupled growth and differentiation, and ends in a tightly integrated,

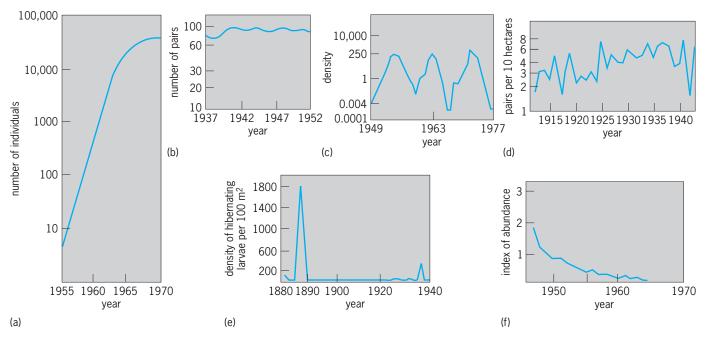


Fig. 1. Graphs of dynamical behaviors showing the diversity of population dynamics. (a) Collared dove; a phase of exponential growth asymptoting toward carrying capacity. (b) Yellow-eyed penguin; steady state. (c) Budmoth; regular cycles. (d) Songbird; irregular fluctuations. (e) Dendrolimus moth; outbreaks. (f) Blue whale; decline toward extinction.

unitary, adult organism. Population size can be measured by merely counting adult units and their juvenile prologues. But in most plants and some colonial animals, growth and differentiation proceed in a modular fashion; growth involves the replication of a basic body unit, so that a fertilized egg generates a spatially distributed "population" of connected modules. Modular organisms show tremendous plasticity in size and form. The connections between modules sometimes break, leading to a dispersed clone of physiologically independent units. For example, one clone of quaking aspen (which spreads by root buds) occupies 200 acres (80 hectares) and is more than 10,000 years old. To a geneticist, this clone constitutes a single huge individual; to an ecologist, each aspen trunk may be considered to be an individual. Modular organisms have an additional dimension of complexity that must be quantified to understand their dynamics. In this article, the population concepts presented apply broadly to both unitary and modular organisms.

There are only four ways a population can change in size: birth, death, immigration, and emigration. If immigration and emigration are negligible, the population is closed, and the difference between birth and death rates drives its dynamics. Terrestrial animals on islands often have closed populations. If immigration and emigration are important, however, the population is open, and its abundance may be substantially influenced by spatially distant events. For example, the number of barnacles that are found on a rocky coastline often reflects the density of setting larvae, which in turn is governed by events in offshore waters. If a population that is under study is found to be highly open, the spatial scale of the study may be too narrowly circumscribed to capture the important mechanisms of its population dynamics.

Basic population models. Populations exhibit a great variety of dynamical patterns, ranging from explosive outbreaks, to local extinctions, to regular cycles or relatively constant abundances (Fig. 1). To help describe and explain these patterns, ecologists rely on population models. Simple life cycles and closed populations provide a useful starting point in developing population models. Many temperatezone insects have one annual generation, and so at any given time all individuals are at the same stage of life. For a population with discrete, nonoverlapping generations, if N(t) is the number of adults censused in generation t, and R(t) is the number of adult offspring in generation t + 1 produced per adult in generation t, the number of individuals in the next generation is given by Eq. (1). The quantity R(t) is

$$N(t+1) = N(t)R(t)$$
(1)

the growth rate of the population for generation t. Iterating this discrete time growth model for subsequent generations allows one to project population numbers through time. In the limit of very small census intervals, population growth is described by the differential equation (2), where r(t), the instan-

$$\frac{dN}{dt} = r(t)N(t) \tag{2}$$

taneous per-capita growth rate, is the difference between per-capita birth and death rates. Equation (2) is literally true only if populations grow continuously with overlapping generations (as is approximately true for some microbes). Constant values for *R* or *r* imply exponential growth. If R > 1 (r > 0), a population grows without bounds; if R < 1 (r < 0), it declines to extinction. The theoretical framework of population ecology largely consists of elaborations of these basic growth models, including extensions to more complicated life cycles and multiple species.

Exponential growth has a snowballing effect: if the growth rate is constant, then the more individuals there are, the faster the population grows. Even low growth rates eventually lead to populations of enormous sizes. Some natural populations show transient phases of exponential growth, particularly in colonizing episodes: for instance, the collared turtledove invaded Great Britain in 1955 and increased exponentially for nearly a decade (Fig. 1*a*). The per-capita rate of growth during exponential growth is called the intrinsic rate of increase, r_0 . The value of r_0 quantitatively expresses the interplay of individual traits, such as life history strategies, with the environment.

Exponential growth during colonization often involves expansion across space as well as an increase through time. Equation (2) can be expanded to include immigration and emigration, as given in Eq. (3).

$$\frac{dN}{dt} = rN + (\text{net change due to dispersal})$$
(3)

When individuals move down spatial density gradients, which is analogous to chemical diffusion along concentration gradients, their rate of movement is characterized by a diffusion coefficient, *D*. In a homogeneous environment, this model predicts that the expanding population wave advances at a velocity $2(rD)^{1/2}$, implying a linear expansion in range area with time. Data from a number of colonizing populations show exactly this relationship.

Population heterogeneity. Models such as those given in Eqs. (1) and (2), when interpreted literally, assume that all members of a population are identical. This is rarely true. Birth and death rates typically vary as a function of age, body weight, and genotype. A great deal of work in population ecology is devoted to elucidating age-specific schedules of mortality and fecundity, using these patterns to predict population growth, and interpreting these patterns in the light of evolutionary theory. To study age-structured population dynamics, the number of individuals in each age class must be monitored. The two ingredients needed to project changes in population size and age structure are the mortality schedule or survivorship curve, which describes the fraction of newborns surviving to each age, and the fecundity schedule, which describes the rate of female births per female at each age. It is a formidable task to measure complete fecundity and mortality schedules in natural populations, but if these schedules are given, the geometric growth model of Eq. (1) can be generalized to a matrix model, as in Eq. (4), where N(t) is a vector in

$$N(t+1) = A(t)N(t) \tag{4}$$

which each element is the number of individuals in an age class, and A(t) is a matrix incorporating the fecundity and mortality schedules. When individuals can be categorized into discrete stages such as body size (for example, number of connected modules in a clonal organism) in addition to age, more general matrices can describe population growth; the matrix elements are rates of transition between each pair of stages. [The continuous time model of Eq. (2) can be similarly generalized by using partial differential equations.] A fundamental principle of demographic theory is that if these schedules are constant, a population will (with rare exceptions) eventually settle into a stable age distribution in which each age class comprises a constant fraction of the total population. A population in its stable age distribution grows geometrically, as in Eq. (1), at a rate of increase uniquely determined from the mortality and fecundity schedules. A population displaced from its stable age distribution may exhibit transient phases of growth or decline, divergent from its long-term growth pattern.

All populations are genetically variable. If different genotypes have different fecundity or mortality schedules, genetic variation can influence population dynamics. Models that simultaneously incorporate changes in genetic composition and population growth can be quite complex; however, they may be important in describing some populations and are necessary for linking population ecology with evolutionary theory.

Population limitation and regulation. Populations cannot expand exponentially forever. For instance, the growth rate of the collared dove substantially declined in the second decade of its invasion, probably because mortality rates rose or birth rates declined as a result of competition for limited resources such as food or nest sites. Such mechanisms are called negatively density-dependent factors. The notion of density dependence is a specific example of the more general concept of feedback. An alternative hypothesis is that the environment worsened for reasons unrelated to dove density, such as a shift in the weather. Such causes for variation in birth or death rates are called density-independent factors. Ecologists have long disputed the relative importance of density-dependent and density-independent factors in determining population size. The current consensus is that both are important but to differing degrees in different species and environments. For a population to be regulated, it must tend to increase when below a certain size and decrease when above that size. If growth rates vary with time but in a fashion unrelated to density, closed populations will eventually fluctuate to extinction or expand without limit. If a closed population persists over long periods of time, it must be regulated to some degree.

As illustrated below, population regulation by no means implies population stability. In general, a population is stable if it returns to equilibrium following a perturbation. Moreover, many local populations may not be persistent over long periods of time and so may not be regulated in the usual sense. Open populations, by definition coupled by dispersal with other populations, can become reestablished by immigration following a local extinction. The total population of a species may persist, even though no

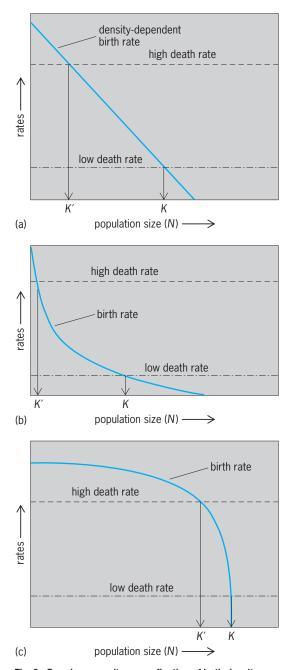


Fig. 2. Carrying capacity as a reflection of both density dependence and the intensity of density-independent mortality. Births are considered to be purely density-dependent, deaths purely density-independent. The three curves show the effects on population size of (a) an increase in density-independent mortality (a high death rate), (b) weak density dependence (near K), and (c) strong density dependence (near K).

single local population survives, because there is a spreading of risk among an ensemble of local populations (which experience somewhat different environmental conditions) that are loosely coupled by dispersal. *See* POPULATION DISPERSAL.

A useful method for considering the interplay of density-dependent and density-independent factors in determining population size is to plot birth and death rates as functions of density (**Fig. 2**). The carrying capacity K of a population in a given en-

vironment is defined to be the largest number of individuals for which the birth rate just matches the death rate. The population decreases above K and increases below K. A given change in density-independent death rates can produce very different changes in population size, depending on the form of the underlying density dependence. If density dependence is weak (Fig. 2*b*), fluctuations in mortality generate large oscillations in population size; if density dependence is strong (Fig. 2*c*), the population readily buffers such fluctuations. Density-dependent factors are necessary to regulate populations, but density-independent factors must also be considered to understand fully what limits populations to a given value of K.

Evidence for density dependence. Statistical analyses of time series of population data can suggest density dependence, but it is difficult to demonstrate density dependence conclusively by using such data. Sometimes density dependence can be shown for particular stages in the life cycle, and in general, density dependence may be observed in births, deaths, or dispersal. However, without examining density dependence at each stage in the life cycle, it is difficult to infer the regulatory importance of any single stage. Stronger evidence comes from manipulative experiments in which control populations are compared with artificially enhanced or depressed populations.

Density-dependent population models. The exponential growth model [Eq. (2)] can be modified to include density dependence by expressing r as a function of N. The simplest model that generalizes Eq. (2) is the logistic equation (5), in which per-

$$\frac{dN}{dt} = r_0 N \left(1 - \frac{N}{K} \right) \tag{5}$$

capita growth rate declines linearly with increasing density. Populations displaced from *K* converge smoothly back to it, without overshoots. Population growth is maximal when N = K/2. This model provides a good fit to some laboratory colonies and captures much of the qualitative behavior of more complex models; it is, however, a somewhat crude first approximation to accurate models of population growth.

The logistic model can be improved upon in numerous ways. The simplest method is to use a nonlinear function for the per-capita growth rate. Further refinements in the model can be achieved by incorporating threshold effects and explicitly incorporating submodels that encapsulate the mechanisms of density dependence. At densities far below carrying capacity, there may be an Allee effect-a positive relationship between density and per-capita growth rates. One intriguing possibility that arises in more complex models is that the population may exhibit alternative stable states; the one it actually occupies will depend upon accidents of history. For instance, insects may be regulated at low densities by bird predation, but at a higher threshold density the birds may be satiated and thus no longer regulate insect density; the insect population will then grow until checked by some other factor.

The discrete-time model [Eq. (1)] can be similarly modified to incorporate density dependence by expressing *R* as a function of *N*. Analysis of such models has led ecologists to reevaluate their traditional assumptions about the causes of population fluctuations in nature and about the relationship between regulation and stability. For instance, an analog of Eq. (5) is Eq. (6), which, along with similar equa-

$$N(t+1) = N(t) \exp\left[r\left(1 - \frac{N}{K}\right)\right]$$
(6)

tions, reveals a rich array of dynamical patterns. If r < 2, the population equilibrates stably at *K*; if 2 < r< 2.7, the population fluctuates cyclically; if r > 2.7, the population exhibits chaotic behavior, with cycles of arbitrary periodicity or even aperiodic fluctuations. Such fluctuations are intriguingly similar to the fluctuations in real-world data, which in the past were assumed to be produced by random environmental noise. The qualitative properties of the model that trigger pronounced population fluctuations are the time lag implicit in the discrete-time formulation, high growth rates at low densities, and strong density dependence at high densities. This suggests that a potential for complex dynamical behavior exists whenever there are time lags in the feedback between population size and population growth rates. Age structure is a ubiquitous source of time lags in populations, simply because time is required to reach reproductive maturity. Similarly, interactions between two, three, or more species can introduce long time lags, together with strong density dependence. It is an open question whether observed variability in natural populations reflects to any significant extent the complex dynamics latent in deterministic growth models instead of the force of fluctuations in the physical environment.

Mechanisms of density dependence. Given that density dependence exists, the mechanisms generating it can be used both to predict the consequences of environmental change for population dynamics and to provide insight into systems where experimental manipulations are difficult. Density dependence often arises from competition, which is said to exist when organisms utilize common limiting resources and thereby negatively affect each other. (A resource is limiting if an increase in its supply increases percapita growth rates.)

There are two principal sorts of competition, interference and exploitative. Interference competition occurs when one individual directly harms another. Interference may be dramatic, as in lethal aggression, or subtle, as when social interactions reduce the time available for gathering resources or increase the risk of predation. A surprising number of animal species are cannibalistic. Large scorpions, for instance, eat with relish their smaller-bodied conspecifics. Because encounter rate increases with increasing population size, cannibalism is likely to be a potent density-dependent factor in scorpion populations. Exploitative competition occurs when one individual consumes a resource such as food that otherwise would have been consumed by another individual. Because exploitative competition is mediated indirectly through a shared resource base, it can be more difficult to demonstrate than interference. In territorial animals, such as many songbirds, less space is available for additional territory holders as population size increases. As a result, competition for space can sharply cap population numbers.

Population regulation and interspecific interactions. Negative density dependence may arise from interspecific interactions. A schematic classification of interactions between two species comes from considering the positive (+) or negative (-) effect that individuals of one species have on the growth rate of the other. In interspecific competition the interaction is (-,-); in mutualism it is (+,+). Natural enemies, defined broadly to include predators, herbivores, and parasites, are often engaged in (+,-) relations with their prey or hosts. Most species are potential prey to one or more natural enemies; even top-level carnivores may be beset by parasites.

Obviously, competitors and mutualists can dramatically affect the size of a given population and so must be considered when studying population limitation. However, natural enemies are far more likely to be regulatory agents than are either competitors or mutualists. If two species are competing and one increases in density, the other will decrease. This will relax the interspecific competition on the first, which can then increase even more. Hence, competitive loops (and for similar reasons, mutualist loops) tend to produce positive feedback and so will not regulate population growth. By contrast, predator-prey interactions may produce negative density dependence acting across several time scales on both the predator and prey. As prey in one habitat patch become more numerous, predators may almost immediately become more active or switch over from other prey types or patches. Predators may also show an intergenerational numerical response to increased prey availability. Because an increase in predator numbers usually decreases prey numbers, this induces delayed density dependence in both the predator and its prey. In hostpathogen systems, these numerical responses may be pronounced, even within a host generation. For example, in tropical trees, seedling mortality from fungal pathogens increases with increasing seedling density, because the rate of spore dispersal increases as the distance between hosts declines. There is mounting evidence that parasites are significant regulatory factors in a wide range of natural populations, including economically important forest insects and game animals.

Several cautionary remarks about natural enemies and density dependence are in order. Predation and intraspecific competition can interact in complex ways. Compensatory responses by prey populations sometimes diminish the impact of enemies. For instance, plants typically produce many more seeds than can possibly become reproductive adults. Predation on some seeds may simply enhance the survivorship of the nonconsumed seeds, leading to no net effect on adult numbers. Conversely, predation can intensify intraspecific competition in mobile animals by restricting them to limited refuges. Generalist natural enemies can greatly depress prey numbers without being important regulatory agents, because by depending on many prey types they are less likely to show a strong numerical response to any one. Indeed, generalist predators may often be destabilizing, driving local prey populations to extinction. Finally, density-dependent responses by natural enemies often involve time lags, setting up the possibility of oscillatory behavior. Host-pathogen systems seem particularly likely candidates for generating strongly cyclic or chaotic dynamics.

Population growth models can incorporate interspecific interactions by taking models such as those in Eqs. (5) and (6) and adding expressions that describe how competitors, mutualists, or natural enemies affect the growth rate of a given species. The dynamics of two or more coupled species is studied in theoretical community ecology, which among other things seeks to understand how the species richness and stability of communities is related to the pattern and strength of interspecific interactions. Analyses of models of interacting species suggest that strong intraspecific density dependence is required for community stability. For two competing species to persist at a stable equilibrium, the negative effect that each species exerts on its own growth must on average exceed the cross-species negative effects. In like manner, predator-prey interactions are most stable when one or both populations experience intraspecific density dependence.

Applied population ecology. In addition to its intrinsic conceptual appeal, population ecology has great practical utility. Control programs for agricultural pests or human diseases ideally attempt to reduce the intrinsic rate of increase of those organisms to very low values. Analyses of the population dynamics of infectious diseases have successfully guided the development of vaccination programs. In the exploitation of renewable resources, such as in forestry or fisheries biology, population models are required in order to devise sensible harvesting strategies that maximize the sustainable yield extracted from exploited populations. Conservation biology is increasingly concerned with the consequences of habitat fragmentation for species preservation. Population models can help characterize minimum viable population sizes below which a species is vulnerable to rapid extinction, and can help guide the development of interventionist policies to save endangered species. Finally, population ecology must be an integral part of any attempt to bring the world's burgeoning human population into harmonious balance with the environment. See ECOLOGY; MATH-EMATICAL ECOLOGY; THEORETICAL ECOLOGY.

Robert Holt

Bibliography. M. Begon, J. L. Harper, and C. R. Townsend, *Ecology: Individuals, Populations, Com*-

munities, 3d ed., 1998; M. Gilpin and I. Hanski, *Metapopulation Dynamics: Empirical and Theoretical Investigations*, 1991; C. J. Krebs, *Ecology: The Experimental Analysis of Abundance and Distribution*, 5th ed., 2000; T. Royama, *Analytical Population Dynamics*, 1992; I. A. Shilov, *Population Ecology and Ecophysiology*, 1989; J. W. Silvertown and J. L. Doust, *Introduction to Plant Population Biology*, 1993.

Population genetics

The study of both experimental and theoretical consequences of mendelian heredity on the population level, in contradistinction to classical genetics which deals with the offspring of specified parents on the familial level. The genetics of populations studies the frequencies of genes, genotypes, and phenotypes, and the mating systems. It also studies the forces that may alter the genetic composition of a population in time, such as recurrent mutation, migration, and intermixture between groups, selection resulting from genotypic differential fertility, and the random changes incurred by the sampling process in reproduction from generation to generation. This type of study contributes to an understanding of the elementary step in biological evolution. The principles of population genetics may be applied to plants and to other animals as well as humans. See MENDELISM.

Mendelian populations. A mendelian population is a group of individuals who interbreed among themselves according to a certain system of mating and form more or less a breeding community. These individuals share a common gene pool which is the total genic content of the group. A mendelian population is the unit of study in population genetics. The population may be very large or very small, and is to be distinguished from species or varieties, which may consist of numerous isolated or partially isolated mendelian populations. Mendelian population is a genetic rather than a taxonomic term. Mendelian populations differ from each other in their genic content or chromosomal organization, not necessarily in their taxonomic features. The term deme, originally defined as an assemblage of taxonomically closely related individuals, has been used as a synonym for mendelian population. Gamodeme, a deme forming a more or less isolated local intrabreeding community, would be a better substitute.

Mutation pressure. Gene mutation arises from time to time in nature. The causes for mutation are not fully known, and thus it can be said that mutations arise "spontaneously." The effect of a new mutant gene is unpredictable and the gene is therefore said to mutate "at random." One property of mutation has been established: It is recurrent. Each type of gene mutates at a certain rate per generation. The rate is usually low—about 1 mutant in 10^5-10^8 genes of a given sort, varying from locus to locus on the chromosomes, even under uniform conditions. Ionizing radiation, certain chemicals, heat, and some

other agents increase the rate of mutation. See MU-TATION.

Let μ be the rate of mutation from an allele A to another form a per generation. If a fraction p of the genes of a population is A in one generation, then in the next generation the frequency of A will be diminished by the amount $p\mu$, so that the new frequency of A will be $p(1 - \mu)$. The amount of change, $p\mu$, is said to be due to the mutation pressure. If this pressure is unopposed generation after generation, the gene A will gradually disappear from the population, as $p_n = p_0(1 - \mu)^n \doteq p_0 e^{-n\mu}$, where p_0 is the initial gene frequency and p_n is the frequency after n generations. Therefore, for all existing genes there must be some kind of compensating mechanism which supports its continuing presence in nature. One important problem in population genetics is the mechanism of maintenance of a gene in a population or its change in frequency from generation to generation.

If, in addition to the mutation from A to a, there is reverse mutation from a to A at the rate v per generation, then the net amount of change in the frequency of *a* is $\Delta q = p\mu - q\nu$. At the time when these opposing changes cancel each other, there will be no change in gene frequency despite the recurrent mutations. This state of affairs is said to be in equilibrium and is obtained when $\Delta q = 0$; that is, $\hat{p} = \nu/(\mu + \nu)$ and $\hat{q} = \mu/(\mu + \nu)$, where q is a frequency of a, \hat{q} is the equilibrium point for a, and \hat{p} is the equilibrium point for A. The equilibrium gene frequencies are determined by the opposing rates of mutation only and are independent of the initial frequencies of the genes in the population. The amount of change in gene frequency per generation is larger when the current q is far away from the equilibrium \hat{q} than when q is close to \hat{q} . Substitution gives $\Delta q =$ $-(\mu + \nu)(q - \hat{q})$, indicating that the amount of change per generation is proportional to the deviation $(q - \hat{q})$. It also shows that if $q > \hat{q}$, q decreases, and if $q < \hat{q}$, q increases, or that q will approach \hat{q} from either side. Such an equilibrium is said to be stable. The changes in q described above are independent of the mating system practiced in the population.

In nature, and under artificial conditions, the mutation rates may not remain constant in all generations but may fluctuate within a certain range from time to time. In such cases, instead of a single fixed equilibrium point \hat{q} , there will be an equilibrium distribution of q within a certain range, and the apparent change in gene frequency from one generation to the next may be purely a stochastic phenomenon without necessarily having long-term significance. The same remark applies to all equilibria to be established in subsequent paragraphs. *See* STOCHASTIC PROCESS.

Migration and intermixture. If a fraction *m* of a population with a gene frequency *q* consists of immigrants from outside and the immigrant group has a gene frequency \overline{q} , the new gene frequency of the population will be $q_1 = (1 - m)q + m\overline{q} = q - m(q - \overline{q})$. The amount of change in gene frequency in one generation is thus $\Delta q = q_1 - q = -m(q - \overline{q})$, showing that the change is proportional

to the deviation $(q - \overline{q})$. This expression for Δq is of the same form as that for mutation. If the immigrants have the same gene frequency as the population, there will be no change in gene frequency in spite of the migrations. The continued intermixture of neighboring populations will eventually make them homogeneous in terms of gene frequencies. Thus, if a large population is divided into a number of partially isolated subpopulations, migrations between the groups will eventually make all subpopulations have the same gene frequency \bar{q} , which then denotes the average for the entire population in the absence of other disturbing factors. If the local populations are differentiated genetically, there must be some mechanism (for example, local selection) to counteract the pooling effect of migrations so that an equilibrium condition may be reached. The change in gene frequency due to migration is independent of the mating system practiced in the population.

Mating systems. In a gene pool with respect to one locus, if a proportion p of the genes is A and a proportion q of the genes is a, the genotypic proportions in the population are still unknown until the mating pattern is specified. The mating pattern is a system by which the genes are associated into pairs to form the diploid genotypes. The mating systems vary widely in nature for different organisms and populations. Thus, wheat may have 1% crosspollination and 99% self-fertilization, whereas maize practices just the reverse. One of the simplest and most extensively studied systems is random mating, also known as panmixis.

Panmixis. Random mating between individuals is equivalent to a random union of gametes. Thus, if the (pA, ga) gametes of one sex unite at random with the (pA, qa) gametes of the opposite sex, the resulting genotypic array will be p^2AA , 2pqAa, q^2aa . This principle was first discovered by W. E. Castle in 1903. These genotypic proportions will be realized only in very large populations. *See* HARDY-WEINBERG FORMULA; HUMAN GENETICS.

Inbreeding. Inbreeding refers to mating between genetically related individuals; the frequency with which two A gametes unite will be greater than p^2 ; and a similar situation is true for gene a. Consequently, inbreeding leads to an increase of homozygosis at the expense of heterozygosis. Let H be the heterozygosis proportion in a population, H' that in the preceding generation, H' that two generations ago, and so on. On continued systematic inbreeding, the manner in which the value of H decreases is shown in **Table 1**.

Continued close inbreeding, such as those degrees indicated in Table 1 and many others, eventually leads to complete homozygosis. The population will then consist of pAA and q aa; that is, an A gamete will always unite with another A gamete, and an a gamete with another a. Inbreeding between remote relatives does not necessarily lead to complete homozygosis but only decreases the heterozygosis below the random mating level to a certain extent.

Inbreeding system	Manner of heterozygosis (H) decrease	Limiting situation
Self-fertilization	$H = \frac{1}{2}H'$	H = 0.500H'
Same purebred sire $ imes$ successive daughters	$H = \frac{1}{2}H'$	H = 0.500 H'
Brother $ imes$ sister	$H = \frac{1}{2}H' + \frac{1}{4}H''$	H = 0.809 H'
Younger parent $ imes$ offspring	$H = \frac{1}{2}H' + \frac{1}{4}H''$	H = 0.809H'
Half brother $ imes$ half sisters	$H = \frac{3}{4}H' + \frac{1}{8}H''$	H = 0.890H'
Half brother $ imes$ full sisters	$H = \frac{1}{2}H' + \frac{1}{4}H'' + \frac{1}{16}H'''$ $H = \frac{1}{2}H + \frac{1}{4}H'' + \frac{1}{6}H'''$	H = 0.870 H'
Double first cousins	$H = \frac{1}{2}H + \frac{1}{4}H'' + \frac{1}{8}H'''$	H = 0.920H'

The inbreeding coefficient is an index intended to measure the amount or degree of inbreeding that has been accomplished in a population. Various indices may be constructed. One that has been proved highly useful in both theoretical investigation and practical breeding work is the inbreeding coefficient F defined as the correlation coefficient between the uniting gametes. The value of F ranges from 0 for random mating to 1 for inbreeding in a homozygous populations, as shown below.

Random mating:				
	A	а		
A	p^2	pq	p	
а	pq	q^2	q	
	þ	q	1	
	Correla	tion = 0)	

Inbred population:

	A	а	
A	$p^2 + Fpq$	pq – Fpq	p
а	pq – Fpq	$q^2 + Fpq$	q
	þ	q	1
	Correl	ation = F	

Inbreeding in homozygous population:

	A	а	
A	p	0	p
а	0	q	q
	p	q	1
(Correla	tion =	1

In an inbred population where the correlation between the uniting gametes is F, the genotypic array in the population will be as in Eqs. (1)-(3). These equa-

$$AAp^{2} + Fpq = (1 - F)p^{2} + Fp$$
 (1)

$$Aa2pq - 2Fpq = 2(1 - F)pq \tag{2}$$

$$aq^{2} + Fpq = (1 - F)q^{2} + Fq$$
 (3)

tions show that the population may be mathematically considered as having two separate components, (1 - F) panmictic and *F* fixed. If the mating system is such that *F* remains constant (instead of increasing) from generation to generation, the population will reach an equilibrium state with the genotypic array shown above.

The correlation between uniting gametes is due to the correlation between mating individuals. In an equilibrium population, if *M* denotes the correlation between mates, then M = 2F/(1 + F) or F = M/(2 - M).

Genotype selective values. Within a large population not all individuals produce the same number of offspring. In the situation to be considered, the average number of living offspring born to each of the genotypes in the population is studied, while the random fluctuation in the number of offspring from family to family is ignored. Furthermore, it is assumed that the population is so large that only the relative frequencies of the various genotypes and genes in the population are of interest. Suppose that the average number of offspring for each genotype is as follows: AA: 2.00; Aa: 2.50; aa: 1.50. Given these differential rates of reproduction, the new gene frequency of the next generation may be calculated. Inasmuch as it is only their relative magnitude that matters, these reproductive rates may be simplified into the ratio $W_{11}: W_{12}: W_{22} = 1:1.25:0.75$ or alternatively into 0.80:1:0.60. In order to standardize the description, it is convenient to take one of the three reproductive values as unity. In the previous example, depending upon whether the reproductive value of AA or Aa is taken as unity (the standard), that of *aa* is 0.75 = 1 - 10.25 or 0.60 = 1 - 0.40 = 1 - s in general. The value of s is known as the selection coefficient against the genotype aa. When a selection coefficient is used, it should always be stated which genotype has been employed as the standard.

Natural selection. The doctrine of the survival of the fittest needs clarification from the genetic viewpoint. The relative genotypic reproductive values $(W_{11},$ W_{12} , W_{22} of the preceding paragraph) simply given an ex post facto description, by which the genetic composition of the offspring generation may be related to that of the parent generation. These W values include all causes for differential reproduction such as fecundity, fertility, sexual maturity and capacity, survival and viability, length of reproductive life, and many others depending on the details of the life cycle of the organism. The W value, sometimes briefly referred to as relative fitness, is not necessarily correlated with any observable morphological characteristics, no matter how desirable they may seem to humans. From the genetic viewpoint, only those who reproduce count. Thus, natural selection has no

TABLE 2. Ef	TABLE 2. Effect of selection in a random mating population				
Genotype	Frequency,	Fitness, W	Frequency after selection, <i>fW</i>		
AA Aa aa Total	$\frac{p^2}{2pq}$ $\frac{q^2}{1.00}$	W ₁₁ W ₁₂ W ₂₂	$\frac{p^2 W_{11}}{2pq W_{12}}$ $\frac{q^2 W_{22}}{\overline{W}}$		

particular purpose except to perpetuate those who are fit to reproduce under the given conditions. Only when a characteristic lowers the organism's reproductive capacity does it have a genetic effect on the subsequent generations.

Selection pressure and equilibrium. The effect of selection may be described in terms of changes in gene frequency. In a random mating population with respect to one gene locus, the situation is as shown in **Table 2**.

The population after selection is the parental population of the next generation through random mating. The value \overline{W} is the total of the selected parental population, but may also be regarded as the average fitness of the original unselected population. The new frequency of gene *a* among the selected is $q' = (pqW_{12} + q^2W_{22})/\overline{W}$ and therefore the amount of change per generations is $\Delta q = q' - q$, or more explicitly, Eq. (4). This represents the effect of se-

$$\Delta q = \frac{pq}{2\overline{W}} - \frac{d\overline{W}}{dq} \tag{4}$$

lection pressure on gene frequency. When *p* or *q* is zero, there is no change in gene frequency; there can be no selection in the absence of alternatives. Therefore, all selection effects involve the factor *pq*. Further, when *p* or *q* is very small, the selection is ineffective whether it is for or against a gene. Besides these terminal conditions, if there exists a *q* value such that $\Delta q = 0$, it is called the equilibrium value of gene frequency, because a population with that particular gene frequency will remain unchanged in spite of the selection pressure. When such a *q* value exists, it must be the solution of the relations $d\overline{W}/dq = 0$ shown in Eq. (5). In order that \hat{q} be

$$\hat{q} = \frac{W_{11} - W_{12}}{(W_{11} - W_{12}) + (W_{22} - W_{12})}$$
(5)

a positive fraction, the differences $W_{11} - W_{12}$ and $W_{22} - W_{12}$ must be both positive or both negative; that is, the selective value of the heterozygote must be lower or higher than those of both homozygotes. For all other cases, there will be no equilibrium except when q = 0 or 1.

Stability of an equilibrium. The value of $\overline{W} = p^2 W_{11} + 2pqW_{12} + q^2W_{22}$ may be plotted against the value of q or p. When W_{12} is greater than W_{11} and W_{22} . The \overline{W} curve has a maximum point (**Fig. 1**). The q value corresponding to the maximum value of \overline{W} is the stable equilibrium point. This means that whether it is smaller or larger than \hat{q} , the q value will ap-

proach \hat{q} as selection proceeds from generation to generation. A stable equilibrium of this type leads to balanced genetic polymorphism, that is, to the coexistence of alleles in a population. Conversely, if q_{11} is lower than both W_{11} and W_{22} , the \overline{W} curve has a minimum point yielding an unstable equilibrium; the selection pressure will make the q value move away from the equilibrium value toward either 0 or 1, depending upon which side of the equilibrium the q happens to be. Consequently, selection against the heterozygote leads to the elimination of one of the alleles. In more complicated situations, there could be more than one stable or unstable equilibrium value in a population, or both.

The genotype selective values W_{11} , W_{12} , W_{22} have been assumed to be fixed for each genotype, but in nature they may vary in a number of ways. In addition to the omnipresent random fluctuations, the selective values may vary with the gene frequency itself. For instance, a genotype favored by selection when it is rare in the population may suffer a disadvantage when it is too common. For such cases, there will be an equilibrium yielding genetic polymorphism. Let U_{11} , a function of q, be the varying selective value of genotype AA, and so on. Then the equilibrium value of gene frequency is given by the appropriate solution of Eq. (6). The study of selection effects may be

$$q = \frac{U_{11} - U_{12}}{(U_{11} - U_{12}) + (U_{22} - U_{12})}$$
(6)

extended to cases with multiple alleles, sex-linked alleles, autopolyploids, and inbreeding populations.

Gamete selection. The effective rate at which the *A* and *a* gametes function may not be the same; that is, selection may operate in the gametic stage instead of in the diploid genotypic stage. If the selective actions for the genotype *aa*, for example, and gamete *a* are in opposite directions, an equilibrium may result.

Balance between selection and mutation. There are many different types of genotypic selection. Two simple cases will illustrate the principle of balance between selection and mutation pressures.

Selection against recessives. Suppose that the selective values of *AA*, *Aa*, and *aa* are 1, 1, and 1 - s, where *s* is a positive fraction known as the selection coefficient. Then the new gene frequency will be $q' = (q - sq^2)/(1 - q^2)$ in the next generation, so that the amount of change per generation is $q' - q = -sq^2(1 - q)/(1 - sq^2)$. At the same time, if μ is the

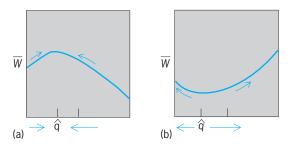


Fig. 1. Diagram of the relationship between the average fitness value \bar{W} and \hat{q} , the equilibrium point for *a*. (*a*) Stable equilibrium. (*b*) Unstable equilibrium.

mutation rate from *A* to *a*, the value of *q* will be increased by the amount $\mu(1 - q)$ per generation. At equilibrium, the forces to increase and to decrease the gene frequency must cancel each other; that is, $\mu(1 - q) = sq^2(1 - q)/(1 - sq^2)$. By solving, $sq^2 = \mu/(1 + \mu)$ is obtained, which closely approximates μ . Hence, $a^2 = \mu/s$ and $q = \sqrt{\mu/s}$, which usually is a small quantity. When *aa* is lethal or unable to reproduce, s = 1 and $q = \sqrt{\mu}$. This explains the persistence of deleterious recessive genes in a population in spite of continuous selection.

Selection against dominants. If the selection is against homozygous dominants only, the situation is the same as in the previous instance except for substitution of p for q and v (mutation rate from a to A) for μ . To bring out the distinction between selection against dominants and that against recessives, take the extreme case in which AA is lethal, the selective value of Aa is 1 - s, and aa is the norm. The value of p will then be so low that the usual genotypic proportions p^2 , 2pq, q^2 will take the limiting form 0, 2p, 1 - 2p, as q is very close to unity. The increase in q through selection is approximately *sp*, whereas the loss through mutation from *a* to *A* is qv = v. Hence, at equilibrium, p = v/s. This value is much lower than $q = \sqrt{\mu/s}$ for selection against recessives. Thus, selection against dominant alleles is more effective than selection of the same intensity against recessives. Selection against heterozygotes will eventually lead to the same limiting situation. All the equilibrium values supported by mutation pressure are low but stable. Mutations prevent complete extinction of an allele.

Random drift. The random drift of gene frequencies in finite populations is often called the Sewall Wright effect because of his analysis of its significance. The gene frequency of any generation is determined by the uniting gametes produced by the parents of the preceding generation. If the number of parents is limited and constitutes a random sample of the entire population, the gene frequency of the next generation will not remain exactly the same as that of the previous generation but will be subject to a random fluctuation on account of the sampling process. In a random mating population of N individuals, one-half of whom are males and one-half females, and maintaining the same population size, the variance of the gene frequency based on 2N gametes is q(1-q)/2N. The gene frequency may become a little higher or a little lower in the following generation. The smaller the population, the greater is the variance. This random process will continue to operate in all generations. In a sufficiently long time, the value of q will reach either the terminal value 0 or 1. Hence the random drift leads eventually to complete homozygosis for small populations. It can be shown that the limiting rate of reaching the state 0 or 1 is each $\frac{1}{4}$ per generation, so that the total rate of "decay" of genetic variability is $1/_2N$ per generation. Naturalists have found numerous small isolated colonies (for example, snails in mountain valleys) with characteristics uncorrelated with the environmental conditions to substantiate the theory of random (nonadaptive) fixation.

The effective size of a population is the actual number of individuals producing offspring and thereby responsible for the genetic constitution of the next generation. The random mating population with onehalf males and one-half females, and producing the same number of offspring, is an idealized model. Any deviation from the ideal situation will have a different sampling variance and a different rate of decay. Equating these to the "standard" variance q(1-q)/2N or the ideal decay rate 1/2N, an equivalent N is obtained for the ideal population. The latter number is known as the effective size of a population. It is convenient to use in mathematical descriptions of the genetic behavior of a population. Some of the factors that tend to make the effective size smaller than the actual breeding size are given below.

1. Unequal number of males and females. If *M* and *F* are the respective numbers, the effective size N_e is not simply M + F but is defined by $1/N_e = \frac{1}{4}M + \frac{1}{4}F$ and is equal to $N_e = \frac{4MF}{(M+F)}$. The larger the difference between *M* and *F*, the smaller the number N_e as compared with M + F.

2. Unequal size of families. If the gametes are drawn wholly at random from the parents, the number of gametes k contributed by a parent will form a Poisson distribution. In such a case, the effective size is the same as the actual breeding size. However, without perfect random sampling, if the mean number of gametes per parent is $\bar{k} = 2$, the effective size is equal to N_e as expressed in Eq. (7), where σ is presumably larger than 2.

$$N_e = \frac{4N-2}{\sigma_k^2 + 2} \tag{7}$$

3. Inbreeding. If *F* is the inbreeding coefficient of a population, then the effective size is $N_e = N/(1 + F)$.

Periodic change in population size. If N_1, N_2, \ldots , N_t are the respective sizes of the *t* generations, the average effective size for the period is approximately equal to the harmonic mean of the *t* sizes. The harmonic mean is much closer to the smallest number of a series than to the largest one.

Distribution. A stationary distribution of gene frequencies results from two opposing forces: the systematic pressures (mutation, migration, selection) which tend to make the gene frequency attain a certain fixed value, and the random variation due to sampling which tends to make the gene frequency drift away from any fixed value. The result of these opposing tendencies is not a single equilibrium value of gene frequency but a stationary distribution of gene frequencies. This distribution may be viewed in three different ways: as the distribution of q for a particular locus in a population in a long period of time; as the distribution of the allelic frequencies of all loci subject to the same pressures in one population at any given time; and finally as the distribution of q of one locus among a large number of populations of the same size and with the same pressures at a given time. See ALLELE.

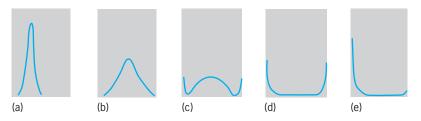


Fig. 2. Distribution of gene frequency in various populations under differing conditions. (a) Large population, distribution clustered close to equilibrium value. (b) Intermediate population, distribution of wider spread. (c) Intermediate small population, distribution with some terminal fixation. (d) Small population, most genes close to fixation or extinction in the absence of systematic pressures. (e) Small population with strong selection pressure.

Under selection pressure and the sampling variation, the distribution function $\phi(q)$ is $\phi(q) = C\bar{W}^{2N}/q(1-q)$, where *C* is a constant, \bar{W} the average fitness of the population, and *N* the effective size of the population. The exact form of the distribution depends upon the value of \bar{W} , which is a function of the selection coefficients as well as the gene frequency.

If there is mutation, or migration pressure, or both, the distribution of gene frequency is simply a β distribution: $\phi(q) = Cq^{U-1}(1-q)^{V-1}$, where $U = 4N\mu$, $V = 4N\nu$ if there is only mutation pressure; U = $4Nm\bar{q}$, $V = 4Nm\bar{p}$ if there is only migration pressure; and $U = 4N(\mu + m\bar{q})$, $V = 4N(\nu + m\bar{p})$ if there are both. The exact form of the β distribution depends upon the values of U and V. When they are smaller than unity, the population is considered small; when they are close to unity, the population is intermediate in size; when they are much larger than unity, the population is considered large. When the effects of mutation, migration, and selection are combined, the distribution function is as shown in Eq. (8). **Figure 2** shows some of the forms of the

$$\phi(q) = C \overline{W}^{2N} q^{U-1} (1-q)^{V-1}$$
(8)

distribution under various conditions. The joint distribution of more than one locus is naturally very complicated. Furthermore, if the mutation rates and selection coefficients also vary instead of being constant, the mathematical description of the stochastic process becomes very laborious and the forms shown in Fig. 2 give only a first approximation to the real situation.

The distribution forms of the gene frequencies depend upon the relative magnitudes of the various factors which bring about populational changes. In large random-mating populations all gene frequencies remain close to their stable equilibrium values, which are determined by the counteracting but systematic pressures of mutation, selection, and migration. There will be no further genetic change unless the environmental conditions change so as to define new equilibrium points. Evolution is such large populations is guided essentially by intragroup selection, and progress is very slow.

In small and completely isolated populations, most of the gene frequencies are close to 0 or 1 because of the random drift process which dominates the situation. Selection is ineffective. The loci are prevented from being completely fixed only by occasional mutations or immigrants. The ultimate fate of such small homozygous populations is probably extinction because they are nonadaptive and unable to respond to new conditions.

In populations of intermediate size, all factors, both random and systematic, come into play and the population is more responsive to evolutionary change. If a large population is subdivided into many partially isolated groups with migrations between them, there will be some differentiation among the groups, some of it adaptive and some nonadaptive, but there is very little fixation. The selection effect, varying from one locality to another, then operates largely on an intergroup basis which is more efficient than the intragroup selection within one single large population. If the groups are small, some of them will be eliminated by selection while others flourish. This provides the most favorable condition for evolutionary success for the species as a whole. The conclusion is that there is no one all-important factor in evolution. Evolutionary advance depends upon the interplay and balance of all factors.

Two-locus selection. All of the properties described above apply strictly to one locus, or to independently distributed loci. When selection acts upon a two-locus genotype directly, the genes of two loci, linked or unlinked, are in general not independently distributed, even in a large random mating population. In selection involving more than one locus, it is inadequate to consider the gene frequencies at the various loci separately and it is necessary to consider the gamente frequencies. As the simplest example, consider two unlinked loci (A,a and B,b). Let the four gamete frequencies (x) be as follows:

	В	b	
A	x_1	x_2	p
а	<i>x</i> ₃	x_4	q
	и	v	1

Random union of these gametes yields the following offspring genotype frequencies:

	BB	Bb	bb	
AA	x_{1}^{2}	$2x_1x_2$	x_{2}^{2}	<i>p</i> ₂
Aa	$2x_1x_3$	$2(x_1x_4 + x_2x_3)$	$2x_2x_4$	2pq
aa	x_{3}^{2}	$2x_3x_4$	x_{4}^{2}	q_2
	u^2	2 <i>uv</i>	v^2	. 1

Suppose that the nine genotypes have the relative fitness values (*W*) indicated below:

		BB	Bb	bb
	AA	1	1	0
<i>W</i> :	Aa	1	2	1
	aa	0	1	1

The symmetry of the selection pattern ensures the

gene frequencies to be $p = q = \frac{1}{2}$ at the (*A*,*a*) locus and $u = v = \frac{1}{2}$ at the (*B*,*b*) locus. But it may be verified that the equilibrium gamete frequencies are

	В	b	
A a	.28493 .21508	.21508 .28492	.50 .50
	.50000	.50000	1.00

and not

	В	b	
A a	.25 .25	.25 .25	.50 .50
	.50	.50	1.00

The alleles of the two loci are positively correlated. The deviation from random distribution is D = .28492 - .25000 = .03492. If these two loci were linked, the deviation D would be even more pronounced. Thus, selection leads to association (positive or negative) among the alleles of different loci. Also, the ultimate result of selection is not necessarily the maximization of the average fitness (\overline{W}) of a random mating population as is the case with only one locus (Fig. 1). The lack of independence of genes at different loci renders the problem of natural selection and its evolutionary consequences far more complicated than is depicted by simple mathematical models. *See* BIOMETRICS; GENETICS; MUTATION. C. C. Li

Bibliography. G. DeJong, *Population Genetics and Evolution*, 1988; J. S. Gale, *Theoretical Population Genetics*, 1989; J. B. S. Haldane, *The Causes of Evolution*, 1966, reprint 1990; N. D. Henderson and R. Benno (eds.), *Developmental Behavior Genetics: Neural, Biometrical and Evolutionary Approaches*, 1990; T. Nagylaki, *Introduction to Theoretical Population Genetics*, 1992.

Population viability

The ability of a population to persist and to avoid extinction. The viability of a population will increase or decrease in response to changes in the rates of birth, death, and growth of individuals. In natural populations, these rates are not stable, but undergo fluctuations due to external forces such as hurricanes and introduced species, and internal forces such as competition and genetic composition. Such factors can drive populations to extinction if they are severe or if several detrimental events occur before the population can recover. The concept of population viability gained importance with the National Forest Management Act of 1976, which charged the U.S. Forest Service with maintaining "viable populations" of all species of vertebrates found on its properties. See ECOLOGY; POPULATION ECOLOGY.

Factors affecting population viability. The viability of a population depends on the number of individuals in the population and on how the population size is changing.

Intrinsic rate of increase. Simple deterministic models can be used to assess how well a population is doing. In these models, the current number of individuals in the population and the expected number of births, deaths, immigrants, and emigrants determine the number of individuals found in the future. To evaluate a population's viability, ecologists calculate the intrinsic rate of increase (r), which is the change in population size per individual per unit time. When r > 0, the population is increasing in size, and when r < 0, it is decreasing in size. Viable populations are those with intrinsic rates of increase greater than or equal to 0. See DETERMINISM.

Stochasticity. In natural populations, viability depends both on the intrinsic rate of increase and on the sources and magnitude of stochasticity (chance variation) in this growth rate. Three different causes of stochasticity are usually recognized: environmental, catastrophic, and demographic.

1. Chance variation in a population's growth rate caused by random variation in temperature, rainfall, competitors, or other conditions is called environmental stochasticity. As a result of these factors, the population may grow faster or slower than the longterm average. The variation may be due to small-scale fluctuations in, for example, resources or predators, and runs of "bad" or "good" years can occur randomly.

2. Chance variation in a population's growth rate resulting from large-scale or severe disturbances such as hurricanes, fires, or floods is known as catastrophic stochasticity. In general, it involves factors that reduce a population's size. For example, in 1994 the golden toad went extinct because the ephemeral breeding pools in Costa Rica where the toads normally develop, dried up before metamorphosis took place.

3. Chance variation in a population's growth rate that arises from random variation in birth and death rates is known as demographic stochasticity. Traits such as survival and reproduction are not perfectly predictable. For example, a bird may have a 50% probability of producing female offspring, but 50% of her offspring will not necessarily be female. This example of predicted sex ratios had real consequences for the dusky seaside sparrow-all five remaining individuals were males, so the species went extinct in 1987. Demographic stochasticity is especially a problem for small populations, in which low population density can result in lower rates of mating and, therefore, lower birth rates and population growth. See ECOLOGICAL SUCCESSION; POPULATION DISPERSION; STOCHASTIC PROCESS.

Genetic variability. In addition to the sources of variability above, changes in the frequency, number, or distribution of genes in a population may put populations at risk of extinction. When populations are small, genetic drift (changes in genetic diversity due to chance mating and survival) can result in the loss of genetic diversity. Low genetic diversity often produces a number of deleterious effects, including inbreeding depression and a lack of potential for evolutionary response to changing environmental conditions. One example is the well-studied lion

population from the Ngorongoro Crater in Tanzania. This population was dramatically reduced by biting flies (catastrophic stochasticity), and the outcome was a dramatic loss of genetic diversity and increased inbreeding. As a result, reproductive success has declined, and males show high levels of sperm abnormalities.

Population viability analysis. One of the most important uses of population viability models comes from modern conservation biology, which uses these models to determine whether a population is in danger of extinction. This is called population viability analysis (PVA) and consists of demographic and genetic models that are used to make decisions on how to manage populations of threatened or endangered species. The National Research Council has called population viability analysis "the cornerstone, the obligatory tool by which recovery objectives and criteria [for endangered species] are identified." *See* ECOLOGICAL MODELING.

Population viability analysis treats extinction as a random process and therefore quantifies viability in terms of the probability of extinction. The viability of a population can be evaluated in two principal ways. Scientists can estimate (1) the mean time to extinction of a population of a given size, given specified stochastic factors, and (2) the minimum population size necessary for a population to persist for a specified period of time (for example, the population size necessary to be 95% confident that the population will persist for 100 years). This population estimate is referred to as the minimum viable population size (MVP). The MVP is the smallest population size that will persist for some specified length of time with a specified probability. In practice an MVP is defined as the smallest population size that has a 95% probability of persisting for 100 years.

The earliest population viability models were based on principles of population genetics. Scientists realized that the number of individuals present in a population was not equal to the number of individuals breeding in the population. Most estimates of the number of individuals breeding in a population suggest it is usually 10-33% of the true population size. The difference between the true and breeding population sizes can be as simple as the difference between the number of reproductively active individuals and the total number of individuals, or it can be due to more subtle factors such as variation in population size, sex ratio, or among-individual variance in reproduction. Early population-genetic models predicted that a population size of 50 would reduce the inbreeding rate to 1% per generation and would maintain genetic variation over the short term, whereas a population size of 500 would be large enough to maintain genetic variation in the population over the long term. These early benchmarks have been roundly criticized for a number of reasons, but particularly because they are based on the genetic data from a single population of fruit flies that may not be representative of the genetics of other species. See POPULATION GENETICS.

Deterministic models had been used since the early 1970s in the management of endangered species. The first true application of PVA was with stochastic models developed to investigate grizzly bear management in Yellowstone National Park. M. L. Shaffer developed a stochastic population simulation that incorporated chance events (demographic and environmental stochasticity) and produced estimates of extinction and population sizes needed to prevent extinction. *See* ENDANGERED SPECIES.

The components of PVA are not strictly defined; each PVA must be tailored to the biology of the study species. Therefore, it must incorporate the best available information on the study species. For example, many small organisms are restricted to a few habitat patches, but within those patches they often have high population densities. For these species, PVAs should focus on environmental and catastrophic stochasticity. A good PVA is designed to determine (1) whether the population is viable in both the short term and long term, (2) the factors that are currently putting the population at risk, and (3) how these risks can be reduced or eliminated so that the population can both survive and recover.

Population risk analysis. Over the past decade, the use of PVA for population risk analysis has become accepted. The development of easy-to-use computer programs has also contributed to the proliferation of PVA models. Although use of these models has increased, the method has come under increasing criticism. The two main sources of criticism are the data used to calibrate the models, and the methods used to confirm and structure models. For example, demographic data, the principal inputs of the models, are often based on studies too limited in duration to reasonably estimate the variance in birth rates and survival due to environmental or catastrophic stochasticity.

Population viability analysis has both strengths and weaknesses. It does allow a prediction of how long a population of a given size will persist, given the magnitude and frequency of different stochastic disturbances, which can, in turn, be useful in making predictions for endangered species or for designing preserves. However, any given population may go extinct within the first year, or it may persist for over 10 years. Therefore, by its very nature, the prediction of stochastic behavior of a particular population may not be very accurate.

Despite these methodological concerns, population viability will remain an important concept in conservation biology and population ecology. Habitat loss, habitat fragmentation, introduced species, and other recent environmental changes have resulted in the loss or serious decline in a number of plant and animal populations. Our ability to preserve and protect these species will depend on understanding the forces that determine population viability. Gretchen LeBuhn; Thomas E. Miller

Bibliography. M. S. Boyce, Population viability analysis, *Annu. Rev. Ecol. Systemat.*, 23:481-506, 1992; W. Morris et al., *A Practical Handbook for* Population Viability Analysis, Nature Conservancy, Arlington, VA, 1999; M. L. Shaffer, Minimum population sizes for species conservation, *BioScience*, 31:131–134, 1981; M. E. Soulé, *Viable Populations for Conservation*, Cambridge University Press, 1987; B. A. Wilcox, Extinction models and conservation, *Trends Ecol. Evol.*, 1:46–48, 1986.

Porcelain

A high-grade ceramic ware characterized by high strength, a white color (under the glaze), very low absorption, good translucency, and a hard glaze. Equivalent terms are European porcelain, hard porcelain, true porcelain, and hard paste porcelain. *See* GLAZING; POTTERY.

Porcelain is distinguished from other fine ceramic ware, such as china, by the fact that the firing of the unglazed ware (the bisque firing) is done at a lower temperature $(1800-2200^{\circ}F \text{ or } 1000-1200^{\circ}C)$ than the final or glost firing, which may be as high as $2700^{\circ}F (1500^{\circ}C)$. In other words, the ware reaches its final state of maturity at the maturing temperature of the glaze.

The white color is obtained by using very pure white-firing kaolin or china clay and other pure materials, the low absorption results from the high firing temperature, and the translucency results from the glass phase.

The term porcelain has been applied to such items as electrical insulators and bathroom fixtures. Very often these are made in a one-fire process, the glaze being applied to the green or unfired ware; where this is the case and high-grade materials are used in compounding the body, the term porcelain may be correctly applied. However, the pieces have no translucency because of their great thickness. On the other hand, the term porcelain is often applied to quite different ware. For example, zircon porcelain is used to describe a material made largely of zircon (ZrSiO₄), with small amounts of fluxes to yield a low absorption. *See* CERAMICS.

J. F. McMahon

Bibliography. J. Doherty, *Porcelain*, 2002; J. S. Reed, *Principles of Ceramics Processing*, 2d ed., 1995.

Porcupine

A mammal with spines or quills in addition to regular hair. The 23 species are classified in the order Rodentia and consist of two families: the Hystricidae, or Old World porcupines (11 species), and the Erethizontidae, or New World porcupines (12 species). Although traditionally considered to be related, the two families of porcupines may have evolved independently from different, unrelated ancestors, The infraorbital foramen is usually larger than the foramen magnum, and the zygomatic arch is also very large. Members of both families have a dental formula of I 1/1, C 0/0, Pm 1/1, M 3/3 \times 2 = 20. Porcupines are herbivorous, terrestrial, and generally nocturnal. *See* DENTITION; MAMMALIA; RODENTIA.

Quills. Quills are present on the back, sides, and tail of porcupines. Each quill consists of long sharp bristles of hair that are fused together. Porcupines defend themselves by striking attackers with their quilled tails. When a porcupine is disturbed, it rattles the quills on its tail. These produce a warning signal that resembles the noise created by a rattlesnake's tail. If the sound does not make the would-be aggressor retreat, the porcupine turns its back on it, erects its rapierlike quills, and charges backward. The quills come out easily and stick in the attacker's flesh. The porcupine grows new quills to replace the ones lost. Porcupines cannot "shoot" quills at their enemies. In New World porcupines, the tip of each quill is covered with tiny backward-pointing projections called barbs. The barbs hook into the flesh, and the quills are almost impossible to remove. Because of the barbed tip, the quill works continuously deeper into the wound and can even cause death. Quills progress at the rate of 1 millimeter (0.04 in.) or more per hour. Porcupine victims may die from infections caused by germs on the quills or from damage to a vital organ. Quills may even stick in an attacker's jaw, making the animal unable to open its mouth and causing starvation. Fishers (Martes pennanti), large members of the weasel family, attack porcupines by flipping them over onto their backs and attacking their unprotected ventral surface.

Hystricidae. Old World porcupines live in Africa, southeastern Asia, India, and southern Europe. Most species attain a length of about 3 ft (90 cm), including the tail. They are terrestrial or semifossorial (live both above and below ground) and make their homes in tunnels in the ground. Often six or seven porcupines share a communal burrow. They do not climb trees.

The crested porcupine (Hystrix) is the largest living porcupine. It may reach a length of over 3 ft (90 cm) and weight of 40-60 lb (18-27 kg). Its tail is just a few inches long. A mass of needle-pointed hollow quills extending from the nape of its neck and down its back is mixed with much longer and more slender flexible spines, or guards, measuring up to 21 in. (53 cm) in length. The guards protect the sharp points of the quills when they are not in use. This species lives in arid mountains, rocky outcrops, wooded hills, scrub-covered plains, and thick forest. These animals feed on bulbs, the bark of trees, and fallen fruit, and are destructive to farm crops, including root tubers, pumpkins, sweet potatoes, Irish potatoes, and maize. The annual litter usually consists of two young that are born after a gestation of 6-8 weeks. The nest consists of leaves, dry grass, and root fibers. The precocial young are well developed at birth with their eyes open and their bodies covered with soft flexible spines. The young remain in the den until the spines have grown and hardened; they are already quite dangerous at the end of 10 days. Average life expectancy in the wild is 12-15 years, although captive animals have lived for over 20 years.

prcupines and their distribution		
Name	Number of Species	Range
Hystricidae		
Long-tailed porcupine (Trichys)	1	Malaysia, Sumatra, Borneo
Brush-tailed porcupine (Atherurus)	2	China, Thailand, Malaysia, Gambia, Kenya, Zaire
Old World porcupine (Hystrix)	8	Philippines, Sumatra, Borneo, China, Malaysia, India, Africa
Erethizontidae		
Upper Amazonian porcupine (Echinoprocta)	1	Colombia
Prehensile-tailed porcupine (Coendou)	4	Central and South America
Hairy dwarf porcupine (Sphiggurus)	6	South America
North American porcupine (Erethizon)	1	North America

The brush-tailed porcupine (*Atherurus*) is only 20 in. (51 cm) in length, not including the 9-in. (23-cm) tail. It has short soft spines on the head, legs, and ventral surface, and very long grooved spines on the back. The long spiny tail terminates in a tuff of stiff bristles. The odd thing about this brush is that the bristles are composed of alternating thick and thin regions, appearing rather like a string of beads. The long-tailed porcupine (*Trichys*) has weak spines, or bristles, without quills and a long tail.

Erethizontidae. New World porcupines (Erethizontidae) are found in mixed coniferous forests, tropical forests, grasslands and deserts. Despite their generally chunky heavyset bodies, erethizontids can be arboreal unlike Old World porcupines. The North American porcupine (*Erethizon*) [see illustration], which weighs an average of 15 lb (7 kg) but may reach 40 lb (18 kg), is both terrestrial and semiarboreal, whereas South American species (Coendou and Sphiggurus) have prehensile tails and spend most of their time aboveground. Erethizon has a well-developed big toe on its hindfoot, whereas the hindfoot of the South American species has a broad pad opposable to the four toes. Heavy spines with a barbed tip, embedded singly and not in clusters as in Old World porcupines, occur over much of the dorsum and sides of the body. The ventral surface has coarse long hair and no spines. The claws are long and curved.

The staple winter food of North American porcupines is the cambium layer of many woody plants.



North American porcupine (Erethizon dorsatum). (Photo from National Park Service)

During other seasons, they feed on buds, twigs, sedges, and young pine needles. *Erethizon* characteristically takes shelter in rock piles, beneath overhanging rocks, or in hollow logs, but does not dig burrows as do Old World porcupines. These porcupines are normally solitary animals except during the breeding season and in winter denning concentrations. Breeding occurs from September through November. After a gestation period of 205-215 days, exceptionally long for a rodent, a single young is born.

Prehensile-tailed porcupines, or coendous, are arboreal, with short limbs modified for climbing. They have a thick-set short body covered with long hairs mixed with spines, which increase in number toward the posterior and dorsal part of the body and the tail. These spines, or quills, can be neither "dropped" nor "shot" but can be made loose. Although not usually found great distances from trees, these porcupines can exist in areas of jumbled rock and shrubs.

Although porcupines are not endangered, they are being affected by loss of habitat, especially in South America. One species (*Sphiggurus pallidus*) may have become extinct in historic times. It was reported from the West Indies in the midnineteenth century, but no porcupines occur there now.

Donald W. Linzey

Bibliography. J. A. Chapman and G. A. Feldhamer (eds.), *Wild Mammals of North America: Biology, Management, and Conservation*, Johns Hopkins University Press, 2003; D. Macdonald (ed.), *The Encyclopedia of Mammals*, Andromeda Oxford, 2001; R. M. Nowak, *Walker's Mammals of the World*, 6th ed., Johns Hopkins University Press, 1999; J. O. Whitaker, Jr., and W. J. Hamilton, Jr., *Mammals of the Eastern United States*, Comstock Publishing, 1998.

Porifera

The sponges, a phylum of the animal kingdom which includes about 5000 described species. The body plan of sponges is unique among animals. Currents of water are drawn through small pores, or ostia, in the sponge body and leave by way of larger openings called oscula. The beating of flagella on collar cells or choanocytes, localized in chambers on the interior of the sponge, maintains the water current. Support for the sponge tissues is provided by calcareous or siliceous spicules, or by organic fibers, or by a combination of organic fibers and siliceous spicules. Some species have a compound skeleton of organic fibers, siliceous spicules, and a basal mass of aragonite or calcite. The skeletons of species with supporting networks of organic fibers have long been used for bathing and cleaning purposes. Because of their primitive organization, sponges are of interest to zoologists as an aid in understanding the origin of multicellular animals. *See* ANIMAL KINGDOM; PARAZOA.

The Porifera have a fossil record extending from the Precambrian to Recent times. More than 1000 genera of fossil sponges have been described from the Paleozoic, Mesozoic, and Cenozoic eras.

Taxonomy. The living Porifera are divided into four classes on the basis of their skeletal structures. A taxonomic scheme of the Porifera follows. See separate articles on each class.

Class Hexactinellida Subclass Amphidiscophora Order: Amphidiscosa Hemidiscosa Reticulosa Subclass Hexasterophora Order: Hexactinosa Lynchniscosa Lyssacinosa Class Calcarea Subclass Calcinea Order: Clathrinida Leucettida Subclass Calcaronea Order: Leucosoleniida Sycettida Subclass Pharetronida Class Demospongiae Subclass Tetractinomorpha Order: Homosclerophorida Choristida Spirophorida Hadromerida Axinellida Subclass Ceractinomorpha Order: Dendroceratida Dictyoceratida Halichondrida Haplosclerida Poecilosclerida (Permosphincta, extinct) Class Sclerospongiae

General structure and cell types. The structure and functioning of a sponge is most easily studied in young fresh-water sponges as they develop from dormant bodies called gemmules. Such young sponges possess two discrete canal systems, an inhalant and an exhalant system, which are in communication with each other by way of numerous chambers lined with flagellated collar cells (**Fig. 1**). The outer or dermal epithelium is made up of flattened polygonal

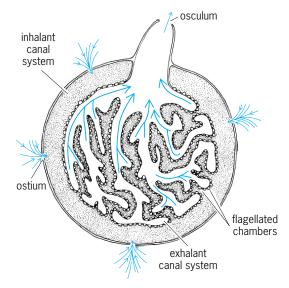


Fig. 1. Diagram of the canal system of a young fresh-water sponge which was cultured between cover slip and microscope slide.

cells called pinacocytes, and is perforated at intervals by pores, or ostia, through which water enters the inhalant canal system, usually by way of large subdermal cavities. The exhalant canals are also lined by pinacocytes.

All other cell types in the sponge (**Fig. 2**) are located in the inhalant canals. These cells may be divided into (1) those with large nucleolate nuclei and (2) those with smaller nuclei lacking nucleoli or with very small nucleoli. In the former category are the archeocytes, which are ovoid in shape and have

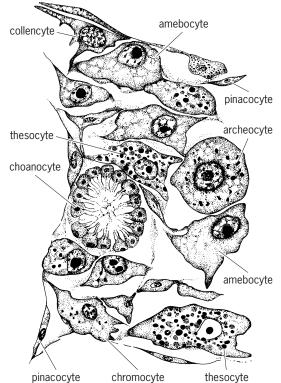


Fig. 2. Cell types found in a fresh-water sponge as seen in a cross section through the interior of the sponge.

blunt pseudopodia and hyaline cytoplasm, and many types of wandering amebocytes, which are named according to their functions. They are termed chromocytes if they bear either pigment granules or zoochlorellae (unicellular green algae); thesocytes or spherular cells if they are filled with food reserves; and trophocytes if they are acting as nutritive cells for developing eggs. Mucus-secreting gland cells, presumably derived from amebocytes, are of common occurrence in sponges. Scleroblasts, derived from archeocytes, secrete spicules, and spongioblasts secrete spongin fibers.

The anucleolate cells with small nuclei include the pinacocytes, collencytes, and choanocytes. In addition to their function as epithelial cells, pinacocytes may be differentiated into spindle-shaped contractile cells, called myocytes, which occur in circlets under the epithelium around ostia and oscula. Collencytes are called desmacytes or fiber cells and are found in abundance in the cortex of some species of sponges. Star-shaped collencytes with a vesicular or vacuolate cytoplasm are called cystencytes. The choanocytes are the most highly differentiated cells of the sponge. Each consists of a small spherical cell body surmounted by a collar. In freshwater sponges and certain marine Demospongiae and Calcarea, cells called lophocytes (Fig. 5b) occur beneath the dermal membrane. These cells bear a process terminating in a tuft of fine cytoplasmic filaments that secrete fibrils of collagen.

The mesenchymal cells and the skeleton of the Demospongiae and Calcarea are surrounded by a colloidal gel, called mesoglea, secreted by amebocytes and possibly choanocytes. Elastin fibers help provide support for the sponge and occur in the mesoglea of most species.

Water-current-system physiology. The flagellated chambers are hemispherical with a diameter of 40-60 micrometers in fresh-water sponges. They are composed of choanocytes firmly held together. The uncoordinated beating of the flagella of the choanocytes creates the flow of water through the sponge. Water enters the flagellated chambers from the inhalant canals by way of openings between choanocytes. Two or three such small openings (prosopyles) lead into each chamber. Water leaves each chamber and enters the exhalant system by way of a single larger pore (apopyle) through the epithelial lining of the exhalant canal. The lumen of an apopyle is 10 times greater than that of all the prosopyles leading into the chamber; thus water enters the chambers at a velocity 10 times greater than that at which it leaves. The collars of the choanocytes are so oriented that each directs its current toward the apopyle.

A fall in pressure takes place between apopyles and prosopyles, and therefore the prosopyles exert a suction effect on the water in the inhalant canal. The effect of this lowered pressure is transmitted to the subdermal cavity, and water is therefore drawn into the ostia forcibly. Food particles that may be at considerable distances from the pores are sucked in along with the water. The incurrent canals are filled with a network of free cells which tend to decrease the velocity of the water, but the retarding effect of these frictional forces is compensated by the resultant decrease in diameter of the streams of water passing through the inhalant canals. The velocity of current flow through the inhalant canals is twice as great as that through the exhalant canals, which have smooth walls but are also greater in diameter. An excess pressure exists in the exhalant canals as a result of the pumping action of the flagellated chambers and the changing size of the oscular opening. Hence food-free water laden with excretory products is carried far away from the sponge as it leaves the osculum, and is unlikely to reenter through the ostia.

Feeding and digestion. Feeding activities are best known in fresh-water sponges in which the dermal ostia, the inhalant canals, the flagellated chambers, and the choanocyte collars constitute a set of sieves of decreasing mesh size. The dermal ostia are 50 μ m in diameter, the prosopyles are 5 μ m in diameter, and the spaces between the cytoplasmic tentacles of the collars vary from 0.10 to 0.15 μ m (Fig. 3). Only the smallest particles can enter the flagellated chambers and come in contact with the collars of the choanocytes. The latter trap the small particles, which are passed to the cell body and ingested. Larger particles that have been able to pass through the dermal pores are ingested by the archeocytes and collencytes that form a reticulum in the inhalant canals or by the cell bases of the choanocytes (Fig. 4). Particles too large to enter the ostia can be ingested by the pinacocytes of the dermal epithelium. Studies of tropical, coral-reef-inhabiting demosponges indicate that they are highly efficient in filtering out bacteria and unarmored flagellates and moderately efficient in retaining diatoms, dinoflagellates, and

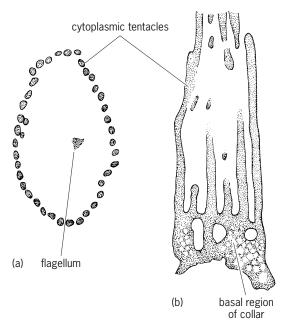


Fig. 3. Sections through the collar of a choanocyte of a fresh-water sponge. (a) Transverse. (b) Tangential.

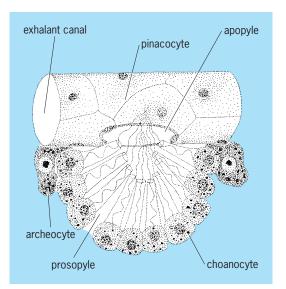


Fig. 4. Diagram of a flagellated chamber of a fresh-water sponge. The chamber is shown in optical section; exhalant canal, in three dimensions.

coccolithophores. Microscopically resolvable detritus and unresolvable particulate organic matter are additional sources of nutrients for these sponges, with the latter constituting by far the most important food source. Sponges are uniquely adapted to exploit food sources of minute particle size, and as a result can inhabit nutrient-poor waters of tropical reefs, polar shelves, and submarine caverns.

Direct observations of living fresh-water sponges indicate that the archeocytes are the chief cells responsible for digestive processes. After food has entered the sponge, choanocytes take up and then rapidly lose the particles which they have ingested, and it has been assumed that the particles are transferred to archeocytes and amebocytes for digestion. Among simple calcareous sponges and homosclerophorid demosponges with large choanocytes that make up a great proportion of the total mass of sponge cells, digestion appears to be confined to the choanocytes. Studies of the activity of digestive enzymes in the marine siliceous sponge Halichondria indicate higher proteolytic, lipolytic, and carbohydrate-decomposing activities in choanocytes than in archeocytes. Following the digestion of the particles in the archeocytes, they migrate to the walls of the exhalant canals or to the dermal epithelium, where indigestible material is voided from blisterlike vacuoles

Reactions of sponges. Studies of a shallow-water marine siliceous sponge, *Hymeniacidon beliophila*, have shown that the oscula close upon exposure to air and when the sponge is in quiet seawater, although they may remain open when the sponge is in silt-laden water. Touching or stroking either oscula or ostia does not induce closure. The reactions of this sponge are thought to be solely the result of the contractions of myocytes reacting directly to external stimuli. Movements are slow and transmission of stimuli is limited. Reactions of oscula in the freshwater sponge *Ephidatia fluviatilis* differ somewhat from those of the previous species. Oscula remain open in running or still water. Oscular tubes elongate in still water and flatten out in strong currents. If a needle is rubbed around the edge of the oscular opening, the orifice contracts immediately. Weak electrical stimuli applied to an osculum tip cause it to close, and a wave of contraction runs down the entire oscular tube. Transmission of the electric stimuli to neighboring regions of the sponge was not observed.

Although neither of the siliceous sponges described above reacted to light, it has been reported that in the calcareous sponge *Leucosolenia* bright light causes oscular constriction and a change in tube shape.

The marine demosponge *Tetbya aurantia* reacts to needle pricks by general contraction of the body and closure of the oscules. The reactions begin within 30 s and are not completed until 3–7 min have elapsed. The slow behavioral responses of sponges fail to support the existence of nerve cells, although these have been reported by some zoologists from histological preparations (**Fig. 5***a*).

Studies of field populations of sponges demonstrate that in some species the rhythmic cessation of water pumping occurs. All oscules of such sponges close simultaneously when the water currents cease; a coordination mechanism for this is indicated.

Skeleton. Characteristic of sponges is the presence of a skeleton of sclerites or of organic fibers or both. Only a few genera, such as *Halisarca* and *Oscarella*, lack skeletal elements. The skeleton is of primary importance in the classification of sponges; indeed, the four classes of the Recent phylum are separated on the basis of the skeletal structure.

The shapes of the sclerites or spicules vary greatly (**Fig. 6**), and an elaborate terminology has been

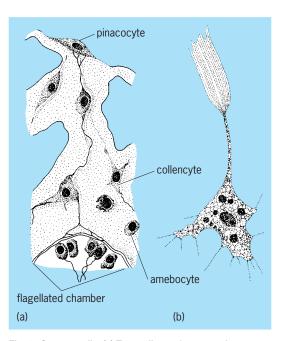


Fig. 5. Sponge cells. (a) Two-cell arc of supposed nerve cells joining pinacocyte and choanocyte chamber in a calcareous sponge. (b) Lophocyte of a fresh-water sponge.

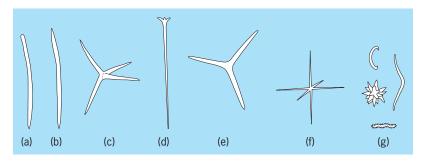


Fig. 6. Spicule types. Monaxons: (a) monactinal and (b) diactinal. Tetraxons: (c) equirayed and (d) triaene. (e) Triradiate. (f) Triaxon or hexactine. (g) Microscleres.

developed for them by taxonomists. An initial subdivision into megascleres and microscleres is made on the basis of size. A more detailed nomenclature is based on the number of axes or rays present; an appropriate numerical prefix is added to the endings -axon (referring to the number of axes) or -actine (referring to the numbers of rays or points). Spicules formed by growth along a single axis are known as monaxons. Such spicules are monactinal (pointed at one end only) if growth occurs in one direction, or diactinal (pointed at each end) if growth occurs in both directions. Tetraxons are spicules with four axes or rays, each pointing in a different direction. All four rays may be of equal length, but often one ray is longer than the other three, and those spicules are called triaenes. Triradiate (three-rayed) spicules are commonly found in calcareous sponges. Hexactinellid sponges are characterized by the presence of spicules with three axes (triaxons) meeting at right angles. Such spicules have six rays and are also called hexactines.

The reticulate fibrous skeletons of bath sponges are composed of collagenous protein called spongin, which also occurs as fine fibrils in the mesoglea. Spongin is also found in many species of Demospongiae with siliceous spicules, where it serves as an interspicular cement or forms fibers in which the spicules are embedded.

Reproduction. Both sexual and asexual reproduction occurs in sponges. The formation of asexual bodies, gemmules, is characteristic of most fresh-water species.

Sexual reproduction. The germ cells may arise from the flagellated cells of the larva, from choanocytes, from archeocytes, or from pinacocytes. The sperm enters the egg not directly, but through the intermediary of a choanocyte which loses its collar and flagellum and migrates from the flagellated chamber to a position adjacent to the egg. This process has been observed in several species of Calcarea and Demospongiae and is probably of general occurrence in these groups. The sperm, which loses its tail and enlarges, is eventually transferred to the egg, presumably following the fusion of the carrier cell with the egg (**Fig. 7**). *See* FERTILIZATION.

Subsequent development of the embryo and larva varies somewhat in the several classes. The freeswimming flagellated larvae are either hollow amphiblastulae or solid parenchymellae. *See* CALCAREA; DEMOSPONGIAE; HEXACTINELLIDA. Asexual reproduction. Many fresh-water and marine sponges disintegrate with the onset of unfavorable environmental conditions, leaving behind reduction bodies which are compact masses of cells, chiefly amebocytes, covered by an epidermal layer. In some shallow-water marine sponges, the entire sponge remains alive during the winter months but changes internally into a mass of amebocytes. Both reduction bodies and the partially dedifferentiated adult sponges can develop into functional individuals upon return of favorable conditions.

Most fresh-water and many marine sponges produce asexual reproductive bodies, called gemmules, as part of their life cycle (Fig. 8). Gemmules are formed of masses of archeocytes laden with food reserves in the form of lipids and proteins. In freshwater sponges the inner mass of cells is surrounded by a layer of columnar cells which secretes a double layer of spongin around the gemmule. Characteristic spicules are deposited between those layers. The gemmules of many species of marine sponges are also enclosed in a spongin coat, which may or may not be provided with spicules. The gemmules of fresh-water sponges and a few marine species carry those species over periods of unfavorable environmental conditions, such as drought or low temperatures, during which times the adults degenerate. Germination of the gemmules occurs upon return of suitable conditions. In many marine sponges, gemmules are formed at all seasons, and the adult sponges show no seasonal degeneration.

In a few species of Demospongiae and Hexactinellida, asexually produced gemmules, lacking spongin coats, have been described as developing into flagellated larvae identical in structure to those formed as a result of sexual processes.

Budding is another common type of asexual reproduction in sponges. In *Tethya*, for example, groups of archeocytes migrate to the tips of spicule bundles projecting from the surface (**Fig. 9**). These

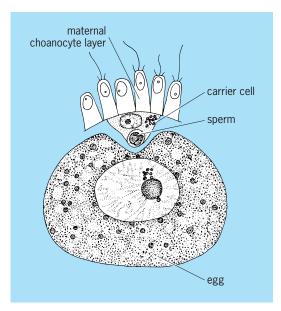


Fig. 7. Carrier cell transmitting sperm to egg in a calcareous sponge.

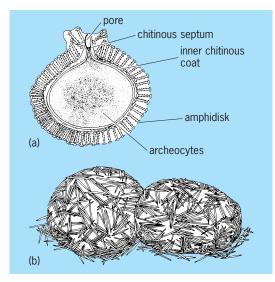


Fig. 8. Sponge gemmules (asexual) reproductive bodies. (a) Fresh-water sponge gemmules in optical section. (b) Marine siliceous sponge gemmules, surface view.

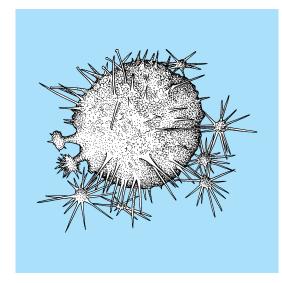


Fig. 9. Colony of *Tethya*, a marine siliceous sponge, with buds located at the tips of projecting spicule bundles.

buds fall off and form new individuals. *See* REPRO-DUCTION (ANIMAL).

Regeneration. Fragments cut from sponges or pieces which break off accidentally in nature can reattach to the substrate and reconstitute functional individuals. Suspensions of cells prepared by squeezing fragments of sponges through fine silk bolting cloth into seawater are also capable of reorganizing into functional individuals. All types of cells normally present in the adult sponge, except scleroblasts, have been identified in the cell suspensions. Reorganization of the cells to form functional sponges thus involves chiefly a migration of the cells to take up their appropriate positions in the reconstituted organism. The pinacocytes form a peripheral epithelium around the aggregated cells, and the various mesenchymal cells form the central mass. The collencytes are active in forming the inhalant and exhalant canals; the latter become lined with pinacocytes. Choanocytes group together to reconstitute flagellated chambers between the inhalant and exhalant canals. Scleroblasts form anew from archeocytes. Some workers have reported that collar cells dedifferentiate or are phagocytized by pinacocytes or amebocytes and are formed anew during the reorganization of the sponge.

The initial aggregation of the dissociated cells to form masses results from random movement of archeocytes, amebocytes, and other cell types, including choanocytes that have been observed to put forth pseudopodia and also move short distances by means of their flagella. As cells come into contact with one another, they adhere, possibly as a result of antigen-antibody-like forces. The presence of homologous antibodies inhibits reaggregation. Neighboring cell aggregates often adhere as well, if they come into contact with one another.

Mixed cell suspensions of two species of sponges of different colors may result in an initial intermixture of cells of the two species, but these later sort out so that the cell aggregates are eventually composed exclusively or predominantly of cells of one species or the other.

Commercial sponge fisheries. Although plastic sponges now offer competition, there is still a demand for natural sponges for use by various artisans, for surgical purposes, and for cleaning automobiles. Commercially valuable sponges are harvested in the eastern part of the Mediterranean Sea, off the western coast of Florida and off the Florida Keys, in the West Indies, and to a limited extent off the Philippines. They are gathered by hooking or harpooning from a boat in shallow waters; by nude diving, a method used especially for exploiting cave populations in the Aegean Sea; by machine diving with the aid of diving suits attached by a life line to an air pump; and by dredging, a wasteful method prohibited on most sponge grounds.

Artificial propagation of sponges has proven to be feasible when serious efforts have been made to do so. In the Bahamas, sponge cuttings planted on natural bottoms or on concrete disks in carefully selected areas yielded a harvest of 140,000 sponges during the years 1936–1939. Regrettably a disease, presumably caused by a fungus, spread through the West Indies in 1938–1939, put an end to the cultivation experiments, and killed a large percentage of the natural population of sponges.

Before World War II the Japanese had achieved considerable success in cultivating sponges in the Marshall and Caroline islands. Their methods included planting directly on concrete disks or stringing cuttings on aluminum wire anchored to rafts of various sorts or provided with floating buoys.

Organisms associated with sponges. Sponges are preyed upon regularly by chitons, snails, and nudibranchs, and occasionally by fish and turtles and probably other types of animals. The larvae of neuropteran insects of the family Sisyridae (spongilla flies) live in and feed upon colonies of fresh-water sponges. Sponges also play host to myriads of animals that live in the natural cavities of the sponge or form burrows of their own. Sea anemones, zoan-thids (**Fig. 10**), polychaetes, octopuses, copepods,

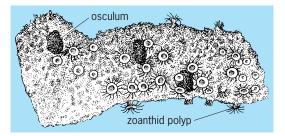


Fig. 10. Zoanthids living in a sponge.

barnacles, amphipods, shrimps, brittle stars, and fish live in sponges, but the natures of their relationships with their hosts have been little explored. The siliceous sponge *Suberites domunculus* always lives on shells occupied by hermit crabs and provides the crabs with a continually growing house. Certain crabs plant or hold pieces of sponge on their backs and thus blend into their environment. Other encrusting animals which compete for space with sponges, such as hydroids, bryozoans, and ascidians, often overgrow or are overgrown by sponges in the struggle.

Zoochlorellae commonly live in the amebocytes of fresh-water sponges and many shallow-water marine species. Because the algal cells fail to develop starch reserves when living in sponge cells, it seems likely that the sugars synthesized by the algae are utilized by the sponge cells. Dying algal cells are consumed by the amebocytes. Apparently both the algae and sponges can survive independently of each other. Young sponges are reinfected with algae from the water in which they live. Certain species of marine Demospongiae harbor algal symbionts in the form of zooxanthellae (dinoflagellates) or blue-green algae.

Certain marine sponges are regularly filled with filamentous green or coralline algae. The blue color of some sponges has been attributed to bacteria living in the cells. *See* ARCHAEOCYATHA. Willard D. Hartman

Fossils. Fossil sponges are relatively rare when compared with other invertebrate fossils, but they do occur in profusion in some areas. Upon death the spicular skeletons of sponges are commonly disarticulated or disjointed, except in those forms where the spicules are articulated or fused to one another to construct a resistant skeleton. These latter forms are most frequently preserved in the fossil record. Dissociated sponge spicules are common in marine sedimentary rocks, but once isolated from the skeletal fabric they are difficult to identify, and assignment to species or genera is usually impossible and certainly not recommended unless the spicule has unique characteristics. One sponge may have several kinds of spicules, and the same kind of spicule may be present in several genera.

Relatively solid skeletons occurred in lithistid and sphinctozoid Demospongiae; dictyid, brachiospongid, hexactinosid, and lychniscosid Hexactinellida; Sclerospongia; and pharetronid Calcarea. As a result, these sponges have left a moderately common fossil record.

Geologic record. Precambrian sponges are known from a few localities in the world, such as in the Ediacaran beds of Australia or roughly equivalent beds in China. Lower Cambrian sponges are also known from a few localities, such as Hunan Province in China. It was not until the Middle Cambrian, however, that sponges left even a modest record. The Burgess Shale of the Canadian Rocky Mountains has produced the most extensive Cambrian sponge fauna known to date, although scattered sponges of that general age are known from many other areas of the world. Hexactinellida, Demospongiae, Sclerospongea, and Heteractinida are known from Cambrian rocks and, except for the Heteractinida, range up to the Recent. Ordovician and Silurian sponge faunas are dominated by a great variety of demosponges and are moderately common throughout the Appalachian Mountains, in western Tennessee, Utah, Nevada, Texas, and Oklahoma, in Arctic Canada in North America, and in limited localities in northern Europe, Australia, and China. Devonian fossil sponges are dominated by hexactinellid forms in North America, but lithistid sponges helped produce reefs during that period in western Australia. Sclerosponges, represented by stromatoporoids, appear most common in Ordovician through Devonian rocks. Sponges of Mississippian and Pennsylvanian ages occur as scattered rare specimens in many areas, but it was not until the Permian that sponges became common fossils again. In Permian rocks, permosphinctid (sphinctozoan) demosponges and aspiculate inozoan Calcarea are major elements in fossil reefs and are particularly common in West Texas and southern New Mexico, Tunisia Sicily, Timor, and southern China.

Triassic permosphinctid sponges are common in reefs and associated beds in the Alpine region of southern Europe, but are rare in the Americas and most other parts of the world. Jurassic sponges, mainly lithistid and hexactinellid genera, are common in Europe, in reefs of Morocco, and in eastern Mediterranean countries, but are virtually unknown in North America. The Cretaceous was a time when hexactinosid and lychniscosid hexactinellid sponges were common in western and northern Europe, but fossil sponges are relatively rare in Cretaceous rocks in North America. Cenozoic fossil sponges are closely related to Recent sponges, but have left only a moderate record, particularly in North America.

Fossil types. Fossil sponges are divided into five classes: Calcarea (calcisponges), Hexactinellida, Demospongiae, Sclerospongia, and the extinct Heteractinida.

Calcarea. The Calcarea are sponges with spicules of calcite or aragonite. Such spicules in the pharetrone sponges are monaxons, triradiates, or quadriradiates and are commonly fused into a rather rigid network of fibers. Pharetrone sponges first appear in the Permian of West Texas and appear slightly later in Permian rocks in Sicily and Tunisia. They range to the Recent, but only locally are they common in modern seas. *See* CALCAREA.

Hexactinellida. These sponges have skeletons composed of opaline silica and have spicules composed of four or six, rarely more, rays which are at right angles to each other. They appeared in the late Precambrian, and by the Early Paleozoic two large subgroups

are recognizable. One has thin walls in which the spicules are regularly arranged, and the other has thick walls in which spicules are irregularly arranged. In neither early group are the spicules fused together, but the skeletons may have been held together by organic materials which were not preserved. Late Paleozoic forms that are descended from the regularly spiculed protosponges developed thick regular walls and appear transitional to the sponges with fused skeletons that became common Mesozoic and Cenozoic representatives of the class. Abundant and wellpreserved hexactinellid sponges have been recovered from Devonian rocks of southern New York and northern Pennsylvania and from Mississippian rocks of Indiana. Oldest Hexactinosa with fused, rectangularly based, three-dimensional skeletons have been questionably reported from the Permian, but by the Triassic they were certainly present. These sponges reached their peak of diversity in the Cretaceous, along with the Lychniscosa. The latter are similarappearing sponges and also have rectangular-based skeletons. In addition, the lychniscosid sponges have short diagonal braces that form the lanternlike lychnisc structure at the spicule nodes. These Mesozoic hexactinosid sponges reached their maximum diversity and abundance in Jurassic and Cretaceous rocks in Europe. See HEXACTINELLIDA.

Demospongiae. Demosponges have skeletons composed of spongin fibers, siliceous spicules, or mixed spicules and spongin. Only those forms with spicules have left a fossil record. Spicules range greatly in shape from monaxons, to tetraxons where rays do not meet at right angles, to irregular rootlike shapes. The vast majority of known fossil demosponges belong to the order Lithistida, a group that is well preserved because their articulated and fused spicules produced a solid cross-braced skeleton. Most lithistid spicules are irregular, with knobs and spines, or with complexly branching articulating ray tips that united to produce skeletons that could be rather easily preserved. Demosponges first appeared in the geologic record in the Middle Cambrian, but rapidly diversified and increased in numbers in the Early Ordovician to become one of the major reef builders of that period. They continued to diversify and remained common in the Silurian, but these families decreased in relative abundance and diversity in the later Paleozoic.

Permosphinctid (sphinctozoan) demosponges have calcareous beadlike or honeycomblike skeletons. They were considered as calcareous sponges until recently. The sphinctozoans are locally common and are composed of hollow segments arranged like a string of beads or as bubblelike clusters with solid walls in which the spicule structure is not preserved. Earliest known sphinctozoan sponges are from the Cambrian of southern Australia, and by the Ordovician these sponges are known from Australia, California, and Alaska. They remained as relatively minor elements of sponge faunas in the Devonian and through the Carboniferous, but rapidly reached their maximum diversity, abundance, and geographic distribution in tropical realms of the Permian and Triassic. Their diversity was greatly reduced in the Jurassic, and all but one species became extinct by the end of the Cretaceous.

Spiculate demosponges have another peak of abundance and diversity in Jurassic and Cretaceous rocks in Europe, and they are the major sponge elements in modern seas. *See* DEMOSPONGIAE.

Heteractinida. The Heteractinida have skeletons composed of large calcareous spicules. Early representatives of the class have spicules with six rays in one plane; later sponges of the class have octactinebased spicules in which six equally spaced rays occur in one plane that is approximately at right angles to the other two opposed rays. Two broad patterns of spicule arrangement are recognizable in the class: one group is characterized by thin walls composed of regularly arranged spicules, and the other is characterized by thick-walled sponges with irregularly arranged, almost felted spicules. Oldest known Heteractinida are from the Lower Cambrian, and sponges of the class are moderately common in Middle Cambrian rocks, but they are particularly common in Silurian and Carboniferous rocks of North America and Europe. The Heteractinida is the only class of sponges to become extinct, for they died out near the end of the Permian.

Sclerospongiae. The class Sclerospongiae was recently erected to include six genera of living sponges. These modern forms have laminated skeletons of calcium carbonate and are structurally similar to stromatoporoids, which are common fossils in Paleozoic and Mesozoic rocks. Stromatoporoids are now considered by some workers to belong in the Porifera as early sclerosponges. Their laminate skeletons are pierced by astrorhizal canals, and presumably the living tissue was confined to their upper surfaces, as in modern sclerosponges. Until recently, stromatoporoids were included as fossil cnidarians allied to hydrozoans. Sclerosponges are now considered to have first appeared in the geologic record in the Cambrian, and flourished during the Middle Paleozoic when they were important reef builders in the Silurian and Devonian. After that, particularly from near the end-Paleozoic, they became relatively minor elements, even into modern seas. See SCLE-J. Keith Rigby ROSPONGE.

Bibliography. P. R. Bergquist, *Sponges*, University of California Press, Berkeley, 1978; D. Hill, Archaeocyatha, *Treatise on Invertebrate Paleontology*, pt. E, vol. 1, 2d ed., rev. 1972; J. N. A. Hooper and R. W. M. Van Soest (eds.), *Systema Porifera: A Guide to the Classification of Sponges*, vols. 1 and 2, Kluwer Academic/Plenum, New York, 2002; R. L. Kaesler (ed.), *Treatise on Invertebrate Paleontology*, pt. E (rev.), vols. 2 and 3, 2003, 2004.

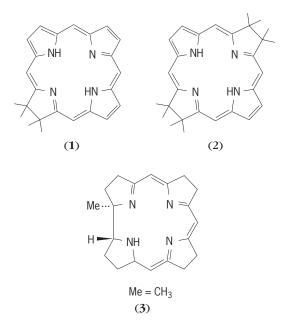
Porocephalida

One of two orders in the class Pentastomida of the phylum Arthropoda. In this order the larvae have four legs. The hooks on the adult articulate on a chitinous base or fulcrum and are arranged in a flattened trapezoidal pattern, or in a curved line or straight line. There are no podial or parapodial lobes. The two suborders are Porocephaloidea and Linguatuloidea. The Porocephaloidea comprise five families—Porocephalidae, Sebekidae, Armilliferidae, Sambonidae, and Subtriquetridae—and include most of the pentastomid species. These animals are cylindrical, with rounded ends. The hooks are sessile; the mouth is anterior between the hooks. The cephalothorax contains large cephalic and hook glands. The female genital pore is at or near the posterior end of the body; the uterus is long and tubular and disposed in loops. Eggs in the terminal part of the uterus contain fully formed larvae. There is a single testis; the male genital pore is ventral, at the anterior border of the abdomen. A cirrus sac is present; the cirrus is long.

Adults occur in the respiratory passages of snakes, lizards, turtles, and crocodiles; the larvae are encysted in fishes, snakes, lizards, crocodiles, and mammals. Linguatuloidea contains a single family, with a single genus, *Linguatula* Fröhlich, 1789. Adults live in the nasal cavities of carnivorous mammals; the larvae are encysted in herbivorous mammals. The body is elongate, flattened ventrally, annulate, and attenuated posteriorly. The females are much larger than the males and may exceed 4 in. (100 mm) in length. In the males the testis is double. The hooks of the adults are simple; those of the larvae are binate. *See* PENTASTOMIDA. Horace W. Stunkard

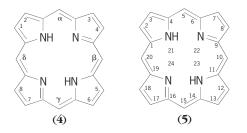
Porphyrin

One of a class of cyclic compounds in which the parent macrocycle consists of four pyrrole-type units linked together by single carbon bridges. Several porphyrins with selected peripheral substitution and metal coordination carry out vital biochemical processes in living organisms. Chlorins (1), bacteriochlorins (2), and corrins (3) are related tetrapyrrolic



macrocycles that are also observed in biologically important compounds.

Nomenclature. The complexity of porphyrin nomenclature parallels the complex structures of the naturally occurring derivatives. Hans Fischer (who won a Nobel prize for his studies on porphyrins) used a simple numbering system for the porphyrin nucleus (4) and a set of common names to identify the different porphyrins and their isomers. A systematic naming based on the 1-24numbering system for the porphyrin nucleus (5)



was later developed by the International Union of Pure and Applied Chemistry (IUPAC) and the International Union of Biochemistry (IUB), and this system has gained general acceptance. The need for common names (see **table**) is clear after examination of the systematic names; for example, protoporphyrin IX has the systematic name 2,7,12,18-tetramethyl-3,8-divinyl-13,17-dipropanoic acid.

Properties. The unique way in which some molecules can delocalize their own electrons make them especially stable. Such molecules (for example, benzene) are said to be aromatic. The aromatic character (hence stability) of porphyrins has been confirmed by measurements of their heats of combustion. In addition, x-ray crystallographic studies have established planarity of the porphyrin macrocycle; this is a basic requirement for aromatic character. *See* DELOCALIZATION; X-RAY CRYSTALLOGRAPHY.

Metal-free porphyrins are usually red-orange in solution. They exhibit characteristic optical spectra (**Fig. 1**), with a very strong absorption band around 400 nanometers (known as the Soret band) and usually four additional bands in the 500–700-nm range. In acidic solution, the inner=N- groups are

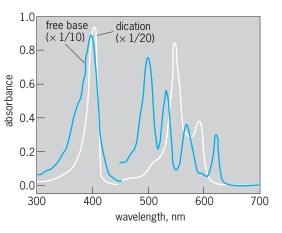


Fig. 1. Optical spectra of a porphyrin and its protonated dication.

		Substituents and locants [†]						
Name*	2	3	7	8	12	13	15	17
Coproporphyrin I	Me	Cet	Me	Cet	Me	Cet	Н	Me
Coproporphyrin II	Me	Cet	Me	Me	Me	Cet	Н	Cet
Coproporphyrin III	Me	Cet	Me	Me	Me	Cet	Н	Cet
Coproporphyrin IV	Me	Cet	Me	Cet	Cet	Me	Н	Cet
Cytoporphyrin	Me	-CH(OH)CH 2R'	Me	Vn	Me	Cet	Н	Cet
Deuteroporphyrin	Me	H	Me	Н	Me	Cet	Н	Cet
Etioporphyrin I	Me	Et	Me	Et	Me	Et	Н	Me
Etioporphyrin II	Me	Et	Et	Me	Me	Et	Н	Et
Etioporphyrin III	Me	Et	Me	Et	Me	Et	Н	Et
Etioporphyrin IV	Me	Et	Me	Et	Et	Me	Н	Et
Hematoporphyrin	Me	-CH(OH)CH 3	Me	—CH(OH)CH₃	Me	Cet	Н	Cet
Mesoporphyrin	Me	Et	Me	Et	Me	Cet	Н	Cet
Phylloporphyrin	Me	Et	Me	Et	Me	Н	Me	Cet
Protoporphyrin	Me	Vn	Me	Vn	Me	Cet	Н	Cet
Pyrroporphyrin	Me	Et	Me	Et	Me	Н	Н	Cet
Rhodoporphyrin	Me	Et	Me	Et	Me	-CO ₂ H	Н	Cet
Uroporphyrin I	Cm	Cet	Cm	Cet	Cm	Cet	Н	Cm
Uroporphyrin II	Cm	Cet	Cet	Cm	Cm	Cet	Н	Cet
Uroporphyrin III	Cm	Cet	Cm	Cet	Cm	Cet	Н	Cet
Uroporphyrin IV	Cm	Cet	Cm	Cet	Cet	Cm	Н	Cet
Phytoporphyrin	Me	Et	Me	Et	Me	—C(O)	$-CH_2-$	Cet
[†] Cet = carboxyethyl = Cm = carboxymethyl = Vn = vinyl = Me = methyl =	-CH ₂ CH ₂ CO ₂	a particular order of peripheral H C = carbon H = hydrogen O = oxygen R = organic group	substituents.					

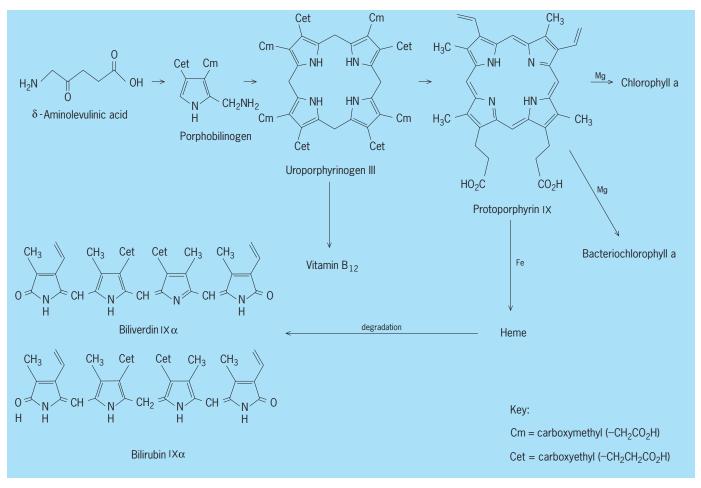
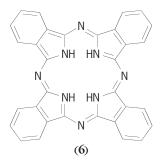


Fig. 2. Biosynthesis and metabolism of porphyrin-related macrocycles.

protonated, giving rise to the dication. This increases the symmetry of the molecule and simplifies the optical spectrum.

Porphyrins, metalloporphyrins, and their metabolic products, the bile pigments biliverdin and bilirubin (**Fig. 2**), are used as pigments in nature. Indeed, many of the rich colorations in birds' eggs are derived from porphyrin. Phthalocyanins (**6**), which

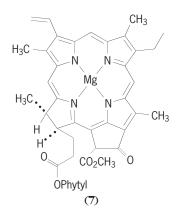


are closely related to porphyrins, are synthetic pigments. Like the porphyrins, they are both colorful and stable and are used industrially to color bluejeans and automobiles. *See* DYE.

Metalloporphyrins. Most metals and metalloids have been inserted into the central hole of the porphyrin macrocycle. The resulting metalloporphyrins are usually very stable and can bind a variety of small molecules (known as ligands) to the central metal atom. Metalloporphyrins also exhibit a strong Soret band, but have a simplified spectrum in the visible region of their absorption spectrum due to increased symmetry. In addition to the iron complexes (heme) and magnesium complexes (chlorophyll and bacteriochlorophyll), other metalloporphyrins occur in nature. In particular, nickel and vanadium complexes occur in oil, oil shales, and other geological deposits During the refining process involved in the conversion of oil to gasoline, these very stable metalloporphyrins pose many problems. See COORDINATION COMPLEXES.

Heme (**Fig. 3**), the iron complex of protoporphyrin IX, is the prosthetic group of a number of major proteins and enzymes that carry out diverse biological functions. These include binding, transport, and storage of oxygen (hemoglobin and myoglobin), electron-transfer processes (cytochromes), activation and transfer of oxygen to substrates (cytochromes P-450), and managing and using hydrogen peroxide (peroxidases and catalases). *See* CYTOCHROME; ELECTRON-TRANSFER REACTION; EN-ZYME; HEMOGLOBIN; IRON METABOLISM; PROTEIN.

Chlorophylls and bacteriochlorophylls are magnesium complexes of porphyrin derivatives known as chlorins and bacteriochlorins, respectively. They are the pigments responsible for photosynthesis. Several chlorophylls have been identified, the most common being chlorophyll a (7),



which is found in all oxygen-evolving photosynthetic plants. Bacteriochlorophyll *a* (8) is found in many photosynthetic bacteria. *See* CHLOROPHYLL; PHOTO-SYNTHESIS.

Coenzyme B_{12} is the biologically active form of vitamin B_{12} (9). The parent macrocycle, referred to as corrin, is a highly reduced tetrapyrrole with one less bridging carbon atom than the porphyrin. The B_{12} coenzyme and related corrinoids represent the most complex nonpolymeric structures found in nature. In addition, with a unique cobalt-carbon bond, they are the only known naturally occurring organometallic complexes. Although many enzymatic reactions depend on coenzyme B_{12} , its mechanism of action is not yet fully understood. *See* COENZYME; VITAMIN B_{12} .

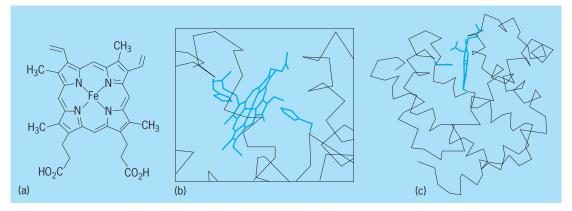
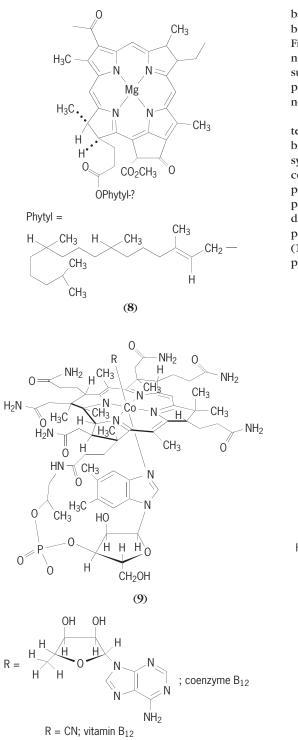
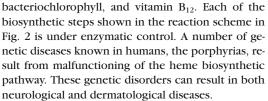


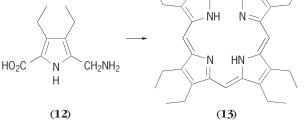
Fig. 3. Heme. (a) Structure. (b) Its bonding to myoglobin. (c) View from side showing that it is essentially flat.



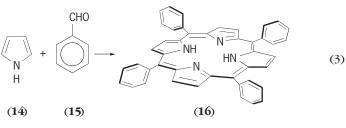


In nature, porphyrins are synthesized by the tetramerization of a single pyrrolic unit (porphobilinogen; Fig. 2). Chemists have developed similar synthetic routes [reactions (1)-(3)]. Thus, the self-condensation of 2-hydroxymethylpyrrole (7) gives porphin (7) [the parent structure in this class of compounds]. Similarly, coupling of 5-aminomethyl-3,4-diethylpyrrole-2-carboxylic acid (9) gives octaethylporphin (10), while the condensation of pyrrole (11) and benzaldehyde (12) readily gives *meso*-tetraphenylporphyrin (13). Many additional synthetic

(10) (10) (10) (11) (11) (11)







Biosynthesis. Porphyrins and their related macrocycles found in nature are derived by the same biosynthetic pathway (Fig. 2). Glycine and succinic acid are condensed to yield δ -aminolevulinic acid, which is dimerized to the pyrrole porphobilinogen. Four units of porphobilinogen are coupled to give uroporphyrinogen III (the Roman numerals here refer to the order of the peripheral substituents). Uroporphyrinogen III is the cornerstone from which all other naturally occurring tetrapyrrolic macrocycles are made, including heme, chlorophyll,

routes to porphyrins have been developed. *See* BIOSYNTHESIS.

Applications. Porphyrins and metalloporphyrins exhibit many potentially important medicinal and industrial properties. Metalloporphyrins are being examined as potential catalysts for a variety of processes, including catalytic oxidations. They are also being examined as possible blood substitutes and as electrocatalysts for fuel cells and for the electrochemical generation of hydrogen peroxide. *See* CATALYSIS.

The unique optical properties of porphyrins make them likely candidates for photovoltaic devices and in photocopying and other optical devices. A major area where porphyrins are showing significant potential is in the treatment of a wide range of diseases, including cancer, using photodynamic therapy (PDT). Because of their unique optical and physical properties, many porphyrins have the capacity to produce singlet oxygen when they are illuminated by visible light. If the porphyrin could be made to accumulate in a tumor, then bathing the tumor in red light might result in its destruction. Photodynamic therapy using a porphyrin-based drug is in clinical trials and may eventually prove a powerful tool in the treatment of cancer and other diseases.

Photochemical techniques have been developed for use in neonatal clinics. Many infants are jaundiced and yellow at birth because of their inability to metabolize bilirubin (Fig. 3), a metabolite of heme. Bilirubin can be neurotoxic. However, by merely placing the jaundiced child under a fluorescent lamp, the buildup of bilirubin is diminished to safe levels until the infant's own bilirubin metabolism increases. *See* BILIRUBIN; COORDINATION CHEMISTRY; PYRROLE. T. Wijesekera; D. Dolphin

Bibliography. P. K. Bondy and L. E. Rosenberg (eds.), *Metabolic Control and Disease*, 8th ed., 1980; D. K. Dixon, *Advances in Porphyrin Chemistry*, I: *Cytochrome b5*, 1991; D. Dolphin (ed.), *The Porphyrins*, vols. 1-7, 1978-1979; IUPAC-IUB, Nomenclature of tetrapyrroles, *Eur. J. Biochem.*, 178:277-328, 1988.

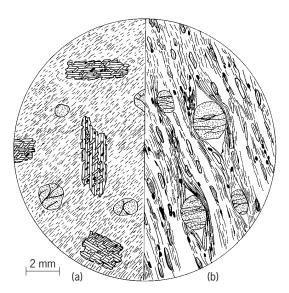
Porphyroblast

A relatively large crystal formed in a metamorphic rock. The presence of abundant porphyroblasts gives the rock a porphyroblastic texture. Minerals found commonly as porphyroblasts include biotite, garnet, chloritoid, staurolite, kyanite, sillimanite, andalusite, cordierite, and feldspar.

Porphyroblasts are generally a few millimeters or centimeters across (see **illus.**), but some attain a diameter of over 1 ft (0.3 m). They may be bounded by well-defined crystal faces, or their outlines may be highly irregular or ragged. Very commonly they are crowded with tiny grains of other minerals that occur in the rock.

Some porphyroblasts appear to have shoved aside the rock layers (foliation) in an attempt to provide room for growth. Others clearly transect the foliation and appear to have replaced the rock. The presence of ghostlike traces of foliation, in the form of stringers and trails of mineral grains, passing uninterruptedly through a porphyroblast is further evidence of replacement.

Porphyroblasts have many features in common with phenocrysts but are to be distinguished from the latter by the fact that they have developed in solid rock in response to metamorphism. Most commonly they develop in schist and gneiss during the late stages of recrystallization. As the rock becomes re-



Porphyroblasts. (a) Porphyroblastic mica schist. Large crystals (porphyroblasts) of biotite mica formed late and replaced the rock except for the enclosed grains (quartz) aligned parallel to foliation. Smaller porphyroblasts of garnet replaced the rock completely. (b) Porphyroblastic quartz-mica schist. Large porphyroblasts of garnet have grown by spreading apart mica-rich layers of schist. Elongate grains (gray) are flakes of biotite mica. Finer flakes of muscovite mica are widespread and give a pronounced schistosity. They are closely packed where crowded by garnet crystals. Elongate grains of iron and titanium oxide (black) are also aligned.

constituted, certain components migrate to favored sites and combine there to develop the large crystals. *See* GNEISS; METAMORPHIC ROCKS; PHENOCRYST; SCHIST. Carleton A. Chapman

Porphyry

An igneous rock characterized by porphyritic texture, in which large crystals (phenocrysts) are enclosed in a matrix of very fine-grained to aphanitic (not visibly crystalline) material.

Porphyries are generally distinguished from other porphyritic rocks by their abundance of phenocrysts and by their occurrence in small intrusive bodies (dikes and sills) formed at shallow depth within the earth. In this sense porphyries are hypabyssal rocks. *See* PHENOCRYST.

Compositionally, porphyries range widely, but varieties may be distinguished by prefixing to the term the common rock name which the porphyry most closely resembles (such as granite porphyry, rhyolite porphyry, syenite porphyry, and trachyte porphyry).

Porphyries are gradational to plutonic rocks on the one hand and gradational to volcanic rocks on the other. In the granite clan, for example, six porphyritic types may be recognized: porphyritic granite, granite porphyry, rhyolite porphyry, porphyritic rhyolite, vitrophyre, and porphyritic obsidian or porphyritic pitchstone.

A rock of granitic composition with abundant large phenocrysts of quartz and alkali feldspar in a very fine-grained matrix of similar composition is a porphyry, or more specifically a granite porphyry. Granite porphyry passes into porphyritic granite as the grain size of the matrix increases and abundance of phenocrysts decreases. Thus the rock becomes a granite, or more specifically a porphyritic granite. Granite porphyry passes into rhyolite porphyry as the grain size of the matrix and abundance of phenocrysts decrease. The main distinction between rhyolite porphyry and porphyritic rhyolite is mode of occurrence. Rhyolite porphyry is intrusive; porphyritic rhyolite is extrusive. Porphyritic rocks with a glass matrix are known as vitrophyres. With decrease in number of phenocrysts, vitrophyre passes into porphyritic obsidian or porphyritic pitchstone. *See* GRANITE.

The phenocrysts of porphyries consist commonly of quartz and feldspar. Quartz occurs as well-formed (euhedral) hexagonal bipyramids, which in thin section under the microscope exhibit a diamond, square, or hexagonal outline. Individual phenocrysts may show more or less rounding or resorption with deep embayments. Alkali feldspar is usually euhedral sanidine, orthoclase, or microperthite. Plagioclase occurs more in association with phenocrysts of hornblende or other dark-colored (mafic) minerals. Porphyritic rocks with predominantly mafic phenocrysts (olivine, pyroxene, amphibole, and biotite) are commonly classed as lamprophyres. *See* LAMPRO-PHYRE.

Porphyries occur as marginal phases of mediumsized igneous bodies (stocks, laccoliths) or as apophyses (offshoots) projecting from such bodies into the surrounding rocks. They are also abundant as dikes cutting compositionally equivalent plutonic rock, or as dikes, sills, and laccoliths injected into the adjacent older rocks. *See* IGNEOUS ROCKS. William Ingersoll Rose, Jr.

Positron

The elementary antiparticle of an electron with mass equal to that of an electron but with positive rather than negative charge. Its existence was theoretically predicted by P. A. M. Dirac in 1928. It was first observed by C. D. Anderson in 1932. Positrons and electrons have the same spin and statistics. Positrons are naturally produced by radioactive decay of many heavier particles; this is referred to as beta-plus decay. Electron-positron pairs are also produced when high-energy photons or electrons interact with matter. *See* ANTIMATTER; ELECTRON; ELECTRON-POSITRON PAIR PRODUCTION; ELECTRON SPIN; ELE-MENTARY PARTICLE; FERMI-DIRAC STATISTICS; PARTI-CLE ACCELERATOR; QUANTUM STATISTICS; RADIOAC-TIVITY; SPIN (QUANTUM MECHANICS).

In addition, quantum field theory predicts that positrons can be created when electrons are subjected to very strong, static electric fields. For this to occur, the field strength must exceed the energy equivalent of twice the rest mass of an electron, $2m_0c^2$, where m_0 is the electron rest mass and *c* is the speed of light. Experimentally, such strong fields can be achieved by using heavy-ion accelerators to collide two heavy atoms and temporarily form a superheavy atom with an atomic number Z > 173. However, to date there has been no definitive observation of positrons produced via this method. *See* NUCLEAR MOLECULE; QUASIATOM; SUPERCRITICAL FIELDS.

A positron is, in itself, stable, but cannot exist indefinitely in the presence of matter where it will ultimately collide with an electron. The two particles will be annihilated as a result of this collision, and photons will be created. However, a positron can become bound to an electron to form a short-lived "atom" termed positronium. *See* POSITRONIUM.

Positrons are of interest or have practical uses in many fields. In the field of physics, they are used to study differences between particles and antiparticles and how particles and antiparticles interact. In 1995 at the CERN laboratory in Geneva, Switzerland, these studies took on a new role when positrons and antiprotons (the antiparticle of a proton) were combined to produce antihydrogen atoms. The original studies produced only a few antihydrogen atoms, but methods have improved and now it is routine to produce a hundred antihydrogen atoms per second. Present and future antihydrogen studies include investigating whether excited antihydrogen produces the same light emission as hydrogen, whether they have the same magnetic properties, and whether they interact with gravity in the same manner. Should differences be found, existing laws of physics will have to be altered, and answers as to why the observable universe appears to be made entirely of matter and not antimatter may emerge. Another interest in the field of physics is with respect to the virtual production of electron-positron pairs by an electromagnetic field. This produces a polarization of the vacuum and results in effects such as the scattering of light by light and modification of the electrostatic Coulomb field at short distances. See CPT THEOREM; QUANTUM ELECTRODYNAMICS.

In medicine, positron emission tomography (PET) is used to produce three-dimensional images of the inside of the body. It is primarily used to image the brain and to identify tumors. In this procedure, a radioisotope that undergoes positive beta decay is injected. After a wait period, images are produced by monitoring the two photons that are produced and emitted in opposite direactions (to conserve momentum) when an emitted positron annihilates with an electron. *See* MEDICAL IMAGING; NUCLEAR MEDICINE.

In the field of material science, positrons are routinely used to study defects in the crystalline structure of solids or the properties of electrons within the solid. To study defects, that is, either missing or replaced atoms in the solid, positrons are injected into the solid and the annihilation photons are measured as a function of time. To annihilate, a positron must spend sufficient time close to an electron; this occurs predominantly in voids in the crystal where the positron can become temporarily "trapped." Thus, shorter lifetimes imply more defects. By precisely measuring the directions of the annihilation photons that are emitted back-to-back, the location of the defects can be determined. This back-to-back emission is also used to provide information about the electron density and momentum distribution within the solid. *See* ANGULAR CORRELATIONS; CRYS-TAL DEFECTS; QUANTUM THEORY OF MATTER.

In astrophysics, positron-filled jets of gas have been observed at the center of the Milky Way Galaxy by National Aeronautics and Space Administration's *Compton Gamma-Ray Observatory (CGRO)*. It is unclear how these antimatter jets are produced, but one suggestion is that they are related to the birth of stars occurring near the large black hole at the center of the Milky Way Galaxy. *See* BLACK HOLE; GAMMA-RAY ASTRONOMY; MILKY WAY GALAXY.

At some future date, positron-electron or other forms of antimatter-matter annihilation might be used as a power source or weapon. However, this is still far away. For example, currently 10^{10} times more energy is required to produce antiprotons than could be derived from them, and it would take all the antimatter produced each year at CERN to power a 100-watt light bulb for 15 minutes. Robert D. DuBois

Bibliography. K. Bethge et al., *Medical Applications of Nuclear Physics*, Springer Series: Biological and Medical Physics, Biomedical Engineering, Springer Verlag, 2004; D. A. Bromley (ed.), *Heavy Ion Science*, vol. 5, 1985; M. Charlton and J. Hangst, Probing the antiworld, *Phys. World*, 18(10):22-27, October 2005; P. G. Coleman and P. Coleman (eds.), *Positron Beams and Their Applications*, World Scientific Publishing, 2000; S. Gasiorowicz, *The Structure of Matter: A Survey of Modern Physics*, 1979; J. M. Jauch and F. Rohrlich, *The Theory of Photons and Electrons*, rev. ed., 1975; C. M. Surko and F. A. Gianturco (eds.), *New Directions in Antimatter Chemistry and Physics*, Springer Verlag, 2001.

Positronium

The bound state of the electron and its antimatter counterpart, the positron. Positronium (abbreviated Ps) is perhaps the simplest bound state, since the constituents are structureless and the electromagnetic binding force is well understood. As in the hydrogen atom, the states of positronium are quantized. The spins of the two spin-1/2 constituents combine to yield total spin zero (singlet state) or one (triplet state). Singlet and triplet states are called parapositronium (p-Ps) and orthopositronium (o-Ps), respectively. States of positronium are labeled by principal quantum number n, orbital angular momentum, total spin, and total angular momentum (orbital plus spin). The energy levels are approximately $-6.80 \text{ eV}/n^2$, with numerous small corrections due to relativity, virtual annihilation, and other electromagnetic effects. Additional corrections from strong and weak interactions are negligible. See ANGULAR MOMENTUM; ATOMIC STRUCTURE AND SPECTRA; ELECTRON; ELECTRON SPIN; POSITRON; SPIN (OUANTUM MECHANICS).

Positronium has a finite lifetime due to electronpositron annihilation into energy in the form of gamma-ray photons. The energetics of this reaction is governed by Albert Einstein's equation $E = mc^2$, where *m* is the total positronium mass (about twice the electron mass), *c* is the speed of light, and *E* is the total photon energy. A selection rule constrains the allowed decays, so that singlet positronium decays into an even number of photons (usually two), and triplet positronium into an odd number (usually three). Decay into a single photon is forbidden by energy-momentum conservation. The lifetimes are approximately 10^{-10} s for the singlet state and 10^{-7} s for the triplet state. *See* CONSERVATION LAWS (PHYSICS); GAMMA RAYS; REST MASS; SELECTION RULES (PHYSICS).

Positronium has been extensively studied since it was first produced by Martin Deutsch in 1951. High precision measurements have been done on the energy levels and decay rates of the lowestenergy states, and on energy levels of the first excited states. Because positronium is relatively free of strong and weak interaction complications, comparison between theory and experiment for these measured quantities provides a stringent check of quantum electrodynamics (QED). Positronium has been used to check for violations in fundamental symmetries and as a probe for physics beyond the standard model. Positronium is an analog for the more complicated quarkonium (quark-antiquark) bound states, particularly charmonium and bottomonium. See ELEMENTARY PARTICLE; MESON; QUANTUM ELEC-TRODYNAMICS; QUARKS; STANDARD MODEL; SYMME-TRY LAWS (PHYSICS).

Annihilation radiation from positronium forms a component of the gamma-ray spectrum observed by astronomers, in particular from the galactic center. *See* GAMMA-RAY ASTRONOMY; MILKY WAY GALAXY.

Formation. Positronium is formed in the laboratory when a positron, usually obtained from the decay of a radioactive isotope such as sodium-22 (22 Na), captures an electron. The energetic positrons emitted in the nuclear decay must first be slowed to a few eV of kinetic energy before positronium will form. This slowing can be accomplished through inelastic collisions with the atoms or molecules of a material, often the same material from which the electron is obtained. Formation media include gases and powders. Positronium can also be formed when a beam of slow positrons is incident on a thin foil, a gas target, or the surface of a solid in vacuum. *See* CHARGED PARTICLE BEAMS.

Spectroscopy. Many of the energy intervals involving n = 1 and n = 2 states of positronium have been measured.

Hyperfine structure. The difference in energy between the n = 1 triplet and singlet states of positronium is traditionally referred to as the hyperfine splitting. The measured value of the corresponding frequency is 203,389 \pm 1 MHz. Theoretical calculations based on QED are of comparable precision, and give a result that differs from this by about 3 MHz. This small difference is not thought to represent a significant problem.

Optical transition. The 5.10-eV interval between the

n = 1 and n = 2 triplet states with zero orbital angular momentum can be bridged by the simultaneous absorption of two photons of visible light. When the photons are arranged to move in exactly opposite directions, the usual frequency spread caused by the Doppler shift is greatly reduced. Measurements done using counterpropagating high-energy laser pulses incident on ground-state positronium in a vacuum give a result for the corresponding transition frequency of 1,233,607,216.4 \pm 3.2 MHz. Theoretical results are in accord with this value.

Fine structure. Relativistic corrections, magnetic effects, and virtual annihilation split the n = 2 level into several sublevels. The frequencies of many of these sublevel differences have been measured by monitoring the absorption of microwave radiation by a positronium beam moving through a microwave cavity. Experiment and theory are in good agreement for these transitions. *See* MOLECULAR BEAMS.

Decay rates. The lifetime of the relatively longlived n = 1 orthopositronium has been measured to high precision. A typical experimental configuration consists of a ²²Na positron source, a gas chamber in which the positronium forms, photomultiplier gamma detectors, and timing electronics. The emission of a position from the source is accompanied by a gamma ray which can be detected and used as a start signal. The stop signal is the detection of the final annihilation radiation. Following the prompt peak caused by direct positron-electron annihilation and the fast decay of parapositronium, there is an extended exponential distribution coming from orthopositronium decay. Since the presence of the gas increases the rate of decay, an extrapolation to zero gas pressure must be done to obtain the vacuum rate. High-precision measurements are also done for positronium formed in powders and in thin coatings in vacuum. The best lifetime measurements give a value of 142.05 ± 0.03 nanoseconds, in excellent agreement with theoretical calculations based on quantum electrodynamics. See GAMMA-RAY DETECTORS; PHOTOMULTIPLIER.

The lifetime of parapositronium is too short for a high-precision direct timing procedure. Instead, positronium is produced in a magnetic field, which mixes the triplet and singlet states and increases a component of the triplet decay rate by an amount that depends on the singlet rate and the field strength. The parapositronium lifetime, measured using this approach, is 125.14 ± 0.03 picoseconds, in good agreement with theoretical expectation.

Materials science probe. Since the formation of positronium requires the close approach of a positron and an electron, beams of slow positrons can be used as probes of the electron density in gases, in insulating solids, or near surfaces. Since the singlet and triplet forms of positronium have very different lifetimes, and transitions between the two states can be induced by environmental factors, study of the decay of positronium can also provide information about electron densities and magnetic fields on a microscopic scale. This is useful in the study of density fluctuations in gases near the critical point for con-

densation into liquids or solids. This technique is also used to study vacancies in metals and semiconductors, and is now being applied to high-temperature superconductors and biological materials. *See* CRIT-ICAL PHENOMENA; CRYSTAL DEFECTS; SUPERCONDUC-TIVITY. Gregory S. Adkins

Bibliography. S. Berko and H. N. Pendleton, Positronium, *Annu. Rev. Nucl. Part. Sci.*, 30:543– 581, 1980; M. Charlton and J. W. Humberston, *Positron Physics*, Cambridge University Press, 2001; T. Hyodo et al. (eds.), *Positron Annibilation ICPA-13*, Trans Tech Publications, Uetikon-Zuerich, 2004 (http://www.scientific.net/0-87849-936-9/); Y. C. Jean, P. E. Mallon, and D. M. Schrader (eds.), *Positron and Positronium Chemistry*, World Scientific, Singapore, 2003; B. Maglic (ed.), Discovery of Positronium, *Adventures Exp. Phys.*, 4:64-127, 1975; A. Rich, Recent experimental advances in positronium research, *Rev. Mod. Phys.*, 53:127-165, 1981.

Postglacial vegetation and climate

The Pleistocene or Glacial Epoch is less remarkable for radical evolutionary changes in plant life than for repeated shifts of position because of climatic changes and the resulting ice movements. The influence of these changes is known to have extended far beyond the limits of glaciation, particularly in mountainous regions, whose belts or zones of vegetation migrated up and down in response to climate. *See* PLEISTOCENE.

Four main glacial stages are recognized in North America within the Pleistocene Epoch: Kansan, Nebraskan, Illinoian, and Wisconsin. The climatic change which initiated the last major ice advance of the late Wisconsin began more than 30,000 years ago. This ice advance reached its maximum extent in southern Illinois, Indiana, and Ohio about 18,000-20,000 years ago. The climate began to warm about 16,000-18,000 years ago and the glaciers melted more rapidly than they advanced. Glacier ice withdrew from the continental United States sometime prior to 12,000 years ago, with the exception of residual ice caps on higher elevations in mountains in New England and in the Rocky Mountains of the West. There was a brief readvance of the ice about 10,500 years ago, known as the Valders readvance

The retreat of the ice was not uniform; it occurred with many pauses and some readvances, each producing an elevated belt of drift called moraines. As the ice retreated, broad expanses of sand, silt, and clay mixed with rocks and boulders of all sizes were uncovered. Bare mineral soil was colonized by plants which had been forced south in advance of the ice some thousands of years before. In New England, leaves and pollen of high-arctic plants have been radiocarbon-dated at more than 15,000 years, marking a time of extremely harsh climate similar to that which is today found north of the Arctic Circle. *See* GLACIOLOGY; RADIOCARBON DATING.

Palynological studies. Attempts to understand and interpret the climatic changes that have occurred since the ice retreat began have involved many different types of study. One of the most fruitful techniques for environmental reconstruction has been the field of pollen analysis, or palynology. The technique, pioneered by the Swedish geologist Lennart von Post in 1916, was first extensively applied in the United States by Paul B. Sears in the early 1930s. The method depends upon the fact that each year flowering plants produce enormous numbers of pollen grains; although many plants are pollinated by insects, most are pollinated by wind. Austrian pine (Pinus sylvestris) produces as many as 1,480,000 pollen grains per male strobilus, and there may be up to 20 strobili in a single inflorescence, with more than 100 inflorescences on a single medium-sized tree. The spruce forests of southern Sweden are estimated to produce up to 75,000 tons (67,500 metric tons) of pollen each spring. Small size (10-250 micrometers) and light weight permit pollen grains to be carried for long distances by wind. Stray pollen grains of alder (Alnus viridis) have been found in the middle of the North Atlantic Ocean, more than 400 mi (650 km) from the nearest source. See PALYNOLOGY.

Eventually, pollen grains fall to the ground, the majority remaining close to the source plants. Those which fall into lakes and bogs are not decomposed, but are preserved in the annual accumulation of sediment and form a record of the numbers and kinds of pollen of the local and regional flora. Pollen grains are distinctive and can be recognized at least to genus level in most cases. A pollen diagram showing the succession of dominant pollen types throughout a stratigraphic section of bog or lake sediments can then be used to interpret the vegetational and climatic changes that have occurred since the deposit was formed. *See* BOG; POLLEN.

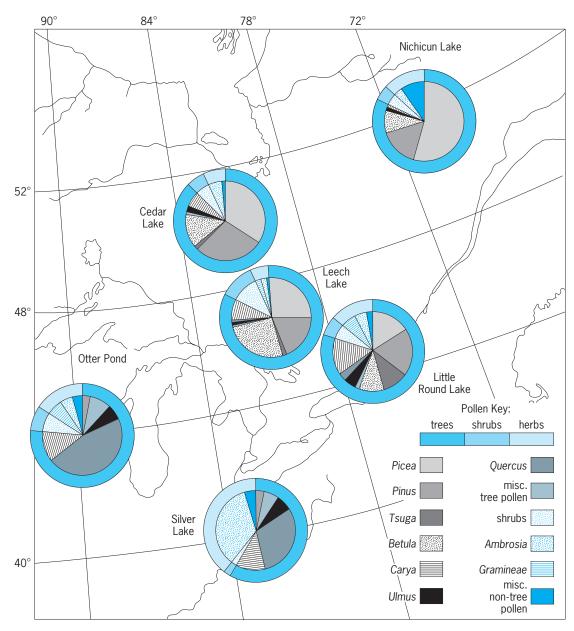
Regional pollen diagram. Interpretation of a pollen diagram is facilitated by an understanding of the kinds of pollen records produced by different environmental types. Surface pollen samples from six different vegetational and climatic regions are shown in the illustration. Nichicun Lake, the northernmost lake, is at the northern edge of Boreal Parkland in northern Quebec, a region characterized by isolated stands of spruce and pine trees with extensive open areas of tundra. Just south of James Bay is Cedar Lake, Ontario, near the northern edge of the Boreal Forest, which consists of dense stands of spruce and pine together with considerable amounts of fir and birch. Further south, Leech Lake is in the Northern Great Lakes-St. Lawrence Forest with dense stands of spruce, pine, and birch. In the southern part of the Great Lakes-St. Lawrence Forest, Little Round Lake is surrounded by stands of spruce and pine with extensive upland areas in oak and hickory and considerable amounts of hemlock and northern white cedar around the lake border. Otter Pond, in the Maple-Basswood region of the Deciduous Forest in southern Wisconsin, is closely similar to Silver Lake, Ohio, which is in the Oak-Hickory region of the Deciduous Forest. The principal difference between these two pollen samples is that there is a larger amount of hickory pollen at Silver Lake and a larger amount of beech pollen at Otter Pond.

At Silver Lake, in sediments which are approximately 12,000 years old, the pollen spectra most closely resemble the environment represented by the Nichicun Lake surface samples. A computer program to calculate product-moment correlations (r) on 25 different pollen types shows a high correlation between the contemporary Nichicun Lake surface sample and 12,000-year-old Silver Lake sediments (r = +.925 for N = 25). This value is not different from a population coefficient of r = +1.0 at the .995 significance level. By 11,000 years ago the climate had warmed sufficiently to permit the establishment of a closed Boreal forest, and a computer-generated search of more than 200 surface pollen samples from all of eastern North America show highest correlation (r = .896) with Cedar Lake in the Boreal Forest. The environment of Silver Lake 6500 years ago was apparently very similar to the present environment around Otter Pond in the Maple-Basswood portion of the Deciduous Forest (r = .912). About 3500-5000 years ago, the climate around Silver Lake was apparently warmer and drier than at present, for the highest correlation with contemporary surface samples is found in southwestern Illinois near the boundary between the Deciduous Forest and the Tall Grass Prairie. This period was first identified in Midwestern pollen diagrams as the "Xerothermic Interval" by P. B. Sears in the early 1940s. A return to somewhat cooler and moister conditions is indicated following the Xerothermic Interval by increases in the amount of beech pollen and a corresponding decrease in oak and hickory pollen in sediment cores from Silver Lake.

A computer search of a matrix formed by 25 different pollen types in more than 200 lakes from eastern North America and 38 pollen spectra from different levels in sediment cores at Silver Lake indicates that at no time in the postglacial history of Silver Lake did the environment resemble very closely the modern pollen records of Little Round Lake or Leech Lake in Ontario.

The appearance of European colonists is indicated in the upper portions of sediment cores throughout North America by the presence of large amounts of the pollen of ragweed (*Ambrosia* sp.) and the presence of European weed pollen such as white-man's foot (*Plantago lanceolata*), dandelion (*Taraxacum* sp.), and species of *Rumex*. These plants were either brought in by the colonists or the permanent clearings made for houses and farms provided ideal conditions for the spread of these and native species.

Environmental reconstruction. Since the beginning of the Christian era, historical documents have recorded the effects of climatic change on the activities of humans. A period of benign climatic conditions around A.D. 1000 favored the westward explorations of the Vikings who colonized Iceland, Greenland, and briefly the coast of North America near L'Anse aux Meadows in Newfoundland. About A.D. 1150 the climate began to turn colder and



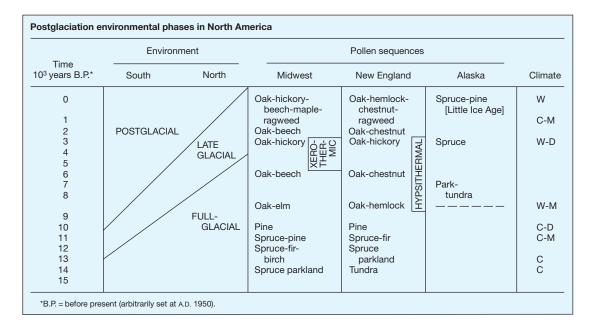
Pollen diagrams from six different vegetational and geographic regions.

winter storminess increased. The climatic change was not the least of the reasons for abandoning the settlements in North America and Greenland.

By the middle of the seventeenth century, glaciers in high mountains in North America and Europe had advanced dramatically, in some cases overwhelming farms or small settlements. Toward the close of the seventeenth century the climate again began to warm and glaciers around the world began to retreat. This warming trend is also recorded by dates of the opening of cherry blossoms in monastery gardens in Asia and has continued until the present. There is some evidence that the rate of warming has decreased since 1940 and that rainfall has been increasing north of about $40^{\circ}N$ latitude.

Although the amount of vegetational change in the last 14,000 years has been great, the actual temperature changes seem rather small. Temperature measurements derived from measurement of the ratios of oxygen-16 to oxygen-18 in the carbonates deposited in marine Foraminiferida indicate that mean annual temperatures were not more than $12-14^{\circ}F$ (6.5–8.5°C) lower than those during the maximum advance of the late Wisconsin ice. Similarly, measurements of ocean water temperature show an increase of only $0.9^{\circ}F$ ($0.5^{\circ}C$) over the last 200 years, covering the period of the retreat of the "Little Ice Age" maximum of the seventeenth century. *See* FORAMINIFERIDA.

Radiocarbon dating of late Pleistocene events indicate that environmental changes were nearly synchronous throughout the world at least for the last 20,000 years. Although the evidence is less clear for earlier periods of ice advance, four main periods of ice advance with similar substages are recognized in Europe and Russia.



In the reconstruction of climatic change within or near the glacier boundary, it is possible to distinguish three major environmental events: Full-glacial, which represents a period of time during which ice either covered an area or its presence prevented the establishment of a vegetational record; Late-glacial, an indefinite period of early colonization by plants following the retreat of the ice but prior to the establishment of closed forests; and Postglacial, marking the development of dense forests. In Europe, the Postglacial is arbitrarily taken as the period following the emplacement of the Fennoscandian end moraines about 10,000-11,000 years ago. In North America, their equivalent is the Valders end moraines in Wisconsin. The distinction between Full-glacial, Late-glacial, and Postglacial in North America has generally been recognized to depend upon particular localities. For example, Postglacial time begins about 12,500 years ago in southern Ohio and Indiana as well as in southern New England. These areas were ice-free several thousand years before upper Michigan, southern Ontario, and the Maritime Provinces. Much of Greenland and the arctic islands of Canada can be considered to be still in Full-glacial.

Within Postglacial time in North America, two subdivisions are generally recognized. The Hypsithermal period, about 9500-2500 years before present, marks a period in Postglacial time during which the climate was at least as warm as it is at present. The warm maximum of Postglacial time which is not everywhere recognizable in northeastern North America has been called the Xerothermic Interval and marks a time of about 5000-3500 years ago during which prairie, today restricted to the region west of Illinois, extended eastward into Ohio.

Postglaciation environmental changes are summarized in the **table**. The area included is limited to the major glaciated portions of North America partly because more work has been done there and partly because the record beyond the glacier boundary is not as clear. Studies in southwestern North America have demonstrated considerable shifts in woodland and forest belts moving up and down the sides of the Rocky Mountains, and these changes seem to be broadly synchronous with the changes observed inside the glacier boundary. Additional work on ancient lake basins will help to establish the environmental changes outside the glaciated regions. J. Gordon Ogden, III

Bibliography. R. J. Dodd and R. J. Stanton, Jr., *Paleoecology: Concepts and Applications*, 2d. ed., 1990; *Symposium on Changes in Climate*, Rome, UNESCO, 1965; H. E. Wright, Jr., and D. G. Frey (eds.), *International Studies on the Quaternary*, 1965.

Posttraumatic stress disorder

An anxiety disorder in some individuals who have experienced an event that poses a direct threat to the individual's or another person's life. The characteristic features of anxiety disorders are fear, particularly in the ongoing absence of a real-life threat to safety, and avoidance behavior.

Diagnosis. A diagnosis of posttraumatic stress disorder requires that four criteria be met. First, the individual must have been exposed to an extremely stressful and traumatic event beyond the range of normal human experience (for example, combat, armed robbery, rape, or natural disasters such as earthquakes). Second, the individual must periodically and persistently reexperience the event. This reexperiencing can take different forms, such as recurrent dreams and nightmares, an inability to stop thinking about the event, flashbacks during which the individual relives the trauma, and auditory hallucinations. Third, there is persistent avoidance of events related to the trauma, and psychological numbing that was not present prior to the trauma.

Individuals often withdraw from others, feel emotionally deadened, lose interest in activities, and avoid contact with anything even remotely similar to the traumatic event. Some children with posttraumatic stress disorder regress to an earlier stage of development. Fourth, enduring symptoms of anxiety and arousal are present. These symptoms can be manifested in different forms, including anger, irritability, a very sensitive startle response, an inability to sleep well, and physiological evidence of fear when the individual is reexposed to a traumatic event. Posttraumatic stress disorder can occur at any age, and the symptoms typically begin soon after the cessation of the traumatic event.

Most knowledge of posttraumatic stress disorder comes from research on combat soldiers and other victims of war; other knowledge comes from rape and assault victims, and survivors of disasters or medical illness. There is a growing appreciation for the conceptualization of posttraumatic stress as a spectrum disorder. At one pole are individuals who are exposed to significant trauma but do not go on to develop symptoms of posttraumatic stress. At the opposite pole are individuals who manifest the full range of posttraumatic stress disorder symptoms. In between these extremes are individuals who develop mild forms of the disorder or manifest only certain symptom complexes which do not qualify as an actual diagnosis of posttraumatic stress disorder. Recent evidence suggests that there may be gender differences in how posttraumatic stress is manifested. Men appear to be significantly more likely than women to experience irritability and to use alcohol to excess. Within both genders, however, symptoms tend to follow an acute stress reaction, occur early, and persist for many months.

Causative factors. In the United States, most adult individuals have experienced at least one event that qualifies as sufficiently traumatic to stimulate the development of posttraumatic stress/posttraumatic stress disorder, and approximately 10% meet diagnostic criteria for posttraumatic stress disorder at some time in their lives. This incidence is probably influenced by the subjective nature of perceived stress. What is mildly stressful for some individuals may be moderately stressful or highly stressful for others. Thus, it can be seen that most individuals do not develop a pathological reaction to psychological trauma. Estimates of occurrence of the disorder in large groups of combat veterans have ranged from 15 to 50%. The degree of the posttraumatic stress response is likely to be influenced by a complex interaction of biological processes, personality, nature of the trauma, and posttraumatic events. High levels of avoidance behavior may actually potentiate and prolong the symptoms. Failure to expose oneself to stimuli that evoke fear allows the fear response to remain strong. Increased exposure, however, diminishes the fear response. In short, many symptoms of posttraumatic stress disorder may be selfperpetuating. Events that occur after the cessation of the traumatic event also seem to be important in whether symptoms become severe and intractable.

The prediction of posttraumatic stress disorder and symptoms of posttraumatic stress following a trauma are also complex and multidimensional. Biological predictors include lower baseline levels of cortisol, hippocampal atrophy, and deficiencies in hypothalamic-pituitary-adrenal activity. Cognitive and personality factors also play a role in predicting the severity of posttraumatic stress following a trauma. Individuals who tend to exaggerate the probability of future negative events, to become somatically focused, to exhibit cognitive biases, and to experience shame following the trauma are at an elevated risk of developing posttraumatic stress disorder. Tendencies to avoid aversive information are also related to the development of chronic posttraumatic stress. Thus, the development of posttraumatic stress disorder is likely influenced by a complex pattern of memory encoding, self-appraisal, coping strategies, cognitive styles, the nature of the trauma exposure, and biological responses that occur soon after the traumatic event and continue in a deteriorating cascade.

Physiological changes. There are also important biological events that accompany stress. Physiological arousal responses in individuals with posttraumatic stress disorder include increases in heart rate, respiration rate, and skin conductivity upon reexposure to traumatic stimuli. Posttraumatic stress disorder may also be associated with structural and physiological changes in the brain. For example, when individuals are placed under high levels of stress, the activity levels of two brain neurotransmitters, norepinephrine and serotonin, are reduced. Reductions in these chemical transmitters are not typically found under mild or moderate stress conditions. Stressful events also affect the activity level of the pituitary and adrenal glands. These fluctuations in adrenergic hormonal activity are likely to be associated with periodic alterations in the brain's endogenous opiates. Under conditions of chronic stress, hormones (glucocorticoids, in particular) contribute to impairment of cognitive functioning and can actually result in damage to brain structures, particularly the hippocampus. All these physiological changes are probably complexly related to the persistence and waxing and waning of symptoms in posttraumatic stress disorder. In addition, extreme and prolonged stress is associated with a variety of physical ailments, including heart attacks, ulcers, colitis, and decreases in immunological functioning. See ADRENAL CORTEX; BRAIN; NEUROBIOLOGY; NOR-ADRENERGIC SYSTEM; SEROTONIN.

Associated disorders. Posttraumatic stress disorder is rarely the only psychological problem that an affected individual manifests. When an individual is diagnosed as having this disorder, particularly after it has been present for a number of years, it is common to also find significant depression, generalized anxiety, substance abuse and dependence, marital problems, and intense, almost debilitating anger. Although the primary symptoms of posttraumatic stress disorder (anxiety, fear, intrusion, and avoidance) are usually amenable to treatment efforts, these secondary problems commonly associated with the chronic disorder are more difficult to treat. *See* AD-DICTIVE DISORDERS; AFFECTIVE DISORDERS; ANXIETY DISORDERS.

Treatment. Posttraumatic stress disorder can be treated by pharmacological means and with psychotherapy.

Psychotherapeutic methods. Most psychological treatments for the disorder involve reexposure to the traumatic event. This reexposure is typically imaginal and can range from simply talking about the trauma to having the person vividly imagine reliving the traumatic event. This latter behavioral procedure is called implosion therapy or flooding. While flooding is not appropriate for all posttraumatic stress disorder cases, especially individuals with substance abuse and interpersonal difficulties, the procedure can dramatically decrease anxiety and arousal, intrusive thoughts, avoidance behavior, and emotional numbing. Along with specific behavior interventions, individuals with posttraumatic stress disorder should become involved in psychotherapeutic treatment for secondary problems. A wide variety of effective treatments for both posttraumatic stress disorder and associated psychiatric and psychological conditions have been demonstrated to be effective. These interventions include cognitive-behavioral therapy, group therapy, psychodynamic treatment, psychosocial rehabilitation, hyponosis, and marital and family therapy. See PSYCHOTHERAPY.

Pharmacological methods. Numerous medications have been demonstrated to be effective in treating posttraumatic stress disorder, including selective serotonin reuptake inhibitors, antiadrenergic drugs, and tricyclic antidepressants. Other medications that are showing promise include mood stabilizers and some antipsychotic drugs. *See* PSYCHOPHARMACOLOGY; STRESS (PSYCHOLOGY). Robert W. Butler

Bibliography. American Psychiatric Association, Diagnostic and Statistical Manual of Mental Disorders, 4th ed., 1994; N. Breslau, Epidemiologic studies of trauma, posttraumatic stress disorder, and other psychiatric disorders, Can. J. Psychiat., 47:923-929, 2002; C. R. Brewin and E. A. Holmes, Psychological therories of posttraumatic stress disorder, Clin. Psychol. Rev., 23:339-376, 2003; E. B. Foa, T. M. Keanne, and M. J. Friedman (eds.), Effective Treatments for PTSD, 2000; B. S. McEwen, The neurobiology and neuroendocrinology of stress: Implications for post-traumatic stress disorder from a basic science perspective, Psychiat. Clinics N. Amer., 25:469-494, 2002; C. Moreau and S. Zisook, Rationale for a posttraumatic stress spectrum disorder, Psychiat. Clinics N. Amer., 25:775-790, 2002.

Postulate

In a formal deductive system, a proposition accepted without proof, from which other propositions are deduced by the conventional methods of formal logic. There is a certain arbitrariness as to which propositions are to be treated as postulates, because when certain proved propositions are treated as such, other propositions which were originally postulates often become proved propositions.

The term postulate is sometimes loosely used in connection with tentative assumptions with regard to matters of fact. In strict usage, the term postulate is nearly equivalent to axiom, although axiom is often loosely used to denote a truth supposed to be selfevident. *See* LOGIC.

Percy W. Bridgman; Henry Margenau

Postural equilibrium

A lifeless object is said to be in equilibrium, or in a state of balance, when all forces acting upon it cancel. The result is a state of rest. In an actively moving animal, internal as well as external forces have to be considered, and the maintenance of a balanced attitude in a body consisting of a number of parts that are loosely connected by movable joints is complex.

Maintenance of equilibrium. The maintenance of equilibrium is relatively easy in limbless animals, such as worms, snails, or starfish, which rest and move on an extensive surface of support. The tactile contact between the normal creeping surface and the substrate furnishes sensory information on the animal's orientation with respect to the plumb line. When the animal is turned on its back, the lack of contact pressure on the creeping surface and the tactile stimulation of the back initiate movements which return the animal to its normal position. This is known as the righting reflex. The accompanying change in the direction of illumination may also contribute to efficient righting.

Free-swimming and flying animals are often in a precariously poised state of equilibrium, so that the normal attitude can be maintained only by the continuous operation of corrective equilibrating mechanisms. The same applies to a lesser degree to longlegged quadrupeds, such as many mammals, and to bipeds, such as birds and some primates (including humans). In these animals the center of gravity lies some distance above the area of support which, in some cases, such as in humans, can be relatively small.

Sensory control. For most species, there is a specific orientation of the whole body or of body segments (such as the head) with respect to gravity. This orientation is based on multisensory inputs and on a set of inborn reflexes acting on the musculature. These postural reflexes also stabilize the genetically defined body orientation against external disturbances. The righting reflexes in vertebrates is an example of this basic organization. When a cat is falling upside down, the head is first reoriented toward the horizontal plane, and then the trunk axis and finally the legs are oriented vertically.

The sensory information relies on a number of sources: (1) static and dynamic information from the eyes; (2) static and dynamic mechanoreceptors

incorporated in the various types of statocysts in invertebrate animals and in the vestibular organ or labyrinth of vertebrates; (3) proprioceptor organs such as muscle spindles, Golgi endings in tendons, Pacinian corpuscles and similar encapsulated endings associated with tendons and joints, and other pressure receptors in supporting surfaces (for example, the soles of feet); and (4) sensory endings in the viscera, capable of being differentially stimulated by changes in the direction of visceral pull on mesenteries, and on other structures, *See* SENSATION; SENSE ORGAN.

Vision plays an important role in orientation in the gravitational field. In daylight, visual fixation of the horizon or reference to the bright sky as opposed to the ground can serve as a means of basic orientation. This is important in the equilibration of freeswimming and flying animals. Visual postural mechanisms contribute in part to balance in aquatic invertebrates such as shrimp and fish. These animals turn their back toward the light, showing the dorsal light reflex. Some shrimp swim belly up, showing the ventral light reflex. Among flying animals, insects in flight rely largely on visual orientation. The compound eye is well suited to the task of keeping the body in a horizontal or nearly horizontal plane. Its subdivision into separate ommatida makes it possible to hold the image of the horizon, or of other horizontal or vertical landmarks, fixed on a certain part of the retina. If the animal deviates from an even keel, the image leaves the fixation region, evoking corrective reflex movements designed to return the image to the same part of the eye. Like other arthropods, insects are equipped with hair receptors located between joints in the armor of limbs and body, monitoring their relative position. See EYE (INVERTEBRATE); INSECT PHYSIOLOGY.

Gravity reception. In many invertebrates and in all vertebrates, specialized receptor organs known as statocysts supply information on the animal's orientation with respect to the plumb line. A statocyst consists of a fluid-filled space, the floor of which is made up of sensory and supporting cells. The sensory cells have hair processes which project into the lumen of the statocyst. In the lumen the hair processes come into close contact with the statolith, a body of higher specific gravity than the surrounding fluid and tissues. The displacement of the statolith caused by the animal's deviation from an even keel deforms the hair processes, initiating afferent impulses in the sensory nerve. Thus, the central nervous system receives information regarding the direction of the gravity vector and the degree of deviation with respect to this vector. The otolith organ of the vertebrate labyrinth functions on the same principles and monitors the head position in space. The semicircular canal of the vertebrate labyrinth is arranged in three different planes of space and constitutes receptor organs for angular accelerations of the head, serving to stabilize the head orientation in space.

Graviception in vertebrates is also monitored by

body graviceptors. For example, experiments under water or in microgravity show that afferent inputs related to body weight as well as receptors around the kidney play an important role in the body orientation with respect to gravity. *See* EAR (VERTEBRATE).

Stretch receptors. Invertebrate stretch receptors are usually not as closely associated with skeletal muscles as in vertebrates. Endings of sensory neurons are applied to connective tissue fibers in the vicinity of the joints, monitoring changes in their length brought about by flexion or extension of the joint. More complex types of stretch receptors are the chordotonal organs. Consisting of nerve endings capped by peglike structures, they are hung on an elastic chord suspended within the cavity of the limbs and trunk. The chordotonal organs are frequently involved in the control of posture and movement.

In vertebrates stretch receptors are mostly associated with muscle stretch, and in mammals they are inserted in specialized organs called the muscle spindles. They monitor muscle length and velocity. Muscle force is measured by Golgi tendon organ receptors. The stretch receptors are involved in the regulation of the segmental muscle tone and in the antigravity postural tone maintained by the extensor muscles opposing the gravity forces. They also contribute to a central representation of the body segment position.

Organization of posture. In higher terrestrial vertebrates, during stance the center of mass (CM) is usually situated high above the ground due to the support of the body by the limbs. A critical aspect of posture in quadrupedal and bipedal stance is equilibrium maintenance which is preserved only when under static conditions the projection of the center of mass remains inside the support base. This positioning of the center of mass is based on two main controls. A "bottom up" control is based on the afferent nerve impulses, cutaneous and proprioceptive, from the feet and the ankle joint muscles. These nerve impulse serve in building up posture from the feet to the head. A "top down" control starts from the head, and is predominant during dynamic activities such as locomotion. Due to labyrinthine afferent nerve impulses, the head axis orientation remains stable with respect to space. The movementrelated visual afferents recorded by the retina monitor the head displacements with respect to space and adjust the body posture as a function of these inputs.

Internal representation. In order to explain the dual mode of postural organization, bottom up and top down, it was proposed that the control of posture by the central nervous system depended on a hierarchical organization. At a higher level, an internal representation of posture or postural body schema was organized. This representation was rather stable, partly genetically determined and partly acquired through learning. This level included (1) a representation of the body-segments geometry from the feet to the eyes mainly based on proprioceptive muscle

afferents; (2) a representation of the segments mass and center of inertia; (3) a representation of the body posture and mass with respect to the external world, based on multisensory inputs (visual, labyrinth, cutaneous, graviceptor). This level of representation has been illustrated, for example, by the illusory perceptions provoked by artificial sensory stimulations such as those of the spindle afferents under the action of tendon vibration. When holding the fingertip on the nose, the subject feels, during arm biceps tendon vibration, an elongation of the nose (Pinocchio nose). Biceps vibration stimulates the spindle afferents which are normally excited during stretching of the muscle (extension of the arm). As the arm is not moving, the subject's interpretation is the elongation of the nose.

A lower level in the central organization is that of the postural reactions. Balance disturbances are corrected by several mechanisms, depending on their intensity: (1) the joint stiffness related to the muscle contraction around the joints which minimize the angular displacements; (2) the postural reactions characterized by plurisegmental muscle activation patterns reestablishing the initial reference posture through a whole-body rotation around the ankle joint (ankle strategy) or through a hip flexion or extension (hip strategy); (3) stepping; (4) arm stretching during falling.

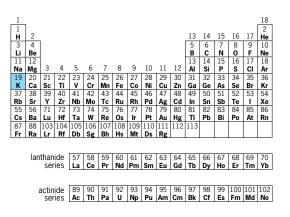
Central organization. Two levels of control are involved in maintaining balance. A first level includes the spinal cord and the brainstem, where a set of inborn reflexes are organized for stance regulation and head orientation. Most postural reflexes rely on networks at that level. The cerebellum is involved in the adaptation of these reflexes to the external constraints. *See* REFLEX.

A second level of control includes cortical areas involved in multisensory integration and control as well as the basal ganglia. The postural body schema and the body orientation with respect to the external world are organized mainly at that level, with a predominant role in the right hemisphere. Coordination between balance control and locomotion or movements also depends on these higher levels. Jean Massion

Bibliography. C. Gans, *Biomechanics: An Approach to Vertebrate Biology*, 1980; F. Horak and J. M. Macpherson, *Postural Orientation and Equilibrium in Handbook of Physiology*, sec. 12, 1996; T. D. M. Roberts (ed.), *Neurophysiology of Postural Mechanisms*, 2d ed., 1978.

Potassium

A chemical element, K, atomic number 19, and atomic weight 39.102. It stands in the middle of the alkali metal family, below sodium and above rubidium. This lightweight, soft, low-melting, reactive metal (see **table**) is very similar to sodium in its behavior in metallic forms. *See* ALKALI METALS; PE-RIODIC TABLE; RUBIDIUM; SODIUM.



Potassium chloride, KCl, finds its main use in fertilizer mixtures. It also serves as the raw material for the manufacture of other potassium compounds. Potassium hydroxide, KOH, is used in the manufacture of liquid soaps, and potassium carbonate in making soft soaps. Potassium carbonate, K₂CO₃, is also an important raw material for the glass industry. Potassium nitrate, KNO₃, is used in matches, in pyrotechnics, and in similar items which require an oxidizing agent. Marshall Sittig

Potassium is a very abundant element, ranking seventh among all the elements in the Earth's crust, 2.59% of which is potassium in combined form. Seawater contains 380 parts per million, making potassium the sixth most plentiful element in solution.

Potassium is even more reactive than sodium. It reacts vigorously with the oxygen in air to form the monoxide, K_2O , and the peroxide, K_2O_2 . In the presence of excess oxygen, it readily forms the superoxide, KO_2 .

Potassium does not react with nitrogen to form a nitride, even at elevated temperatures. With hydrogen, potassium reacts slowly at 200° C (392° F) and rapidly at $350-400^{\circ}$ C ($662-752^{\circ}$ F). It forms the least stable hydride of all the alkali metals.

The reaction between potassium and water or ice is violent, even at temperatures as low as -100° C (-148° F). The hydrogen evolved is usually ignited in reaction at room temperature. Reactions with aqueous acids are even more violent and verge on being explosive.

The potassium ion (K^+) is the most common intracellular cation and is essential for maintaining osmotic pressure and electrodynamic cellular properties in organisms. The intracellular potassium ion concentrations are typically high for most cells, whereas the potassium ion concentrations present in extracellular fluids are significantly lower. The hydrolysis of the coenzyme adenosine triphosphate (ATP) is mediated by the membrane-bound enzyme Na⁺, K⁺-ATPase. This enzyme is called the sodium pump and it is activated by both potassium and sodium ions; however, many enzymes are activated by potassium ions alone (for example, pyruvate kinase, aldehyde dehydrogenase, and phosphofructokinase).

Physical properties of potassium metal							
Property	Temperature			Customary (engineering)			
	°C	°F	SI units	units			
Density	100	212	0.819 g/cm ³	51.1 lb/ft ³			
-	400	752	0.747 g/cm ³	46.7 lb/ft ³			
	700	1292	0.676 g/cm ³	42.2 lb/ft ³			
Melting point	63.7	147	-				
Boiling point	760	1400					
Heat of fusion	63.7	147	14.6 cal/g	26.3 Btu/lb			
Heat of vaporization	760	1400	496 cal/g	893 Btu/lb			
Viscosity	70	158	5.15 millipoises	6.5 kinetic units			
	400	752	2.58 millipoises	3.5 kinetic units			
	800	1472	1.36 millipoises	2 kinetic units			
Vapor pressure	342	648	1 mm	0.019 lb/in. ²			
	696	1285	400 mm	7.75 lb/in. ²			
Thermal conductivity	200	392	0.017 cal/(s)(cm ²)(cm)(°C)	26.0 Btu(h)(ft ²)(°F)			
	400	752	0.09 cal/(s)(cm ²)(cm)(°C)	21.7 Btu/(h)(ft ²)(°F)			
Heat capacity	200	392	0.19 cal/(g)(°C)	0.19 Btu/(lb)(°F)			
	800	1472	0.19 cal/(g)(°C)	0.19 Btu/(lb)(°F)			
Electrical resistivity	150	302	18.7 microhm-cm				
	300	572	28.2 microhm-cm				
Surface tension	100-150	212-302	About 80 dynes/cm				

Potassium deficiency may occur in several conditions, including malnutrition and excessive vomiting or diarrhea, and in patients undergoing dialysis; supplementation with potassium salts is sometimes required. *See* OSMOREGULATORY MECHA-NISMS. Duarte Mota de Freitas

Bibliography. H. J. Adrogue and D. E. Wesson, *Potassium*, 1994; F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; D. R. Lide, *CRC handbook Chemistry and Physics*, 85th ed., CRC Press, 2004; C. A. Pasternak (ed.), *Monovalent Cations in Biological Systems*, 1990.

Potato, Irish

A plant of the genus Solanum in the nightshade family, Solanaceae; it is related to tomatoes and peppers. There are more than 2000 species of Solanum, of which about 150 bear tubers. The potato of commerce, S. tuberosum, originated in South America, probably in the highlands of Peru and Bolivia, where it has been cultivated for several thousand years. Potatoes were introduced into Europe by Spanish explorers in the late sixteenth century and into the United States from Ireland in 1719. The crop became a staple in Europe, was a primary source of food in Ireland, and is known even today as the Irish potato. The potato blight famine of 1845 and 1846, caused by Phytophthora infestans, was responsible for the death of more than a million Irish people, and the emigration of about 1.5 million others. See SOLANALES.

The potato plant (**Fig. 1**) is an annual, herbaceous dicotyledon that is grown primarily for its edible tubers, which are short, thick underground stems that form on the ends of stolons (lateral stems). Lateral buds (eyes) on the mature tuber are the growing points for a new crop, provided that whole tubers or pieces with at least one eye are planted. Because it can reproduce vegetatively by means of its tubers, it is sometimes considered to be a perennial.

Varieties. There are 95-100 varieties certified for seed production in the United States and Canada. In the United States, seven varieties account for more than 70% of the commercial acreage planted. These are Russet Burbank, Norchip, Atlantic, Russet Norkotah, Superior, Centennial Russet, and Kennebec.

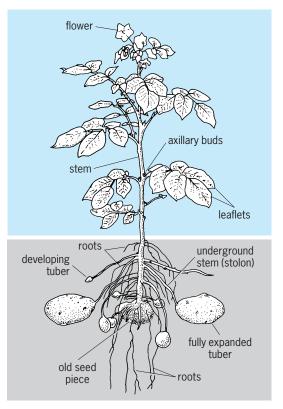


Fig. 1. Diagram of a potato plant. (After Commercial Potato Production in North America, Potato Association of America, 1980)

Production. Potatoes are grown in more countries than any crop except for corn (maize), and they are the fourth-most important food crop in supplying energy in the human diet, following rice, wheat, and corn. The leading potato producers are Russia, China, Poland, and the United States, in descending order. In the United States, potatoes are grown commercially in every state, but most production is in Idaho, Washington, Maine, Oregon, Colorado, Wisconsin, California, North Dakota, Minnesota, Michigan, Florida, New York, Pennsylvania, Texas, and New Mexico, in descending order.

Constituents. Chemical constituents can be affected by variety, production area, cultural practices, maturity at harvest, and storage conditions. The average composition of potatoes is 78–80% water, 14–18% starch, 2% protein, 1% minerals, 0.4% fiber, and 0.1% fat, with some sugars, organic acids, amino acids, and vitamins.

The potato is very nutritious, serving as a nearly complete food itself, and it produces more food per acre than any other crop. It ranks second in carbohydrate and protein production per unit area of land, behind sugarcane and soybean, respectively. The quality of potato protein is excellent. It has the best balance of the eight essential amino acids of any plant food. As a world average, 200 lb/acre (226 kg/hectare) of potato protein is produced, which is greater than that in wheat grain or in rice grain. Potatoes provide the world with more than 6.6×10^6 tons (6 $\times 10^6$ metric tons) of protein. Starch constitutes about 65-80% of the dry weight of the potato and, calorically, is the most important component. The texture of a cooked potato is determined largely by its starch content.

Potatoes contain appreciable amounts of the B vitamins (niacin, thiamine, riboflavin, and pyridoxine), and they are an excellent source of vitamin D. One medium-sized baked potato yields 20 mg of vitamin C, about one-third of the recommended daily allowance; potatoes contribute more vitamin C to the United States food supply than any other single food source. Inorganic constituents or minerals of potatoes are predominantly potassium, phosphorus, sulfur, chlorine, magnesium, and calcium. Potatoes are the best nutritional source of potassium and contain sufficient quantities of iron to be a good nutritional source of this mineral as well. Since sodium content is low, potatoes are an ideal food for individuals on a low sodium diet. J. Creighton Miller, Jr.

Diseases. Diseases of potatoes are especially difficult to control for three reasons: potatoes are propagated vegetatively, using tubers for seed, and thus bacteria, viruses, fungi, and nematodes may easily be transmitted from generation to generation; clonal propagation results in genetically homogeneous populations which favor epidemic diseases; potato tubers are highly perishable and subject to storage diseases.

The most economically important group of diseases are those caused by viruses. Most common are mild, rugose, and latent mosaics, and leaf roll. Spindle tuber viroid is also widespread. Viruses reduce yield and sometimes lower the quality of the tubers. Because they are tuber-borne, viruses are controlled principally by roguing out diseased plants and growing seed crops in isolation. Some potato varieties are resistant or immune to certain viruses. *See* PLANT VIRUSES AND VIROIDS.

Several species of bacteria cause potato diseases. *Corynebacterium sepedonicum* (causing ring rot and wilt) is seed-borne and is controlled by the same methods used for viruses. *Pseudomonas solonacearum* (causing brown rot or southern wilt) is also seed-borne and in the southern United States survives in the soil over the winter, making potato culture unprofitable on fields that have become infested. *Erwinia carotovora* var. *atroseptica* (cause of blackleg) is also seed-borne and is similar to *E. carotovora* var. *carotovora*, which causes soft rot of tubers in storage. Proper storage of tubers and careful handling of seed pieces at planting reduce incidence of these diseases.

The most destructive fungus disease is late blight, caused by *Phytophthora infestans* (**Fig. 2**). Primary infection is from infected seed and typical epidemics occur when the weather is cool and wet. Genes for immunity have been bred into new varieties, but the genetic plasticity of the fungus has resulted in failure

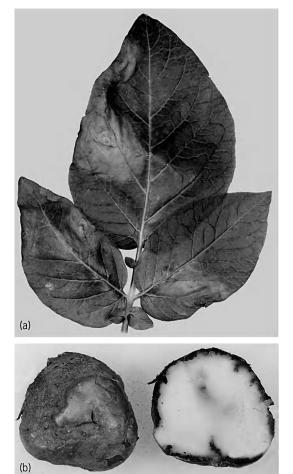


Fig. 2. Late blight (*Phytophthora infestans*) of Irish potato. (*a*) Foliar infection. (*b*) Tuber infection. (*USDA*)

of such resistance. There are, however, potato varieties with generalized but partial resistance which are useful in reducing the spread of epidemics. Fungicides are still the principal means of control. *See* FUNGISTAT AND FUNGICIDE.

Alternaria solani causes early blight, and various species of *Rbizoctonia, Verticillium*, and *Fusarium* cause stem cankers, wilts, and storage rots. Tubers are also subject to diseases that are principally blemishes, such as common scab, caused by *Streptomyces scabies*; silver scurf, caused by *Helminthosporium atrovirens*; and black scurf, caused by *Rbizoctonia*.

In addition to storage diseases caused by organisms, potato tubers are subject to nonparastic disorders such as black heart, hollow heart, and various kinds of of internal necroses and discolorations due to growing or storage conditions.

Nematodes cause root knots and tuber rots. The most destructive is the golden nematode (*Globodera rostochiensis*). At present it is not widespread, but once introduced into the soil, it persists indefinitely. Varieties resistant to *G. rostochiensis* and nematicides are used for control. The lesion nematode (*Pratylenchus penetrans*) and root-knot nematodes (*Meloidogyne* spp.) are serious in some potato-growing areas. *See* NEMATA (NEMATODA); PLANT PATHOLOGY. H. David Thurston

Bibliography. W. G. Burton, *The Potato*, 3d ed., 1989; G. Lisinska and W. Leszcynski (eds.), *Potato Science and Technology*, 1989; R. C. Rowe (ed.), *Potato Health Management*, 1993; J. A. Woolfe, *The Potato in the Human Diet*, 1987.

Potato, sweet

The fleshy root of the plant *Ipomoea batatas*. The sweet potato was mentioned as being grown in Virginia as early as 1648. In 1930 the selection of outstanding strains of the Porto Rico variety, which was introduced into Florida in 1908, was begun in Louisiana, and the best strain, Unit I Porto Rico, was released in 1934. The Unit I Porto Rico is now being replaced by the Centennial, a yam type which has three times its vitamin A content. More than 70% of the commercial crop in the United States is of the Centennial variety. Another variety, Nemagold, is used largely in the Virginia area because of its resistance to nematodes. Most varieties released now have multiple disease resistance, particularly resistance to wilt and soil rot.

In 1937 new techniques were developed for inducing the sweet potato to bloom and set seed. This stimulated a surge of research on breeding for higher yield, greater nutritional value, better shape, storage ability, market and canning quality, greater disease resistance, and new food products and industrial uses for the sweet potato throughout the Southern states and from New Jersey to California. Louisiana is the leading state for commercial production of both the canned and fresh products.





Fig. 1. Sweet potatoes. (a) Porto Rico, a yam variety. (b) Big-stem Jersey type. (USDA)

Types. There are two principal types of sweet potato, the kind erroneously called yam and the Jersey type (**Fig. 1**). The chief difference between the two is that in cooking or baking the yam, much of the starch is broken down into simple sugars (glucose and fructose) and an intermediate product, dextrin. This gives it a moist, syrupy consistency somewhat sweeter than that of the dry (Jersey) type. On cooking, the sugar in the dry type remains as sucrose.

The yam is produced largely in the Southern states; however, because of the breeding of more widely adapted varieties, it is now being grown farther north. The Jersey sweet potato is grown largely along the eastern shore of Virginia. Maryland, Delaware, and New Jersey, and also in Iowa and Kansas. With the development of new varieties, the northern limits of sweet potato production have been extended to Canada and the northern United States. This has come about by breeding early and better varieties. In Michigan and Iowa a number of baby-food manufacturers are growing this crop for processing. Julian C. Miller

Production practices. Technically, the sweet potato is a perennial plant, but in commerical production the growth is terminated by harvest due to impending change of seasons (cold or rainy) or by achievement of optimum storage root size. Plantings are established from sprouts obtained from stored roots or from vine cuttings. Production of storage roots does not require pollination, so that adverse conditions only delay or reduce harvests; crop failures are rare. A variety of soil types may be used; however, light sandy soils are preferred for ease of harvest. The harvested roots may be marketed at once or cured for a week at 85°F (29°C) and then stored for up to a year at 55°F (13°C).

Attributes. A number of factors contribute to the worldwide importance of sweet potatoes. Perhaps foremost is the high energy yield, which exceeds that of Irish potatoes and grains. Sweet potato roots are also high in nutritional quality (see **table**). For example, if the energy requirement of the average man between 23 and 50 years of age were met by sweet potato roots, 73% of the protein and 4160% of the vitamin A requirement would be met. The nutritional quality of the protein is high as measured by amino acid balance.

Utilization. The sweet potato is a versatile crop that is used as a food, an ingredient in other foods, an animal feed, and a feedstock for the production of starch and ethanol.

Food. Sweet potatoes can be baked, boiled, or strained and may be substituted in recipes for carrots, squash, or pumpkins to improve culinary and nutritional properties. They may be preserved by canning, freezing, or dehydration. Sweet potato flour produced from dried sweet potatoes can be substituted for 10-25% of wheat flour (depending on the product) with no loss of culinary quality and sometimes with a considerable increase in nutritional qual-

Nutrient	Value
Protein	73
Vitamin A	4160
Vitamin C	825
Niacin	81
Riboflavin	91
Thiamin	167
Phosphorus	139
Iron	161
Calcium	95

*After Composition of Foods, Handb. 8, USDA, 1975; Recommended Daily Allowances, NAS/NRC, 1980.

[†]Expressed as the percentage ratio of nutrient supplied per unit energy supplied to nutrient required per unit energy required. Based on a male, age 23–50, weight 154 lb (70 kg), height 5 ft 6 in. (1.70 cm). ity. Tender vine tips and leaves are consumed in many parts of the world as a pot herb, providing a good source of pro-vitamin A, calcium, iron, vitamin B_2 , and vitamin C.

Feed. The high energy production and nutritional quality of the sweet potato also contribute to its importance as an animal feed, especially in the tropics, where many temperate-zone feed grains are poorly adapted. Dried roots and dried or fresh vines are appropriate feeds for dairy cows, beef cattle, and sheep. Dried roots are less suitable for swine and poultry and should not exceed 25% of the energy provided to them because of their low levels of trypsin inhibitors and poor digestion of starch. *See* ANIMAL FEEDS.

Industrial products. The energy content of sweet potato roots is largely due to starch, and the crop has been used as a feedstock for the production of starch and ethanol. The starch-extracting procedures are similar to those used for other root crops such as cassava or canna. The starch is of fairly good quality, with a amylose:amylopectin ratio between 20:80 and 30:70. It is suitable for food or for industrial products such as glues and sizing. *See* STARCH.

Subsistence agriculture. Sweet potatoes will generally produce a crop even under adverse conditions. In some cultures, sweet potatoes are considered a survival food-an important staple food in times of war or adverse weather. In the United States, land devoted to sweet potato production reached an all-time high of 1.06×10^6 acres (4.3 \times 10⁵ hectares) in 1932, more than 10 times the current acreage. In some countries, subsistence on sweet potatoes is associated with hard times, leading many people to avoid them as economic conditions improve. The most accepted utilization technologies in developing countries may eventually be in the areas of food ingredients, feeds and feeding systems, and improved efficiencies for industrial production.

Germ-plasm resources. Plant and human factors contribute greatly to a large germ-plasm base. Plants flower sparingly and usually require cross pollination, so progeny rarely resemble parents closely. Either chance seedlings or mutations, if superior to parents, can easily be preserved by vegetative reproduction, resulting in a wider gene pool. The germ plasm has not been adequately evaluated, but the work that has been done suggests considerable variability for most traits studied. The lack of importance of sweet potatoes in many temperate countries and the lack of status in tropical developing countries have contributed to low priorities for the collection and preservation of germ-plasm resources, a situation that has been addressed by the International Board for Plant Genetic Resources and the U.S. Department of Agriculture. Worldwide, over 6000 accessions of sweet potato and 500 of related species have been collected and are being maintained. See BREEDING (PLANT). John C. Bouwkamp

Diseases. Sweet potato diseases are caused by fungi, bacteria, viruses, nematodes, and mycoplasmas, and by nutritional and environmental factors during plant growth and transit and storage of

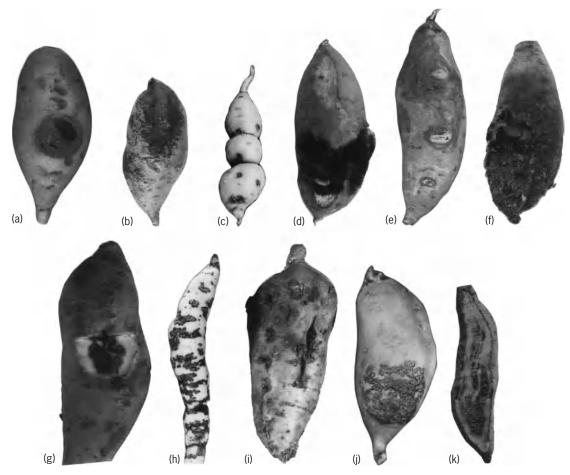


Fig. 2. Common diseases of sweet potato. (a) Black rot. (b) Scurf. (c) Soil pox. (d) Rhizopus soft rot. (e) Surface rot. (f) Java black rot. (g) Internal cork. (h) Russet crack. (i) Root knot. (j) Blister. (k) Internal breakdown.

edible roots (Fig. 2). Sweet potatoes are propagated vegetatively, and the biotic disease agents are transmitted from crop to crop by planting infected or contaminated plants. The fungi that infect roots in transit and storage gain entrance through mechanical wounds made in harvesting and packaging the potatoes. Major fungal diseases in the field are Fusariuminduced wilt, black rot, scurf, stem blight, and soil pox. Minor fungal diseases include mottle root necrosis, two leaf spots, and white rust. Rhizopus-induced soft rot and black rot are important transit diseases that also occur in storage. Other storage diseases are Fusarium-induced rot, charcoal rot, surface rot, Java black rot, and Diaporthe-induced rot. A bacterial wilt and soft rot occur in the United States and China. The viruses of internal cork, russet crack, and dwarf, and the mycoplasma of witches'-broom impair root quality and reduce plant growth and productivity. The importance of viruses such as those of mosaic, feathery mottle, and other diseases is uncertain. Root knot is the important nematode-induced disease. Deficient boron nutrition during plant growth stimulates blister in the stored roots. Exposing stored roots to temperatures below 50-55°F (10-13°C) induces internal breakdown.

Sweet potato diseases are controlled by using disease-free roots or plants; rotating other crops with sweet potatoes in infested land; chemically treating infested land, propagative roots, and storage containers and houses; curing freshly harvested roots; and planting disease-resistant varieties (cultivars). Field diseases are controlled by growing plants from disease-free roots treated with a fungicide. Sometimes soils infested with nematodes or the soil pox organism must be chemically treated to reduce losses. To control storage diseases, freshly harvested roots are placed in new or fungicide-treated containers and transferred immediately to a ventilated room at 86°F (30°C) and 90% relative humidity for 1 week to heal (cure) harvest wounds. Cured roots are stored at a minimum of $55^{\circ}F(13^{\circ}C)$. Cultivars with resistance or tolerance to Fusariuminduced wilt, internal cork, soil pox, and some root knot nematodes have been developed by breeding and selection. Some cultivars do not develop the destructive russet crack virus symptoms, and these are replacing those that develop symptoms. Sources of resistance to other disease are being sought in sweet potatoes from around the world to be used for developing other or improved disease-resistant cultivars. See PLANT PATHOLOGY; POTATO, IRISH.

L. W. Nielson

Bibliography. J. C. Bouwkamp (ed.), *Sweet Potato Products: A Natural Resource for the Tropics*, 1985; R. K. Jansson and K. V. Raman, *Sweet Potato Pest Management: A Global Perspective*, 1991; J. A. Woolfe, *Sweet Potato: An Untapped Food Resource*, 1992.

Potential flow

A fluid flow that is isentropic and that, if incompressible, can be mathematically described by Laplace's equation. For an ideal fluid, or a flow in which viscous effects are ignored, vorticity (defined as the curl of the velocity) cannot be produced, and any initial vorticity existing in the flow simply moves unchanged with the fluid. Ideal fluids, of course, do not exist since any actual fluid has some viscosity, and the effects of this viscosity will be important near a solid wall, in the region known as the boundary layer. Nevertheless, the study of potential flow is important in hydrodynamics, where the fluid is considered incompressible, and even in aerodynamics, where the fluid is considered compressible, as long as shock waves are not present. See BOUNDARY-LAYER FLOW; COM-PRESSIBLE FLOW; ISENTROPIC FLOW.

In the absence of viscous effects, a flow starting from rest will be irrotational for all subsequent time, and according to Kelvin's theorem will be circulation-free. In the absence of circulation, solid bodies (such as a sphere or an airfoil) within the flow cannot experience any net forces. The paradox that an object in steady inviscid flow does not produce drag, whereas the same object in a real flow would do so, is attributed to Jean le Rond d'Alembert. *See* D'ALEMBERT'S PARADOX; KELVIN'S CIRCULATION THE-OREM.

For an irrotational flow, the curl of the velocity is zero ($\nabla \times V = 0$). The curl of the gradient of any scalar function is zero ($\nabla \times \nabla \phi = 0$). It then follows mathematically that the condition of irrotationality can be satisfied identically by choosing the scalar function, ϕ , such that $V = \nabla \phi$. For this reason, this scalar function ϕ has been traditionally referred to as the velocity potential, and the flow as a potential flow. *See* CALCULUS OF VECTORS; POTENTIALS.

By applying the continuity equation to the definition of the potential function, it becomes possible to represent the flow by the well-known Laplace equation ($\nabla^2 \phi = 0$), instead of the coupled system of the continuity and nonlinear Euler equations. The linearity of the Laplace equation, which also governs other important physical phenomena such as electricity and magnetism, makes it possible to use the principle of superposition to combine elementary solutions in solving more complex problems. In a potential flow problem, ϕ is calculated first by solving $\nabla^2 \phi = 0$, subject to appropriate Dirichlet or Neumann boundary conditions. Then, V is determined by evaluating $V = \nabla \phi$. Finally, the pressure field is determined by solving Bernoulli's equation. Many computer codes exist that can solve and demonstrate a general class of potential flow problems. A convenient way to visualize potential flows in two spatial dimensions is by drawing the orthogonal system of potential lines (lines of constant ϕ) versus streamlines (lines of constant stream function ψ). In two spatial dimensions, the potential function is related to the stream function by the Cauchy-Riemann conditions for a complex function w = $+ i\psi$. See BERNOULLI'S THEOREM; COMPLEX NUMBERS AND COMPLEX VARIABLES; EQUATION OF CONTINUITY; LAPLACE'S DIFFERENTIAL EQUATION; LAPLACE'S IRROTATIONAL MOTION; STREAM FUNC-TION.

The numerical solution of potential flows, including those with sources and sinks, can be obtained easily by the use of standard finite-difference, finitevolume, and finite-element methods. Solutions can also be obtained by the use of general-purpose mathematical and symbolic software, as well as by spreadsheet software. Given the elliptical nature of the governing equations, the solution at one point in the flow field depends on the surrounding boundary values. Consequently, boundary integral methods can provide computationally efficient solutions, especially if the interest lies in only a few points within the domain. See DIFFERENTIAL EQUATION; FI-NITE ELEMENT METHOD; FLUID-FLOW PRINCIPLES; IN-TEGRAL EQUATION; MATHEMATICAL SOFTWARE; SYM-BOLIC COMPUTING. Peter E. Raad

Bibliography. G. K. Batchelor, An Introduction to Fluid Dynamics, Cambridge University Press, 1967, reprint 2000; I. G. Currie, Fundamental Mechanics of Fluids, 2d ed., McGraw-Hill, New York 1993; H. Lamb, Hydrodynamics, 6th ed., Cambridge University Press, 1932, reprint 1993; A. R. Paterson, A First Course in Fluid Dynamics, Cambridge University Press, 1983, reprint 1987; F. M. White, Fluid Mechanics, 4th ed., McGraw-Hill, Boston, 1999.

Potentials

Functions or sets of functions from whose first derivatives a vector can be formed. A vector is a quantity which has a magnitude and a direction, such as force.

When the potential is a single function (scalar potential), the vector is obtained by the gradient operation, as in Eq. (10). When the potential itself is a vector (vector potential), the vector is obtained by the curl operation, as in Eq. (21).

Potential theory. The mathematical theory of the potential is basically the study of differential equations (1) and (2), called Laplace's equation and Pois-

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} + \frac{\partial^2 \phi}{\partial z^2} = 0 \tag{1}$$

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} + \frac{\partial^2 \phi}{\partial z^2} = -C\rho \tag{2}$$

son's equation, respectively. Here ϕ is a function defined in part or all of space, *C* is some constant, and ρ is a function characterizing the source of ϕ . The theory is best expressed in terms of vector calculus. The three vectors of unit magnitude which point in the positive directions of the *x*, *y*, and *z* axes of a right-handed cartesian coordinate system are denoted by **i**, **j**, and **k**. The radius vector which points from the origin to an arbitrary point with coordinates *x*, *y*, *z* is then $\mathbf{r} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$. Suppose that to each point in a region of space (or all of space) there is

associated a number ϕ such that the function $\phi(\mathbf{r})$ has partial derivatives with respect to x, y, and z, of first and second order everywhere in that region; and suppose .further that the second derivatives satisfy $\partial^2 \phi / \partial x \partial y = \partial^2 \phi / \partial y \partial x$, and corresponding equations for x and z, and for y and z. Then $\phi(\mathbf{r})$ is called a scalar potential. The vector **V** given by Eq. (3) is defined by

$$\mathbf{V}(\mathbf{r}) = \mathbf{i}V_x + \mathbf{j}V_y + \mathbf{k}V_z$$

= $-\mathbf{i}\frac{\partial\phi}{\partial x} - \mathbf{j}\frac{\partial\phi}{\partial y} - \mathbf{k}\frac{\partial\phi}{\partial z}$ (3)

this equation for every \mathbf{r} in the region and is called the gradient vector field associated with this potential. Equation (3) is abbreviated as $\mathbf{V}(\mathbf{r}) = -\nabla \phi(\mathbf{r})$, in which ∇ is called the gradient operator. It is a vector and a differentiation operator. The minus sign is a matter of convention. The curl of \mathbf{V} , a vector denoted by $\nabla \times \mathbf{V}$ and defined by Eq. (4), satisfies $\nabla \times \mathbf{V} = 0$ if

$$\nabla \times \mathbf{V} = \mathbf{i} \left(\frac{\partial V_z}{\partial y} - \frac{\partial V_y}{\partial z} \right) + \mathbf{j} \left(\frac{\partial V_x}{\partial z} - \frac{\partial V_z}{\partial x} \right) + \mathbf{k} \left(\frac{\partial V_y}{\partial x} - \frac{\partial V_x}{\partial y} \right)$$
(4)

V is defined by Eq. (3). Conversely, whenever a vector **V** satisfies $\nabla \times \mathbf{V} = 0$ in a certain region, there exists a scalar potential ϕ in that region such that $\mathbf{V} = -\nabla \phi$. Then **V** is called the (negative) potential gradient. If, however, the divergence of **V** vanishes, that is, if **V** satisfies Eq. (5), which can be written as

$$\frac{\partial V_x}{\partial x} + \frac{\partial V_y}{\partial y} + \frac{\partial V_z}{\partial z} = 0$$
(5)

 $\nabla \cdot \mathbf{V} = 0$, then ϕ satisfies Laplace's equation, Eq. (1), that is, $\nabla^2 \phi = 0$. In that case ϕ is called a harmonic function. *See* CALCULUS OF VECTORS.

If ϕ satisfies Laplace's equation $\nabla^2 \phi = 0$, in a finite region of three-dimensional space enclosed by the closed surface *S*, then the solution ϕ inside *S* is uniquely determined by specifying either the values of ϕ on *S* or the values of the derivatives of ϕ normal to *S*, $\partial \phi / \partial n$, on *S*. In the former case, ϕ is given by Eq. (6), in the latter case by Eq. (7). Here the normal

$$\phi(\mathbf{r}) = -\frac{1}{4\pi} \int \int \phi(\mathbf{r}') \frac{\partial}{\partial n} \left(\frac{1}{R}\right) dS \qquad (6)$$

$$\phi(\mathbf{r}) = \frac{1}{4\pi} \int \int \frac{\partial \phi(\mathbf{r}')}{\partial n} \frac{1}{R} \, dS \tag{7}$$

derivative is taken with the normal pointing out of the enclosed volume; the quantity R is the distance between the end point of **r** which is inside the surface, and the end point of **r'** which is on the surface,

$$R = \sqrt{(x - x')^2 + (y - y')^2 + (z - z')^2}$$

and the integration is extended over the whole closed surface *S*. The two cases are called the Dirichlet and Neumann boundary conditions, respectively.

It is also possible to specify mixed boundary conditions such as ϕ on part of *S* and $\partial \phi / \partial n$ on the remaining parts of *S*.

If the vector field satisifes $\nabla \times \mathbf{V} = 0$ and $\nabla \cdot \mathbf{V} = C\rho$, where ρ is some function defined in the region and whose integral over the region exists, then ϕ satisfies Poisson's equation, $\nabla^2 \phi = -C\rho$. A solution of this equation is given by Eq. (8). The integral ex-

$$\phi(\mathbf{r}) = \frac{C}{4\pi} \int \int \int \frac{\rho(\mathbf{r}')}{R} dx' dy' dz' \qquad (8)$$

tends over the whole three-dimensional region and R is again the distance from \mathbf{r} to \mathbf{r}' , $R = |\mathbf{r} - \mathbf{r}'|$. The most general solution of Poisson's equation is obtained by adding to the particular solution given by Eq. (8) the general solution of Laplace's equation.

Laplace's equation also has important applications in two dimensions for problems in a plane. *See* LAPLACE'S DIFFERENTIAL EQUATION; LAPLACE'S IRRO-TATIONAL MOTION.

Gravitational potential. According to Isaac Newton's theory of gravitation, every spherical body of mass M produces a force field **F**. Any other body, of mass m, will experience an attractive force, directed toward M, and of the magnitude given in Eq. (9),

$$F = G \frac{Mm}{r^2} \tag{9}$$

where distance between the centers of gravity of the masses *M* and *m* is denoted by *r*; if *M* is at the origin, **r** is the position vector of *m* and *r* is its magnitude; G is Newton's constant of gravitation. Such a force is present at every point in the space surrounding the mass *M* at which one chooses to put the mass *m*. This gravitational force F can be expressed in terms of the gravitational field strength \mathbf{f} defined by $\mathbf{f} =$ \mathbf{F}/m , that is, the force per unit mass. The gravitational potential is that function ϕ whose negative gradient equals $\mathbf{f}, \mathbf{f} = -\nabla \phi$, or $\mathbf{f}_x = -\partial \phi / \partial x$, and so forth. This definition of ϕ is made unique by requiring ϕ to vanish as the distance from M goes to infinity. Then $\phi = -GM/r$ for the gravitational potential of the mass M, for all points outside the mass M. The gravitational force exerted by the mass M on the mass m can be written as Eqs. (10): the differentia-

$$\mathbf{F} = -m\nabla\phi \qquad \phi = \frac{-GM}{r} \tag{10}$$

tion is with respect to the position of the mass m; ϕ is negative because **F** is attractive. The quantity $V = m\phi$ is called the gravitational potential energy. The negative gradient of ϕ gives **f**; the negative gradient of *V* gives **F**. The potential energy of *m* in the field of *M* increases (becomes less negative) as the distance increases. *See* GRAVITATION.

If the body is not spherically symmetric, **F** in Eq. (9) and ϕ in Eqs. (10) will hold only at large distances. But for any distance and arbitrary shape, if the mass *M* is spread out over a certain volume with mass density $\rho(\mathbf{r})$, then ϕ satisfies Poisson's equation (2) in the form $\nabla^2 \phi = +4\pi G \rho$. Its solution is given by Eq. (8) with $C = -4\pi G$.

When there are several masses present, M_1 , M_2 , ..., M_n , each of them will produce a force on a given mass m. The total force on m can be obtained by adding all the forces $\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_n$ by vector addition. However, it is obtained more easily by adding the potentials $\phi_1, \phi_2, \ldots, \phi_n$, produced by these masses and then finding the total force \mathbf{F} from $\mathbf{F} = -m\Delta\phi$, where ϕ is the sum of the potentials. This is the principle of superposition. It is one reason why potentials are so useful.

Those sets of points at which ϕ has a fixed value form two-dimensional surfaces in space. These surfaces are called equipotential surfaces. For a spherically symmetric mass distribution these surfaces are concentric spheres whose centers are at the center of mass. Since the vector $\nabla \phi$ is parallel to the normal of the surface $\phi = \text{constant}$, the lines of force are always perpendicular to the equipotential surfaces. *See* LINES OF FORCE.

Electrostatic potential. An electrically charged object with total electric charge Q will, when at rest, produce an electrostatic force field **F**. If another charge q is placed at a distance r from it which is large compared to the size of the two charges, this force is given by Coulomb's law. Its magnitude is as given in Eq. (11), and it is repulsive or attractive de-

$$\mathbf{F} = \frac{1}{4\pi\epsilon} \frac{Qq}{r^2} \tag{11}$$

pending on whether the two charges Q and q have the same signs (both positive or both negative) or have opposite signs. The constant ϵ can be written as $K\epsilon_0$, where K is the relative permittivity (K equals 1 for vacuum) and ϵ_0 depends on the units. In SI units, $(4\pi\epsilon_0)^{-1} = c^2/10^7$, where c is the velocity of light. The analogy between Eqs. (9) and (11) is obvious; an important difference is that gravitational forces are always attractive. *See* COULOMB'S LAW.

If the charge Q is spherically symmetric, of radius a, and centered at the origin, the electrostatic potential ϕ at a distance r > a from the origin is given by Eq. (12). This formula also holds when Q is not

$$\phi = \frac{1}{4\pi\epsilon} \frac{Q}{r} \tag{12}$$

spherically symmetric, but of characteristic size *a*, provided *r* is much larger than *a*. The electrostatic field strength is then $\mathbf{E} = -\nabla \phi$. It is the analog of **f** in the gravitational case. In general, the electrostatic force on a charge *q* is $\mathbf{F} = -q\nabla \phi$. In the special case when ϕ is given by Eq. (12), this force can be expressed by Eq. (13). The magnitude of this vector is

$$\mathbf{F} = -q\frac{\mathbf{r}}{r}\frac{d\phi}{dr} = \frac{1}{4\pi\epsilon}\frac{Qq}{r^3}\mathbf{r}$$
(13)

given by Eq. (11). See ELECTRIC FIELD.

If the charge Q is spread out over some volume with charge density $\rho(\mathbf{r})$ so that $\int \int \int \rho \, dx \, dy \, dz = Q$, then ϕ satisfies Poisson's equation (2) in the form $\nabla^2 \phi = -\rho/\epsilon$. Its solution is given by Eq. (8) with $C = 1/\epsilon$. The Coulomb potential given by Eq. (12) is a special case of this result. If the charge is not distributed over a volume, but over a surface such as the surface of a sphere [surface charge density $\sigma(\mathbf{r})$], then the triple integral over $\rho dx' dy' dz'$ in Eq. (8) must be replaced by the double integral over σdS , where dS is the differential surface element.

As in the gravitational case, the total force acting on a charge q due to the presence of a set of charges Q_1, \ldots, Q_n can be obtained by adding the potentials of these charges and finding $\mathbf{F} = -q\nabla\phi$; the principle of superposition holds. Equipotential surfaces are defined as in gravitation theory.

The electrostatic potential energy *V* of a charge *q* in a potential ϕ is given by the relation $V = q\phi$. *See* ELECTROSTATICS.

Magnetic scalar potential. If the magnetic induction vector B satisfies $\nabla \times \mathbf{B} = 0$, for example, in the absence of current densities and time-dependent electric fields, then **B** can be derived from a scalar potential ψ by the relation $\mathbf{B} = -\nabla \psi$. Maxwell's equations require $\nabla \cdot \mathbf{B} = 0$, so that the magnetic scalar potential must be a solution of Laplace's equation $\nabla^2 \psi = 0$ which has the form of Eq. (1). *See* MAXWELL'S EQUATIONS.

Logarithmic potential. A straight electrically charged cylinder of circular cross section (typically a wire) and of effectively infinite length poses an electrostatic problem in two dimensions, in the plane perpendicular to the cylinder axis. If λ is the charge per unit length on the cylinder, then the electrostatic field at a distance *r* greater than the radius of the cylinder and measured from the cylinder axis is given by the relation $\mathbf{E} = \lambda/(2\pi\epsilon\mathbf{r})$. It is directed radially outward if λ is positive. The potential ϕ whose negative gradient gives this field **E** is called logarithmic potential and is given by Eq. (14). Here ln is the natural logarithm. The

$$\phi(r) = -\frac{\lambda}{2\pi\epsilon} \ln r + \text{constant} \qquad (14)$$

additive constant can be so chosen that $\phi = 0$ for r = a, where *a* is the radius of the cylinder. Then Eq. (14) can be written as Eq. (15).

$$\phi(r) = -\frac{\lambda}{2\pi\epsilon} \ln \frac{r}{a} \tag{15}$$

Work and potential difference. When a charge q is moved in an electrostatic potential field, work is being done. This work W amounts to a loss of energy and is therefore negative. If q is moved from point 1 to point 2, W will be given by Eq. (16). The

$$W = -\int_{1}^{2} \mathbf{F} \cdot d\mathbf{r} = q \int_{1}^{2} \nabla \phi \cdot d\mathbf{r} = q(\phi_{2} - \phi_{1})$$
$$= V_{2} - V_{1}$$
(16)

work done equals the difference in potential energy between the final and initial positions. Thus, W does not depend on the path along which q is moved, but only on the initial and the final position. This is a direct consequence of the fact that **F** is a gradient vector field. It follows, in particular, that no work is being done if the charge q is moved in an arbitrary path which returns to the initial position, for example, a path which is closed. Mathematically this means $\oint \mathbf{F} \cdot d\mathbf{r} = 0$ whenever \mathbf{F} is a gradient field. Physical systems in which all forces can be derived from potentials are therefore called conservative systems. *See* WORK.

The potential difference between two points 1 and 2 in a potential field is the work per unit charge that needs to be done to move a charge from 1 to 2. The electrostatic potential at **r** given by Eq. (12), in which ϕ is chosen to vanish at infinite distance, is the work per unit charge necessary to move a positive charge from infinity to the position **r**. In an electric network the potential difference between two points is the work per unit charge that must be done to move a positive charge through the network from one point to the other. It is usually measured in volts.

Neither potential nor potential energy has absolute physical meaning. They must be measured relative to some arbitrary reference potential or potential energy. For example, it can be chosen to be zero at infinite distance from a charged body; or in electric networks one point can be connected to the ground and defined to have zero potential. Only voltage difference is a physically meaningful quantity, not absolute voltage.

In gravitation theory the potential difference between a body located at a distance *b* above the Earth's surface and its location on the Earth's surface is, according to Eq. (10), expressed as $\phi(b) - \phi(0) =$ $-GM/(R+b) + GM/R = GMb/R^2$, where the Earth's radius and mass are *R* and *M*, respectively, and *b* is much smaller than *R*. This potential difference is written *gb* where $g = GM/R^2$ is the gravitational acceleration on the Earth's surface. If the body has a mass *m*, the two locations differ in potential energy by *mgb*.

Vector potential. The electromagnetic vector potential **A** is a vector, that is, a set of three functions A_x , A_y , A_z , whose partial derivatives give the magnetic induction vector **B** through the relation **B** = $\nabla \times \mathbf{A}$. The curl operation is defined in Eq. (4). This definition ensures that one of the Maxwell equations, $\nabla \cdot \mathbf{B} = 0$, is identically satisfied. **A** is not uniquely defined by $\nabla \times \mathbf{A}$. Since the curl of a gradient vanishes identically, a gradient added to **A** will not change **B**. This nonuniqueness is called the gauge invariance of **B**.

If the magnetic field strength **H** and the induction **B** are linked by a constant permeability μ such that **B**= μ **H**, then the vector potential **A** satisfies Eq. (17),

$$\nabla^2 \mathbf{A} = -\mu \mathbf{j} \tag{17}$$

where **j** is the current density which produces the magnetic field. Each component of this vector equation is a Poisson equation of the type of Eq. (2). Its solution is given by Eq. (18), where R is the mag-

$$\mathbf{A}(\mathbf{r}) = \frac{\mu}{4\pi} \int \int \int \frac{\mathbf{j}(\mathbf{r}')}{R} \, dx' \, dy' \, dz' \qquad (18)$$

nitude of the vector $\mathbf{r} - \mathbf{r}'$. If the current *I* does not have a density **j** but is a line current, Eq. (16) is

replaced by Eq. (19), the integral being a line inte-

$$\mathbf{A}(\mathbf{r}) = \frac{\mu}{4\pi} \oint \frac{Id\mathbf{s}}{R}$$
(19)

gral along the whole current loop. When this loop is small compared to R, Eq. (17) can be transformed by Stokes' integral theorem into the potential of a magnetic dipole, which is given by Eq. (20). Here **R**

$$\mathbf{A}(\mathbf{r}) = \frac{\mu}{4\pi} \frac{\mathbf{M} \times \mathbf{R}}{R^3}$$
(20)

is the distance of the end point of **r** from the dipole, and **M** is the magnetic dipole moment. When **M** is produced by a current *I* forming a circle of radius *a*, $M = \pi a^2 I$. See STOKES' THEOREM.

Electromagnetic potentials. In the general timedependent case, the electric field strength **E** and the magnetic induction **B** can be expressed in terms of the scalar potential ϕ and the vector potential **A** as in Eqs. (21). This ensures that the homogeneous

$$\mathbf{E} = -\nabla\phi - \frac{1}{c}\frac{\partial \mathbf{A}}{\partial t} \qquad \mathbf{B} = \nabla \times \mathbf{A} \qquad (21)$$

Maxwell equations are identically satisfied. In free space, outside a source distribution of charge density ρ and current density **j**, the potentials satisfy inhomogeneous wave equations (22). The constant

$$\nabla^2 \phi - \frac{1}{c^2} \frac{\partial^2 \phi}{\partial t^2} = \frac{-\rho}{\epsilon_0}$$

$$\nabla^2 \mathbf{A} - \frac{1}{c^2} \frac{\partial^2 \mathbf{A}}{\partial t^2} = -\mu_0 \mathbf{j}$$
(22)

 ϵ_0 was defined above [after Eq. (11)] and μ_0 is in SI units $4\pi \times 10^{-7}$. When the potentials are time independent, Eqs. (22) reduce to the Poisson equations. *See* WAVE EQUATION.

The potentials ϕ and **A** are not uniquely determined by Eqs. (21) if the fields are known. One can always add a gradient $\nabla \Lambda$ to **A** and a term $-\partial \Lambda/c\partial t$ to ϕ without affecting **E** and **B**. Such a transformation of the potentials is called a gauge transformation. The invariance of the fields under this transformation is gauge invariance. This freedom of ϕ and **A** can be restricted in a relativistically invariant way by the Lorentz condition on the potentials, Eq. (23).

$$\nabla \cdot \mathbf{A} + \frac{1}{c} \frac{\partial \phi}{\partial t} = 0$$
 (23)

In the special theory of relativity the four functions ϕ , A_x , A_y , A_z transform under Lorentz transformations like four components of a four-vector. They are sometimes collectively called four-potential. *See* RELATIVITY.

Retarded and advanced potentials. Electromagnetic fields propagate through empty space with a finite velocity *c*, the velocity of light, rather than with infinite velocity. Consequently, it takes a certain time for an electric or magnetic field produced by a charge or current at time t_Q and at position \mathbf{r}_Q to reach another position r_P . This time is $|\mathbf{r}_P - \mathbf{r}_Q|/c$. The time t_Q is the retarded time of the fields which are observed at time $t_P = t_Q + |\mathbf{r}_P - \mathbf{r}_Q|/c$. The fields present at

 t_p and produced at $t_Q < t_p$ are the retarded fields. The potentials from which they are derived are the retarded potentials. They are important in problems of electromagnetic radiation.

If time would run backward, the fields and potentials observed at time t_P at point \mathbf{r}_P would have to be produced at some future time t_Q' and at a position \mathbf{r}_Q' of the source such that $t_P = t_Q' - |r_Q' - r_P|/c$. The potentials at time t_P which are produced at the later time $t_Q' > t_P$ are the advanced potentials. While they have no direct physical meaning, they are important in the mathematical analysis of radiation theory.

Liénard-Wiechert potentials. The retarded and advanced electromagnetic scalar and vector potentials produced by a moving point charge can be expressed in terms of the (retarded or advanced) position and velocity of that charge. This was first done by A. Liénard in 1898 and E. Wiechert in 1900.

Polarization potentials. Two vectors \mathbf{II}_e and \mathbf{II}_m can be defined by $\mathbf{A} = (1/c) (\partial \mathbf{II}_e/\partial t) + \nabla \times \mathbf{II}_m$ and $\phi = -\nabla \cdot \mathbf{II}_e$. The electromagnetic fields \mathbf{E} and \mathbf{B} can then be expressed in terms of \mathbf{II}_e and \mathbf{II}_m . These are the polarization potentials. They are unusual because from them are derived not only a vector, \mathbf{A} , but also a scalar, ϕ . The Lorentz condition is identically fulfilled. The fields \mathbf{E} and \mathbf{B} are now given by second derivatives of \mathbf{II}_e and \mathbf{II}_m , an exception to the usual definition of potentials, which involves only first derivatives. Heinrich Hertz in 1889 was the first to use such potentials. They are therefore often called Hertz potentials.

Debye potential. Radiation or scattering of electromagnetic waves by a distribution of localized sources in a homogeneous isotropic medium leads to fields **E** and **B** which can be expressed in terms of only two scalar potentials II_e and II_m . They were first introduced by Peter Debye in 1909 and are related to the polarization potentials.

Velocity potential. In the theory of fluid dynamics the absence of vortices is expressed by $\nabla \times \mathbf{v} = 0$, where \mathbf{v} is the velocity vector which is a function of position inside the fluid. If that equation holds, there always exists a velocity potential, that is, a scalar function ϕ with the property $\mathbf{v} = -\nabla \phi$. One then has potential flow of that fluid. This potential satisfies Laplace's equation (1) if the fluid is incompressible and in steady motion, so that the continuity (conservation of mass) equation requires that $\nabla \cdot \mathbf{v} = 0$. *See* FLUID-FLOW PRINCIPLES.

Thermodynamic potential. In thermodynamics the Helmholtz function of free energy A (or F) and the Gibbs function G are both referred to as thermodynamic potentials. The reason for this name is the analogy to mechanical potential energy rather than to potential. When an isolated system is in equilibrium under various conditions, these quantities are a minimum. Thus thermodynamic equilibrium can be attained at constant temperature and volume, in which case A is a minimum; alternatively, it can be attained at constant temperature and pressure, in which case G is a minimum. *See* CHEMICAL THERMODYNAMICS; FREE ENERGY.

Chemical potential. A thermodynamic system in general consists of several chemical constituents. Let n_k be the number of moles of constituent k. Then the chemical potential of this constituent is the rate of change of the Gibbs function G with n_k , $\partial G/\partial n_k = \mu_k$. Even when constitutent k is not initially present, μ_k can be different from zero, so that its addition to the system will affect the Gibbs function. Chemical potentials play an essential role in phase transitions and chemical equilibrium.

Bibliography. J. Bliedtner, *Potential Theory*, 1986; M. Born and E. Wolf, *Principles of Optics*, 7th ed., 1999; J. Kral et al. (eds.), *Potential Theory: Surveys and Problems*, 1988; F. Rohrlich, *Classical Charged Particles*, 1965, reprint 1989; W. R. Smythe, *Static and Dynamic Electricity*, 3d ed., 1968, reprint 1989; M. W. Zemansky and R. Dittman, *Heat and Thermodynamics*, 7th ed., 1996.

Potentiometer

An instrument that precisely measures an electromotive force (emf) by balancing it against a known potential drop established by a three-terminal resistive divider (Fig. 1), across which a known voltage exists. A divider can adjust and control the voltage applied to some device or part of a circuit, for example, a volume control. The ratio of output to input voltage is equal to that of the tapped to total divider resistance when the current required by the controlled device is negligible. Potentiometers as voltage-measuring instruments have largely been replaced by the ratio instrument known as a digital voltmeter. The Weston standard cells which supplied a known standard voltage have also been superseded by solid-state voltage reference devices which are usually incorporated in the digital voltmeter.

Another type of potentiometer (Fig. 2) involves a current comparator which senses and corrects inequality of ampere-turns in two windings on a magnetic core. Two matched toroidal cores, wound with an identical number of turns, are excited by a fixedfrequency oscillator. The fluxes induced in the cores are equal and oppositely directed so that they cancel with respect to a winding which links both cores. In the absence of any additional magnetomotive force (mmf), there is no signal in the detector winding which links both cores. If, in another winding A linking both cores, a direct current is injected, its mmf adds to the flux in one core and subtracts from the other at any instant. The resulting net flux in the detector winding induces a voltage in it. This signal is used to control direct current in another wind-

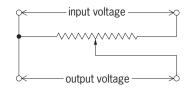


Fig. 1. Three-terminal resistive voltage divider.

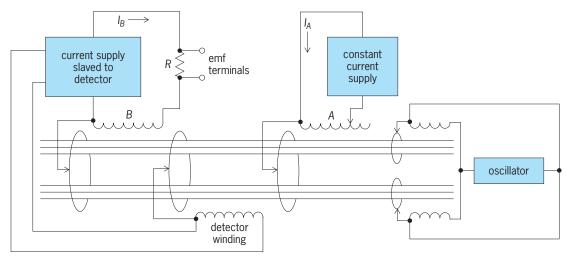
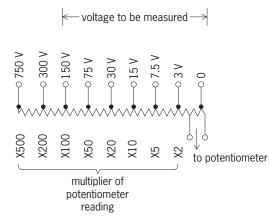
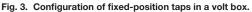


Fig. 2. Circuit diagram of a current-comparator potentiometer, a type of constant-resistance potentiometer.

ing B that also links both cores. Its mmf is opposed to that of A and when they are equal, there is no signal in the detector winding. Thus a constant current in an adjustable number of turns in winding A is matched to a variable current in a fixed number of turns in winding B; that is, their ampere-turns are equal and opposed. The voltage drop $I_B R$ is used to oppose the emf to be measured. The potentiometer is made direct-reading in volts in terms of the turnsratio B/A, by adjusting the constant-current source, using a voltage standard (omitted to simplify the figure). An advantage of this potentiometer over those whose long-term accuracy depends on the stability of a resistance ratio is that the ratio here is the turnsratio of windings on a common magnetic core, dependent solely on conductor position and therefore not subject to drift with time. See CURRENT COMPARATOR; ELECTROMAGNETIC INDUCTION; MAG-NETISM.

Volt box. Volt boxes, used to extend the range of potentiometers (in some instances to 1500 V), are resistive voltage dividers having a limited number of fixed-position taps (**Fig. 4**). Depending on its anticipated value, the unknown voltage is connected across appropriate input terminals (say 0-150 V),





and the potentiometer is connected to the 0-1.5 V output terminals. The potentiometer is balanced in the usual way, and its reading is multiplied by the factor corresponding to the tap point used. While no current is taken by the potentiometer at balance, current is supplied to the volt box by the source, and it is the voltage drop in the divider resistance which is measured, not the open-circuit emf of the source. For a given voltage, power consumption and internal heating are greater in a low- than in a high-resistance circuit, with consequent greater change in resistance ratio. Thus a higher ohms-per-volt design is desirable. However, the quality of insulation becomes more critical in high-resistance designs, since the insulating structures that support the resistance elements are themselves high-resistance leakage paths. As a compromise, the resistance of working volt boxes ranges between 200 and 750 ohms/V. High-accuracy standard volt boxes (1000 ohms/V) have a guard circuit which maintains shields at appropriate potentials along the insulating structure.

Current measurements. Potentiometer techniques may also be used for current measurement, the unknown current being sent through a known resistance and the *IR* drop opposed by balancing it at the voltage terminals of the potentiometer. Here, of course, internal heating and consequent resistance change of the current-carrying resistor (shunt) may be a critical factor in measurement accuracy; and the shunt design may require attention to dissipation of heat resulting from its I^2R power consumption. *See* CURRENT MEASUREMENT; JOULE'S LAW.

Alternating-voltage measurements. Potentiometer techniques have been extended to alternating-voltage measurements, but generally at a reduced accuracy level (usually 0.1% or so). Current is set on an ammeter which must have the same response on ac as on dc, where it may be calibrated with a potentiometer and shunt combination. Balance in opposing an unknown voltage is achieved in one of two ways: (1) a slide-wire and phase-adjustable supply; (2) separate in-phase and quadrature adjustments

on slide wires supplied from sources that have a 90° phase difference. Such potentiometers have limited use in magnetic testing. *See* ALTERNAT-ING CURRENT; ELECTRICAL MEASUREMENTS; VOLTAGE MEASUREMENT.

Voltage divider. A three-terminal resistive voltage divider (sometimes called a potentiometer) is shown schematically in **Fig. 5**. Two terminals are the ends of the resistor (or a series combination of resistors) and the third terminal is an intermediate movable connection. This divider is used to adjust and control the voltage applied to some device or part of a circuit, for example, the volume control on a radio. The ratio of output to input voltage is equal to the tapped to total divider resistance when the current required by the output device is negligible.

Forest K. Harris; Ronald F. Dziuba; Bryan P. Kibble Bibliography. C. F. Coombs, *Electronic Instrument Handbook*, 3d ed., 1999; B. P. Kibble et al., *A Guide To Measuring Direct and Alternating Current and Voltage Below 1 MHz*, Institute of Measurement and Control, U.K., 2003; M. V. Reissland, *Electrical Measurements*, 1989; L. Schnell (ed.), *Technology of Electrical Measurements*, 1993; L. M. Thompson, *Electrical Measurements and Calibration*, 2d ed., 1994.

Pottery

Vessels made entirely or partly of clay, and fired to a strong, hard product; occasionally, the term refers to just the lower grades of such ware. Alternatively, it refers to the manufacturing plant at which such ware is made. An older meaning is the art of making such ware; in this use it becomes synonymous with the older definition of ceramics. Pottery may be glazed or unglazed. *See* CERAMICS; GLAZING.

Grades of pottery are distinguished by their color, strength, absorption (the weight of water soaked up when the piece is submerged, expressed as a percentage of the original weight), and translucency (ability to transmit light). All these properties refer to the material or "body" under any glaze present. *See* PORCELAIN.

China is white in color, strong, has less than 2% absorption, is always glazed, and is translucent in thin ware. Special types are bone china, containing phosphates from calcined bones as a fluxing material; hotel china, made extra thick for maximum strength and therefore not translucent; frit china (also called frit porcelain), containing ground glass to give a translucent body maturing at a moderate firing temperature.

Stoneware has a characteristic cream or brown color, high strength, 0–5% absorption, and no translucency; it is frequently unglazed.

Earthenware (sometimes known as semivitreous china) is white or ivory, has less strength than china or porcelain, 3–10% absorption, and no translucency; it is usually glazed. Most everyday tableware is of this type. Faience is a special type with a soft,

porous, red or yellow body covered by an opaque glaze. Majolica ware is a type having over 15% absorption and an opaque glaze over a relatively weak red or gray body.

Other special types of pottery are Parian ware, a body with a high flux content, usually unglazed, which fires to a smooth, marblelike finish, and terra cotta, a yellow, red, or brown earthenware with no glaze, used for art sculpture and similar purposes.

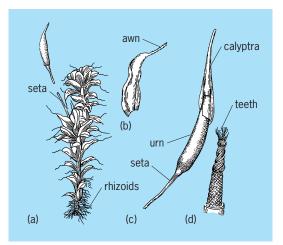
The firing of glazed ware is usually done in two steps; first the unglazed body is fired to give it strength, and then after the glaze is applied, it is refired at a lower temperature (except porcelains, when the second firing is at a higher temperature). *See* KILN.

Absorption is due to the presence of open pores or voids in the fired material into which water can penetrate; in general, the higher the firing temperature, the lower the absorption. Body color is determined mainly by raw-material purity. Strength depends on the porosity and also on the amount and type of glass and crystals developed in the body on firing. Translucency is obtained in products in which there is low porosity and little difference in index of refraction between the glass and crystals in the body. *See* CLAY. J. F. McMahon

Bibliography. J. Colbeck, *Pottery Materials: Their Composition, Preparation, and Use*, 1988; J. Colbeck, *The Technique of Throwing*, 1988; E. Cooper and D. Royle, *Glazes for the Studio*, 1986; P. Rado, *Introduction to the Technology of Pottery*, 2d ed., 1988; C. F. Speight, *Hands in Clay: An Introduction to Ceramics*, 4th ed., 1999.

Pottiales

An order of the true mosses (subclass Bryidae), consisting of three families and about 91 genera. The order is characterized by short, papillose leaf cells and deeply divided, commonly twisted peristome teeth (see **illus**.). Members of the Pottiales typically



Tortula ruralis. (a) Entire plant. (b) Leaf. (c) Urn with calyptra. (d) Peristome. (After W. H. Welch, Mosses of Indiana, Indiana Department of Conservation, 1957)

grow erect, with stems simple or forked by subfloral innovation, and they often produce gemmae (asexual reproductive bodies) on leaves, stems, or rhizoids. The leaves have revolute margins, and are arranged in many rows. The single costa often ends in a hyaline hairpoint. The upper cells are short and thick-walled and often roughened with numerous C-shaped papillae. The alar cells are not differentiated, although the basal cells sometimes extend up the margins as a border. The cylindric capsules are generally exserted and usually erect. An operculum is usually differentiated, and the peristome, usually present, consists of 16 teeth which are often deeply divided into two filiform segments and papillose-roughened. The calyptra is usually cucullate. The chromosome numbers are generally 12, 13, or multiples thereof. See BRYI-DAE; BRYOPHYTA; BRYOPSIDA. Howard Crum

Poultry production

Poultry production comprises two major categories, meat production and egg production. Most poultry produced in North America is grown under close control on highly specialized farms. The evolution from small flocks to large commercial units after World War II was facilitated by rapid advances in the knowledge of nutrition, breeding, housing, disease control, and processing of poultry and eggs, and by improvements in transportation and refrigeration which made possible distant marketing of fresh products.

Incubation. Artificial incubation was a major advance in poultry production because it became possible to hatch large numbers of chicks of the same age for farmers to raise for meat or egg production. Poultry eggs can be successfully incubated in fanventilated incubators by maintaining a temperature of 99.5-100°F (37.5-37.8°C) and 55% relative humidity. Eggs must be turned during the first two-thirds of the incubation period to prevent adherence of the embryo to the embryonic membranes. The duration of incubation, under these standard conditions, for several types of poultry are: chicken, 21 days; turkey, guinea fowl, peafowl, and most ducks, 28 days; most geese, 30 days; pheasants, Chukar partridge, and Bobwhite quail, 24 days; Japanese quail, 17-18 days; and Muscovy ducks and Egyptian geese, 35 days.

Modern incubators are constructed of materials that can be effectively cleaned and disinfected and that provide good insulation of the chamber. Eggs are set in specially designed plastic flats which fit into channels in egg racks that move on wheels. The egg racks are equipped with mechanical systems to tilt the eggs 45°; turning is usually done hourly. Eggs are transferred from the setting trays into hatching trays 3 days before expected hatch. Hatching trays are held in specially designed hatching units much like setter cabinets. After the hatch is completed, the chicks are transferred to a conveyer belt for processing or directly into plastic boxes with absorbent paper pads or into disposable paper boxes with wood fiber pads. Chick servicing often involves sexing, vaccination, and beak trimming. A machine is available for injection of eggs at transfer; vaccines for protection of chicks against Marek's disease can be administered to the 18-day-old-embryos with improved early immunity. Usually, chicks are held at the hatchery prior to shipment to farms, starting early on the day after hatching. Specially designed delivery trucks or buses are used to provide adequate ventilation for chicks during shipment.

Hatcheries must be carefully designed for proper sanitation, and adequate air supply to incubators and hatchers. Ventilation air must be cooled in very hot weather and warmed in cold weather. Most hatcheries add moisture to air with humidifiers. Hatchery ventilation systems are designed to prevent movement of contaminated air from the hatching room, chick processing area, or equipment washing area into the clean egg storage or setting rooms.

Breeding. The genetic stock used for modern poultry production is produced by highly specialized breeding companies. Meat poultry is selected for good meat type, fast growth, disease resistance, and efficient conversion of feed to meat. Different strains of chickens are used for table egg production. These are selected for high egg production, large egg size, and small body weight for better conversion of feed to eggs and good livability. The body weights of meat and egg production strains are dramatically different (**Fig. 1**). The slow growth and low body weight of the commercial Leghorn demonstrate why these strains have low meat value.

Dual-purpose breeds or strains of poultry are no longer used extensively for commercial production. Instead, producers of specialty poultry and eggs (for example, free-range poultry) may still use traditional breeds such as New Hampshire Red or Barred Plymouth Rock which were used in the past for both meat and egg production.

Parent stock. Commercial poultry breeders usually sell parent stock chicks to hatcheries or integrated poultry companies. In most cases the commercial product consists of female and male parent lines which are crosses. The customer must continue to buy parent stock from the breeding company because the second generation from these stocks would not produce the same desirable product.

The parent stocks used to produce commercial chicks for meat or egg production require special

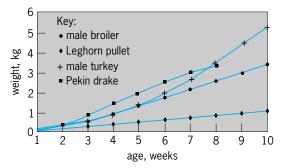


Fig. 1. Comparison of growth of commercial poultry.

care and management. Chicken parent stock is usually kept on litter floors or pens with part litter and part slatted floors. The litter areas are necessary for effective mating. Meat-type chickens must have feed restricted to control body weight, or they rapidly become overweight and mating, fertility, and egg production decline rapidly. Feed restriction is usually started by 3 weeks of age and continues throughout life. Males and females are raised separately for more effective weight control. At sexual maturity, males and females are housed together, but they feed from separate feeders. This system allows the manager to control the body weight of both sexes.

Turkey parent lines are also kept on litter because of their heavy body weight, but all mating is performed by artificial insemination. This procedure is more expensive, but high-quality males can be mated with larger numbers of hens, which is a distinct advantage. Also, the change to artificial mating has allowed geneticists to make more rapid progress in body weight gain and breast yield. Turkey semen cannot be successfully preserved by freezing, but short-term storage is possible. Therefore, toms may be raised on separate stud farms, where semen is harvested, processed, extended, and transported to the location of the hens for insemination. The duration of fertility after one mating in poultry can exceed 3 weeks, but commercial flocks are often inseminated weekly for high fertility.

Controlled lighting programs. Controlled lighting programs are necessary for breeding poultry. They start with short days during the growing phase to prevent the onset of egg production before females reach an appropriate body size. At the correct age for the genetic stock, the daylength is increased to provide a photostimulatory photoperiod (often 14–16 h of light per day). Males must be exposed to photostimulatory daylength at least a week prior to the females because their response to light is slower.

Poultry have a diurnal photosensitive period that occurs 11–16 h after dawn. A significant period of light must be provided within this period to stimu-



Fig. 2. Commercial broiler house showing side curtain and ridge vents.

late a reproductive response. Light penetrates into the brain, where it stimulates cells in the hypothalamus to produce gonadothophin-releasing hormone. The latter stimulates the pituitary gland to produce follicle-stimulating hormone and luteinizing hormone, which initiate and maintain reproductive function. *See* BREEDING (ANIMAL).

Brooding and rearing. Day-old chicks require an ambient temperature of $85-87^{\circ}F$ (29-30°C) for normal growth and health during the first week of life. As the chicks grow and feather, they can tolerate lower temperatures. Brooding heat is often provided by a radiant gas brooder stove. The temperature is typically adjusted to $95^{\circ}F$ ($35^{\circ}C$) at the stove edge during the first week, and the chicks are allowed to find their comfort zone. The stove temperature is often decreased $5^{\circ}F$ ($3^{\circ}C$) per week until the prevailing house temperature is reached. This system provides an easy method of managing the environment for young chicks, with a low risk of overheating or chilling the birds.

If a controlled-environment house is available, the entire brooding area can be heated to the comfort zone, but it is important to provide uniform conditions to prevent piling or chilling mortality. If houses are well insulated, this system reduces the amount of energy required for brooding; with increasing energy costs and improved ventilation controls available, warm room brooding has increased on commercial farms.

Most chicks are started on floors that are covered with 2-4 in. (5-10 cm) of a litter material such as pine shavings, rice hulls, or peanut hulls. Litter materials should be dry and free from harmful chemicals or microbes. If meat flocks are healthy, the litter is often reconditioned by removal of cake, top dressed with new litter, and reused. Complete litter removal may occur annually, depending on flock health and economic conditions.

Feeding is usually done is small troughs or on plastic trays until the chicks learn to eat and drink. Chicks are quickly trained to eat from mechanical feeders and drink from closed water delivery systems to reduce labor. Watering systems have traditionally provided water in small troughs or cups, but nipple drinkers are also popular because they provide little opportunity for bacterial growth. Modern brooding houses have the feeders and waterers suspended from above with a cable-and-winch system. This system allows the height to be adjusted frequently for bird comfort and improved sanitation.

Feeding and nutrition. Poultry diets consist of common grains and protein sources with mineral and vitamin supplements. Animal or vegetable fats may be added to increase energy and reduce dustiness. Corn, grain sorghum, wheat, oats, and barley are often used for poultry feeding in the United States. Soybean meal is widely used as a protein supplement. Other important protein supplements are meat meal, fish meal, safflower meal, feather meal, and canola meal. By-product ingredients (for example, corn gluten meal or brewers grains) are also used in poultry diets when warranted by availability and price. Rations

for poultry are typically formulated by computer programs which compare ingredient costs with dietary requirements to provide least-cost formulas.

Ingredient quality is important in realizing optimum flock performance. Some feeds such as soybeans contain antinutritional factors which must be destroyed by heat treatment before they are useful for poultry feed. The ingredients must be free from mold toxins, harmful bacteria, and chemical contaminants. Moreover, quality control at the feed mill is extremely important to assure optimum flock performance and freedom of poultry and eggs from undesirable residues. *See* ANIMAL FEEDS.

Housing. The purpose of a poultry house is to confine the birds; to protect them from predators and environmental extremes which would cause mortality or reduce growth, feed efficiency, immuno-competence, fertility or egg production; to facilitate light control; and to facilitate bird management (**Fig. 2**).

Poultry houses can be constructed from locally available building materials. Smooth interior surfaces are preferred for effective sanitation. Houses are usually a maximum of 40 ft (12 m) wide to facilitate more uniform ventilation. House length is approximately 500 ft (152 m); most houses are constructed with a gable roof (**Fig. 3**).

In mild climates, poultry houses with open sides covered with poultry netting are often used. The side openings are usually covered with a curtain which can be raised or lowered to control the internal environment. If ambient temperatures rise above 95°F (35° C) or fall below 32° F (0° C), the roof of the house is insulated. When temperatures over 100° F (38° C) are expected, the house should also be equipped with a fogging system and fans (or both) to move air over the birds during high-temperature periods.

Mechanically ventilated houses allow more complete control of the environment within. For example, a well-constructed house in a mild climate might be able to maintain temperatures of $65-80^{\circ}$ F (18- 27° C), which reduce stress and provide efficient conversion of feed to meat or eggs. In hot climates, incoming air can be cooled with evaporative pad coolers or fogging systems.

Production systems. Chickens for table egg production are often housed in cages to provide cleaner eggs and protect the birds from disease agents which are recycled to birds from the manure (**Fig. 4**). Cages also allow more birds to be placed within a house because the cages can be arranged in tiers. Increasing the bird density in the house increases the overall bird heat which is available, and houses can be kept warmer in cold climates. Laying cages also allow eggs to be collected by belt conveyors.

Meat chickens and turkeys are usually grown in litter-floor houses because heavier poultry experience more lameness, breast blisters, and weaker bones and joints when grown in cages. Litter floors require careful management and ventilation to prevent problems with wet or caked litter. Meat chickens are typically grown in the same house from 1-



Fig. 3. Interior of commercial broiler house showing pan feeder, bell-type waterers, and radiant gas brooders.

day-old until market (6-9 weeks of age). Meat turkeys are typically kept in a brooder building until about 6 weeks of age and then transferred to a growing house until marketing (12-25 weeks of age).

Ducks and geese. With some modification of husbandry, ducks and geese can be successfully raised in confinement. They do not require water for swimming and can be grown in litter-floor houses similar to meat chickens or turkeys. Young ducklings are sometimes started on slatted or raised wire floors because of the wetness of the droppings. Pekin and Muscovy ducks are the most common breeds used for meat production, but Indian Runner and Khaki Campbell breeds are used for both eggs and meat.

Embden geese are most often used for meat production. The White Chinese breed is often used for weeding in commercial crops or backyard gardens. Toulouse geese are common on small farms and are sometimes used for commercial meat production.

Health maintenance. The production of commercial poultry in large flocks requires well-designed disease-control programs. The first requirement is

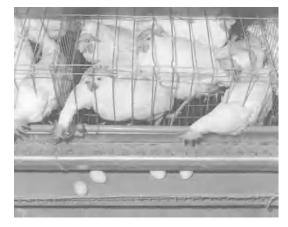


Fig. 4. Multiple-hen laying cage showing feeder, water cup, and egg collection belt.

for maintenance of biosecurity in production units. This means that entrance of contaminated workers and visitors, birds, feed, and equipment must be prevented. Equipment must be cleaned and disinfected before it is allowed into a production unit. Farms are operated on an all-in all-out basis with only one age of poultry at any given time. This prevents the vertical transmission of disease between age groups. Houses and equipment are often cleaned and disinfected between flocks, and always if a serious disease problem has been experienced.

Some poultry diseases are more effectively or economically controlled by vaccination. Examples are Marek's disease, Newcastle disease, infectious bronchitis, avian pox, infectious bursal disease, and often several others, depending on the disease history of the farm and the area where the poultry are raised. Some virus vaccines (for example, Newcastle, infectious bronchitis) can be administered by spray or in drinking water, while others (Marek's disease) must be injected. *See* NEWCASTLE DISEASE.

Processing and marketing. Shell eggs are often processed in plants located on the farm. With successful mechanization of egg collection, the eggs are often carried from the house directly into the processing machinery on a bar conveyor (in-line processing). This reduces processing labor and egg handling and breakage. By this procedure, eggs must be processed daily. Off-line processing provides more flexible processing hours, and eggs can be processed at a remote site; this system also allows one-age farms to be used for production.

A mechanical system can detect and separate cracked eggs. Egg processing machines can process up to 300 cases of eggs (360 eggs per case) per hour. Machines are also available to separate the yolk and albumen from the egg shell. Liquid yolk, albumen, or whole egg is pasteurized and sold for production of many food products. It can also be dried for use in cake mixes and similar products. Hard-cooked eggs are produced in mechanized plants and sold to restaurants and institutional food establishments. Poultry are usually processed at large central plants under inspection by the U.S. Department of Agriculture. The meat birds are loaded on trucks at farms during the night, and processing typically begins at midnight, continues for two shifts, and is followed by extensive cleaning and disinfection of the plant. Poultry are removed from transport racks or coops and hung on shackles. After stunning, bleeding, scalding, and feather removal, the carcasses must be transferred to a second line or table in a different room for evisceration, chilling, cutting, and packaging. Many poultry processing plants cut up poultry before packaging, and they may also separate meat from bone or skin. Special cuts include nuggets, breast strips, and cuts for fried chicken outlets. Meat may be mechanically removed from bone of less desirable cuts (for example, backs) to be used in production of hot dogs, sausages, and other products. See AGRICULTURAL SCI-ENCE (ANIMAL); EGG (FOWL). Ralph A. Ernst

Bibliography. B. W. Calnek, (ed.), *Diseases of Poultry*, 10th ed., 1997; R. D. Crawford, ed., *Poultry*

Breeding and Genetics, 1990; S. Leeson and J. D. Summers, Commercial Poultry Nutrition, 2d ed., 1997; L. S. Mercia, Raising Poultry the Modern Way, 1993; M. O. North and D. D. Bell, Commercial Chicken Production Manual, 4th ed., 1990; M. L. Scott and W. F. Dean, Nutrition and Management of Ducks, 1991.

Powder metallurgy

A metalworking process used to fabricate parts of simple or complex shape from a wide variety of metals and alloys in the form of powders. The process involves shaping of the powder and subsequent bonding of its individual particles by heating or mechanical working. Powder metallurgy is a highly flexible and automated process that is environmentally friendly, with a low relative energy consumption and a high level of materials utilization. Thus it is possible to fabricate high-quality parts to close tolerance at low cost. Powder metallurgy processing encompasses an extensive range of ferrous and nonferrous alloy powders, ceramic powders, and mixes of metallic and ceramic powders (composite powders). *See* METALLURGY.

Metal powders. Regardless of the processing route, all powder metallurgy methods of part fabrication start with the raw material in the form of a powder. A powder is a finely divided solid, smaller than about 1 mm (0.04 in.) in its maximum dimension. There are four major methods used to produce metal powders, involving mechanical comminution, chemical reactions, electrolytic deposition, and liquid-metal atomization. The method selected is dictated primarily by composition, intended application, and cost. Typically, metal powders for commercial usage range from 1 to 1200 micrometers. Depending on the method of production, metal powders exhibit a diversity of shapes ranging from spherical to acicular. Commercial atomized powders are shown in Fig. 1. Particle shape is an important property, since it influences the surface area of the powder, its permeability and flow, and its density after compaction. Chemical composition and purity also affect the compaction behavior of powders. For most applications, powder purity is higher than 99.5%.

Processes. Powder metallurgy processes include pressing and sintering, powder injection molding, and full-density processing.

Pressing and sintering. The basic powder metallurgy process is two-step involving powder compaction followed by sintering. First the rigid steel die (mold) is filled with powder. Pressure is then applied uniaxially at room temperature via steel punches located above and below the powder, after which the compact is ejected from the mold (**Fig. 2**). Commercial compaction pressures are normally in the range from 140 to 900 megapascals ($2-13 \times 10^4$ lb/in.²). The powder compact is porous. Its density depends on compaction pressure and the resistance of the powder particles to deformation; compact densities about 90% of the theoretical level are common.

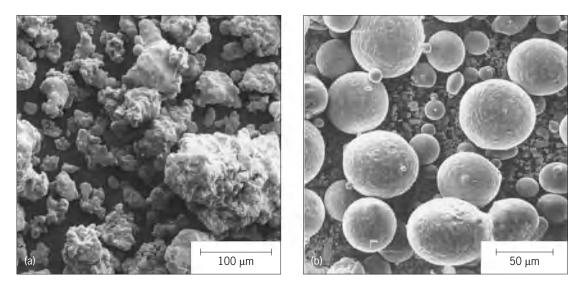


Fig. 1. Typical commercial iron-based powders: (a) water-atomized, (b) gas-atomized.

In the sintering process, the powder compact is heated below its melting point in order to promote bonding of the solid powder particles. The major purpose of sintering is to develop strength in the compact; normally an increase in the density of the compact also occurs, but the compact still contains pores. The internal architecture (microstructure) of the material is developed during sintering. As a rule of thumb, the sintering temperature must be higher than one-half the melting temperature (Kelvin scale) of the powder. This condition results in atomic diffusion and neck formation between the powder particles in the solid state. Commercial sintering furnaces are designed to provide controlled atmospheres during sintering. The atmosphere may be chemically neutral, reducing, or oxidizing with reference to the powder compact.

Normally, parts made by pressing and sintering require no further treatment. However, properties, tolerances, and surface finish can be enhanced by secondary operations. Examples of secondary or finishing operations are repressing, resintering, machining, heat treatment, and various surface treatments such as deburring, plating, and sealing. Warm compaction is a new process for enhancing powder density, whereby the powder is treated with an organic binder and then compacted at 100–150°C (212–302°F). *See* SINTERING.

Powder injection molding. This process builds on established injection molding technology used to fabricate plastics into complex shapes at low cost. The metal powder is first mixed with a binder consisting of waxes, polymers, oils, lubricants, and surfactants. After granulation, the powder-binder mix is injection-molded into the shape desired. Then the binder is removed and the remaining metal powder skeleton sintered. Secondary operations may be performed on the sintered part, similar to those for conventional press-plus-sinter parts. The viscosity of the powder-binder mix should be below 100 Pa \cdot s (1000 poise) with about 40% by volume of binder, and spherical powders less than 20 μ m in diam-

eter are preferred. Powder injection molding produces parts which have the shape and precision of injection-molded plastics but which exhibit superior mechanical properties such as strength, toughness, and ductility. *See* PLASTICS PROCESSING.

Full-density powder processing. Parts fabricated by pressing and sintering are used in many applications. However, their performance is limited because of the presence of porosity. In order to increase properties and performance and to better compete with products manufactured by other metalworking methods (such as casting and forging), several powder metallurgy techniques have been developed that result in fully dense materials; that is, all porosity is eliminated. Examples of full-density processing are hot isostatic pressing, powder forging, and spray forming.

In the hot isostatic pressing process, the metal (or ceramic) powder is sealed in a metallic or glass container, which is then subjected to isostatic pressure at elevated temperature. Temperatures up to 2200° C (4000° F) and pressures up to 250 MPa (3.6×10^{4} lb/in.²) are possible in modern presses. After

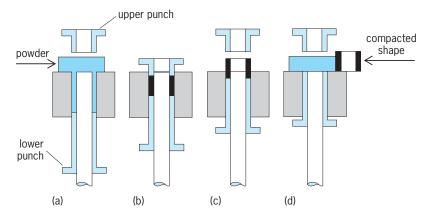


Fig. 2. Uniaxial powder compaction. (a) Die cavity is filled with powder. (b) Upper and lower punches simultaneously compact the powder in the die. (c) Upper punch is withdrawn and the powder compact is ejected by upward movement of the lower punch. (d) Compact is removed in order to repeat operating cycle.

complete densification of the powder, the compact is removed and the container stripped from the densified compact. Hot isostatic pressing is applied to the consolidation of nickel and titanium alloys, tool steels, and composites.

Powder forging is an adaptation of conventional forging in which a compact is prepared that is then sintered to retain about 20% by volume of porosity. This preform is transferred to a forging press and formed in a closed die cavity to its final shape with one stroke of the press. All porosity is removed in the forging operation. Powder forging is used primarily to fabricate components from low-alloy steels. *See* FORGING.

In the spray forming process, a stream of liquid metal is gas-atomized into a spray of droplets that impinge on a substrate. The spraying conditions are such that the droplets arrive at the substrate in the semisolid state. By control of the geometry and motion of the substrate, it is possible to fabricate sheet or plate, tubes, and circular billets from a wide range of ferrous and nonferrous alloys. It is a rapid process with a deposition rate up to about 2 kg (4 lb) per second.

Characteristics and applications. Powder metallurgy competes with several more conventional metalworking methods in the fabrication of parts, including casting, machining, and stamping. Characteristic advantages of powder metallurgy are close tolerances, low cost, net shaping, high production rates, and controlled properties. Other attractive features include compositional flexibility, low tooling costs, available shape complexity, and a relatively small number of steps in most powder metallurgy production operations. Control of the level of porosity is the intrinsic feature of powder metallurgy that enables the parts producer to predict and specify physical and mechanical properties during fabrication. Broad areas of usage for powder metallurgy parts include structural (load-bearing) components; and controlled-porosity, electrical, magnetic, thermal, friction, corrosion-resistant, and wear-resistant applications. Industries that make extensive use of powder metallurgy parts are aerospace, agriculture, automotive, biomedical, chemical processing, and electrical. In addition, the fabricators of domestic appliances and office equipment are dependent on the availability of a wide range of sizes and geometries of powder metallurgy parts exhibiting unique combinations of physical and mechanical properties.

Ferrous parts fabricated by pressing and sintering make up the largest segment of the powder metallurgy industry, and a majority of the applications are for the automotive market. In 1999 the typical United States passenger car contained more than 16 kg (35.2 lb) of powder metallurgy parts. Representative powder metallurgy parts include gears, bearings, rod guides, pistons, fuel filters, and valve plates. In the nonautomotive sector, examples of parts include gears, levers, and cams in lawn tractors and garden appliances; and gears, bearings, and sprockets in office equipment.

Press-plus-sinter fabrication techniques are used to create steel main bearing caps for 3.1- and 3.8-liter



Fig. 3. Powder metallurgy main bearing end caps made by pressing and sintering. (*Metal Powder Industries Federation*)



Fig. 4. Automobile connecting rods made by powder forging. (Metal Powder Industries Federation)

V-6 engines (**Fig. 3**). The 3.1-liter engine contains three caps weighing about 2.2 kg (4.8 lb), and the 3.8-liter engine contains four caps weighing slightly more than 3.6 kg (7.9 lb). With a density of 6.6 g/cm³ (3.8 oz/in.³), the caps have a tensile strength of 448 N/mm² (64,000 lb/in.²), an apparent hardness of 70 units using the Rockwell method on the B scale (HRB), an elongation of 3%, a fatigue endurance limit of 159 N/mm² (23,000 lb/in.²), and a Young's modulus of 107 kN/mm² (15.5 × 10⁶ lb/in.²). A recent development is the aluminum cam bearing cap.

Full-density processing has been used in the powder forging of steel automobile connecting rods (**Fig. 4**). The iron-copper-carbon rod is forged to a minimum density of 7.84 g/cm³ (4.54 oz/in.³), and has a tensile strength of 758 N/mm² (110,000 lb/in.²), a yield strength of 551 N/mm² (79,900 lb/in.²), and an endurance limit of 275 N/mm² (39,900 lb/in.²). Important characteristics of powder forging include straightness, elimination of surface defects inherent in conventional forging, uniform microstructure, dimensional control, and superior machinability. Secondary operations include deflashing, shot peening, and machining.

An example of a part manufactured by powder injection molding is the miniature read/write latch arm made from stainless steel powder and used in a small high-capacity hard-disk drive (**Fig. 5**). The geometric form of the entire outside profile of the part is very critical relative to the pivot hole. Coining

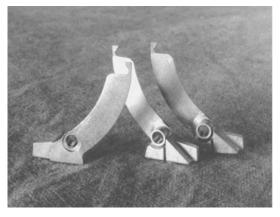


Fig. 5. Metal-powder injection-molded latch arms used in a hard-disk drive. (*Metal Powder Industries Federation*)

is the only secondary operation. See COINING.

Safety and health considerations. Because metal powders possess a high surface area per unit mass, they can be thermally unstable in the presence of oxygen. Very fine metal powders can burn in air (pyrophoricity) and are potentially explosive. Therefore, the clean handling of powder is essential; methods include venting, controlled oxidation to passivate particle surfaces, surface coating, and minimization of sparks or heat sources. Some respirable fine powders pose a health concern and can cause disease or lung dysfunction: the smaller the particle size, the greater the potential health hazard. Control is exercised by the use of protective equipment and safe handling systems such as glove boxes. There is no recognized hazard associated with the normal handling of the common grades of metal and alloy powders such as copper and iron. See INDUSTRIAL HEALTH AND SAFETY. Alan Lawley

Bibliography. ASM Handbook: Metal Powder Technologies and Applications, vol. 7, 1998; R. German, A-Z of Powder Metallurgy, 2006; L. F. Pease III and W. G. West, Fundamentals of Powder Metallurgy, 2002; Powder Metallurgy Design Manual, 3d ed., 1998.

Power

The time rate of doing work. Like work, power is a scalar quantity, that is, a quantity which has magnitude but no direction. Some units often used for the measurement of power are the watt (1 joule of work per second) and the horsepower (550 ft-lb of work per second).

Usefulness of the concept. Power is a concept which can be used to describe the operation of any system or device in which a flow of energy occurs. In many problems of apparatus design, the power, rather than the total work to be done, determines the size of the component used. Any device can do a large amount of work by performing for a long time at a low rate of power, that is, by doing work slowly. However, if a large amount of work must be done rapidly, a high-power device is needed. High-power machines are usually larger, more complicated, and more expensive than equipment which need operate only at low power. A motor which must lift a certain weight will have to be larger and more powerful if it lifts the weight rapidly than if it raises it slowly. An electrical resistor must be large in size if it is to convert electrical energy into heat at a high rate without being damaged.

Electrical power. The power *P* developed in a direct-current electric circuit P = VI, where *V* is the applied potential difference and *I* is the current. The power is given in watts if *V* is in volts and *I* in amperes. In an alternating-current circuit, $P = VI \cos \phi$, where *V* and *I* are the effective values of the voltage and current and ϕ is the phase angle between the current and the voltage. *See* ALTERNATING CURRENT.

Power in mechanics. Consider a force F which does work W on a particle. Let the motion be restricted to one dimension, with the displacement in this dimension given by x. Then by definition the power at time t will be given by Eq. (1). In this equation W can be

$$P = \frac{dW}{dt} \tag{1}$$

considered as a function of either t or x. Treating W as a function of x gives Eq. (2). Now dx/dt represents

$$P = \frac{dW}{dt} = \frac{dW}{dx}\frac{dx}{dt}$$
(2)

the velocity v of the particle, and dW/dx is equal to the force F, according to the definition of work. Thus Eq. (3) holds. This often convenient expression for

ŀ

$$P = Fv \tag{3}$$

power can be generalized to three-dimensional motion. In this case, if ϕ is the angle between the force **F** and the velocity **v**, which have magnitudes *F* and *v*, respectively, Eq. (4) expresses quantitatively the ob-

$$P = \mathbf{F} \cdot \mathbf{v} = Fv \cos \phi \tag{4}$$

servation that if a machine is to be powerful, it must run fast, exert a large force, or do both. *See* WORK. Paul W. Schmidt

Power amplifier

The final stage in multistage amplifiers, such as audio amplifiers and radio transmitters, designed to deliver appreciable power to the load. *See* AUDIO AMPLIFIER; RADIO TRANSMITTER.

Power amplifiers may be called upon to supply power ranging from a few watts in an audio amplifier to many thousands of watts in a radio transmitter. In audio amplifiers the load is usually the dynamic impedance presented to the amplifier by a loudspeaker, and the problem is to maximize the power delivered to the load over a wide range of frequencies. The power amplifier in a radio transmitter operates over a relatively narrow band of frequencies with the load essentially a constant impedance.

The usual mode of operation of power amplifiers is denoted by class A, AB, B, and C. Class C operation is limited to radio frequencies with a tuned load. The other classes may be used for audio and high-frequency operation. For discussion of the modes of operation *see* AMPLIFIER

Class A. Class A operation is used when the amount of power transferred to the load is relatively small, say, less than 2 W. The amount of harmonic distortion introduced into the load voltage can be kept small by operating transistors within the nearly linear region of their characteristics. Class A operation is limited to low-power applications, such as small radios, cassette players, and compact-disk (CD) players, because the conversion efficiency (the efficiency of a power amplifier) is low. The maximum possible efficiency is 50%. However, for the usual operating conditions the efficiency is on the order of 10%. If the power amplifier were required to deliver 10 W with 10% efficiency, the transistor would have to be capable of dissipating an average power of 100 W. Furthermore, the power supply must be capable of supplying the power dissipated as heat plus the useful power delivered to the load. This poses an unnecessary burden upon the power supply. Other classes of operation have a higher conversion efficiency and are therefore used for higher-power applications.

Class AB. An improvement in the conversion efficiency can be obtained by using class AB operation. However, while a class A amplifier can be operated single-ended (one output transistor), a class AB amplifier must be operated push-pull. In class AB operation the transistor current does not flow for the complete cycle of the input voltage. In a single-ended circuit this would introduce excessive distortion. *See* PUSH-PULL AMPLIFIER.

Class B. This class is often used for the power amplifier in an audio amplifier. The amplifier in this class must be a push-pull circuit. Theoretically, with ideal transistors, the class B amplifier can have a conversion efficiency of 78.5%; practically, the efficiency is on the order of 50%, an appreciable improvement over that of class A operation.

Audio applications constitute a significant percentage of the applications of the class B amplifier. The requirements may range from a power output of a few watts for a small system, to 100 watts per channel for a home stereo system, to much more for a music concert sound system. The design of the loudspeaker loading is challenging, since the loading consists of three speakers (the woofer, midrange, and tweeter) operating over different frequency ranges. The rated loading is typically 8 Ω . The load may be transformer coupled to the output transistors, but it may also be connected directly in a complementary symmetry circuit. Since the operation for peak power output is over a wide range of the transistor characteristics, the design has to be done carefully to reduce the total harmonic distortion (THD) that can be generated by the nonlinear transistors. Typical figures for a high-quality home stereo system are 0.09% total harmonic distortion or less. If a load other than the rated value is used (say, a $4-\Omega$ speaker system), the operating region of the output transistors is shifted, and increased total harmonic distortion may result. Operation at maximum power output with a load less than rated may also cause other damaging effects. *See* DISTORTION (ELECTRONIC CIRCUITS).

The amplifier bandwidth is another important parameter of an audio amplifier system. Typical figures for a home stereo system are 40 Hz for the lower half-power frequency to 20,000 Hz for the upper half-power frequency. At these limits, the power output is half the output at midrange for the same input signal. The frequency response characteristics of the speakers have to be considered in the overall picture (woofers for low frequencies and tweeters for high frequencies). High-quality systems give the user the ability to adjust with equalizers the system frequency response for personal preferences. *See* EQUALIZER; RESPONSE.

Class C. Because the collector current flows for less than one-half cycle of the input sinusoidal signal, this class of operation is restricted to radio-frequency operation where a tuned load is employed. The load is usually the input impedance of an antenna or of an antenna matching network. The load voltage will be nearly sinusoidal, even though the current flows in pulses, because of the relatively sharp tuning of the load. This phenomenon allows the amplification of large amounts of power at conversion efficiencies as high as 80%. This is extremely important for applications requiring delivery of large amounts of power to the load.

The driving source must usually be called upon to deliver power to the base circuit of the power amplifier, in many cases as much as 10% of the power delivered to the load. This requirement is not excessive. A class B power amplifier can be used in the driving circuit to obtain an efficient combination of driver and final amplifiers. Harold F. Klock

Bibliography. G. Ballou (ed.), *Handbook for Sound Engineers*, 2d ed., 1991; P. Chirlian, *Analysis and Design of Integrated Electronic Circuits*, 2d ed., 1987; A. Sedra and K. Smith, *Microelectronic Cir cuits*, 4th ed., 1997; R. Smith, *Electronics: Circuits and Devices*, 3d ed., 1987.

Power factor

In sinusoidal alternating-current theory, the cosine of the phase angle between the voltage across and the current through electrical circuitry. It equals the ratio of the mean power dissipated to the product of the root-mean-square voltage and current, and is unity for a circuit containing only pure resistances. It is less than unity if inductances and capacitances are involved. *See* ALTERNATING-CURRENT CIRCUIT THEORY. Bryan P. Kibble

Power-factor meter

An instrument used to indicate whether load currents and voltages are in time-phase with one another. In an alternating-current circuit, the current *I* may lead or lag the voltage *E* by some angle ϕ , and the average power *P* would be the average of either $(E_{\text{max}} \cos \omega t) [I_{\text{max}} \cos (\omega t \pm \phi)]$ or $P = EI \cos \phi$. The factor $\cos \phi = P/EI$ is called the power factor of the load. *See* ALTERNATING-CURRENT CIRCUIT THEORY.

The formula $P = EI \cos \phi$ shows that, if two customers use equal currents, the energy loss in the line is the same for the two. If one of them has a lower power factor, that customer uses less energy, and a rate based only on energy delivered would involve an unequal charge for service.

Single-phase meter. This instrument contains a fixed coil that carries the load current, and crossed coils that are connected to the load voltage. There is no spring to restrain the moving system, which takes a position to indicate the angle between the current and voltage. The scale can be marked in degrees or in power factor.

The angle between the currents in the crossed coils is a function of frequency, and consequently each power-factor meter is designed for a single frequency and will be in error at all other frequencies. For low harmonic content the indication of the power-factor meter is usually accepted for contract purposes.

Polyphase meters. These meters are usually designed differently from single-phase meters. For balanced polyphase loads their application is straightforward. For a quarter-phase load, two voltages 90° out of phase are present and each can supply one of the crossed coils of the power-factor meter. If the voltages are unequal or their relative phase angle is different from 90° , the indication will no longer have meaning.

Three-phase power-factor meters connect the fixed coil in one line and the crossed coils between the other two lines. The meter is constructed with a 60° angle between the crossed coils. An unbalance in either voltage or angle will cause erroneous readings.

In the case of balanced four-phase and balanced six-phase loads, single-phase power-factor meters can be used directly without any correction.

In the case of unbalanced polyphase circuits with harmonics, various portions of the load may have different power factors and for the combined load there is no common phase angle; consequently the statement that the power factor is equal to $\cos \phi$ becomes meaningless. *See* NONSINUSOIDAL WAVEFORM. Harry Sohon; Edward C. Stevenson

Bibliography. S. D. Prensky and R. Castellucis, *Electronic Instrumentation*, 3d ed., 1982; M. B. Stout, *Basic Electrical Measurements*, 2d ed., 1960.

Power integrated circuits

Integrated circuits that are capable of driving a power load. The key feature of a power integrated circuit that differentiates it from other semiconductor technologies is its ability to handle high voltage, high current, or a combination of both.

In its simplest form, a power integrated circuit may consist of a level-shifting and drive circuit that translates logic-level input signals from a microprocessor to a voltage and current level sufficient to energize a load. For example, such a chip may be used to operate electronic displays, where the load is usually capacitive in nature but requires drive voltages above 100 V, which is much greater than the operating voltage of digital logic circuits (typically 5 V). At the other extreme, the power integrated circuit may be required to perform load monitoring, diagnostic functions, self-protection, and information feedback to the microprocessor, in addition to handling large amounts of power to actuate the load. An example of this is an automotive multiplexed bus system with distributed power integrated circuits for control of lights, motors, air conditioning, and so forth. See AUTOMOTIVE ELECTRICAL SYSTEM; ELEC-TRONIC DISPLAY; MICROPROCESSOR.

Technology. In addition to high-voltage and highcurrent operation, power integrated circuits are often also capable of performing many of the tasks associated with analog, digital, and sensor chip technologies.

Sensors. Due to the need to carry out a variety of diagnostic functions, power integrated circuits usually incorporate the current-, voltage-, and temperaturesensing functions that are normally performed by analog circuits. The sensor elements are placed on the power integrated circuit to provide local feedback at the point of use. Power modules can be made immune to short circuits or to misapplication of loads because the sensing function is performed locally. This enables local power interruption before the current in the system can rise to destructive levels.

Digital logic. Digital logic circuits incorporated on the power integrated circuit enable it to interface with microprocessors. The logic is implemented with complementary metal-oxide-semiconductor (CMOS) technology. Because of the high temperatures and rapid increase of voltage (voltage transients) encountered in typical applications for power integrated circuits, special measures are required during design and processing to make this CMOS logic immune from latch-up. The logic circuits may also need to be placed in special electically isolated silicon regions that can be floated up to high voltages. *See* LOGIC CIRCUITS.

Analog control. The control loop between the sensor and the protection circuit is usually implemented with bipolar analog circuits. Thus, bipolar and CMOS transistors must be fabricated simultaneously (BI-CMOS technology) during the manufacture of power integrated circuits. In some applications the bipolar transistors in the analog loop must have a very high frequency response to enable rapid implementation of control strategies for power regulation and integrated circuit protection.

High-voltage operation. Since the power integrated circuit usually operates at high voltages, unique process considerations become necessary. The design of

	Power rec	quirements	Operating frequency*	Circuit configuration
Application	Voltage, V	Current, A		
Display drives	100-400	0.01-0.1	High	dc
Computer power supplies	50	10–50	High	dc
Motor drives	300-600	10–50	Medium	dc
Factory automation	300-600	10–50	Low	ac
Telecommunications	100-600	1–10	High	ac
Appliance control	300-600	1–20	Low	ac
Consumer electronics	50	0.1-10	Medium	dc
Lighting	300-600	1–10	Medium	ac
Smart house	300-600	5–50	Low	ac
Aircraft electronics	100-600	10–50	Medium	dc
Automotive electronics	50-100	1–20	Low	dc

high-voltage transistors requires proper termination to achieve large blocking voltage capability. The need to bus high voltages around the chip requires special metal crossover design methodology and unique processes.

High-current operation. The fabrication of high-current devices on the same chip with high-density digital logic circuits requires the use of a double metal architecture in addition to the polysilicon gate technology. Further, if a high switching speed for the power device is desirable, the application of a refractory metal-silicide process becomes mandatory to obtain uniform turn-on and turn-off in the largearea power devices. Without the refractory gate process, the power dissipation in the power devices due to nonuniform switching can become prohibitive, leading to problems with reliability.

Dielectric isolation. The incorporation of multiple high-current bipolar devices on a monolithic chip becomes impossible with junction isolation technology because of the interaction between the adjacent devices. This problem can be solved by utilizing dielectric isolation technology. As the cost of dielectrically isolated silicon wafers is reduced by the development of improved techniques of wafer preparation, it is expected that most power integrated circuits will be fabricated with this starting material because it can simplify circuit design by eliminating circuit parasitics.

Requirements. The combination of digital logic circuits, analog sensing functions, high-voltage capability, and high-current devices must operate under harsh environmental conditions such as elevated ambient temperatures and high-voltage transients. All of these criteria must be satisfied under severe cost constraints. This demands continual advancements in device and process technology.

In most power integrated circuits, the power devices occupy a major portion of the chip area. Consequently, advances in the power-device structures, designs, and processes will dictate the pace of improvements in this technology.

Applications. Power integrated circuits are expected to have an impact on all areas in which power

semiconductor devices are presently being used. In addition, they are expected to open up new applications based upon their added features. The wide spectrum of voltages and currents over which power semiconductor devices are utilized are summarized in the **table**, where the power requirements, system configuration, and operating frequency for the various applications are listed. At one extreme are display drives that require relatively low currents and moderate voltages. These applications are already served by smart power chips. At the other extreme are chips, under development for automotive electronics, that demand very high levels of current. *See* INTEGRATED CIRCUITS. B. Jayant Baliga

Bibliography. P. Antognetti, *Power Integrated Circuits*, 1986; B. J. Baliga, *Higb Voltage Integrated Circuits*, 1988; B. J. Baliga, *Power Semiconductor Devices*, 1996.

Power line communication

A technology that uses electrical wiring as a medium to transmit digital signals from one device that has been rendered capable of employing this technology to another. Power line communication (PLC) enables the communication of data as well as multimedia messages by superimposing digital signals over the standard 50- or 60-Hz alternating current (AC). Low-speed power line communication has been used widely for some time in electrical utilities to provide voice communication, to control substations, and to protect high-voltage transmission lines. High-speed data transmission over power lines began in the late 1990s and was targeted to provide broadband Internet services and utility applications such as remote meter reading, as well as in home networking for data and multimedia communications. A short-range form of power line communication is also used for home automation and controls. See DATA COMMUNICATIONS; ELECTRIC POWER SYS-TEMS; ELECTRIC POWER TRANSMISSION; INTERNET.

While high-speed power line communication indeed is used in developed countries to provide such services, it is of greater interest in developing countries because of the lack of alternative broadband infrastructure such as cable television (CATV) or telephone lines [for digital subscriber lines (DSL)]. Using the existing electrical power distribution infrastructure, power line communication can provide areawide and even nationwide digital communication without the need to lay new cabling. The benefits are obvious, and thus many countries are testing and deploying power line communication technology, despite concerns that power line communication can introduce electromagnetic interference (EMI). Many countries are beginning to define regulations addressing spectrum usage and coexistence of power line communication with other existing and anticipated applications. See ELECTRICAL INTERFERENCE.

In-home power line communication provides another networking option in addition to the popular wireless communication technology and other wired technologies like Ethernet, Home PNA (Phone Networking Alliance), and fiber optics. The major benefit of in-home power line communication is that people are able to use existing power plugs as access points for computer communications and other applications like surveillance video and home automation.

Background. Traditional power circuits were designed for distributing electrical power to and within buildings. The existence of a plethora of sources of interference from nearby electrical devices, motors, and lights makes it a challenge to use the harsh environment of existing power lines as a high-frequency communication channel. The channel characteristics, such as noise, impedance, and attenuation of power lines, are found to be highly unpredictable and variable with time, frequency, and location. Fortunately, the rapid advances in digital communications, signal processing, and error-control coding enable quite successful power line communication systems.

Power line communication devices operate by injecting a carrier wave into the electrical wires at the transmitter. The carrier is modulated by digital signals. Each receiver in the system has a unique address and can receive and decode the signals transmitted over the wires. These devices may be either plugged into regular power outlets or permanently wired in place.

However, the harsh environment requires that power line communication devices use advanced signal modulation, digital signal processing, and error detection and correction techniques. These technologies convert the digital 0 and 1 into a useful carrier waveform that becomes the underlying physicallayer signals. The frequency band used in power line communication depends on the developer and technology used. Typically, the wider the frequency band, the higher the raw bit rate. For example, the X10 home automation system uses power line communication at the zero crossing voltage points in the AC wave (the two instants in each cycle when the AC voltage vanishes), and thus has a frequency of 120 Hz (for 60-Hz systems) and a relatively low data rate, good only for sending short bursts of data for controlling devices in the home. However, current (2006) high-speed power line communication devices use frequency bands in the range from 1 to 30 MHz and are able to provide raw data rates up to 200 megabits per second (Mbps). Power line communication chips operating at a raw data rate of 1 gigbit per second (Gbps) are contemplated, using higher bandwidths and more sophisticated communication techniques.

To provide simultaneous communication between power line communication devices, rules or protocols must be adopted to enable the effective use of the PLC channel for delivering digital data. These rules regulate how long and when a device can use the channel, and define the medium access (MAC) protocols of power line communication. Typical MAC protocols used in power line communication devices are carrier sense multiple access/collision avoidance (CSMA/CA), time-division multiple access (TDMA), and frequency-division multiple access (FDMA). While several proprietary systems (often backed by industry consortia) are available, there is an ongoing effort in international standards organizations to propose and adopt uniform standards for coexistence and interoperability of different power line communication products, for both in-home and Internet access domains.

Types of technology. Power line communication technologies may be distinguished by their voltage levels.

Low-voltage power line communication. Low-voltage power line communication devices usually are used in home networks. The benefit of using power lines in-home network infrastructures is the availability of power outlets almost everywhere in a home. Unlike wireless networks, power line communications do not require a specific access point since power outlets are the access points. Since the power line communication channel is essentially a broadcast channel, no switches are required in the communication; all devices communicate directly with each other. Because of these properties, to set up an in-home network based on power line communication requires no new wires, no drilling of walls, and no rewiring-the procedure is literally plug-and-play.

Low-voltage power line communication may also be used for temporary local-area network (LAN) setups for conferences, temporary medical centers, and other situations that require simple and fast setups. To set up a power line communication-capable computer, a user has only to find a power outlet and plug the device to it using either an Ethernet cable or a USB (Universal Serial Bus) connection, and then the computer is ready for communication. While most Ethernet or USB power line communication adapters are still separate from computer power supplies, manufacturers also embed the power line communication capabilities directly into device power

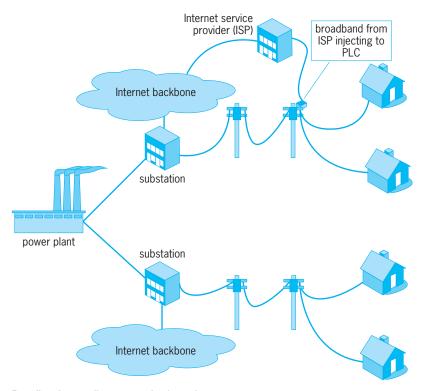
Comparison of throughputs of power line and wireless communication technologies				
Technology	Power line	Wireless: 802.11n	Wireless: 802.11g	
Theoretical maximum throughput	First-generation: 14 Mbps Turbo: 85 Mbps HomePlug AV/UPA: 200 Mbps	540 Mbps	54 Mbps	
Actual expected throughput	First-generation: 10 Mbps Turbo: 40 Mbps HomePlug AV/UPA: 100 Mbps	150–200 Mbps	15–20 Mbps	

supplies, so that the wires that provide electrical power to the device are also used to connect the power line communication system.

While low-voltage power line communication is also used for "last mile" Internet access, mediumand high-voltage power line communication systems are used to provide power line communication over much larger distances.

Medium- and high-voltage power line communication. Medium- and high-voltage power line communications usually are used for broadband over power line (BPL) or home electricity usage metering. Mediumand high-voltage power line communication systems can often cover longer distances with less attenuation and higher data rates because of the channel conditions and the marked absence of interfering sources common in low-voltage and in-home power line communication channels.

BPL provides broadband Internet access through ordinary power lines. It offers obvious benefits over regular cable or DSL connections; the extensive infrastructure already available should allow more peo-



Broadband power line communication to homes.

ple in more locations to have access to the Internet. The BPL Internet service provider (ISP) usually uses a coupler near a house to superimpose broadband signals into a low-voltage electric power line, delivering high-speed broadband capabilities to the end user. BPL can be used not only for computer communication, but also for voice over the Internet (VoIP) services and streaming video services such as Internet Protocol Television (IPTV). *See* VOICE OVER IP.

Comparisons between technologies. Power line communication may be compared with alternative technologies.

Wireless and power line communication. Wireless communication and power line communication both make it possible to set up a local-area network with no new wires. The advantage of using one or the other technology is not as considerable as might be supposed. The two technologies are comparable in cost, and their throughputs are compared in the **table**. *See* WIRELESS FIDELITY (WIFI).

While power line communication may in some cases provide an alternative for wireless networking, or vice versa, the two technologies also can be used cooperatively. While power line communication has the advantage of using any power socket for power and communication, it is not operable at some locations that lack power sockets. On the other hand, while wireless certainly has the mobility advantage, it has been shown that wireless signals are not able to penetrate multiple concrete walls, and mobile devices eventually must connect to a power socket to recharge batteries. By combining power line networking technology with wireless access points, it is possible to reduce the "dead zone" problems and also provide another broadband access option to users

BPL, DSL, and cable modem. Supporting broadband Internet access for remote areas has been a problem for a long time. Existing DSL and digital cable (CATV) providers are hesitant to set up infrastructure for these areas for fiscal reasons, and wireless access technology like satellite is expensive. The emerging WiMax technologies seem promising but are still questionable, especially in mountainous areas. Broadband over power line provides a viable solution in solving this "last mile" problem (see **illustration**).

Prospects. In-home power line communication is promising as it provides a convenient way for setting up a local-area network with good throughput and coverage. BPL for high-speed Internet access is emerging as a viable option. However, BPL has several unsolved problems. Radio-frequency users reported severe interference during field tests of BPL equipment by some manufacturers. Regulations are being crafted to address these issues. In addition, international standards urgently need to be adopted to avoid incompatibilities in terms of both coexistence and interoperability for equipment from different manufacturers and for both in-home and access applications. The future of power line communications hinges on the ability of the power line communication community to resolve effectively these issues in the short term.

Haniph A. Latchman; Yu-Ju Lin; Sunguk Lee Bibliography. H. A. Latchman and L. W. Yonge, Power line local area networking, IEEE Commun. Mag., 41(4):32-33, April 2003; Y. Lin et al., A power line communication network infrastructure for the smart home, IEEE Wireless Commun., 9(6):104-111, December 2002; Y. Lin et al., A comparative performance study of wireless and power line networks, IEEE Commun. Mag., 41(4):54-63, April 2003; A. Majumder and J. Caffery, Jr., Power line communications, IEEE Potentials, 23(4):4-8, October-November 2004; D. Ruiz et al., In-home AV PLC MAC with neighboring networks support, pp. 17-21 in 2005 International Symposium on Power Line Communications and Its Applications, April 6-8, 2005.

Power plant

A means for converting stored energy into work. Stationary power plants such as electric generating stations are located near sources of stored energy,

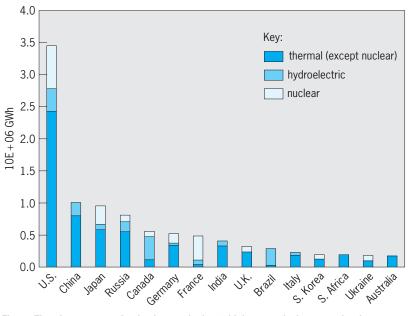


Fig. 1. Electric energy production by type in the 15 highest-producing countries, in gigawatt-hours. Nearly 27% of world electrical energy is generated in United States plants. (U.S. DOE Publ. DOE/EIA-0219 (97), April 1999)

such as coal fields or river dams, or are located near the places where the work is to be performed, as in cities or industrial sites. Mobile power plants for transportation service are located in vehicles, as the gasoline engines in automobiles and diesel locomotives for railroads. Power plants range in capacity from a fraction of a horsepower (hp) to over 10^6 kW in a single unit (**Table 1**). Large power plants are assembled, erected, and constructed on location from

Туре	Unit size range, kW [†]	Fuel [*]	Plant weight, lb/kW [†]	Plant volume, ft ³ /kW [‡]	Heat rate, Btu/kWh [§]
Central station					
Hydro	10,000-700,000				
Steam (fossil-fuel-fired)	10,000-1,300,000	COG		20-50	8,200-15,00
Steam (nuclear)	500,000-1,300,000	N			10,000-12,00
Diesel	1,000-5,000	DG			10,000-15,00
Combustion turbine, simple cycle	2,000-250,000	D'G			8,600-15,00
Combustion turbine, simple cycle with recuperator	4,000-5,000	D'G			8,300–9,500
Combustion turbine, combined cycle	7,500-800,000	GD′			5,700-9,000
Industrial (by-product) steam	1,000-25,000	COGW		50-75	4,500-6,000
Diesel locomotive	1,000-5,000	D	100-200	2-3	10,000-15,0
Automobile	25-300	G′	5-10	0.1	15,000-20,0
Outboard motor	1-50	G′	2-5	0.1-0.5	15,000-20,0
Truck	50-500	D	10-20		12,000-18,0
Merchant ship, diesel	5,000-20,000	D	300-500		10,000-12,0
Naval vessel, steam	25,000-100,000	DON	25-50		12,000-18,0
Airplane, reciprocating engine	1,000-3,000	G	1–3	0.05-0.10	12,000-15,0
Airplane, turbojet	3,000-10,000	D′	0.2-1		13,000-18,0

C, coal; D, diesel fuel; D, distillate; G, gas; G, gasoline; N, nuclear; O, fuel oil (residuum); W, w

 $^{+1}$ 1 lb/kW = 0.45 kg/kW. *1 ft³/kW = 2.83 × 10⁻² m³/kW.

 $^{\$}1$ Btu/kWh = 1.055 kJ/kWh = 2.93 × 10⁻⁴ J(heat)/J(output).

equipment and systems made by different manufacturers. Smaller units are produced in manufacturing facilities.

Most power plants convert part of the stored raw energy of fossil fuels into kinetic energy of a spinning shaft. Some power plants harness nuclear energy. Elevated water supply or run-of-the-river energy is used in hydroelectric power plants. For transportation, the plant may produce a propulsive jet, as in some aircraft, instead of the rotary motion of a shaft. Other sources of energy, such as fuel cells, winds, tides, waves, geothermal, ocean thermal, nuclear fusion, photovoltaics, and solar thermal, have been of negligible commercial significance in the generation of power despite their potential in the future. *See* ENERGY SOURCES.

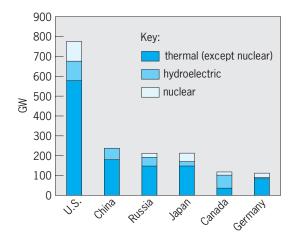


Fig. 2. The six countries with the highest capacities to generate electric power. (U.S. DOE Publ. DOE/EIA-0219 (97), April 1999)

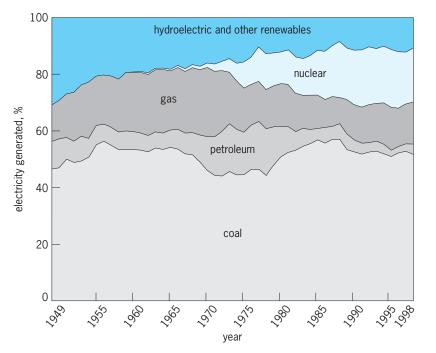


Fig. 3. Percentage of electrical energy generation in the United States utility industry (including Alaska and Hawaii since 1963), by fuel source. (U.S. DOE Publ. DOE/EIA-0384 (98), July 1999)

TABLE 2. Range of capacity factors for selected power plants

Power plants	Factor, %
Public-utility systems, in general	50–70
Industrial cogeneration, three-shift operation	80-95
Seagoing ships, long voyages	70-80
Seagoing ships, short voyages	30-40
Airplanes, commercial	20-30
Private passenger cars	1–3
Main-line locomotives	30–40
Interurban buses and trucks	5–10

If prices of conventional fuels (fossil and nuclear) rise faster than the general inflation rate, alternative energy sources become more significant. While they will not likely supply a large percentage of society's energy needs in the foreseeable future, they could someday become a significant part of the overall generating capacity (**Figs. 1–3**).

Figure 4 shows rudimentary flow or heat-balance diagrams for important types of practical power plants. **Figure 5** is a diagram for a by-product-type industrial steam plant (also known as cogeneration) which has the double purpose of generating electric power and simultaneously delivering heating steam by extraction or exhaust from the prime mover. *See* COGENERATION SYSTEMS; HEAT BALANCE.

Plant load. There is no practical way of storing the mechanical or electrical output of a power plant in the magnitudes encountered in power plant applications, although several small-scale concepts have been researched. As of now, however, the output must be generated at the instant of its use. This results in wide variations in the loads imposed upon a plant. The capacity, measured in kilowatts or horsepower, must be available when the load is imposed. Some of the capacity may be idle during extended periods when there is no demand for output. Hence a signifant portion of the potential output, measured as kilowatt-hours or horsepower-hours, may not be generated because there is no demand for output. Kilowatts cannot be traded for kilowatt-hours, and vice versa. See ENERGY STORAGE.

The ratios of average load to rated capacity or to peak load are expressed as the capacity factor and the load factor, Eqs. (1) and (2), respectively. The range

$$Capacity factor = \frac{average load for the period}{rated or installed capacity} (1)$$

$$Load factor = \frac{average load for the period}{peak load in the period}$$
(2)

of capacity factors experienced for various types of power plants is given in **Table 2**.

Variations in loads can be conveniently shown on graphical bases as in **Figs. 6** and 7 for public utilities and in **Fig. 8** for air and marine propulsion. Rigorous definition of load factor is not possible with vehicles like tractors or automobiles because of variations in the character and condition of the running surface. In propulsion applications, power output

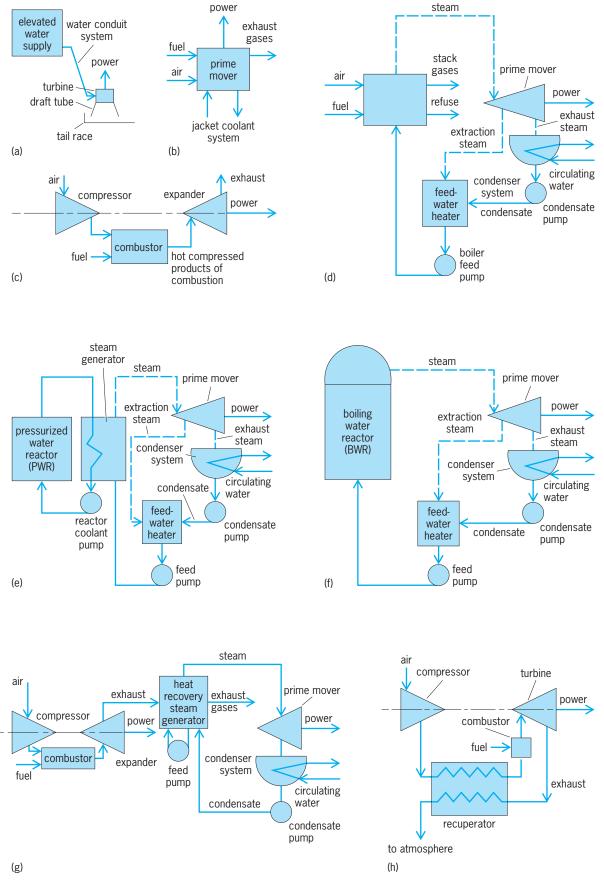


Fig. 4. Rudimentary flow or heat-balance diagrams for power plants. (a) Hydro. (b) Internal combustion. (c) Combustionturbine. (d) Fossil-fuel-fired. (e) Nuclear steam (pressurized water reactor, PWR). (f) Nuclear steam (boiling water reactor, BWR). (g) Combustion turbine combined cycle. (h) Combustion turbine with recuperator.

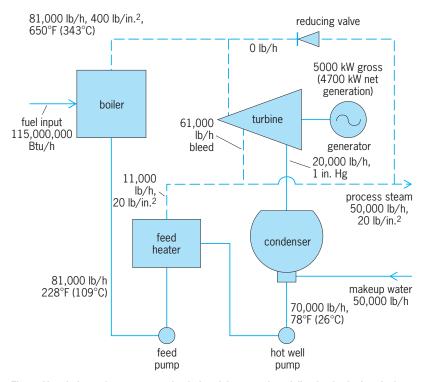


Fig. 5. Heat balance for a cogeneration industrial power plant delivering both electrical energy and process steam. 1 lb/h = 1.26 \times 10⁻⁴ kg/s; 1 Btu/h = 0.293 W; 1 lb/in.² = 6.895 kPa; 1 in. Hg = 3.386 kPa.

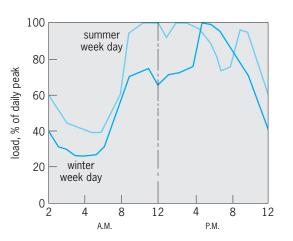


Fig. 6. Daily-load curves for urban utility plant.

may be of secondary import: performance may be based on tractive effort, drawbar pull, thrust, climb, and acceleration.

Plant efficiency. The efficiency of energy conversion is vital in most power plant installations. The theoretical power of a hydro plant in kilowatts is QH/11.8, where Q is the flow in cubic feet per second and H is the head (height of water intake above discharge level) at the site in feet. In metric units, the theoretical power of the plant in kilowatts is 9.8 Q'H', where Q' is the flow in cubic meters per second and H' is the head in meters. Losses in headworks, penstocks, turbines, draft tubes, tailrace, bearings, generators, and auxiliaries will reduce the salable output 15–20% below the theoretical in modern in-

stallations. The selection of a particular type of waterwheel depends on experience with wheels at the planned speed and on the lowest water pressure in the water path. Runners of the reaction type (high specific speed) are suited to low heads (below

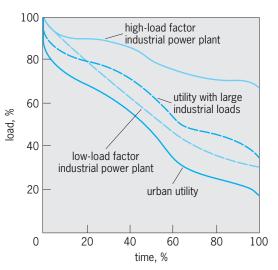


Fig. 7. Annual load-duration curves for selected stationary public utility power plants.

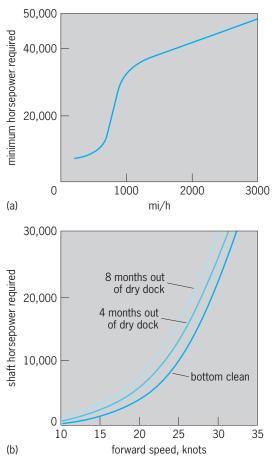


Fig. 8. Air and marine power. (a) Minimum power required to drive a 50-ton (45-metric-ton) well-designed airplane in straight level flight at 30,000 ft (9.1 km) altitude. (b) Power required to drive a ship, showing effect of fouling. 1 hp = 746 W; 1 mi/h = 0.447 m/s; 1 knot = 0.514 m/s.

500 ft or 150 m) and the impulse type (low specific speed) to high head service (about 1000 ft or 300 m). The lowest heads (below 1000 ft or 300 m) are best accommodated by reaction runners of the propeller or the adjustable blade types. Mixed-pressure runners are favored for the intermediate heads (50-500 ft or 15-150 m). Draft tubes, which permit the unit to be placed safely above flood water and without sacrifice of site head, are essential parts of reaction unit installations. *See* HYDRAULIC TURBINE;; WATERPOWER.

With thermal power plants the basic limitations of thermodynamics fix the efficiency of converting heat into work. The cyclic standards of Carnot, Rankine, Otto, Diesel, and Brayton are the usual criteria on which heat-power operations are variously judged. Performance of an assembled power plant, from fuel to net salable or usable output, may be expressed as thermal efficiency (%); fuel consumption (lb, pt, or gal per hp-h or per kWh); or heat rate (Btu supplied in fuel per hp-h or per kWh). American practice uses high or gross calorific value of the fuel for measuring heat rate or thermal efficiency and differs in this respect from European practice, which prefers the low or net calorific value.

Tables 1 and 2 give performances for selected operations. Figures 9 and 10 reflect the improve-

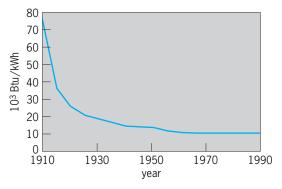


Fig. 9. Thermal performance of fuel-burning electric utility power plants in the United States. 1 Btu/kWh = 1.055 kJ/kWh = 2.93 \times 10⁻⁴ J(heat)/J(output). (*Edison Electric Institute*)

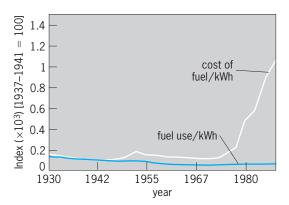


Fig. 10. Effect of efficiency of fuel use upon cost of fuel per kilowatthour generated in the United States utility industry (including Alaska and Hawaii since 1963). (*After Edison Electric Institute and Mon. Energy Rev., no. 4, December 1983*)

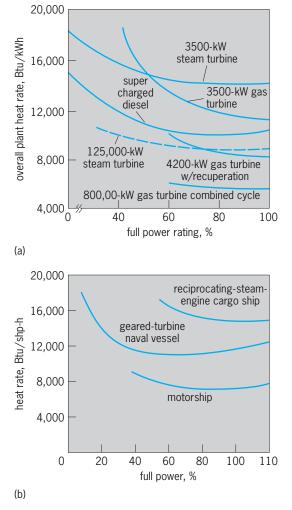


Fig. 11. Comparison of heat rates. (a) Stationary power plants. (b) Marine propulsion plants. 1 Btu/kWh = $1.055 \text{ kJ/kWh} = 2.93 \times 10^{-4} \text{ J(heat)/J(output); 1 Btu/hp-h} = 1.415 \text{ kJ/kWh} = 3.93 \times 10^{-4} \text{ J(heat)/J(output).}$

ment in fuel utilization of the United States electric power industry since 1900, although there have been some minor decreases since 1970. Figure 10 is especially significant, as it shows graphically the impact of technological improvements on the cost of producing electrical energy despite the harassing increases in the costs of fuel during the same period. **Figure 11** illustrates the variation in thermal performance as a function of load for an assortment of stationary and marine propulsion power plants. *See* BRAYTON CYCLE; CARNOT CYCLE; DIESEL CYCLE; OTTO CYCLE; RANKINE CYCLE; THERMODYNAMIC CYCLE.

Most of the improvements in heat rates were achieved by increasing steam temperatures and pressures, thus improving the Carnot cycle efficiency. The slight rise in heat rates since 1970 is due in large part to pollution control equipment, including the increased use of closed cooling water systems. The pollution control equipment is detrimental to heat rates because it uses large quantities of electric power. The closed cooling water systems raise heat rates by increasing the steam cycle condensing temperature.

In a steam plant the steam is condensed back to water after passing through the turbine (Fig. 4d-g). The lower the temperature at which this steam is condensed, the higher the Carnot efficiency. The circulating water used to condense the steam has traditionally come from large bodies of water such as rivers, lakes, and seas, but because of thermal pollution concerns many plants switched to a closed evaporative system with the circulating water being cooled by cooling towers or cooling ponds. The cooling towers and ponds are usually incapable of providing the low condensing temperatures of natural bodies of water. Further pollution requirements have dictated the use of plume-free cooling towers (where the saturated air exhausting from the tower has to be heated), with the corresponding loss of cycle efficiency, or the use of air-cooled condensers, resulting in usually higher back pressure and a considerable amount of electricity used in running their fans.

In scrutinizing data on thermal performance, it should be recalled that the mechanical equivalent of heat (100% thermal efficiency) is 2545 Btu/hp-h and 3413 Btu/kWh (3.6 megajoules/kWh). Modern steam plants in large sizes (75,000-1,300,000 kW units) and internal combustion plants in modest sizes (1000-20,000 kW) have little difficulty in delivering a kilowatt-hour for less than 10,000 Btu (10.55 MJ) in fuel (34% thermal efficiency). For condensing steam plants, the lowest fuel consumptions per unit output (8200-9000 Btu/kWh or 8.7-9.5 MJ/kWh) are obtained in plants with the best vacuums, regenerative-reheat cycles using eight stages of extraction feed heating, two stages of reheat, primary pressures of 4500 lb/in.² gage or 31 megapascals gage (supercritical), and temperatures of 1150°F (620°C). An industrial plant cogenerating electric power with process steam is capable of having a thermal efficiency of 5000 Btu/kWh (5.3 MJ/kWh).

In the 1990s, combustion turbines used in combined cycle configurations began to take a dominant role in new power generation capacity. The primary reasons are the higher efficiency and lower emissions of the power plant in this arrangement. The rapid pace in advances in combustion turbine technology (such as higher firing temperatures that improve the Brayton cycle efficiency) has driven combined cycle efficiency to nearly 60% when using natural gas as fuel, while attaining low emission rates. Low fuel consumption (5700-6000 Btu/kWh or 6.0-6.3 MJ/kWh) is obtained by using higher firing temperatures, steam cooling on the combustor and gas turbine blades, a reheat steam cycle with a threepressure heat recovery steam generator, and higher pressure and temperature of the steam cycle. These conditions are balanced with the need to keep the exhaust flue gas temperature as low as practical to achieve low emissions.

Gas turbines in simple cycle configuration are used mostly for peaking service due to their fast startup capabilities. The advances in the gas turbines have also increased the efficiency of simple cycle operations. Recuperation of the classic Brayton cycle gas turbine (simple cycle) is an accepted method of improving cycle efficiency that involves the addition of a heat exchanger to recover some portion of the exhaust heat that otherwise would be lost. *See* GAS TURBINE.

The nuclear power plant substitutes the heat of fission for the heat of combustion, and the consequent plant differs only in the method of preparing the thermodynamic fluid. It is otherwise similar to the usual thermal power plant. The pressure of a light-water reactor core is limited by material and safety considerations, while the temperature at which the steam is produced is determined by the core pressure. Because a nuclear reactor does not have the capability to superheat the steam above the core temperature, the steam temperature in a nuclear cycle is less than in a fossil cycle. *See* NUCLEAR REACTOR.

Power economy. Costs are a significant, and often controlling, factor in any commercial power plant application. Average costs have little significance because of the many variables, especially load factor. Some plants are short-lived and others long-lived. For example, in most automobiles, which have shortlived power plants, 100,000 mi (160,000 km) and 3000-4000 h constitute the approximate operating life; diesel locomotives, which run 20,000 mi (32,000 km) a month with complete overhauls every few years, and large seagoing ships, which register 1,000,000 mi (1,600,000 km) of travel and still give excellent service after 20 years of operation, have long-lived plants; electric central stations of the hydro type can remain in service 50 years or longer; and steam plants run around the clock and upward of 8000 h a year with very high reliability even when 25 years old. Such figures greatly influence costs. Furthermore, costs are open to wide differences of interpretation.

In the effort to minimize cost of electric power to the consumer it is essential to recognize the difference between investment and operating costs, and the difference between average and incremental costs. Plants with high investment (fixed) costs per kilowatt should run at high load factors to spread the burden. Plants with high operating costs (such as fuel) should be run only for the shortest periods to meet peak loads or emergencies. To meet these short operating periods various types of peaking plants have been built. Combustion (gas) turbines and pumped-storage plants serve this requirement. In the latter a hydro installation is operated off-peak to pump water from a lower reservoir to an elevated reservoir. On-peak the operation is reversed with water flowing downhill through the prime movers and returning electrical energy to the transmission system. High head sites (for example, 1000 ft or 300 m), proximity to transmission lines, and low incremental cost producers (such as nuclear or efficient fossil-fuel-fired plants) are necessary. If 2 kWh can thus be returned on-peak to the system, for an

input of 3 kWh off-peak, a pumped-storage installation is generally justifiable. *See* PUMPED STORAGE.

In any consideration of such power plant installations and operations it is imperative to recognize (1) the requirements of reliability of service and (2) the difference between average and incremental costs. Reliability entails the selection and operation of the proper number and capacity of redundant systems and components and of their location on the system network. Emergencies, breakdowns, and tripouts are bound to occur on the best systems. The demand for maximum continuity of electrical service in modern civilization dictates the clear recognition of the need to provide reserve capacity in all components making up the power system.

Within that framework the minimum cost to the consumer will be met by the incremental loading of equipment. Incremental loading dictates, typically, that any increase in load should be met by supplying that load with the unit then in service, which will give the minimum increase in out-of-pocket operating cost. Conversely, for any decrease in load, the unit with the highest incremental production cost should drop that decrease in load.

In the 1990s, an international trend toward privatization of generating plants and deregulation of the electric power industry emerged. In the United States, this is being adopted on a state-by-state basis and is expected to be widespread in the 2005-2010 time frame. The impact of these initiatives has been a significant increase in the construction of combustion turbine combined cycle plants relative to other types of power plants, such as coal and nuclear. Reasons for this selection include higher efficiency, shorter construction schedules, lower capital and operating costs, widespread fuel availability, and lower emissions. *See* ELECTRIC POWER GENERATION. K. Keith Roe; Reginald S. Gagliardo

Bibliography. E. A. Avallone and T. Baumeister III (eds.), Marks' Standard Handbook for Mechanical Engineers, 10th ed., 1996; Babcock and Wilcox Co., Steam: Its Generation and Use, rev. 40th ed., 1992; Combustion Engineering, Inc., Combustion: Fossil Power Systems, 3d ed., 1981; D. G. Fink and H. W. Beaty (eds.), Standard Handbook for Electrical Engineers, 14th ed., 1999; M. C. Going, 1984 annual statistical report, Elec. World, 198(4):49-72, April 1984; R. L. Harrington, Marine Engineering, 1971; L. C. Lichty, Combustion Engine Processes, 7th ed., 1967; U.S. Department of Energy Publication, DOE/EIA-0219 (97), April 1999; U.S. Department of Energy Publication, DOE/EIA-0384 (98), July 1999.

Power shovel

A power-operated digging machine consisting of a lower frame and crawlers, a machinery frame, and a gantry supporting a boom which in turn supports a dipper handle and dipper (see **illus.**). Dipper sizes

range from 1 to 180 yd³ (0.7 to 137 m³). The machines are powered by on-board diesel engines or by electric motors. Diesel-powered machines utilize a series of clutches and brakes that allow the operator to control various motions. Electric motor machines generally have individual motors for each motion, but occasionally clutches and brakes are used allowing one motor to drive two motions.

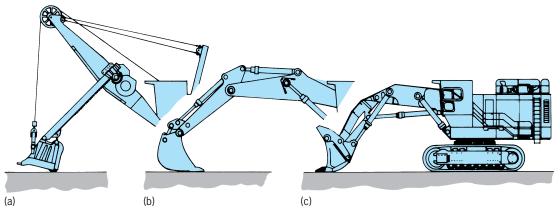
Machine components. The lower frame and crawlers serve to support the machine during digging and provide self-propulsion. The machinery frame rotates upon the lower frame by a large rolling-element thrust bearing. Machinery to raise and lower the dipper and to cause rotation are mounted on the machinery frame along with support points for the boom and gantry. The gantry protrudes above the cab and provides a means for connecting cables to support the boom. The boom is mounted on the front of the machinery frame. The dipper handle is mounted near the midpoint of the boom, where machinery is placed to cause the dipper handle to go in and out (crowd). The dipper is attached to the end of the dipper handle, and these two components are raised and lowered by a system of cables leading over a sheave at the point of the boom and back to the hoist machinery.

Machine operation. Four primary motions of the machine are propel, hoist, crowd, and swing. The controls for these motions are contained in the operator's station. The operation starts with the dipper positioned in front of the crawlers. The operator causes the dipper to enter the material to be dug by using the hoist and crowd motions. Once the dipper is full, it is retracted from the material and the machine is caused to rotate. To discharge the material, the bottom of the dipper is released. This digging cycle continues until no more material can be reached. The operator then propels the machine forward toward the bank.

Hydraulic excavators. A major variation of the power shovel is a machine whose motions are controlled and driven by hydraulic pumps, motors, and cylinders. The machine is operated similar to a mechanical machine, but the hoist and crowd motions are caused by hydraulic cylinders that activate the boom and arm. The boom is approximately half the length of a normal boom, and at its upper end the arm is pivotally connected (illus. *c*). At the other end of the arm, the dipper is pivotally connected. The operation of the machine is the same as that of a mechanical power shovel.

Backhoe machines. Mechanical power shovels and hydraulic excavators frequently use a backhoe front end (illus. *b*). The basic machine remains the same, but the lever arrangement of the front end is reversed so that the dipper movement starts extended away from the machine and is caused to dig downward into the material as it moves toward the machine.

Stripping shovels. Power shovels greater than 40 yd^3 (30 m^3) and equipped with long booms are used to remove overburden from underlying coal



Comparison of the boom, dipper handle, and dipper for (a) conventional power shovel, (b) hydraulic backhoe, and (c) hydraulic front shovel. A typical lower frame, crawlers, and machinery frame are shown for the hydraulic shovel.

seams. These power shovels weigh up to 13,500 tons (12,000 metric tons). To compensate for this weight, four pairs of crawlers are used instead of the normal single pair. These machines are fully electrically operated.

Lattice boom machines. Occasionally the power shovel is converted to another material-handling device that uses a long lattice boom. Cranes, clamshells, grapples, orange peels, and dock or portal cranes are used but do not efficiently use the high-speed functions of the power shovel. When the power shovel is operated as a dragline, the speed of the motions are effectively used. From these draglines, a special device has evolved for operating on soft ground. The lower frames of the power shovel are replaced by a large circular frame that sits on the ground. To propel, the crawlers are replaced by two large shoes that are activated by a propel linkage and machinery mounted on the upper frame. See BULK-HANDLING MACHINES; CONSTRUCTION EQUIPMENT; HOISTING MACHINES. E. W. Sankey

Bibliography. E. A. Avallone and T. Baumeister III (eds.), Marks' Standard Handbook for Mechanical Engineers, 10th ed., 1996; K. Haddock, A History of Marion Power Shovel's First 100 Years, 1984; J. Hollingsworth, History and Development of Strip Mining Machines, Bucyrus Erie Co.; T. Shaw and V. Pavlovic, Surface Mining and Quarrying: Mechanization, Technology, and Capacity, 1992.

Poynting's vector

A vector, the outward normal component of which, when integrated over a closed surface in an electromagnetic field, represents the outward flow of energy through that surface. It is given by Eq. (1),

$$\Pi = \mathbf{E} \times \mathbf{H} = \mu^{-1} \mathbf{E} \times \mathbf{B} \tag{1}$$

where **E** is the electric field strength, **H** the magnetic field strength, **B** magnetic flux density, and μ the permeability. This can be shown with the aid of

Maxwell's equations (2), where D is the electric dis-

$$\mathbf{H} \cdot (\nabla \times \mathbf{E}) - \mathbf{E} \cdot (\nabla \times \mathbf{H}) = \nabla \cdot (\mathbf{E} \times \mathbf{H})$$
$$= -\mathbf{i} \cdot \mathbf{E} - \mathbf{E} \cdot \frac{\delta \mathbf{D}}{\delta t} - \mathbf{H} \cdot \frac{\delta \mathbf{B}}{\delta t} \quad (2)$$

placement and *i* the current density. Integration over any volume v and use of the divergence theorem to replace one volume integral by a surface integral give Eq. (3), where **n** is a unit vector normal to *dS*. In the

$$\int (\mathbf{E} \times \mathbf{H}) \cdot \mathbf{n} \, dS = \int_{v} \left[\frac{\delta}{\delta t} (^{1}/_{2} \mathbf{B} \cdot \mathbf{H}) + \frac{\delta}{\delta t} (^{1}/_{2} \mathbf{D} \cdot \mathbf{E}) + \mathbf{E} \cdot \mathbf{i} \right] \, dv \quad (3)$$

volume integral, $\frac{1}{2}\mathbf{B} \cdot \mathbf{H}$ is the magnetostatic energy density, and $\frac{1}{2}\mathbf{D} \cdot \mathbf{E}$ is the electrostatic energy density, so the integral of the first two terms represents the rate of increase of energy stored in the magnetic and electric fields in v. The product of $\mathbf{E} \cdot \mathbf{i}$ is the rate of energy dissipation per unit volume as heat; or, if there is a motion of free charges so that \mathbf{i} is replaced by ρv , ρ being the charge density, it is the energy per unit volume used in accelerating these charges. The net energy change must be supplied through the surface, which explains the interpretation of Poynting's vector.

It should be noted that this proof permits an interpretation of Poynting's vector only when it is integrated over a closed surface. In quantum theory, where the photons are localized, it could be interpreted as representing the statistical distribution of photons over the surface. Perhaps this justifies the common practice of using Poynting's vector to calculate the energy flow through a portion of a surface.

When an electromagnetic wave is incident on a conducting or absorbing surface, theory predicts that it should exert a force on the surface in the direction of the difference between the incident and the reflected Poynting's vector. *See* ELECTROMAGNETIC RADIATION; MAXWELL'S EQUATIONS; RADIATION PRESSURE; WAVE EQUATION. William R. Smythe

Prairie dog

A type of ground squirrel belonging to the family Sciuridae in the order Rodentia. These stout, shorttailed, short-legged ground squirrels inhabit open plains, short-grass prairies, and plateaus in the western part of North America from Canada to Mexico. The short coarse fur is grayish-brown. They have small beady eyes, pouched cheeks, and a short flat tail. Adults have a head-body length of 280–330 mm (11–13 in.), a tail length of 30–115 mm (1–4.5 in.), and weigh 0.7–1.4 kg (1.5–3 lb). *See* RODENTIA; SQUIRREL.

Classification. Prairie dogs are classified in the genus *Cynomys*. Obviously not dogs, they received this name because of their shrill doglike bark. Five species are currently recognized (see **table**).

Behavior. Prairie dogs are diurnal, terrestrial, social mammals. They once lived predominantly in colonies known as towns, with miles of well-worn tunnels and dens extending in every direction beneath plateaus and upland prairies. A large town may have had millions of inhabitants. In 1901, a prairie dog town north of San Angelo, Texas, was estimated to be 100 mi (160 km) in width and 250 mi (400 km) in length. If the population averaged 25 prairie dogs to the acre, then the expanse of 16,000,000 acres contained 400 million animals. Today, few dog towns of any size remain outside of refuges, parks, or sanctuaries. *See* SOCIAL MAMMALS.

The home is constructed by digging a tunnel straight down for 12 ft (3.7 m) or more. At the bottom of the tunnel, the prairie dog hollows out several rooms, including a sleeping room and a room for storing food. At the entrance to the hole, a mound of hard-packed dirt keeps water from entering the burrow. Most burrows have only one or two entrances, but some have as many as five or six. The prairie dog spends the winter in its home but often comes out on sunny days when the wind is not blowing. True hibernation has been reported for C. leucurus, C. gunnisoni, and C. parvidens but not for C. mexicanus or C. ludovicianus. At high altitudes and in the colder parts of their range, they sleep through the winter, but in southern localities they are active year-round. Prairie dogs feed on grasses, herbs, roots, and insects such as locusts and grasshoppers when they are abundant. Their feeding habits eliminate the taller plants. This deprives predators of cover, improves visibility around the burrow, and encourages fast-growing weeds without abundant seeds and fruits. Tall vegetation in the middle of the town is removed entirely. *See* BURROWING ANIMALS.

Breeding usually occurs during late winter and early spring. The single annual litter of four to six young are born in an underground nest in early May after a gestation period of 28–37 days. During the breeding season, females may defend individual territories around their burrows. The young are weaned at about 4 weeks of age. Sexual maturity is reached at 1 year of age. A black-tailed prairie dog has lived 8.5 years in captivity.

If an enemy, such as a coyote, should come into sight, prairie dogs warn each other with loud chirps or barks. They have a "vocabulary" of 10 different calls, ranging from a commonly used warning bark to a chuckle, a "fear" scream, and a fighting snarl. Each call is the result of a specific stimulus and elicits a characteristic action by nearby individuals. When the warning bark is sounded, they rush into their homes deep underground, where they are out of danger. Besides coyotes, prairie dogs are hunted by badgers, foxes, ferrets, and wild cats. Eagles and hawks swoop down and grab them. Rattlesnakes often hibernate in prairie dog burrows, and burrowing owls live in vacant burrows. *See* ANIMAL COMMU-NICATION.

Habitat loss. Within their geographic range, suitable habitat for the prairie dog was maintained by grazing bisons and prairie fires. By the early 1900s, however, prairie dogs were considered a serious pest and economic menace to farmers and cattle ranchers in the West because of their feeding on crops and pastures. Also, their open burrows could lead to a running horse or cow breaking a leg. At one time, millions of prairie dogs lived in the western United States, but intensive public and private poisoning programs in the United States have greatly reduced their numbers. What was the largest single town remaining in the United States, covering approximately 110,000 hectares (425 mi²) in South Dakota, was practically eradicated by poisoning in 1987-1988. The largest town now known, in northwestern Chihuahua, Mexico, occupies 35,000 ha (135 mi²) and is part of a complex comprising about 55,000 ha (212 mi²) and more than 1 million animals, but that population is jeopardized by habitat destruction and poisoning.

Relationship with black-footed ferret. The fate of the black-footed ferret (*Mustela nigripes*) is closely tied to prairie dog populations. At one time, the

Species and range of prairie dogs (genus Cynomys)				
Species	Common name	Range		
C. ludovicianus	Black-tailed prairie dog	Great Plains from Montana and southern Saskatchewan to northern Mexico		
C. mexicanus	Mexican prairie dog	Northeastern Mexico		
C. leucurus	White-tailed prairie dog	South-central Montana, western Wyoming, northeastern Utah, northwestern Colorado		
C. parvidens	Utah prairie dog	South-central Utah		
C. gunnisoni	Gunnison's prairie dog	Southeastern Utah, southwestern Colorado, Arizona, western New Mexico		



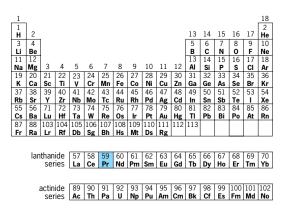
Prairie dogs. (Photo by Ron Singer/U.S. Fish and Wildlife Service)

geographic range of the ferret coincided with the range of the prairie dog. The greatest cause of death of black-footed ferrets has been human activity in control of prairie dogs. The widespread use of sodium fluoroacetate ("1080") to poison prairie dogs has resulted in secondary poisoning of black-footed ferrets that fed on them. However, probably the greatest mortality results from the large-scale eradication of prairie dogs from the range of the blackfooted ferret, thus greatly reducing the food supply. The black-footed ferret was never numerous enough to pose a serious threat to domestic stock, but it benefited early settlers by helping to keep populations of prairie dogs from expanding on grazing lands. Ironically, humans have driven the black-footed ferret to near extinction by reducing the habitat and number of prairie dog towns to levels where effective breeding populations no longer exist. This has resulted in the black-footed ferret becoming one of the most endangered mammals in North America. As of March 1997, the wild and captive ferret population numbered only approximately 500 animals. Since then ferrets have been reintroduced into suitable sites in Montana (1994), South Dakota (1994), and Arizona (1996). See ENDANGERED SPECIES; FER-RET; POPULATION VIABILITY. Donald W. Linzey

Bibliography. R. M. Nowak, *Walker's Mammals of the World*, 6th ed., Johns Hopkins University Press, 1999; D. E. Wilson and S. Ruff (eds.), *The Smithsonian Book of North American Mammals*, Smithsonian Institution Press, 1999.

Praseodymium

A chemical element, Pr, atomic number 59, and atomic weight 140.91. Praseodymium is a metallic element of the rare-earth group. The stable isotope 140.907 makes up 100% of the naturally occurring element. The oxide is a black powder, the composition of which varies according to the method of



preparation. If oxidized under a high pressure of oxygen it can approach the composition PrO₂. The black oxide dissolves in acid with the liberation of oxygen to give green solutions or green salts which have found application in the ceramic industry for coloring glass and for glazes. *See* PERIODIC TABLE; RARE-EARTH ELEMENTS. Frank H. Spedding

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; K. A. Gschneidner Jr., J.-C. Bünzli, and V. K. Pecharsky (eds.), *Handbook on the Physics and Chemistry of Rare Earths*, 2005.

Prasinophyceae

A small class of mostly motile, photosynthetic, unicellular algae in the chlorophyll *a-b* phyletic line (Chlorophycota). It is segregated from the Chlorophyceae primarily on the basis of ultrastructural characters, especially the possession of one or more layers of polysaccharide scales outside the plasmalemma. Prasinophytes are mainly members of the marine plankton, but they are also found in brackish- and fresh-water habitats. A few are benthic, with both coccoid and colonial forms known, while others live symbiotically within dinoflagellates, radiolarians, and turbellarian worms. Approximately 180 species are known in 13 genera.

Cells of Prasinophyceae range in maximum dimension from 1 to more than 800 micrometers. Small cells are bean-shaped or ellipsoid, while large cells are spherical. Some genera lack cell walls, but others have thick walls composed of a noncellulosic material similar to sporopollenin or to pectin. Motile cells have from one to eight flagella usually covered with scales and hairs. The form of the flagellar and body scales is useful in classification. Flagella of uniflagellate and biflagellate cells usually arise laterally or apically, while those of multiflagellate cells usually arise from an apical depression. The flagellar root system is composed of a rhizoplast, an array of microtubular rootlets, and often a microbody. Flagellate cells have a nucleus located beneath the flagellar root system and a large, lobed, parietal chloroplast usually encompassing a thylakoid-traversed pyrenoid and an eyespot. Chloroplast ultrastructure and the photosynthetic pigments are similar to those of other Chlorophycota. Some species have siphonein and siphonaxanthin, xanthophylls that otherwise are characteristic of the chlorophycean order Bryopsidales. Starch, which accumulates in the chloroplast, often as a sheath around the pyrenoid, is the main storage product, but mannitol, a sugar alcohol common in Chromophycota, also occurs in some species. Unlike other Chlorophycota, prasinophytes do not produce sucrose.

Sexual reproduction has not been observed with certainty. Binary fission is the only form of reproduction known for many species. Some benthic species and the plankton species that comprise very large, nonmotile, floating cells (phycomata) produce zoospores. Resting stages and cysts are known for some fresh-water species.

These scaly green flagellates are considered primitive in relation to other Chlorophycota. *See* ALGAE; PHYTOPLANKTON. Paul C. Silva; Richard L. Moe

Bibliography. R. E. Norris, Prasinophyceae, in S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, vol. 1, pp. 162–164, 1982; R. E. Norris, Prasinophytes, in E. R. Cox (ed.), *Phytoflagellates*, pp. 85–145, 1980.

Preamplifier

A voltage amplifier intended for amplifying a lowlevel input signal. The output is the input to another amplifier with a higher input level. For example, if the second amplifier needs a 500-mV signal to produce the desired power output and one has a microphone with a maximum output of 10 mV, one would use a preamplifier with a gain of 50 to achieve the desired power output. The term preamplifier is not precise; it has been applied to the portion of an amplifier preceding the power output stage and is not an additional system component. The term can also apply to a radio-frequency amplifier to amplify weak signals. The output is the radio receiver input. *See* AUDIO AMPLIFIER; RADIO-FREQUENCY AMPLIFIER; VOLTAGE AMPLIFIER.

The design of preamplifiers is critical because the input signal level is low. The noise introduced by transistors, hum from the power supply, and pickup from the leads must be minimized. Audio preamplifiers use bipolar or field-effect transistors, whereas radio-frequency preamplifiers use field-effect transistors only. The input impedance is generally high, and the output impedance is low. A frequencycompensation network may be incorporated into the design to compensate for the frequency characteristics of the input transducer to make the frequency response of the preamplifier-amplifier combination uniform. The bandwidth of the preamplifier must be sufficient to avoid reduction in performance. Operational-amplifier circuits can make good audio preamplifiers. The power output level is low since preamplifiers are not intended to drive loads such as that of a loudspeaker. See AMPLIFIER; ELECTRICAL NOISE; ELECTRONIC POWER SUPPLY; OP-ERATIONAL AMPLIFIER; TRANSISTOR. Harold F. Klock Bibliography. G. Ballou (ed.), *Handbook for Sound Engineers*, 2d ed., 1991; N. J Sclater and S. Gibilisco, *Encyclopedia of Electronics*, 2d ed., 1990; A. Sedra and K. Smith, *Micro-electronic Circuits*, 4th ed., 1997.

Prebiotic organic synthesis

The plausible pathways by which the molecular precursors of life may have formed on the primitive Earth. Amino acids, the nitrogenous bases, and ribose phosphates can be prepared under conditions that might have prevailed on the primitive Earth. The linking together of amino acids to form polypeptides, and of nucleotides to form polynucleotides, has in principle been established.

Harold C. Urey's model of the primordial Earth postulates an atmosphere rich in methane, ammonia, water, and hydrogen. When this gas mixture is subjected to an electrical spark, analogous to the way that lightning may have initiated such syntheses 4 billion years ago, the identified products included several amino acids (glycine, alanine, and aspartic acid), the building blocks of proteins. This novel result lent credibility to a theory in which the origin of life was viewed as a cumulative, stepwise process, beginning with the gaseous synthesis of small molecules, which rained down into oceans, lagoons, and lakes. With water as a ubiquitous solvent, organic molecules could then react with one another to form larger molecules (biopolymers) and finally to assemble into primitive cells. This general scenario has guided the design of prebiotic simulations.

However, an accumulation of geophysical data and computational models has cast doubt on the relevance of the synthesis of amino acids to the primordial Earth. Hydrogen probably escaped rapidly as the Earth cooled, leaving an atmosphere in which methane and ammonia were virtually absent. As the input of hydrogen is diminished, the formation of biomolecules is inhibited. This problem has led some scientists to look for extraterrestrial sources of organic matter. For example, meteorites are known to contain a rich source of amino acids and other small biomolecules, and perhaps the infall of such cosmic bodies onto the young Earth gave life its start. Alternatively, there may have been localized environments on the Earth where methane and other hydrogen-rich precursors were abundant, such as deep-ocean hydrothermal vents, which would have been favorable for the formation of life. See AMINO ACIDS; HYDROTHERMAL VENT.

Role of hydrogen cyanide. Alternative routes to amino acids and other biomolecules begin with hydrogen cyanide (HCN), which can be synthesized from a variety of gas mixtures and energy sources. In a slightly alkaline solution of hydrogen cyanide in water; cyanide reacts with itself to form a pentamer known as adenine—an essential constituent of deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). In fact, adenine can be synthesized from

hydrogen cyanide by two distinct routes, one occurring in the dark and the other proceeding by a light-induced process. *See* DEOXYRIBONUCLEIC ACID (DNA); RIBONUCLEIC ACID (RNA).

Hydrogen cyanide in water can provide a suite of molecular constituents found in proteins and nucleic acids. Following hydrolysis of the soluble products from hydrogen cyanide polymerization, the amino acids glycine, alanine, and aspartic acid were identified along with adenine and orotic acid (which undergoes a light-induced reaction to yield uracil, a component of RNA). Uracil can be formed by a direct pathway that does not require light. Guanine, another constituent of nucleic acids, can be formed by the reaction of urea with an imidazole derivative obtained from hydrogen cyanide. That such a diverse array of important biomolecules can be synthesized from such simple precursors has led many scientists to conclude that this compound may have played a central role in the origins of life. See CYANIDE.

Proteins and nucleic acids. Proteins consist of long chains of amino acids linked together by bonds known as peptides. Amino acids can be joined together by a variety of activating agents, among which the most plausibly prebiotic is a hydrogen cyanide derivative called cyanamide. Cyanamide has been used with amino acids to form small proteinlike structures with four peptide units, although the strategy could be applied to synthesize larger products. Polypeptides containing up to 50 amino acids have been prepared by using adenosine triphosphate (ATP) as the activating agent with clay minerals as catalysts. *See* ADENOSINE TRIPHOSPHATE (ATP).

A question regarding origins of life has been centered on whether proteins came before nucleic acids or vice versa. The sensible answer is "both," since each class of molecules was believed to serve a separate and unique purpose: proteins act as catalysts known as enzymes, accelerating biochemical reactions, while nucleic acids (polynucleotides) encode the information required to make those proteins. Ribonucleic acid can catalyze bond cleavage and bond formation. Thus, it is possible that nucleic acids could have served as both genes and enzymes in the first living things. *See* ENZYME.

If RNA was the first genetic material, there must have been a mechanism for it to make copies of itself without the aid of modern enzymes. In nucleic acid chemistry, the molecular copy is analogous to a photographic negative, since the pairing between nitrogen-containing bases requires a specific complementary sequence in the product: adenine directs the incorporation of uracil, guanine directs the incorporation of cytosine, and vice versa. Nonenzymatic replication is indeed possible, provided the procedure is begun with a set of activated precursors. Although there are some constraints on the range of polynucleotides that can serve as templates, replication proceeds with high fidelity-the products almost always contain the base expected from the pairing principle. The template can discriminate against nonbiological "left-handed" nucleotides, but these

components also inhibit the reaction. However, the newly synthesized polynucleotide does contain a predominance of the same type of sugar-phosphate linkage as natural RNA strands. While these models rely on artificial starting materials that may not have been present on the primitive Earth, they do demonstrate that replication can proceed in the absence of highly evolved enzymes. *See* OLIGONUCLEOTIDE.

Polynucleotides. A more fundamental problem is where the first polynucleotide came from. By starting with an adenine nucleotide and a diphosphatelinked dimer, clay minerals can catalyze a synthesis up to the decanucleotide (10 bases long) containing predominantly the biologically correct ribose-phosphate bond. However, the assembly of a polynucleotide from the constituent nitrogenous bases, ribose and phosphate, is not an easy task. Baking the ingredients as a dry film can cause the ribose to join to the phosphate and to adenine or guanine, but the reaction is inefficient. Because ribose (a five-carbon sugar) is plagued by its instability and tendency to undergo side reactions, it has been proposed that a three-carbon alcohol, glycerol, fulfilled the linking role that was subsequently taken over by ribose; such nucleotide analogs can also undergo template-directed polymerization.

Six-carbon sugars can also serve the same structural function as ribose; a selective route was developed to phosphate-substituted sugars, starting from glycoaldehyde phosphate in alkaline solution. A major advantage to this pathway is that the phosphate group protects the products from decomposition. In the presence of formaldehyde, the major product is ribose 2,4-diphosphate (accounting for one-third conversion of the starting material), a remarkably efficient synthesis. More plausible routes to the constituent nucleotides will be articulated as ribose phosphates undergo greater scrutiny in the laboratory. *See* NUCLEIC ACID; PRO-TEIN. William J. Hagan, Jr.

Bibliography. C. Chyba, The cosmic origins of life on Earth, *Astronomy*, 20:28–35, 1992; A. Eschenmoser and E. Lowenthal, Chemistry of potentially prebiological natural products, *Chem. Soc. Rev.*, 21:1–16, 1992; J. P. Ferris and W. J. Hagan, Jr., HCN and chemical evolution: The possible role of cyano compounds in prebiotic synthesis, *Tetrabedron*, 40:1093–1109, 1984; G. F. Joyce, RNA evolution and the origins of life, *Nature*, 338:217–224, 1989; S. F. Mason, *Chemical Evolution: Origins of the Elements, Molecules and Living Systems*, 1991; L. E. Orgel, Molecular replication, *Nature*, 358:203– 209, 1992.

Precambrian

A major interval of geologic time between about 540 million years (Ma) and 3.8 billion years (Ga) ago, encompassing most of Earth history. The Earth probably formed around 4.6 Ga and was then subjected to a period of intense bombardment by

PHANEROZOIC	Paleozoic	
PRECAMBRIAN	Proterozoic	
PRECAMBRIAN	Archean	

meteorites so that there are few surviving rocks older than about 3.8 billion years. Ancient rocks are preserved exclusively in continental areas. All existing oceanic crust is younger than about 200 million years, for it is constantly being recycled by the processes of sea-floor spreading and subduction. The name Hadean has been proposed for the earliest turbulent part of Earth's history. Development of techniques for accurate determination of the ages of rocks and minerals that are billions of years old has revolutionized the understanding of the early history of the Earth. *See* DATING METHODS; GEOLOGIC TIME SCALE; HADEAN; ROCK AGE DETERMINATION.

Detailed sedimentological and geochemical investigations of Precambrian sedimentary rocks and the study of organic remains have facilitated understanding of conditions on the ancient Earth. Microorganisms are known to have been abundant in the early part of Earth history. The metabolic activities of such organisms played a critical role in the evolution of the atmosphere and oceans. There have been attempts to apply the concepts of plate tectonics to Precambrian rocks. These diverse lines of investigation have led to a great leap in understanding the early history of the planet. *See* PLATE TECTONICS.

Subdivisions. Early attempts to subdivide the Precambrian were mainly based on geological characteristics such as rock structure and metamorphic grade. Meaningful subdivision of the Precambrian has become possible only with the development of sophisticated techniques for deciphering the ages of rocks. The Subcommission on Precambrian Stratigraphy of the International Union of Geological Sciences (IUGS) has suggested a subdivision based on specific time intervals (Fig. 1). Some have claimed that subdivisions of the Precambrian should correspond to specific evolutionary changes that can be observed in the rock record. Many of these changes occurred at widely different times in different parts of the planet so that "arbitrary" time subdivisions are somewhat artificial. To overcome some of these conceptual problems, a simple subdivision using geons, which are time units of 100 million years' duration, has been suggested (Fig. 1). The era names suggested by the IUGS have been widely accepted, whereas the smaller subdivisions (equivalent to periods in younger rocks) have not. Subdivision of the Precambrian has been contentious. Most agree on names for major subdivisions, but the concept of rigidly applicable global subdivisions based on arbitrary times poses philosophical problems for some.

Archean. Rocks of the Archean Eon (2.5-3.8 Ga) are preserved as scattered small "nuclei" in shield areas on various continents. The Canadian shield contains perhaps the biggest region of Archean rocks in the world, comprising the Superior province. Much of the Archean crust is typified by greenstone belts, which are elongate masses of volcanic and sedimentary rocks that are separated and intruded by greater areas of granitic rocks. The greenstones are generally slightly metamorphosed volcanic rocks, commonly extruded under water, as indicated by their characteristic pillow structures (Fig. 2). These structures develop when lava is extruded under water and small sac-like bodies form as the lava surface cools and they are expanded by pressure from lava within. Such structures are common in Archean greenstone assemblages in many parts of the world. See ARCHEAN; METAMORPHIC ROCKS.

Ultramafic magnesium-rich lavas, known as komatiites, are considered to have formed at high

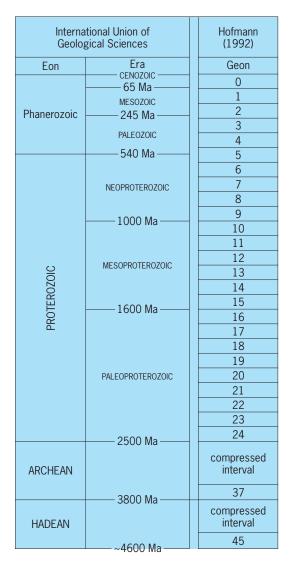


Fig. 1. Major subdivisions of geologic time. The Precambrian comprises the Archaean and Proterozoic eons.



Fig. 2. Neoproterozoic pillow lavas from Anglesey, North Wales.

temperatures (about 3000°F or 1650°C). These unusual lavas have been interpreted to mean that the interior of the Earth (mantle) was much hotter in the Archean than it is now. Geochemical and other investigations of lavas preserved in greenstone belts have led to the interpretation that they formed in a number of environments, analogous to present-day ocean floors, oceanic and continental volcanic arcs, and rifts.

Many of the associated sedimentary rocks are "immature" in nature, having formed by rapid erosion and deposition, from volcanic and, in some cases, older continental crustal sources. Such sedimentary rocks are commonly known as turbidites, formed as a result of gravity-controlled deposition in relatively deep water. Relatively rare "supermature" sediments formed in shallow-water environments as the result of extreme weathering. They are composed almost entirely of quartz and are known as orthoquartzites. In addition, iron-rich sediments formed by chemical precipitation from seawater (Fig. 3). These sediments indicate the existence of liquid water on the Archean Earth's surface and provide important clues concerning the nature of the ancient atmosphere and oceans. See QUARTZ; SEDIMENTARY ROCKS; TUR-BIDITE

Many Archean shield areas also contain terranes composed of highly metamorphosed rocks (highgrade gneiss terranes). The origin of these gneissic terranes is controversial, but some may represent subduction zones located at ancient continental margins where sedimentary rocks were carried to great depths, metamorphosed, and intruded by igneous rocks. There has been a growing (although not universal) tendency to interpret Archean shield areas as the result of some form of plate tectonic activity, probably involving accretion of island arcs and collision of small continental nuclei. Theoretical considerations and the presence of magnesium-rich komatiitic lavas have led to the suggestion that the interior of the early Earth was much hotter than at present. The end of the Archean is marked in many places by amalgamation of large numbers of volcanic arcs and crustal fragments, which together formed the world's first large continental masses. This process appears to have occurred over a long period of time, beginning at about 3.0 Ga in South Africa and Australia and culminating at about 2.5 Ga in North America. The latter date has been adopted as the demarcation point between the Archean and the Proterozoic. *See* CONTINENTAL MARGIN; CONTI-NENTS, EVOLUTION OF; IGNEOUS ROCKS; LAVA.

Proterozoic. The Proterozoic Eon extends from 2.5 Ga until 540 Ma, the beginning of the Cambrian Period and Phanerozoic Eon. Proterozoic successions include new kinds of sedimentary rocks, display proliferation of primitive life forms such as stromatolites, and contain the first remains of complex organisms, including metazoans (the Ediacaran fauna). Sedimentary rocks of the Proterozoic Eon contain evidence of gradual oxidation of the atmosphere. Abundant and widespread chemical deposits known as banded iron formations (BIF) make their appearance in Paleoproterozoic sedimentary basins. *See* BANDED IRON FORMATION; PROTEROZOIC.

Paleoproterozoic. In most Proterozoic depositional basins there is a higher proportion of sedimentary than volcanic rocks. The world's first widespread glacial deposits are preserved in Paleoproterozoic successions. The existence of glacial conditions so long ago (\sim 2.3 Ga) is indicated by the presence of unusual conglomerates (diamictites or tillites) with large rock fragments set in an abundant fine-grained matrix (Fig. 4a). These peculiar conglomerates consist of a variety of rock fragments set in an abundant finer-grained matrix. Such conglomerates may form as a result of deposition of unsorted sediments from glaciers, in which case they are called tillites. Additional evidence of glaciation includes glacially scratched (striated) rock surfaces, striated and faceted rock fragments and "dropstones" (rock



Fig. 3. Banded iron formation (BIF) from the Archean of Finland. The light layers are silica-rich (chert), and the dark layers consist of iron oxides. The highly contorted nature of many such iron formations is due to early slump movements of these chemical sediments prior to complete consolidation.

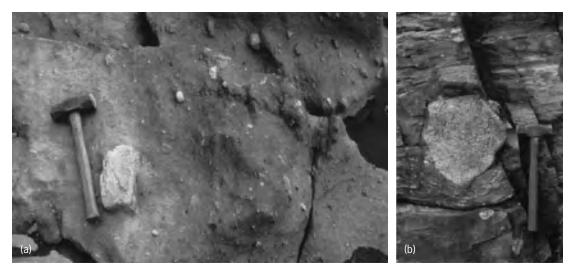


Fig. 4. Glacial sedimentary rocks. (a) Diamictites forming part of the Paleoproterozoic Gowganda Formation (\sim 2.3 Ga) on the north shore of Lake Huron, Ontario, Canada. (b) Dropstones from the Gowganda Formation. Such large rock fragments in otherwise fine-grained bedded sedimentary rocks are thought to have been "vertically" emplaced from melting icebergs.

fragments that are interpreted as having been released, by melting, from floating icebergs into finely bedded sediments; Fig. 4*b*). Possibly contemporaneous Paleoproterozoic glacial rocks are known from five continents. *See* DEPOSITIONAL SYSTEMS AND EN-VIRONMENTS.

Paleoproterozoic sedimentary rocks contain some of the world's greatest accumulations of valuable minerals, such as the uranium deposits of the Elliot Lake region in Ontario, Canada, and the great banded iron formations of the Lake Superior region. The uranium deposits are believed to have resulted from deposition of mineral grains in ancient river systems. Because uranium minerals (and associated pyrite) were not oxidized, it has been inferred that the Earth's atmosphere had a very low free oxygen content. By contrast, the somewhat younger (~2.0 Ga) banded iron formations are thought to indicate an oxygen-rich hydrosphere/atmosphere system.

The Paleoproterozoic also contains abundant carbonate rocks. Some of these contain spectacular organo-sedimentary structures called stromatolites (Fig. 5). The organo-sedimentary nature of these structures is confirmed by the preservation of microscopic primitive prokaryotic cells of photosynthetic cyanobacteria. These photosynthetic microorganisms trapped, or in some cases precipitated, carbonates from seawater, resulting in the construction of stromatolitic mounds of highly varied morphology. The metabolic activity of the microorganisms responsible for building these structures contributed to oxygenation of the atmosphere. Identical structures occur in some areas of the modern ocean, such as shallow-water saline environments of Shark Bay in Western Australia. Widespread development of stable continental shelves at the beginning of the Proterozoic (in contrast to the unstable, dominantly volcanic settings of much of the Archean) may have favored stromatolitic growth. See STROMATOLITE.

Two theories have emerged to explain oxygenation of the atmosphere. Increased photosynthetic activity, due to proliferation of stromatolites, is one. A second theory suggests that oxygen produced by marine photosynthetic microorganisms was consumed by reducing reactions related to widespread hydrothermal circulation of seawater at mid-ocean ridges and in other oceanic volcanic systems. Although there was significant production of photosynthetic oxygen, little was released to the atmosphere. This idea is supported by the preservation of different proportions of carbon isotopes (12C and 13C) in ancient organic carbon and carbonate rocks. Removal of ¹²C during photosynthesis results in seawater that is enriched in the heavier isotope. The carbon-isotopic composition of seawater may, under certain conditions, be preserved in carbonate rocks. Differences in the ratios of carbon



Fig. 5. Stromatolites in the upper part of the Paleoproterozoic Snowy Pass Supergroup in southeastern Wyoming, United States.

isotopes of such carbonates and of organic carbon have been detected in many Archean sedimentary rock successions. These data suggest that photosynthetic activity was fairly widespread even during the early history of the Earth, but that release of significant amounts of free oxygen to the atmosphere was delayed until the tectonic regime changed from being ocean-dominated at the end of the Archean. *See* ATMOSPHERE, EVOLUTION OF; CYANOBACTERIA; MID-OCEANIC RIDGE; PHOTOSYNTHESIS; SEAWATER.

Another aspect of sedimentary geology that supports an increase in free oxygen during the Paleoproterozoic Era is the appearance of redbeds, sedimentary rocks that contain finely disseminated hematite. Many sedimentary rocks of this kind formed in subaerial or shallow-water settings where they were in close contact with the atmosphere, which must have contained at least some free oxygen, to impart the characteristic red or purple coloration of the highly oxidized iron. *See* REDBEDS.

Another indication of atmospheric change is the preservation of widespread banded iron formations at around 2.0 Ga. The major problem with these economically important and enigmatic chemical sedimentary rocks is that iron is virtually insoluble in present-day rivers and oceans. Under reducing conditions, such as those postulated for the Archean, the solubility of iron would have been greatly increased. The development of widespread banded iron formations in relatively shallow depositional settings in the Paleoproterozoic is thought to represent a great "flushing out" of iron from the oceans as a result of changes in the partial pressure of oxygen in the atmosphere/hydrosphere system. Paleoproterozoic limestones in various parts of the world display a period of exceptional ¹³C enrichment at around 2.0 Ga, approximately the same time as the peak in banded iron formations production. One explanation for the high ¹³C content of these carbonates is that photosynthetic organic activity was particularly high, so that considerable amounts of carbon were sequestered from atmospheric carbon dioxide (CO₂), leading to high ¹³C content in the oceans and also to release of oxygen as a by-product.

By the end of the Paleoproterozoic, many of the oceans formed by breakup of the earliest(?) end-Archean "supercontinent," known as Kenorland, appear to have closed as a result of subduction and collisions to produce new large cratonic areas. For example, much of what is now the North American continent amalgamated at this time. There is plentiful evidence of widespread Mesoproterozoic igneous activity, in the form of intrusive bodies such as anorthosites and as abundant volcanic episodes, thought to be related to subduction and continental growth on the margins of large continental masses.

A notable feature of the Mesoproterozoic era is the dearth of evidence of glaciation during this long time interval. The Mesoproterozoic era was brought to a close by the Grenville orogeny, a widespread mountain-building episode that is named from the Grenville tectonic province in eastern North America. This orogenic period marks the construction of yet another supercontinent, known as Rodinia. Much of the subsequent Precambrian history in the Neoproterozoic is concerned with the complicated disintegration of Rodinia and culminates with the opening of a precursor to the presentday Atlantic Ocean, known as the Iapetus Ocean. *See* OROGENY.

Neoproterozoic. The period between the approximate end of the Grenville orogeny (~1.0 Ga) and the beginning of the Cambrian (~540 Ma) is known as the Neoproterozoic. This period is extremely important. It contains evidence of several glacial episodes that may have been the greatest that the world has known. Rocks belonging to this period preserve evidence of the first complex fossil forms, the controversial Ediacaran fauna, that is widely believed to represent the first animals with differentiated cells. These enigmatic fossils are believed by some to represent the ancestors of subsequent animal phyla. Others regard them as a "failed experiment" and assign them to completely separate, extinct taxa. Most Ediacaran forms occur in sandy rocks formed subsequent to deposition of the younger of two widespread glacial units, which is thought to have been deposited at about 620 Ma. See CAMBRIAN; EDIACARAN BIOTA.

As many as four or five glaciations have been postulated in the Neoproterozoic, but at most locations it is possible to document only two. The two most widespread glacial events are dated at about 750 Ma and 620 Ma and are respectively named Sturtian and Marinoan, from localities in Australia. The glacial deposits of the Neoproterozoic Era have recently come under close scrutiny for two main reasons. First, careful studies of the magnetic signature preserved in some of these glacial deposits (at least in parts of Australia), indicate deposition at nearequatorial latitudes. Second, some carbonate rocks associated with the glacial diamictites have very low ¹³C/¹²C ratios. This has been interpreted to mean a drastic reduction in photosynthetic activity, which would normally have led to preferential removal of the light carbon isotope from oceanic waters. It has been speculated that, during this period, life was almost eliminated from Earth and that ice extended into tropical latitudes-a condition referred to as the snowball earth. Alternatively, it has been suggested that increased obliquity of the Earth's ecliptic (greater inclination of the spin axis, relative to the orbital plane) could have led, during cold periods, to preferential buildup of snow and glacial ice in low latitudes. The ultimate cause of these great Proterozoic glaciations may lie in fluctuations in atmospheric greenhouse gases (particularly carbon dioxide). The reasons for such fluctuations remain obscure, but it has been proposed that increased weathering, as a result of uplift of the land surface during orogenic (mountain-building) episodes or greatly increased organic productivity, may have led to such climatic changes. See CLIMATE HISTORY.

Other aspects. Both water and life appear to have existed on Earth from at least 3.8 Ga. This evidence and the dearth of glacial deposits in the Archean have led to the "faint young sun paradox." In spite of much reduced solar radiation inferred by astrophysicists for the early part of Earth history, there is little evidence of low temperatures at the Earth's surface in Archean times. This has been explained by invoking much higher partial pressures of carbon dioxide during the early part of Earth history. The reasons for widespread Paleo- and Neoproterozoic glaciations, some of which may have occurred at low paleolatitudes, are not well understood. The theory of plate tectonics, which provides an elegant explanation of most of the features on the present-day surface of the Earth, appears to be applicable, perhaps in modified form, to rocks of both Archean and Proterozoic eons. The gradual buildup of free oxygen in the atmosphere/hydrosphere system was particularly important, for it permitted new metabolic pathways leading to the plethora of species that inherited the Neoproterozoic Earth and whose descendants survive to the present day. Grant M. Young

Bibliography. P. E. Cloud, *Oasis in Space*, 1988; P. F. Hoffman, United plates of America: Early Proterozoic assembly and growth of Laurentia, *Annu. Rev. Earth Planet. Sci.*, 16:543-603, 1988; H. J. Hofmann, New Precambrian time scale: Comments, *Episodes*, 15:122-123, 1992; A. H. Knoll, End of the Proterozoic Eon, *Sci. Amer.*, 262:64-72, 1991; S. R. Taylor and S. M. McLennan, The origin and evolution of the Earth's continental crust, *AGSO J. Austral. Geol. Geophys.*, 17:55-162, 1997.

Precast concrete

Concrete that is cast into forms and cured at a location other than its final position in a structure. Products that are commonly fabricated in this manner include beams and joists, slab units, wall panels, and utility items such as pipe sections and access holes for workers. Cast stone that is made to simulate cut and dressed stone is a form of precast. Concrete block is also sometimes considered a form of precast. Precast concrete can be unreinforced or reinforced, and prestressed by the pretensioning or posttensioning method. Precast concrete is used with castin-place, structural steel, or load-bearing masonry construction. Field connections are more critical for precast than for monolithic cast-in-place concrete, because the connections must provide structural continuity and load transfer between elements.

Construction products can be fabricated at a precast concrete plant or in a convenient location on the project site. On-site methods of precasting are usually limited to tilt-up and lift-slab construction. Precast concrete is economical because it permits the efficient mass production of repetitious members and the reuse of forms; reduces the cost of onsite formwork, falsework, and shoring; and speeds construction. Precast concrete sections, particularly when they are prestressed, are often lighter weight than the equivalent cast-in-place section would be, representing a savings in materials and permitting a reduction in supporting foundations. Plant casting also provides better quality control over concrete mixing and curing procedures, facilitates assemblyline production methods, and permits work to proceed expeditiously in any weather. Some savings can be offset, however, by transportation and handling costs if hauling distances are long. *See* PRESTRESSED CONCRETE; REINFORCED CONCRETE.

Wall, floor, and roof systems. Complex concrete wall-panel shapes cannot be easily or economically cast in place. Precasting methods, however, make this cladding system competitive. Wall panels, sills, lintels, and decorative elements such as balusters and arches are often precast in special colors and finishes for exposed architectural applications. Wall panels are typically connected to structural supports by welded or bolted connections. Bolted connections and wedge inserts allow field adjustment for construction tolerances. Expansion joints are used to allow thermal expansion and contraction between panels, and exterior joints are protected against water penetration with elastomeric sealants.

Floor and roof slabs can be constructed with precast hollow-core planks, tees, double-tees, or channels. These are typically supported on masonrybearing walls, on steel framing, or on cast-in-place or precast concrete beams and joists. Precast double-tees provide spans of 60 ft (18 m) or more, and are frequently used for parking structures. For shear transfer through diaphragm action, reinforcing steel from supporting elements can be tied into a relatively thin, reinforced or unreinforced topping slab cast in place on top of the precast members. *See* FLOOR CONSTRUCTION; ROOF CON-STRUCTION; WALL CONSTRUCTION.

Tilt-up walls. Large wall panels that are of simple design can be economically precast on site. The panels are cast in a horizontal position immediately adjacent to their wall location and then are tilted up into the final vertical position. This system is popular for warehouse and industrial buildings of simple rectangular plan because of its speed and economy. Doors and windows are easily installed by simply blocking out the proper size of opening. Walls are typically 5-8 in. (13-20 cm) thick and can function as deep, thin beams spanning horizontally between columns to eliminate the need for a continuous footing. The floor slab is frequently used as a casting platform, coated with a bond-breaking agent or covered with sheet materials to prevent bonding with the wall panel. Only side forms are needed, and the upper (outside) surface may be finished in the same way that a typical concrete slab is finished by floating or troweling. More often, however, tilt-up panels are given a decorative finish such as exposed aggregates.

When the wall panel has attained sufficient strength, it is lifted by embedded plates and temporarily braced until permanent attachments are made. The lifting and handling process induces significant stress on the panels, so the lifting plates must be carefully located and installed to prevent damage. Permanent welded connections can be made to steel columns or to precast concrete columns with embedded steel plates. Alternatively, cast-inplace columns can be poured after the wall panels have been erected, with reinforcing steel from the panels extending into the column for structural connection.

Lift-slab construction. Floor and roof slabs for multistory buildings can be cast on a base slab at or near ground level and then lifted into their final position. The slabs are cast in succession one atop the other, with the roof slab cast last, on top. A bond-breaking compound or membrane is used to keep the slabs from bonding to one another. Columns are erected first, usually to the full height of the structure, and the slabs are cast around them. An embedded steel collar is placed around each column opening for subsequent lifting and attachment. Jacks mounted to the columns are used to lift the slabs into place, beginning with the roof slab, and followed by successive floor slabs, working from top to bottom. The slabs are permanently welded to steel supporting plates attached to each column.

Lift-slab construction eliminates most on-site formwork, reduces transportation and handling costs, and greatly reduces construction time. This method is also sometimes called the Youtz-Slick method after its two inventors. *See* BUILDINGS; CONCRETE SLAB.

Civil engineering applications. Precast concrete pipe sections, box culverts, access holes, and other shapes make the construction of utility, highway, and bridge projects faster and more economical. Precast items can be manufactured with steel or fiber reinforcing to produce the thinnest and lightestweight sections to reduce transportation costs. Field connections of various types are required to provide structural or watertight jointing. *See* CONCRETE. Christine Beall

Bibliography. G. Balaguru, *Concrete Handbook*, 1997; F. S. Merritt, *Building Construction Handbook*, 6th ed., 2000.

Precession

The motion of an axis fixed in a body around a direction fixed in space. If the angle between the two is constant so that the axis sweeps out a circular cone, the motion is pure precession; oscillation of the angle is called nutation. An example of precession is the motion of the Earth's polar axis around the normal to the plane of the ecliptic; this is the precession of the equinoxes. A fast-spinning top, with nonvertical axis, which precesses slowly around the vertical direction, is another example. In both examples the precession is due to torque acting on the body. Another kind of precession, called free or fast precession, with a rate which is comparable to the

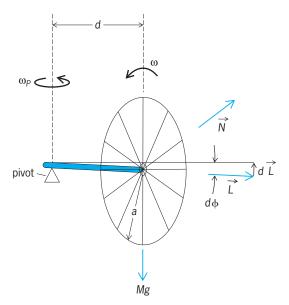


Fig. 1. Simple precession of a rapidly spinning wheel with a horizontal axis supported by a pivot.

rotation rate of the body, is seen, for instance, in a coin spun into the air. *See* NUTATION (ASTRONOMY AND MECHANICS); PRECESSION OF EQUINOXES.

Gyroscopic motion. As a simple example of gyroscopic motion, consider a rapidly spinning wheel with a horizontal axis supported at a distance *d* from the plane of the wheel (**Fig. 1**). The angular momentum \vec{L} is along the wheel symmetry axis and is approximately given by the angular momentum of the wheel about this axis; in the simple precession approximation to the motion, the angular momentum associated with precessional motion is neglected. The external torque \vec{n} due to the gravitational force is perpendicular to the wheel axis in the horizontal plane. The change in the angular momentum \vec{L} in an infinitesimal time interval dt is given by the rotational equation of motion in Eq. (1).

$$d\vec{L} = \vec{N} \, dt \tag{1}$$

For the derivation of Eq. (1) *see* RIGID-BODY DY-NAMICS.

Since $d\vec{L}$ and \vec{L} are perpendicular, the length \vec{L} is unchanged to first order in dt. The direction of \vec{L} is rotated counterclockwise in the horizontal plane through an angle $d\phi$ given by Eq. (2). The angular

$$d\phi = \frac{dL}{L} = \frac{N}{L}dt \tag{2}$$

velocity of precession ω_P about the vertical axis is then given by Eq. (3).

$$\omega_P = \frac{d\phi}{dt} = \frac{N}{L} \tag{3}$$

The magnitude of the gravitation torque is *Mgd*, where *M* is the mass of the wheel and *g* is the gravitational acceleration. The spin angular momentum is $L = I\omega$, where ω is the spin angular velocity and

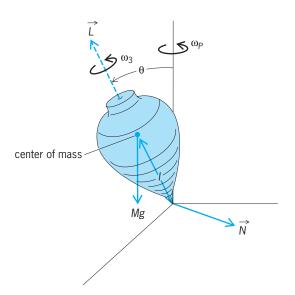


Fig. 2. Simple precession of a fast-spinning top about the vertical axis.

 $I = Ma^2$ is the moment of inertia of the wheel of radius *a*. Thus ω_P , as given in Eq. (4), is independent

$$\omega_P = \frac{gd}{\omega a^2} \tag{4}$$

of *M* and inversely proportional to ω . For ω very large, the precession rate ω_P is quite slow. *See* GYRO-SCOPE.

Spinning top. The precessional motion of a top inclined at an angle θ from the vertical and having a fixed point of contact with a supporting surface (**Fig. 2**) can be readily deduced from the preceding discussion. For a fast-spinning top the angular momentum is nearly along the symmetry axis of the top, and is given by Eq. (5) where I_3 is the moment of

$$L \simeq I_3 \omega_3 \tag{5}$$

inertia about this axis and ω_3 is the angular velocity. The change in \vec{L} is given by Eq. (1) with the torque \vec{n} along the azimuthal direction. Again the magnitude of \vec{L} is unchanged to first order in dt, and from Fig. 2 the change in azimuthal angle ϕ is given by Eq. (6). The magnitude of N is $Mg\ell \sin \theta$, where ℓ

$$d\phi = \frac{dL}{L\sin\theta} = \frac{Ndt}{L\sin\theta} \tag{6}$$

is the distance of the center of mass from the pivot. Hence the precession angular velocity of Eq. (7) is

$$\omega_P = \frac{d\phi}{dt} \simeq \frac{Mg\ell}{I_3\omega_3} \tag{7}$$

obtained. In addition to the slow precession motion of a fast top, the polar angle θ of the symmetry axis may also undergo an oscillating motion, nutation.

Rising top. When a spinning top with a blunt peg end is set down on a rough surface, the top usually

slips initially. Due to the frictional force at the peg, there is a frictional torque about the center of mass of the top of approximate magnitude $N \approx \mu Mg\ell$, which is perpendicular to the peg for a thin peg. Here μ is the coefficient of friction and ℓ is the distance from the center of mass to the peg. The torque is perpendicular to the symmetry axis, and hence to \vec{L} for a fast top, in the direction of decreasing polar angle θ . Thus the frictional torque causes the top to precess toward the vertical with angular velocity θ given by Eq. (8).

$$\dot{\theta} = -\frac{N}{L} \approx -\frac{\mu M g \ell}{I_3 \omega_3} \tag{8}$$

Boomerang. The return of a boomerang is explained by precession. The cross-blade boomerang consists essentially of two perpendicular fan blades that intersect at the boomerang's center of mass. As the boomerang is thrown in a vertical plane, the blades rotate rapidly about the center of mass with the center of mass moving parallel to the ground. Due to its spin the boomerang has an angular momentum about the center of mass that is perpendicular to its plane.

The aerodynamic force on the airfoils of the blades gives a net force and torque on the boomerang. (The torque comes from the higher forces on the upper blades which are moving faster through the air.) The force accelerates the boomerang perpendicular to the plane of rotation, in the direction of \vec{L} . The torque causes the plane of the boomerang to precess about the vertical. The net effect of the force and torque is thus to make the center of mass of the boomerang move in a circle.

Free or fast precession. An axial symmetric body has two of its principal moments of inertia equal, say $I_1 = I_2$; \vec{n}_3 is chosen to be the principal axis which is the symmetry axis. Then the angular momentum \vec{L} about the center of mass of the body takes the form Eq. (9), expressed in terms of the body axes \vec{n}_1 ,

$$\vec{L} = I_1(\omega_1 \vec{n}_1 + \omega_2 \vec{n}_2) + I_3 \omega_3 \vec{n}_3 \tag{9}$$

 \vec{n}_2 , \vec{n}_3 and the body-axis components of the angular velocity ω ($\omega \cdot \vec{n}_1 = \omega_1$, and so forth). The motion of \vec{n}_3 is given by Eq. (10), where the first equality is

$$\dot{\vec{n}}_3 = \vec{\omega} \times \vec{n}_3 = \left(\frac{\vec{L}}{I_1}\right) \times \vec{n}_3 \tag{10}$$

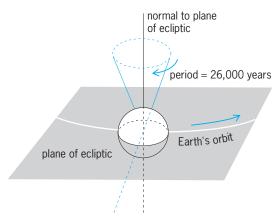
by definition of ω , and the second is true because ω and \vec{L}/I_1 differ only by a vector parallel to \vec{n}_3 . If the body is free, \vec{L} is constant, and so Eq. (10) shows that the symmetry axis \vec{n}_3 precesses around a fixed direction, namely the direction of \vec{L} , at a rate L/I_1 . As an example consider a coin thrown in the air, spinning nearly in its plane. The precession of its symmetry axis, that is, the normal to the plane of the coin, is seen as a wobble; the ratio of the rate of the wobble to the rate of spin is $(L/I_1)/(L/I_3) =$ $I_3/I_1 = 2$. An ordinary top, when first thrown, may exhibit fast precession, similar to what its motion would have been if free, because the effect of gravity and the contact with the ground produces only an additional slow precession. So-called fast precession will not be fast if the spin of the body is not fast. An example is given by a rolling coin in the late stage of its motion when it is "lying down." It exhibits a fast wobble (infinitely fast in the limit of lying flat) which is not a fast precession, being much faster than the spin rate of the coin (as can be seen from the rotation of the emblem on its face), and is instead a precession driven by torque. *See* ANGULAR MOMEN-TUM; RIGID-BODY DYNAMICS. Vernon D. Barger

Bibliography. A. P. Arya, *Introduction to Classical Mechanics*, 2d ed., 1997; V. Barger and M. Olsson, *Classical Mechanics: A Modern Perspective*, 2d ed., 1995; H. Goldstein, C. P. Poole, and J. L. Safko, *Classical Mechanics*, 3d ed., 2002; D. Halliday, R. Resnick, and K. Krane, *Physics*, 5th ed., 2002; K. Symon, *Mechanics*, 3d ed., 1971.

Precession of equinoxes

A slow change in the direction of the axis of rotation of the Earth and also a very small change in the orientation of the Earth's orbit around the Sun. Thus, both the celestial equator (the projection of the terrestrial equator onto the sky) and the ecliptic (the projection of the Earth's orbit onto the sky) move slowly in time and, consequently, so do their intersections, the vernal and autumnal equinoxes. *See* ECLIPTIC; EQUATOR; EQUINOX.

Earth's axis. The change in direction of the Earth's axis of rotation is caused primarily by the Moon and secondarily by the Sun. The Earth is an oblate spheroid, bulging outward around the Equator by about 1 part in 300. Because the Earth's axis of rotation is tilted by about 23.4° to the plane of its orbit about the Sun, this bulge is also tilted. The extra gravitational pull of the Moon and Sun on this bulge produces a revolution, or precession, of the axis of rotation about the perpendicular to the orbit (see **illustration**); one complete revolution requires just



Motion of Earth's axis of rotation.

under 26,000 years. This has been recognized for over 2000 years and was formerly called lunisolar precession. *See* PRECESSION.

In addition to uniform revolution, there are very small amplitude periodic oscillations; the primary one has a period of 18.6 years. These oscillations are collectively called nutation. They are caused by the same forces as precession, but are treated observationally and computationally as separate phenomena. *See* NUTATION (ASTRONOMY AND MECHAN-ICS).

Earth's orbit. The slow change in the Earth's orbital plane (ecliptic) is caused by the gravitational perturbations of the other planets in the solar system. The principal effect is to produce an extremely small rotation of the axis of revolution of the Earth around an axis parallel to the total-angular-momentum axis of the solar system. This therefore produces a very small motion of the ecliptic itself and thus an additional small change in the apparent position of the equinox. This precession was formerly called planetary precession.

Effects. Formulas for precession and nutation are used to give the direction of the Earth's spin axis (pole) in space, while the sidereal time formula is used to give the Earth's rotation about the pole. The combination orients the Earth in inertial space and thus determines the positions of observing telescopes, antennas, and so forth. *See* SIDEREAL TIME.

Such information is necessary for telescope pointing; it is crucial for spacecraft navigation. In fact, even the extensive formulations for precession, nutation, and sidereal time are not adequate for the most precise navigation. The orientation of the Earth is not fully predictable due to earthquakes, volcanoes, atmospheric loading, polar cap melting, and so forth. As a result, it is necessary to continually measure the Earth's exact orientation on a daily basis, collect the measurements, and immediately distribute the information worldwide.

In former times, the Earth's pole, equator, and mean orbital path (ecliptic) were extended out into space in order to define the basic celestial coordinate system. Nowadays, the basic coordinate system, though oriented close to the Earth's equator in the year 2000, is actually defined by the International Celestial Reference Frame (ICRF), a specific catalog of radio sources (quasars and so forth). *See* CELESTIAL REFERENCE SYSTEM.

People in the Northern Hemisphere have been lucky over the past few centuries, for they have had a "north star." The north celestial pole, the direction of the Earth's north pole in space, has been close to the star Polaris. It was 2.3° away in A.D. 1700, was 0.7° away in 2000, and will pass Polaris at about 0.5° in 2100. In the future, the north celestial pole will not come within 5° of any bright star (magnitude <3) until it completes its 26,000-year circuit and comes again near to Polaris. For the south pole, it is just the opposite. At present, the south celestial pole is not within 12° of any bright star, but it will pass within 1° of bright stars during the (approximate) years 6720-6970, 7930-8290, 9070-9430, and 24160-24470. *See* ASTRONOMICAL COORDINATE SYSTEMS; EARTH ROTATION AND ORBITAL MOTION; POLARIS. E. Myles Standish; Robert S. Harrington

Bibliography. K. Lambeck, *The Earth's Variable Rotation*, 1980; H. Moritz and I. I. Mueller, *Earth Rotation*, 1987; P. K. Seidelmann (ed.), *Explanatory Supplement to the Astronomical Almanac*, 1992; F. R. Stphenson, *Historical Eclipses and the Earth's Rotation*, Cambridge University Press, 1997.

Precious stones

The materials found in nature that are used frequently as gemstones, including amber, beryl (emerald and aquamarine), chrysoberyl (cat's-eye and alexandrite), coral, corundum (ruby and sapphire), diamond, feldspar (moonstone and amazonite), garnet (almandite, demantoid and pyrope), jade (jadeite and nephrite), jet, lapis lazuli, malachite, opal, pearl, peridot, quartz (amethyst, citrine, and agate), spinel, spodumene (kunzite), topaz, tourmaline, turquois, and zircon.

The terms precious and semiprecious have been used to differentiate between gemstones on a basis of relative value. Because there is a continuous gradation of values from materials sold by the pound to those valued at many thousands of dollars per carat, and because the same mineral may furnish both, a division is essentially meaningless. All but a few of the gems listed cost many dollars per carat in fine quality. Thus, almost all gemstones merit the use of the term precious. *See* GEM. Richard T. Liddicoat, Jr.

Precipitation (chemistry)

The process of producing a separable solid phase within a liquid medium. In a broad sense, precipitation represents the formation of a new condensed phase, although other terms are often used to describe the process. Thus (1) a vapor or gas condenses to liquid droplets, or more specifically as in meteorology, water vapor in the atmosphere precipitates to form rain, snow, or ice; (2) a substance in the liquid state freezes or solidifies; (3) a dissolved component crystallizes from a supersaturated solution; (4) a new solid phase gradually precipitates within a solid alloy as the result of a slow, inner chemical reaction; or (5) a metal electrodeposits upon the passage of an electrical current through a solution.

In analytical chemistry, precipitation is widely used to effect the separation of a solid phase in an aqueous solution. For example, the addition of a water solution of silver nitrate to a water solution of sodium chloride results in the formation of insoluble silver chloride. Quite often, one of the components in the solution is thus virtually completely separated in a relatively pure form. It can then be isolated from the solution phase by filtration or centrifugation, and the substance determined by weighing. This procedure is known as gravimetric analysis. Precipitation may also be used merely to effect partial or complete separation of a substance for purposes other than that of gravimetric analysis. Such purposes might involve either the isolation of a relatively pure substance or the removal of undesirable components of the solution. *See* GRAVIMETRIC ANALYSIS.

Solubility product constant. The extent to which a component can be separated from solution can be determined from the solubility product constant obtained by determining the quantity of dissolved substance present in a known amount of saturated solution. This value is known as the solubility. The solubility can be drastically altered merely by adding to the solution any of the ions that make up the precipitate, for example, by adding varying quantities of either silver nitrate or sodium chloride to a saturated solution of silver chloride. Although solubility can be altered over a wide range, the solubility product itself remains practically constant over this same range.

The solubility product constant can be used to ascertain the quantity of dissolved component remaining unprecipitated in the presence of known concentrations of the ions common to the precipitate. By proper adjustment of the concentration of the added common ion, it is possible to reduce the quantity of dissolved component to a negligible value, although never to zero. It is an extremely important criterion of a method of gravimetric analysis that the quantity of unprecipitated component be negligible, particularly in comparison with the quantity of precipitate formed. The analytical chemist uses the word quantitative to describe such a chemical reaction. *See* SOL-UBILITY PRODUCT CONSTANT.

Impurities. The purity of the precipitated solid phase is of major concern, both for the preparation of a desired chemical compound and for a quantitative method of gravimetric analysis. It is not possible for a precipitate to be formed as an absolutely pure compound by chemical reaction within the solution phase. Other soluble substances present in the solution, such as the ions not involved in the structure of the precipitate, tend to accompany the solid phase in varying amounts. This phenomenon is known as coprecipitation. The fraction of the total quantity of such foreign ions coprecipitating may be quite small. Although this fraction depends on experimental variables, it is highly dependent upon the relationship between the solubility characteristics of the desired chemical precipitate and the foreign substance. As a specific example, partial precipitation of iodide (as silver iodide) using silver nitrate would coprecipitate (as silver chloride) only a small fraction of any chloride present, whereas it would coprecipitate (as silver bromide) a larger fraction of any bromide present. The iodide is more insoluble than the bromide, which is more insoluble than the chloride. Knowledge of the relative solubility characteristics of the chemical species present is thus extremely desirable to the chemist, who needs to know whether foreign substances are being collected with the solid or are being left in solution.

A precipitated phase may incorporate foreign ions within its structure in several ways. Best understood of these is isomorphous mixed-crystal formation. Radium and barium sulfates form isomorphous mixed crystals because the two compounds have the same crystal structure and the ionic radii of radium and barium are not greatly different. Thus, radium and barium ions are interchangeable within the crystal lattice to a considerable degree. Silver bromide and silver chloride also form isomorphous mixed crystals. Because of the ease with which interchange can take place within the crystal lattice, isomorphous mixed crystal systems should be avoided if a good separation of the two different ionic species is desired. On the other hand, the property of isomorphism may also be put to good use in concentrating minute traces. For example, barium sulfate precipitated in the presence of minute traces of radium carries with it almost all of the radium. Such procedures are frequently used in the collection of minute traces of radioactive species.

An ion present at high dilution may sometimes be incorporated, apparently by mixed crystal formation, even though such formation would not be predicted on the basis of crystallography and ionic radii. An example of this is the coprecipitation of traces of lead with potassium chloride. This phenomenon is known as anomalous mixed crystal formation, or isodimorphism.

When foreign-ion incorporation cannot be ascribed to isomorphism, coprecipitation may occur by adsorption. Residual charges at the surface of a precipitate attract charged ions in the solution. The adsorption process results in a greater concentration of foreign ions near the precipitate surface than exists in the main body of the solution. Adsorbed foreign ions may remain quite firmly attached to the solid. In fact, they may be covered as succeeding layers of the crystal are deposited and cause imperfections within it. This phenomenon is known as occlusion, although it arises as a result of adsorption. The term occlusion should be distinguished from inclusion, which refers to the mechanical trapping of pockets of solution (and solutes in it) within the precipitate.

Sometimes one substance in a mixture precipitates rapidly, but a second foreign substance then precipitates slowly as a second solid phase. This is not generally considered to be coprecipitation, but is referred to as postprecipitation. The postprecipitation of zinc sulfide with copper sulfide is a typical example.

Methods for reducing contamination. In an effort to reduce contamination by foreign ions, the chemist resorts to various techniques. Precipitation from dilute solution is often effective. Heating the reaction mixture, that is, digesting the solid in contact with

the liquid phase, speeds recrystallization processes by which incorporated foreign ions may be returned to the solution phase. Precipitation from homogeneous solution, a technique in which the desired precipitating reagent is formed internally within the solution by chemical synthesis, results in the slow formation of large crystals of small surface area and hence lessens coprecipitation.

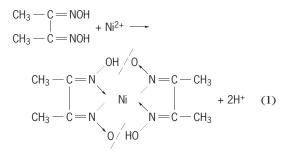
If all these methods fail to reduce adequately the quantity of foreign ions incorporated in the solid phase, then reprecipitation is applied. The precipitate is dissolved and reprecipitated by the previous procedure. As in the initial precipitation, most of the foreign ions remain in solution. The process of reprecipitation must be repeated until the quantity of foreign ions present in the precipitate can be disregarded. *See* ADSORPTION; CHEMICAL SEPARATION TECHNIQUES; CRYSTALLIZATION; NUCLE-ATION.

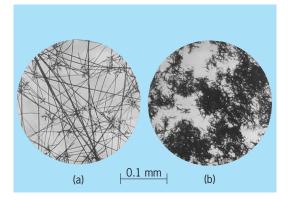
Precipitation from homogeneous solution. In this technique the precipitant is generated in place within the solution phase instead of being added directly, as in the conventional manner.

The substances required to effect precipitation from homogeneous solution can be generated within a solution phase in a variety of ways. For example, if a source of hydroxyl ions is needed to precipitate metallic ions, urea in solution can be hydrolyzed to produce ammonium hydroxide to act as the source. Other anions needed as precipitants can often be formed by the hydrolysis of appropriate esters. Thus, dimethyloxalate can serve as a source of oxalate ions to precipitate thorium, and dimethyl sulfate can be the source of sulfate ions to precipitate lead. There also are methods for the generation of cations. For example, silver ions can be slowly released from a complex ion such as $Ag(NH_3)_2^+$ to precipitate chloride ions. Various investigators have developed ingenious methods for producing the necessary precipitant to effect precipitation from homogeneous solution.

The formation of larger, more perfect, and purer crystals occurs almost without exception when precipitation from homogeneous solution is used in place of direct addition of the precipitant.

The **illustration** shows the difference in appearance of nickel dimethylglyoximate precipitated by two different methods. The precipitate shown in illus. *b* was formed by the direct addition of dimethylglyoxime to a solution containing nickel ions, as in reaction (1). The more perfect crystals shown in





Photomicrographs of nickel dimethylglyoximate precipitated by two methods. (a) Precipitation from homogeneous solution. (b) Direct addition to solution. (From L. Gordon and E. D. Salesin, Precipitation from homogeneous solution, J. Chem. Educ., 38(1):16, 1961)

illus. *a* were precipitated from homogeneous solution by synthesizing the necessary dimethylglyoxime from biacetyl and hydroxylamine by reaction (2).

$$\begin{array}{c} \mathsf{CH}_3-\mathsf{C}=\mathsf{0} \\ | \\ \mathsf{CH}_3-\mathsf{C}=\mathsf{0} \end{array} \xrightarrow{\begin{array}{c} \mathsf{CH}_3-\mathsf{C}=\mathsf{NOH} \\ | \\ \mathsf{CH}_3-\mathsf{C}=\mathsf{NOH} \end{array}} + 2\mathsf{H}_2\mathsf{O}$$

Because the rate at which a precipitant is generated can be controlled very closely, precipitation from homogeneous solution is used as a research technique in studies of the mechanisms of nucleation, precipitation, and coprecipitation. The technique has many applications in gravimetric analysis, in the production of pure chemicals, and in the control of particle size of crystals. Louis Gordon; Royce W. Murray

Bibliography. Center for Occupational Research and Development Staff, *Gravimetric and Volumetric Analysis*, 1985; A. G. Walton, *The Formation and Properties of Precipitates*, 1967, reprint 1979.

Precipitation (meteorology)

The fallout of waterdrops or frozen particles from the atmosphere. Liquid types are rain or drizzle, and frozen types are snow, hail, small hail, ice pellets (also called ice grains; in the United States, sleet), snow pellets (graupel, soft hail), snow grains, ice needles, and ice crystals. In England sleet is defined as a mixture of rain and snow, or melting snow. Deposits of dew, frost, or rime, and moisture collected from fog are occasionally also classed as precipitation.

All precipitation types are called hydrometeors, of which additional forms are clouds, fog, wet haze, mist, blowing snow, and spray. Whenever rain or drizzle freezes on contact with the ground to form a solid coating of ice, it is called freezing rain, freezing drizzle, or glazed frost; it is also called an ice storm or a glaze storm, and sometimes is popularly known as silver thaw or erroneously as a sleet storm. *See* CLOUD. Most precipitation particles carry an electrostatic charge, either positive or negative, but the origin of the charge and its relation to other problems of atmospheric electricity are not completely understood. *See* ATMOSPHERIC ELECTRICITY; THUNDERSTORM.

Rain, snow, or ice pellets may fall steadily or in showers. Steady precipitation may be intermittent though lacking sudden bursts of intensity. Hail, small hail, and snow pellets occur only in showers; drizzle, snow grains, and ice crystals occur as steady precipitation. Showers originate from instability clouds of the cumulus family, whereas steady precipitation originates from stratiform clouds.

The amount of precipitation, often referred to as precipitation or simply as rainfall, is measured in a collection gage. It is the actual depth of liquid water which has fallen on the ground, after frozen forms have been melted, and is recorded in millimeters (mm) or hundredths of inches. A separate measurement is made of the depth of unmelted snow, hail, or other frozen forms. *See* PRECIPITATION MEASURE-MENT.

Liquid types. Rain and drizzle are somewhat arbitrarily differentiated. Raindrops are generally over 0.02 in. (0.5 mm) in diameter and differ from drizzle mainly in having larger sizes. Drizzle may also be distinguished from rain by the meteorological conditions of its formation; drizzle falls usually from fog or thick stratus clouds, whereas rain comes from clouds of cumulus type or clouds extending well above the freezing level. Raindrops are rarely larger than 0.2 in. (5 mm) in diameter because larger sizes tend to break up.

The small raindrops are approximately spherical, but large falling drops are flattened (**Fig. 1**), especially on the bottom side. The following speeds of falling drops, in meters per second, were measured in still air by R. Gunn (drop diameters are in millimeters):

Speed	0.27	2.06	4.03	6.49	8.06	8.83
Diameter	0.1	0.5	1.0	2.0	3.0	4.0

Frozen precipitation. Snowflakes are branched, sixpoint star crystals, irregular forms such as matted ice needles, or combinations of both; often they are coated by rime. The speed of falling flakes is variable; it is greater when the flakes are rimed and is generally 2-4 mi/h (1-2 m/s). *See* SNOW.

Hail, mostly seen in thunderstorms, forms with the aid of warm, moist air when clouds build to great heights. Hailstones usually have concentric layers of rime and hard ice, typically 0.4 in. (1 cm) in diameter but ranging up to more than 4 in. (10 cm). *See* HAIL.

Ice pellets are of hard ice, either clear or opaque, and may be spherical or irregular. They are formed by the freezing of drops of rain or drizzle.

Snow pellets are of soft, opaque ice, irregularly shaped, sometimes with scalloped edges and often studded with a few oblong or branched crystals. Typical diameters are 0.08–0.2 in. (2–5 mm). They fall



Fig. 1. Various forms of precipitation.

usually in showers, sometimes even when ground temperatures are a few degrees above freezing. U. Nakaya found their specific gravity to be about 0.12. Snow grains are similar but smaller and flatter, with diameters around 1 mm or less; they fall generally in small amounts from stratus clouds or fog.

Ice needles are narrow, pointed crystals, roughly

0.04-0.012 in. (1-3 mm) long and about 0.01 in. $(^{1}/_{4} \text{ mm})$ in diameter, which fall singly or in clusters, mostly at temperatures near or a little below freezing.

Ice crystals are small rods or plates. They fall in cold, stable air that is often without clouds, and glisten in light as they settle to the ground.

Geographical distribution. Precipitation is part of the hydrologic cycle, the continuous interchange of water between sea, land, and atmosphere, but its distribution over the Earth is uneven (**Fig. 2**).

Some world-record rainfall amounts are, for 1 min, 1.23 in., Unionville, Maryland, 1956; 1 h, 12.00 in. (305 mm), Holt, Missouri, 1947; 12 h, 52.76 in. (1340 mm), Belouve, La Réunion, 1964; 24 h, 73.62 in. (1870 mm), Cilaos, La Réunion, 1952; 2 days, 98.42 in. (2500 mm), Cilaos, La Réunion, 1952; 31 days, 366.14 in. (9300 mm), Cherrapunji, India, 1861; and 1 year, 1041.78 in. (26,461 mm), 2 years, 1605.05 in. (40,768 mm), both Cherrapunji, India, 1860-1861.

Precipitation and weather. Precipitation occurs most often in cyclones or tropical disturbances. In weather forecasting, several synoptic (weather map) types of precipitation are recognized: warm front, warm moist air rising over a wedge of cold air; cold front, cold air undercutting and lifting warmer air; convective, caused by local updrafts of moist air; convergent, general lifting of air caused by convergence at low levels; and orographic, moist air forced upward on mountain slopes. Any type may act alone; however, convection or convergence may occur with other types. Lake snow falls when air, at first extremely cold, blows off a bay or off lakes such as the Great Lakes. Minor causes of precipitation

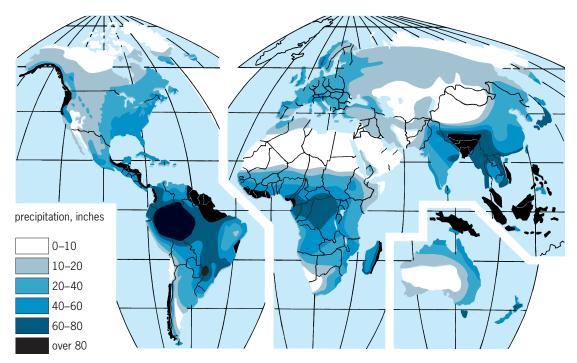


Fig. 2. Map showing the distribution of average annual precipitation over the land areas of the world. 1 in. = 25 mm. (After N. A. Bengtson and W. Van Royen, Fundamentals of Economic Geography, 4th ed., Prentice-Hall, 1956)

are turbulence, contact cooling (air moving over a colder surface), and nighttime radiation from cloud tops, but the amounts of precipitation from these sources are small.

Condensation in the atmosphere. Basically, this requires that air be cooled to and below its dew point. Then minute waterdroplets form by condensation, or ice crystals form by sublimation. These droplets, or ice crystals, are cloud particles from which, by a growth mechanism, the larger precipitation particles may form. The only cooling process sufficient to produce appreciable precipitation is adiabatic expansion, which occurs in air rising toward lower pressures. The rate of condensation in rising saturated air is greater the warmer the air and is directly proportional to the speed of ascent. For discussions of condensation nuclei, formation of precipitation *see* CLOUD PHYSICS; WEATHER MODIFICATION.

For discussions of other topics related to precipitation *see* ATMOSPHERIC GENERAL CIRCULATION; DEW; DEW POINT; FOG; HUMIDITY; HYDROLOGY; HY-DROMETEOROLOGY; METEOROLOGY; RAIN SHADOW; TROPICAL METEOROLOGY; VAPOR PRESSURE.

J. R. Fulks

Bibliography. W. R. Cotton and R. A. Anthes, *Storm* and Cloud Dynamics, 1989; F. K. Lutgens and E. J. Tarbuck, *The Atmosphere: An Introduction to Me teorology*, 7th ed., 1997; B. J. Mason, *Clouds, Rain,* and Rainmaking, 2d ed., 1976; A. Miller and J. C. Thompson, *Elements of Meteorology*, 6th ed., 1999; J. M. Moran and M. D. Moran, *Meteorology: The At* mosphere and the Science of Weather, 5th ed., 1996; H. R. Pruppacher and J. D. Klett, *Microphysics of Clouds and Precipitation*, 1980; C. Riegel, *Funda*mentals of Atmospheric Dynamics and Thermodynamics, 1993.

Precipitation measurement

Instruments used to measure the amount of rain or snow that falls on a level surface. Such measurements are made with instruments known as precipitation gages. A precipitation gage can be as simple as an open container on the ground to collect rain, snow, and hail; it is usually more complex, however, because of the need to avoid wind effects, enhance accuracy and resolution, and make a measurement representative of a large area. Precipitation is measured as the depth to which a flat horizontal surface would have been covered per unit time if no water were lost by runoff, evaporation, or percolation. Depth is expressed in inches or millimeters, typically per day. The unit of time is often understood and not stated explicitly. Snow and hail are converted to equivalent depth of liquid water.

Wind effects. In strong winds the gage catch is reduced by turbulence around the opening. The ideal exposure for a precipitation gage is an area that is free from obstructions that would create large eddies that deflect the flow, and where the winds are



Fig. 1. Automatic gages. (a) Tipping-bucket rain gage. (b) Weighing rain gage. (Belfort Instrument)

light to allow rain and snow to fall vertically. These conditions can be approximated reasonably well for large raindrops but are nearly impossible to achieve for snow. A gage in the lee of an obstruction could over- or undercollect precipitation, depending upon the obstruction, the wind speed, and the particle size. The usual exposure is to set the gage orifice a few feet above ground and well away from obstructions. Sometimes a special wind screen placed around the gage creates turbulence in a form tending to minimize systematic wind bias. Precipitation gages should never be placed around or on a building.

Accuracy and resolution. The funnel of a manual gage directs the precipitation into a tube whose cross-sectional area is one-tenth that of the gage orifice. The depth of precipitation is exaggerated and can be measured more readily with a special dip stock having an expanded scale. Automatic gages use a weighing mechanism or a tipping bucket as shown in Fig. 1. The tipping bucket fills and tips for each 0.01 in. (0.25 mm). The greatest source of error, excluding representativeness of the sample, is exposure, that is, wind and obstruction effects that can reduce the gage catch 5-80%. Exposure problems are compounded with snow, which may block the gage orifice or freeze the mechanism. Gage accuracy specifications ignore problems of exposure and representativeness.

Representativeness. Rain and snow showers are local phenomena; a heavy rain shower may completely miss a nearby precipitation gage. The estimate of precipitation over a large area may be improved by building a dense network. An alternative is to use a radar to supplement ground-based gages. A weather radar can detect precipitation over a large area, typically over a circular area having a 190-mi (300-km) radius. The optimum combination is a weather radar operating in conjunction with a network of gages that report in real time. This would allow prompt reporting of representative precipitation over a large area, which is essential in

predicting flash floods. *See* METEOROLOGICAL IN-STRUMENTATION; PRECIPITATION (METEOROLOGY); SNOW SURVEYING. Fred V. Brock

Radar. Accurate quantitative precipitation measurement is probably the most important weather radar application. It is extremely valuable for hydrological applications such as watershed management and flash flood warnings. Radar can make rapid and spatially contiguous measurements over vast areas of a watershed at relatively low cost. However, radar detection of two equal-diameter raindrops will return less echo power than one raindrop containing the same amount of water. Thus, even though echo power is correlated with precipitation amounts, there is no unique relation between the two; other data are needed to adjust this relation. For example, different precipitation types (such as stratiform rain, thunderstorm rain, and snow) require different relations between the precipitation rate Q and reflectivity factor Z, a radar-derived parameter calculated from measurements of echo power. The relation between rainfall rate $R \equiv Q \pmod{h^{-1}}$ and Z (mm⁶ m⁻³), expressed as $Z = 200 R^{1.6}$, has proven to be quite useful for stratiform rain, where Z is proportional to the rain's backscatter cross section per unit volume. Other relations are used to relate radar measurements of Z to the fall rate of other precipitation types. Rain gages, sparsely deployed over a watershed, can also be used to tune these relations to improve the quantitative measurement of precipitation. See HYDROMETEOROLOGY; METEOROLOGICAL RADAR; PRECIPITATION (METEOROLOGY); RADAR ME-TEOROLOGY; STORM DETECTION.

Nevertheless, Q, Z relations are subject to many sources of error, including (1) radar calibration, (2) attenuation of the echoes, (3) blockage of the radar beam (by buildings, trees), (4) echoes from ground objects, (5) variations in the pedicle size distribution, and (6) beam filling (the extent that the radar beam is uniformly filled with precipitation). Although considerable effort has been expended to improve radar estimates of precipitation, the basic way in which precipitation is measured has not changed. All of these error sources either go away or can be mitigated by the use of polarimetric radar.

Polarimetric weather radar transmits and receives power in two states of wave polarization (horizontal and vertical), whereas weather radars, operated by weather services into the twenty-first century, transmit and receive waves only in a single polarization state (for example, horizontal). The relation between the specific differential phase K_{DP} (the differential propagation phase shift per unit length between vertically and horizontally polarized waves) and rain rate measured with a polarimetric radar is relatively insensitive to variations in drop size distribution and can also form the basis of rain measurement using an R, K_{DP} relation.

The addition of polarimetric capability to operational radars will provide a revolutionary approach to the radar measurement of rainfall. Polarimetric radars may significantly increase the accuracy of moderate-to-heavy rainfall rate measurements, lead

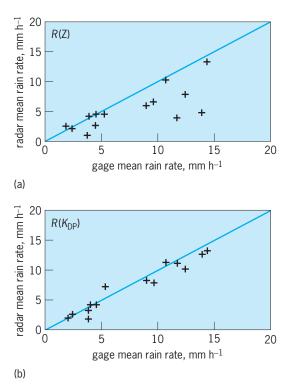


Fig. 2. Comparison of rainfall rates measured with gages and radar using (a) the R, Z relation and (b) the R, K_{DP} relation. Each data point (+) is the rainfall rate for a single storm.

to issuance of more timely and accurate flash flood warnings, and provide other improved hydrological products to agricultural and commercial enterprises.

An *R*, K_{DP} relation has been tested on data collected with a research polarimetric weather radar operated by the National Severe Storms Laboratory in Norman, Oklahoma. Comparisons were made with rainfall R(G) measured by a dense network of 42 rain gages for 15 rain events; R(G) is the rain gage estimate of the accumulated rainfall over the area of the gages. The rainfall rates were averaged over the duration of the events lasting from about 1 to 4 hours. Rainfall measured using the R, K_{DP} relation showed significant accuracy improvement over that measured using the R, Z relation (**Fig. 2**). Richard J. Doviak

Bibliography. F. Dobson, L. Hasse, and R. Davis (eds.), Air-Sea Interaction: Instruments and Methods, 1980; R. J. Doviak and D. S. Zrnić, Doppler Radar and Weather Observations, 2d ed., 1993; D. H. Lenschou (ed.), Probing the Atmospheric Boundary Layer, American Meteorological Society, 1986; Meteorological Office, Handbook of Meteorological Instruments, vol. 5: Measurement of Precipitation and Evaporation, 2d ed., 1981; M. Sachidananda and D. S. Zrnić, Differential propagation phase shift and rainfall rate estimation, Radio Sci., 21:235-247, 1986; J. T. Snow and S. B. Harley, Basic meteorological observations for schools: Rainfall, Bull. Amer. Meteorol. Soc., 69(5):497-507, 1988; D. S. Zrnić and A. Ryzhkov, Advantages of rain measurements using specific differential phase., J. Atm. Oceanic Technol., 13(2):454-464, 1996.

Precipitin

The visible result of the chemical interaction of antigen and antibody. Such precipitate was first described in 1897, and it was later noted that the optimal ratio of antigen to antibody concentrations resulted in the visible precipitate. Not all antibodies will result in precipitation, yet they may participate in agglutination reactions or add onto particulate antigens, and evidence for their occurrence together with precipitating antibody can be obtained for most sera. Precipitins may be noted qualitatively or be quantified by noting the end-point dilution (titer) of serum required to give a precipitate at the threshold of visiblity, or the amount of antibody may be determined in milligrams or micrograms by analysis of the precipitate with correction for the antigen contained therein. See ANTIBODY; ANTIGEN; PRECIPITIN **Donald Raum** TEST.

Bibliography. J. M. Cruse and R. E. Lewis, Jr. (eds.), *The Year in Immunology 1989-90*, 1990; F. Dixon (ed.), *Advances in Immunology*, vol. 48, 1990; J. W. Kimball, *Introduction to Immunology*, 3d ed., 1990; D. M. Weir, *Immunology: An Outline for Students of Medicine and Biology*, 6th ed., 1987.

Precipitin test

A test to measure a specific reaction between antigen and antibody which results in a visible precipitate. *See* PRECIPITIN.

Qualitative tests. Two types of qualitative tests are commonly used. The first is the ring test. A solution of antigen is laid over a solution of antibody or antiserum so that a sharp interface is formed. A precipitate at the interface is a positive reaction and occurs as antigen and antibody diffuse toward one another, creating a zone in which the ratio of their concentrations is optimal for precipitation. As little as 1.0 microgram of protein may be detected. This is an easy, rapid test. The second qualitative test involves mixing solutions of antigen and antibody. This is done by varying either the concentration of antibody or the concentration of antigen. When the amount of antigen is varied, three zones of precipitation are noted: antigen excess, equivalence, and antibody excess.

Quantitative determination. The quantitative determination of precipitating antibodies depends upon the two-stage nature of antigen-antibody interaction. The first stage of antigen-antibody combination reaches completion within a few minutes, while the second stage, in which aggregation of complexes occurs, may take days. To standardize these tests, it is necessary to consider time, temperature, and hydrogen ion and electrolyte concentration. Precipitates are digested and analyzed for nitrogen content, and the amount of antibody calculated. Since complement solubilizes antigen-antibody complexes, spuriously low values will be obtained if the test is performed in the presence of fresh serum. *See* COM-PLEMENT.

Modifications. A number of modifications of the precipitin test allow one to test multiple samples and to determine whether the antigen-antibody system is homogeneous.

In the Oudin method, antigen and antibody are each held in an agar matrix in a test tube and allowed to diffuse toward one another. The Oudin technique can be used as a measure of antigen concentration, and can be read as the position of the precipitin band after prolonged incubation or as the change in position early during diffusion.

In the Ouchterlony method, antigen and antibody are placed in multiple wells cut into agar on a flat slide and allowed to diffuse toward one another. If the system is homogeneous, one line results; if there are multiple components, there will be multiple lines when two samples are placed in juxtaposition to one another and an antiserum. The identity of the two antigens is indicated by a clear corner on the intersecting lines; a spur indicates partial identity; and crossing of the intersection indicates nonidentity.

The Mancini, or single-radial-immunodiffusion, method for quantitating soluble proteins involves placing the solution to be measured into a well cut into an agar or agarose gel containing antiserum and measuring the area subtended by the precipitin ring and relating this to known standards. To improve resolution, the antigen can be made to migrate through a gel matrix (agarose) by an electric current (electrophoresis) or in the presence of a pH gradient (isoelectric focusing). Then a trench can be cut in the gel, and specific antibody placed in the trough and allowed to diffuse into the gel. This procedure can be reversed to locate immunoglobulins.

In Laurell's rocket technique, antibody may be dissolved in the gel, and the antigen sample then driven through the gel by an electric field. The resulting height of the precipitin peak will bear a specific relation to the quantity of antigen in the sample, allowing accurate estimation of antigen concentration in many samples. *See* IMMUNOASSAY; IMMUNOCHEM-ISTRY; IMMUNOLOGY. Donald Raum

Bibliography. F. Dixon (ed.), *Advances in Immunology*, vol. 48, 1990; W. E. Paul (ed.), *Annual Review of Immunology*, vol. 5, 1987; D. M. Weir and J. Stewart, *Immunology*, 8th ed., 1997.

Precision agriculture

The application of technologies and agronomic principles to manage spatial and temporal variability associated with all aspects of agricultural production for the purpose of improving crop performance and environmental quality. The intent of precision agriculture is to match agricultural inputs and practices to localized conditions within a field (sitespecific management) and to improve the accuracy of their application. The finer-scale management of precision agriculture is in contrast to wholefield or whole-farm management strategies, where management decisions and practices are uniformly applied throughout a field or farmstead.

Successful implementation of precision agriculture requires three basic steps. First, farmers must obtain accurate maps of the spatial variability of factors (soils, plants, and pests) that determine crop yield and quality and/or factors that cause environmental degradation. Second, once known, variability can be managed using site-specific management recommendations and accurate input control technologies. Third, precision agriculture requires an evaluation component to understand the economic, environmental, and social impacts on the farm and adjacent ecosystems and to provide feedback on cropping system performance.

Precision agriculture is technology-enabled, information-based, and decision-focused, because it relies on an increasing level of detail in information acquired with technology to improve decision making in crop production. Consequently, precision agriculture will evolve as technology, information management, and decision tools emerge in this era of rapid technological advancement. *See* AGRI-CULTURAL SOIL AND CROP PRACTICES; AGRONOMY; DECISION SUPPORT SYSTEM; INFORMATION SYSTEMS ENGINEERING.

Enabling technologies. The major technologies collectively enabling precision agriculture include computers, geographic information systems (GIS), global positioning systems (GPS), sensors, and application control.

Computers, in terms of both speed and storage, paved the way for precision agriculture; and computer enhancements will drive significant technological development to enable precision agriculture for some time.

A geographic information system is an organized collection of computer hardware, software, geographic data, and personnel designed to efficiently capture, store, update, manipulate, analyze, and display all forms of geographically referenced information. It has been referred to as the brain of a precision farming system. A geographic information system is important because the value of precision agriculture is derived only when farmers use detailed information on spatial variability in their management decisions. The payoff to farmers is when these decisions increase profitability, benefit the environment, or provide some other value to the farm.

Location control is essential to precision agriculture for assessing spatial variability and for sitespecific application control, and a global positioning system fulfills those needs for precision agriculture. The United States NAVSTAR (Navigation System with Time and Ranging) and the Russian GLONASS (Global Navigation Satellite System) global positioning systems have been fully operational since the mid-1990s; together they provide highly reliable and accurate positioning data. A differential correction to the GPS signal is usually required to achieve position accuracies needed in production agriculture. Differential GPS (DGPS) involves the transmission of a differential correction, that is, the difference between actual and predicted position at the base GPS receiver, to rover GPS receivers, which then apply the corrections to received GPS signals to solve for a more accurate position.

Sensors are devices that transmit an impulse in response to a physical stimulus such as heat, light, magnetism, motion, pressure, or sound. With computers to record the sensor impulse, a global positioning system to measure position, and a geographic information system to map and analyze the sensor data, any sensor output can be mapped at very fine scales. Sensors are critical because they have fixed costs, can sample at very small scales of space and time, and facilitate repeated measures. Sensors have been developed to measure machinery, soil, plants, pests, atmospheric properties, and water by sensing motion, sound, pressure, strain, heat, light, and magnetism, and relating these to properties such as reflectance, resistance, absorbance, capacitance, and conductance.

For precision agriculture, control must be achieved in space and time for varying single or multiple agronomic inputs at different rates, at varying soil depths, and in an accurate and location specific manner within fields. Control systems are available at varying degrees of precision for variable seeding, metering granular fertilizers and pesticides, changing crop varieties on-the-go, anhydrous application of nitrogen, spraying, irrigation, manure application, and various methods of tillage. *See* EXPERT SYSTEMS; GEOGRAPHIC INFORMATION SYSTEMS; SATELLITE NAV-IGATION SYSTEMS.

Mapping. Mapping is a fundamental tool for understanding and managing variability, and three types of maps are useful in precision agriculture. Condition maps relate to assessing variability and represent the spatial state of a condition of interest, for example, soil pH or pest distribution. Management maps relate to managing variability and represent the prescription of inputs needed to manage a particular condition site-specifically, for example, application of lime or a pesticide. Management maps are used in application control systems to regulate application rates and products and to form the basis for variable-rate technology. Performance maps are related to evaluation and record either inputs (such as fertilizers, pesticides, seeds, and energy) or outputs (crop yield, quality) and include derivatives of performance maps, such as profit maps (outputs-inputs). Combinations of condition, management, and performance maps can be used to create other maps in each category, and a geographic information system is an excellent tool for this purpose.

Yield mapping serves all three map functions. It requires a yield-monitoring sensor, a measure of the area of harvest, some measures of crop quality, and the location for each sensed measurement. Sensors also are available to measure crop quality, including moisture, density, and protein content, with more sensors under development. A yield monitor measures crop yield by sensing crop flow. A crop-flow sensor measures either the volume or the mass of the harvested portion of a crop using a variety of engineering principles, including light interception, radiation absorption, measurement of impact force, and directly weighing the crop. Yield sensors are available or under development for grain crops (such as corn, soybeans, and wheat), potatoes, sugarbeets, tomatoes, forages, and cotton. *See* CARTOGRAPHY.

Assessing variability. The basis for precision agriculture is found in the notion that the processes and properties that regulate crop performance and yield vary in space. For many variables, particularly pests but also soil variables such as nitrogen and water, the spatial variability can vary over short time periods. Adequately quantifying the variability of those processes and properties and determining when and where different combinations are responsible for the spatial and temporal variation in crop yield or environmental contamination is the challenge facing precision agriculture. Therefore, assessing variability is the first and most important step in precision agriculture, because farmers cannot manage what they do not know. Techniques for assessing spatial variability are readily available and have been applied extensively in precision agriculture, particularly geostatistics. Geostatistics is a major tool in the assessment of spatial variability, and a number of software packages are available. Commercial mapping and analysis software for use on the farm or in agribusiness offer a variety of interpolation and mapping routines of varying difficulty and quality. The most important issue for precision agriculture is that variability within a field is spatially structured (nonrandom) and conditions maps are accurate. See SOIL.

Managing variability. Once variation is adequately assessed, farmers must match agronomic inputs to known conditions using management recommendations that are site-specific, and use accurate application control equipment. Precision agriculture has operated for some time on the assumption that best management practices developed over decades of agronomic research are applicable at any scale of management. These assumptions will be tested as knowledge of spatial variability and its management evolves. While many aspects of production agriculture could respond to precision management, each aspect differs in the type and amount of variability and in the ways in which precision management might be applied. Major aspects of crop production include nutrient management, crop management, pest management, and water management.

Precision nutrient management. Significant opportunities exist for precision management of soil fertility, because the ability of soil to supply nutrients to plants and the plant demand vary in space and time, and nutrient losses through leaching, erosion, and runoff also vary temporally and spatially. One of the first applications of precision agriculture involved the management of fertilizers and lime according to fertilizer application maps, a practice called variablerate technology (VRT). Soils are sampled within a field using various grid or zone sampling schemes, and are analyzed for fertility using established soil testing and analysis procedures. Fertilizer or lime recommendation maps are created by making recommendations for each sample location, interpolating values for unsampled locations within the field using various techniques (for example, geostatistics, inverse distance), and mapping the results. Fertilizers are applied using equipment capable of varying application rates based on a digitized version of the fertilizer application map as input to a computercontrolled application system. In some cases, sensors are available to assess a particular nutrient, and fertilizer applications can be made in real time in response to the sensed value as the applicator moves across the field. Another approach bases fertilizer applications on maps derived by monitoring the nutrient status of a crop. For nitrogen management, for example, the idea is to map plant concentrations of nitrogen by monitoring plant or canopy reflectance of light as a measure of plant nitrogen content, derive a nitrogen fertilizer application map using established relationships between reflectance and nitrogen content, and fertilize the crop using VRT equipment. Results with VRT and other precision nutrient management studies have been mixed. Success here relies to a large extent on the accuracy of the nutrient map, the appropriateness of the fertilizer recommendation, application control, and the extent to which nutrients limit crop yield or quality in a given growing season. See FERTILIZER; FERTILIZ-ING; SOIL FERTILITY.

Precision crop management. Precision crop management could be achieved by varying species or cultivars, and/or by manipulating planting geometry (populations, row spacing, seeding depth) according to variation in environments within fields. The success of precision crop management depends on whether species or cultivars are available that are adapted to specific environments. In some cases, cultivars are bred to adapt to multiple environments, suggesting that these might not respond to precision management. The technologies needed to vary seed type or planting geometry are available. What appears to limit precision crop management are lack of accurate maps of the environments within fields and lack of agronomic decision rules for varying genetics or geometry across environments encountered within fields and across growing seasons.

Precision pest management. Pests are spatially aggregated and not randomly distributed within most agronomic fields. Furthermore, the efficacy, efficiency, and fate of pest-control inputs vary with pest and crop conditions and with physical and chemical properties of the soil, all of which can vary spatially and to some extent temporally. The application of precision agriculture to pest management is potentially beneficial to agriculture not only because it offers an opportunity to reduce chemical/nonchemical inputs into crop production and improve pest management, but also because the acquisition of spatial and temporal information on pest occurrence and distribution that has been made possible with precision-agriculture technologies will lead to an improved understanding of the pest biology and ecology needed to develop more effective pest management strategies. Thus, an important dimension of precision pest management is that new pest management strategies will evolve because of it. A major difficulty with the management of some pests is that their populations are highly dynamic, and prediction of pest density is difficult or uncertain; both of these make it necessary to collect field estimates of pest density in order to monitor density over time. The difficulty and cost in mapping pest density is considered by some to be an insurmountable barrier to the use of precision pest management in established integrated pest management (IPM) programs. New techniques to measure and manage pest variability are needed to support precision pest management.

Precision water management. There are three approaches to precision water management: variable rate irrigation, matching agronomic inputs to water availability defined by soil and/or landscape properties, and drainage. Like other factors, variable irrigation management requires sufficient knowledge about the spatial and temporal variability of the factors that regulate water availability to plants and water use in plants in order to achieve precision management. Differences in water availability within a field are favored by the occurrence of dissimilar soil types, the presence of soil degradation processes (such as erosion), and variation in landscapes that affect watershed hydrology. Variable-rate irrigation systems are available, but cost-effective systems to measure plant water availability within fields will likely emerge with future advances in remote sensing. Poor drainage is often cited by farmers as a source of yield variability within fields. Since a number of options for drainage currently exist and can be applied sitespecifically, there is little need to design site-specific drainage practices. Site-specifically, the cost of draining portions of fields or small isolated areas may be higher because costs to connect to drain outlets are proportional to the distance to the outlet. See IRRI-GATION (AGRICULTURE).

Evaluation. Three important evaluation issues surround precision agriculture: economics, environment, and technology transfer. The economic evaluation focuses on whether the documented agronomic benefits-translated into value through market mechanisms-exceed the technological and service costs. The most important fact regarding the analysis of the profitability of precision agriculture is that the value comes from the application of the data, not from the use of the technology itself. This can be contrasted with traditional agricultural innovations, where the value is derived from the use of the new technology itself, for example, new seed genetics that increase yield or a new herbicide that reduces yield loss. If any value is to be derived from precision agriculture, it will come from the resulting management decisions, not through the use of the enabling technologies.

Environmental evaluation focuses on whether precision agriculture can improve soil, water, and the general ecological sustainability of agricultural systems. Reduced use of agrochemicals, more efficient use of nutrients, increased efficacy of managed inputs, and increased protection of soils from degradation are frequently cited as potential benefits to the environment. While the impacts of precision agriculture on the environment are assumed positive, proof that it is even benign in its environmental impacts is generally undocumented. Studies evaluating the environmental benefits of precision agriculture are limited, and it is quite possible that negative impacts of precision agriculture could derive from such things as increased aggression by farmers on site-specific potentials or problems. Furthermore, a number of factors may limit reductions in chemical applications in some situations and encourage increased use in others.

Finally, there is the question of whether this bundle of enabling technologies and agronomic principles will work on individual farms. The impacts of precision agricultural systems will extend beyond crop production to the environment and to the very structure of the agriculture system. Consequently, evaluation needs to involve all sectors of agriculture. *See* AGRICULTURE. F. J. Pierce; P. Nowak

Bibliography. P. A. Burrough and R. A. McDonnell, *Principles of Geographic Information Systems*, Oxford University Press, New York, 1998; Environmental Systems Research Institute (Redlands, CA), *Understanding GIS: The ARC/INFO Method*, John Wiley, New York, 1997; M. Morgan and D. Ess, *The Precision-Farming Guide for Agriculturalists*, John Deere Publications, Moline, IL, 1997; F. J. Pierce and P. Nowak, Aspects of precision agriculture, *Adu*. *Agron.*, 67:1–85, 1999; F. J. Pierce and E. J. Sadler (eds.), *The State of Site-Specific Management for Agriculture*, American Society of Agronomy, Crop Science Society of America, and Soil Science Society of America, Madison, WI, 1997.

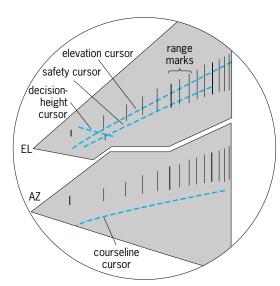
Precision approach radar (PAR)

A tracking system that provides a ground control approach (GCA) air-traffic controller with a precise display of an aircraft's position relative to a runway final-approach course. To ensure absolute safety, precise information is displayed on a plan position indicator (PPI). This display provides the controller with aircraft position information for control of heading and rate of descent. To accomplish this and maintain the required precision for a final-approach aid, the display shows the aircraft position in relation to range, azimuth, and elevation. The information presented on the precision approach radar display allows an air-traffic controller to direct a pilot down along a runway approach course to a precision landing. Precision radar approaches are accomplished in most weather conditions and do not require any on-board avionics equipment, such as an instrument landing system (ILS).

Antenna systems. There are two ways to provide the required range, azimuth, and elevation information on the plan position indicator: use of two antennas, one scanning elevation and the other azimuth; and a single computer-controlled phased-array antenna that can provide pencil-beam

tracking for both elevation and azimuth positions. The dual-antenna system uses a single transmitting magnetron that supplies radio-frequency power at a frequency of approximately 9 GHz to both antennas. A revolving mechanical switch then alternately feeds radio-frequency energy to each antenna. The azimuth antenna beam width scans 20° , and the elevation beam scans 8° . For the single antenna, radio-frequency energy is developed by a traveling-wave tube-crossed-field amplifier. The phasing and direction of the transmitted beam are controlled by a computer through phase shifters. This scan coverage is also 8° in elevation and 20° in azimuth. *See* ANTENNA (ELECTROMAGNETISM).

Display. The plan position indicator is divided into two sections (see illus.). The upper section is used to display the aircraft's elevation on an elevation cursor. The lower display shows the aircraft's azimuth by using a course-line cursor. This combination of azimuth and elevation display is commonly referred to as the AZ-EL indicator. Range marks are also displayed on the AZ-EL indicator at 1-mi intervals. The range marks are spaced logarithmically to provide increased detail at critical close ranges. To increase safety even further, an elevation safety cursor is displayed. The elevation scan requires increased accuracy because of an aircraft's inherent instability while descending compared to steady course. This safety cursor ensures that an aircraft does not fly too low for a safe approach. The safety cursor is distinguished from the elevation cursor by dashing or blinking. A decision-height indicator cursor also appears on the elevation display. Decision height is the point at which the pilot must decide, based on the ability to acquire runway visual references, either to continue the approach to land, or to execute a missed approach.



Standard plan position indicator (PPI) display for precision approach radar, with azimuth (AZ) and elevation (EL) displays. (After Air Traffic Control Radar Specialist, PAR Indicators, Performance Monitoring and Assessment, and Auxiliary Systems, Career Development Course, vol. 4, Extension Course Institute Air University, Gunter Air Force Station, Alabama, 1982)

Configuration and use. Precision approach radar systems are configured as both permanent and mobile facilities. Their use is particularly valuable for locations where terrain, excessive site preparation costs, and airspace restrictions prevent installation of an instrument landing system. The use of precision approach radar systems is heavily dependent on personnel and human factors. Differences in human reaction times can limit the effectiveness of such systems. After observing a target deviation, the controller issues a heading correction or an advisory in relation to the course or elevation cursor. Depending on the controller's reaction time and the pilot's response to instructions, there may be a several-second delay before the aircraft actually alters course. See INSTRUMENT LANDING SYSTEM (ILS).

Precision approach radar systems, which are located next to the runway, rotate to more than one runway to accommodate wind direction and changes to the runway in use. In some cases, the shelter that houses the air-traffic controllers and the plan position indicator display is in a location remote from the precision approach radar equipment. Also, some precision approach radar systems can data-link control instructions to the pilot without radio communication. *See* ELECTRONIC NAVIGATION SYSTEMS; RADAR. Mark A. DePlasco

Bibliography. D. Rider (ed.), *Jane's Airport and ATC Equipment*, 12th ed., 1993–1994; M. I. Skolnik, *Radar Handbook*, 1990; G. J. Sonnenberg, *Radar and Electronic Navigation*, 6th ed., 1988.

Predator-prey interactions

Predation occurs when one animal (the predator) eats another living animal (the prey) to utilize the energy and nutrients from the body of the prey for growth, maintenance, or reproduction. In the special case in which both predator and prey are from the same species, predation is called cannibalism. Sometimes the prey is actually consumed by the predator's offspring. This is particularly prevalent in the insect world; for example, a wasp may paralyze and lay eggs on a butterfly caterpillar, leaving its offspring to consume the prey. Insect predators that follow this type of lifestyle are called parasitoids, since the offspring grow parasitically on the prey provided by their mother.

Predation is often distinguished from herbivory by requiring that the prey be an animal rather than a plant or other type of organism (bacteria). To distinguish predation from decomposition, the prey animal must be killed by the predator. Some organisms occupy a gray area between predator and parasite. For example, lamprey are generally considered parasites that attach to the sides of fish and suck blood. However, lamprey can reach sufficient size to kill their hosts, and then they act as predators. Finally, the requirement that both energy and nutrients be assimilated by the predator excludes carnivorous plants from being predators, since they assimilate only nutrients from the animals they consume. *See* FOOD WEB.

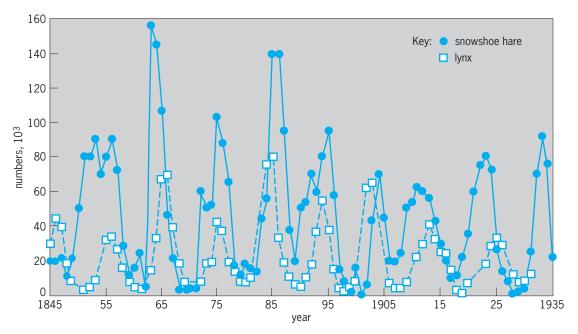
Population dynamics. Population dynamics refers to changes in the sizes of populations of organisms through time, and predator-prey interactions may play an important role in explaining the population dynamics of many species. Predators affect the population dynamics of prey, because increases in the density of the predator population lead to increased predation on the prey and a consequent decrease in the prey's population growth rate. Conversely, increases in the density of prey mean more potential food for predators and a consequent increase in the predator's population growth rate. Predatorprey interactions are thus a type of antagonistic interaction, in which the population of one species has a negative effect on the population of a second, while the second has a positive effect on the first. For population dynamics, predator-prey interactions are similar to other types of antagonistic interactions, such as pathogen-host and herbivore-plant interactions.

The characteristic population dynamics associated with predator-prey interactions are predator-prey cycles. Simple mathematical models of predator-prey interactions exhibit cycles; cycles are also found in simple laboratory experiments. In nature, several examples of cycles have been observed, the most famous of which is the snowshoe hare cycle (see illus.). How cycles arise can be understood by considering what happens when both prey and predator densities are low. With few predators, the density of the prev increases. This in turn increases the population growth rate of the predator. The resulting increase in the predator's density decreases the population growth rate of the prey. When the prey's population growth rate decreases sufficiently, the prey density decreases, and this in turn leads to a decrease in predator density. The cycle then repeats itself.

Even though snowshoe hares and their predators show cyclic dynamics, extensive experiments and observations have shown that the cycles are not simply driven by interactions between hares and their predators. In concert with the predator-hare cycles, there is also a hare-plant cycle in which plant abundance and quality decrease when hare densities are high, resulting in reduced body condition and reproduction of hares. Furthermore, there is a synergistic effect between plant quality and predation, with reduced food availability leading to increased predation rates, possibly because low food availability forces hares to feed in areas that are more exposed to predation. Thus, the cyclic dynamics of snowshoe hares could be the product of simultaneous predator-prey and herbivore-plant interactions.

Stabilization of predator-prey dynamics. Although there are several well-documented cases of cyclic population dynamics, predator-prey cycles are very rare when considering all the possible predator-prey systems in nature. Nonetheless, predator-prey interactions are thought to increase fluctuations in the population dynamics of predator and prey, with these fluctuations having the potential to cause local extinctions. This has been shown by many studies on organisms that can be maintained in the laboratory; when predator and prey are placed together, the population dynamics show cycles of increasing amplitude, eventually leading to the extinction of the predator, or both prey and predator.

Given the contrast between persistent predatorprey systems in nature and the difficulty of constructing persistent predator-prey systems in the lab, much theoretical and experimental work has been



Snowshoe hare and lynx populations inferred from the numbers of pelts sold to the Hudson's Bay Company. (After D. A. MacLuclich, 1937)

devoted to explaining what stabilizes predator-prey dynamics in nature. Most of the numerous possible explanations involve heterogeneity, either in the habitat in which the species live or in the susceptibility of prey individuals to predation. If, for example, the habitat contains a refuge in which the prey are safe from predation, the prey population cannot be driven extinct, and this results in a stable predator-prey system. This appears to occur frequently in insect predator-prey systems, in which the herbivorous prey hide in the nooks and crannies of plants where they cannot be found by predators. Similarly, if a certain fraction of the prey population is invulnerable to attack, then again the predator cannot drive the prey population extinct. A possible example is rhinos, which are vulnerable to predators when they are young, but once they reach full size they can defend themselves against any predator. Research has provided a wealth of rich examples showing how heterogeneity may stabilize predator-prey interactions.

Biological control. Biological control is the use of natural enemies to control pests, and demonstrates how ecological knowledge can be used to address an applied problem. In classical biological control, an exotic natural enemy is introduced into a new country to control a pest, normally a pest that also is exotic. After introduction and establishment of the natural enemy, no further intervention is conducted. In contrast, biological control can be conducted actively, for example by continuously breeding and releasing natural enemies into targeted areas. Repeated, inundative releases are most often used in greenhouses to control pests on vegetable and flower crops.

The first biological control program was conducted in 1888, when the vedalia (a type of ladybird beetle) was released to control scale insects on citrus crops in California. Since then, many hundreds of biological control programs have targeted large numbers of insect and plant pests. For example, prickly pear cacti (Opuntia stricta) were introduced into Australia as garden plants in the nineteenth century, but in the absence of native flora similar to cacti, the prickly pear population exploded to cover much of the rangeland of eastern Australia. In 1926 Australian entomologists introduced the cactus moth (Cactoblastus cactorum) from Argentina, and within 15 years the prickly pear population was reduced to a few patches spread throughout the eastern part of the country. In practice, introduced natural enemies are selected to specialize only on the target pest, and for this reason many control agents are wasp parasitoids, since these tend to attack a narrow range of host species.

The basic problem of what stabilizes predator-prey dynamics figures prominently in understanding the success of classical biological control programs. In successful programs, the natural enemy suppresses the pest density to low levels, yet does not lead to total extinction. If extinction were to occur, the reintroduction of the pest would lead to an outbreak in the absence of the natural enemy. For example, the cactus moth has not completely eradicated the prickly pear from Australia, which still persists in small patches. Even though these patches provide a source of prickly pears that could spread to other areas, they also provide a source of cactus moths that can spread and control incipient prickly pear outbreaks.

Predation and community structure. Community structure refers generally to how species within an ecological community interact. The simplest conception of a community is as a food chain, with plants or other photosynthetic organisms at the bottom, followed by herbivores, predators that eat herbivores, and predators that eat other predators. This simple conception works well for some communities. For example, deer and moose eat plants, and wolves eat deer and moose. Nonetheless, the role of predator species in communities is often not clear. Many predators change their ecological roles over their lifetime. For example, pike may reach 50 lb (23 kg) and are generally viewed as predators at the top of the food chain in lakes, yet many are killed as young fingerlings by diving beetles. Many insect predators that share the same prey species are also quite likely to kill and devour each other. This is called intraguild predation, since it is predation within the guild of predators. Furthermore, many species are omnivores, feeding at different times as either predators or herbivores; for example, brown bears feed on both plant roots and salmon. Therefore, the role of particular predator species in a community is often complex.

Predator-prey interactions may have a large impact on the overall properties of a community. For example, most terrestrial communities are green, suggesting that predation on herbivores is great enough to stop them from consuming the majority of plant material. In contrast, the biomass of herbivorous zooplankton in many aquatic communities is greater than the biomass of the photosynthetic phytoplankton, suggesting that predation on zooplankton is not enough to keep these communities green. As a more specific example, the diversity of species occupying the rocky intertidal area along the coast of the Pacific Northwest is largely determined by the presence or absence of a large predatory starfish (Pisaster ochraceus). When starfish are absent, the community becomes dominated by mussels that are competitively superior to other sessile intertidal organisms. When starfish are present, however, they preferentially eat mussels, thereby creating vacant areas that are colonized by less competitive organisms such as barnacles. See POPULATION ECOLOGY. Anthony R. Ives

Bibliography. M. Begon, J. L. Harper, and C. R. Townsend, *Ecology: Individuals, Populations and Communities*, Blackwell Scientific Publications, Boston, 1990; S. I. Dodson et al., *Ecology*, Oxford University Press, New York, 1998; D. A. MacLuclich, Fluctuations in numbers of the varying hare (*Lepus americanus*), *Univ. Toronto Studies Biol. Ser.*, 43:1-13, 1937; R. J. Taylor, *Predation*, Chapman and Hall, New York, 1984.

Pregnancy

The period during which a developing fetus is carried within the uterus. In humans, pregnancy averages 266 days (38 weeks) from conception to childbirth. Traditionally, pregnancy duration is counted from the woman's last menstrual period, which adds roughly 2 weeks to gestational age. This is how physicians arrive at a pregnancy length of 40 weeks (280 days).

Establishment of pregnancy. The critical events that establish pregnancy are ovulation, fertilization, and implantation. In women of childbearing age (15-44), the release of an egg from the ovary (ovulation) occurs approximately 14 days after the beginning of a menstrual period. Successful ovulation requires the coordinated interplay of hormones produced by the ovary, pituitary gland, and hypothalamus. Conception or fertilization, the union of a single sperm with the egg, usually takes place in the Fallopian tube within a day or two of ovulation. The fertilized egg (zygote) next passes into the uterine cavity and, on about day six, attaches (implants) into the lining tissue or endometrium. This tissue, which normally sloughs off each month in the nonpregnant female, is instead maintained and nurtured by placental hormones. As pregnancy progresses, the placenta mediates the inflow of nutrients and removal of waste products while synthesizing increasing amounts of estrogen and progesterone, hormones that influence nearly every tissue and organ system in the body. The 9 months of pregnancy are typically divided into three periods (trimesters) of 3 months. See FERTIL-IZATION (ANIMAL); MENSTRUATION; OVUM; PLACEN-TATION; SPERM CELL.

The first sign of pregnancy is often the absence of an expected menstrual period. Common symptoms include nausea, breast tenderness, fatigue, and frequent urination. The diagnosis of pregnancy can be made as early as 10 days after fertilization by means of blood tests. By 6 weeks (from the last menstrual period), the uterus feels soft and is palpably enlarged. Pregnancy can be positively confirmed by observing cardiac motion of the fetus by ultrasound scanning (8 weeks) or by hearing fetal heart "tones" by using a Doppler detection instrument (10–12 weeks).

Embryonic development. Early in the first trimester, the embryo's germ layers differentiate into organs and systems, a process that is nearly completed by the twelfth week. It is during this critical period of development that the fetus is most vulnerable to the adverse effects of drugs and other teratogenic influences. The second and third trimesters of pregnancy are characterized by increased fetal growth and gradual physiologic maturation of fetal organ systems. During this time, the maternal changes of pregnancy are greatest. The enlarging uterus encroaches on the abdominal region by the fourth month and at term nearly reaches the diaphragm. The breasts gradually enlarge in preparation for lactation. Striking cardiovascular changes, including nearly a 50% increase in cardiac output, provide the increased blood flow to accommodate the growing fetoplacental unit. Other changes in the renal, digestive, pulmonary, and endocrine systems reflect the numerous maternal adaptations that eventually must occur in a healthy pregnancy. *See* EMBRYOLOGY.

Prenatal care. Early, regular prenatal care is associated with improved pregnancy outcome and seeks to identify risk factors in the pregnancy that may apply to mother or fetus. At 6-8 weeks, a complete physical examination, along with blood and urine analyses, should be performed. In addition to undergoing traditional tests, patients are now routinely screened for hepatitis B at the beginning of pregnancy, for fetal neural-tube defects such as open spine at 16 weeks, and for gestational diabetes at about 28 weeks. In addition to these blood tests, many physicians offer a sonogram at 16-18 weeks to establish gestational age, check for a multiple pregnancy, and screen for birth defects. During prenatal visits, a physician can evaluate nutrition, blood pressure, and fetal growth. Weight gain in the last two trimesters should normally be 1 lb (0.45 kg) per week, for a total of 25-35 lb (11-16 kg). During the pregnancy, common symptoms are noted, such as edema, varicose veins, hemorrhoids, backache, fatigue, heartburn, and vaginal discharge. Ideally, at the end of the third trimester, the process of labor begins. The muscles of the uterus contract, dilating the cervix and allowing the baby to begin moving into the vagina or birth canal. Continued contractions push the baby out of the mother's body. In the final stage of labor, the placenta detaches from the uterine walls and is expelled as the afterbirth. An alternative to vaginal delivery is the cesarean section, in which the baby is removed surgically through an abdominal incision. See PRENA-TAL DIAGNOSIS.

Birth rates and postnatal care. The number of United States births in 1988-82 million-was a modern record, the largest number of births in any single year since 1954. This appears to be due to a slight increase in the fertility rate (the number of births in a given period in relation to the number of women of reproductive age) along with a 1% increase in the number of women of childbearing age. Pregnancy and childbirth in the United States have become safer. The rate of maternal deaths dropped from 37 per 100,000 live births in 1960 to 8 per 100,000 in 1985. During this time, perinatal mortality-the sum of stillbirths and neonatal deaths (deaths within the first month of life)also decreased from 34 per 100,000 births in 1960 to 15 per 100,000 in 1985. Factors credited with improving pregnancy outcome usually include increased hospital births (now over 99%), improved childbirth pain relief (such as epidural anesthesia), and increased reliance on specialists for mother-tobe and baby, including perinatologists and neonatologists. Regionalization of obstetric facilities and referral of certain complicated pregnancies to centers with neonatal intensive care units have improved the outlook for infants of low birth weight, those who weigh less than 5 lb (2500 g). Survival of even extremely premature infants with birth weights of less than 2 lb (1000 g) is often possible, although these

infants have increased rates of mental and physical handicaps. Better monitoring of physicians through hospital quality assurance programs has been identified as another factor that has contributed to safer health care for mother and baby. *See* REPRODUCTIVE BEHAVIOR.

Termination of pregnancy. The legal status of pregnancy termination (therapeutic abortion) varies from country to country, but about two-thirds of women in the world have access to legal abortion. Abortion became legalized throughout the United States in 1973, and since that time, abortion rates have risen steadily, with some leveling off between 1980 and 1985, at about 1.3 million abortions per year. Over 90% of abortions in the United States are performed in the first trimester by suction curettage, a technique that uses suctioning and removal of the uterine contents through the vagina with surgical instruments. Later pregnancies are terminated by a procedure called dilatation and evacuation (D&E) or by administration of drugs to stimulate uterine contractions. Medical and psychological sequelae to abortion are few, and are fewest for terminations in the first trimester. See PREGNANCY DISORDERS. Bruce D. Shephard

Bibliography. W. G. Birch, *Guide to Pregnancy and Childbirth*, 1991; F. G. Cunningham, P. C. McDonald, and N. F. Gant, *Williams Obstetrics*, 20th ed., 1997; N. F. Maklad, *Ultrasonic Diagnosis of Ectopic Pregnancy*, 1990; R. H. Petrie (ed.), *Perinatal Pharmacology*, 1989; B. Shephard, *The Complete Guide to Women's Health*, 3d ed., 1997.

Pregnancy disorders

Physical disorders that can arise as a consequence of pregnancy, ranging from mild to life-threatening. Extreme conditions can result in termination of the pregnancy or death of the mother.

The most common disorder of early pregnancy, persistent vomiting, is without known cause and usually subsides spontaneously within a few weeks. A more serious threat to pregnancy, vaginal bleeding (within the first 20 weeks), may be a sign of miscarriage (spontaneous abortion) or much less often, ectopic pregnancy (embryonic development outside the uterus, usually within the Fallopian tube). Approximately 10-15% of pregnancies reportedly end in miscarriage, although the actual rate is closer to 40%, as some early miscarriages are not recognized. Over 50% of miscarriages are due to a spontaneous chromosomal abnormality in the sperm, egg, or developing embryo. Ectopic pregnancy occurs about once in every 50 pregnancies and represents a leading cause of maternal mortality in the United States. Ectopics are nearly always removed surgically; the involved Fallopian tube usually can be preserved.

Some abnormal conditions in pregnancy are identified as a result of routine blood tests. Examples include rubella (German measles) and syphilis, both of which occur relatively infrequently and can cause birth defects. Rh disease is also uncommon because of prevention by prenatal blood tests and treatment with immunoglobulins in Rh-negative mothers. Routine screening is also recommended for hepatitis B, an increasingly prevalent liver condition that can cause maternal liver damage or infection in the newborn. Although no specific treatment for hepatitis B is available for the mother, it is possible to prevent infection in the newborn by using vaccine and immunoglobulin immediately after birth. *See* HEPATITIS; PRENATAL DIAGNOSIS; RH IN-COMPATIBILITY; RUBELLA; SYPHILIS.

Maternal conditions that may worsen during pregnancy include some forms of heart disease, seizure disorders, hypertensive disease, and acquired immune deficiency syndrome (AIDS). Transmission of the AIDS virus to the newborn occurs in up to 50% of cases and can take place at any stage of pregnancy. Mortality among HIV-infected infants is about 45%, and many die before the age of 2. *See* ACQUIRED IM-MUNE DEFICIENCY SYNDROME (AIDS).

Among the most important disorders during the second half of pregnancy are those associated with low birth weight due to either premature labor or fetal growth problems. Low birth weight is associated with twins, hypertensive disorders, smoking, and inadequate nutrition. The hypertensive disorders include pregnancy-induced hypertension (formerly known as toxemia) as well as chronic hypertension that exists before the pregnancy. Pregnancy-induced hypertension occurs in about 5% of pregnancies, mostly among women who have not given birth previously. Signs and symptoms can include swelling of the hands and face, headaches, and a sudden weight gain of 5 lb (2 kg) or more in 1 week. Pregnancy-induced hypertension can be assessed by in-office fetal monitoring tests as well as by sonography to evaluate fetal growth. These tests are important for evaluating many other complications, including bleeding in late pregnancy, which is a potential emergency condition. Vaginal bleeding often results from a placental problem; either the placenta is abnormally located near or over the cervix (placenta previa), or the placenta separates prematurely from the uterus (abruptio placenta). See HYPERTENSION.

Far less dramatic than these conditions, but equally important as a cause of fetal distress, is the "post dates" pregnancy, which occurs once pregnancy has extended 2 weeks beyond the date of expected delivery. When a pregnancy reaches 42 weeks, delivery is attempted as soon as possible as the aging placenta may lose its ability to provide adequate oxygen and nutrition.

Increased rest and proper nutrition are especially important in pregnancies complicated by high blood pressure, fetal growth problems, and twins. More immediate management often centers on the timing of delivery. Certain abnormal test results may indicate the need for delivery before term, often by cesarean birth. Many potentially difficult pregnancies can be conservatively managed until near term, however. If preterm labor threatens, several promising drugs are available to treat this condition, but this situation as well as other problems such as persistent bleeding may require prolonged hospitalization. *See* PREGNANCY. Bruce Shephard

Bibliography. R. S. Abrams, Handbook of Medical Problems During Pregnancy, 1988; G. N. Burrow and T. R. Ferris, Medical Complications During Pregnancy, 5th ed., 1999; D. F. Hawkins (ed.), Drugs and Pregnancy: Human Teratogenesis and Related Disorders, 2d ed., 1987; C. Stern (ed.), Immunology of Pregnancy and Its Disorders 1988.

Prehistoric technology

The set of ideas that prescribe the manufacture and use of implements before written history. Technology is the principal means through which the human species has succeeded in occupying most of the world. The prehistoric record documents that for over 2 million years the human lineage has been making and using implements. Archeologists tend to use the term "artifact" for any material that was modified by ancient humans, whether this material was used or not, and the term "tool" for any material that was used by ancient humans, whether it was modified or not.

The archeological evidence for prehistoric technologies may be biased because of the nature of the raw materials that were used and the conditions of burial and preservation of these materials. In general, artifacts of stone, bone, pottery, and metal preserve fairly well in many areas, whereas artifacts of wood and other vegetable materials, skin, and horn tend to decay fairly rapidly and are normally found only in exceptional prehistoric contexts.

In 1816, C. Thompsen, director of the Danish National Museum, began to chronologically order the museum's prehistoric collections into three major groups, based upon technology, and now called the Three Age System. The earliest was a Stone Age, followed by a Bronze Age, and finally an Iron Age. As

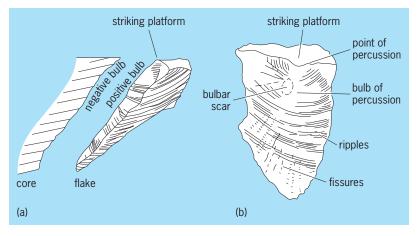


Fig. 1. Elements of stone flaking. (a) Relationship between the core and the flake. (b) Major feature of the ventral or release surface of the flake. (*After I. Tattersall, E. Delson, and J. A. Van Couvering, eds., Encyclopedia of Human Evolution and Prehistory, Garland, 1988*)

time went on, the prehistory of Europe was further divided, based on regional sequences that could be documented through excavation. In other places, such as the Americas, Australia, Oceania, and sub-Saharan Africa, different nomenclature was often used, since the regional technological sequences differed from that of western Europe.

New technological traits can be introduced into a society in a number of ways: (1) innovation or invention, which is the development of new ideas, including new technological characteristics; (2) diffusion, which involves the spread of ideas, including technological knowledge, from one group to another; and (3) migration, which is the spread of peoples, often with new technologies, into new areas. One of the goals of a prehistorian is to try to ascertain, based on archeological evidence, which factors best explain the technological changes seen in the archeological record.

Stone Technology

Archeological evidence has demonstrated that the human lineage has been making and using identifiable tools for at least 2.5-2.6 million years, beginning with stone implements and associated prehistoric remains. Although animals may use unmodified objects as tools (for example, the California sea otter cracks open shellfish with a stone hammer or anvil), most of this behavior appears to be instinctive. It is likely that early hominids (proto-humans) went through rudimentary technological stages that predated the earliest stone artifacts. Materials that could have been used for tools include wood, bone, horn, and shell. Unfortunately, these materials either tend to be perishable or do not usually leave clear indications of modification or use in such a primitive technology

Principles of stone fracture. For most of the Stone Age, rocks were fashioned into various forms by flaking, that is, removing spalls (flakes) from a stone by percussion or pressure. The ideal raw materials for such flaking are hard, fine-grained rocks that fracture well in any direction and have no preferential cleavage planes, such as flint, chert, chalcedony, jasper, agate, quartzite, lava, and obsidian. In flaking stone, a glancing blow at an angle of less than 90° with a hammer of stone or other material normally produces a controlled fracture pattern that results in a parent piece of rock (the core) from which sharp fragments are detached (flakes and fragments). Such conchoidal fracture normally produces flakes that exhibit a bulb of percussion, ripples, and fissures or hackle marks (Fig. 1). Such features help identify rocks that have been modified by humans. A number of techniques can be used for working stone, each of which produces diagnostic flaking characteristics and by-products (Fig. 2). These include hardhammer percussion, anvil technique, bipolar technique, soft-hammer percussion, indirect percussion (punch technique), and pressure flaking.

Cultural-historical outline. Much of the nomenclature used in Stone Age studies dates from the nineteenth century and applies to western Europe. It later became clear that the technological progression of populations in prehistoric Europe was not necessarily universal and that any given geographical region has its own prehistoric sequence of development.

Early Paleolithic. The first definite stone tools come from African prehistoric sites, the earliest in Ethiopia dated to approximately 2.5-2.6 million years ago (Ma). These technologies have been named Oldowan, after the famous site of Olduvai Gorge in northern Tanzania, and consist of simple flaked cobbles or blocks of stone made into a range of core forms, battered stones, retouched flake scrapers, and flakes and fragments produced from these core forms as well. The primary techniques used were hard-hammer percussion (Fig. 2*a*), which consisted of direct blows to the core with a stone hammer, and bipolar technique (Fig. 2*c*), which involved setting the core on a stone anvil and striking the core from above with a hammerstone.

These earliest industries are known to be contemporaneous with at least two major fossil hominid lineages. The first is the smaller-brained *Australopithecus boisei* in eastern Africa and *robustus* (also known as *Paranthropus robustus*) in southern Africa. The second lineage encompasses the largerbrained forms, which are believed to be ancestral to modern humans, *Homo habilis* and *rudolfensis* and subsequently early representatives of *H. erectus* (*H. ergaster* in Africa). Although it cannot be conclusively demonstrated which hominid forms were the major toolmakers, the larger-brained genus *Homo* was probably the more habitual toolmaker. *See* FOS-SIL HUMANS.

Early archeological sites tend to be situated in tropical or subtropical grassland environments, often near streams or lakes, and are associated with animal bones, some of which bear cut marks from sharpedged stone tools. Early stone tools may have served as hammers for cracking open long bones to obtain marrow and as chopping or scraping tools for making wooden spears or digging sticks.

A new industrial stage, the Acheulean, began about 1.5 Ma in Africa and appears to have spread to many parts of Europe, the Near East, and the Indian subcontinent. Artifacts include hand axes, cleavers, and picks, either made on large flake blanks struck from boulder cores or worked down from large cobbles or nodules of rock. These large flakes could have been detached by hard-hammer percussion, with a large hand-held stone hammer or thrown boulder hammer, or by anvil technique (Fig. 2b), in which a large core was struck against a stationary stone anvil to detach flakes. Later Acheulean industries often exhibit finer, more controlled flaking and thinner, more refined hand axes and other tool forms, which suggests that soft-hammer percussion (Fig. 2d) was employed. By this technique, a material that is softer than the stone being worked, such as bone, antler, wood, or ivory, is used as a percussor.

Acheulean and contemporaneous industries are associated with the fossil forms of *H. erectus* and

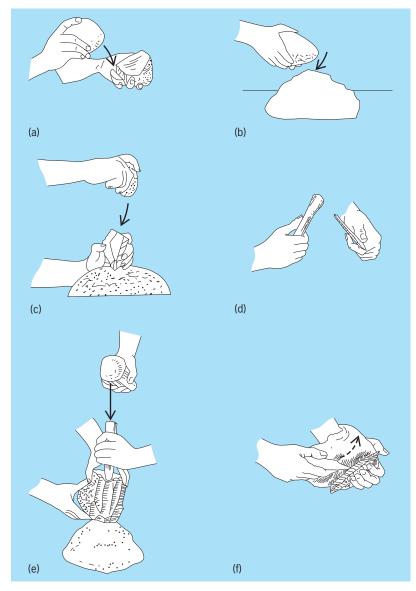


Fig. 2. Principal techniques of stone-tool manufacture. (a) Hard-hammer, with free-hand percussion. (b) Anvil technique. (c) Bipolar technique. (d) Soft-hammer percussion. (e) Indirect percussion or punch technique. (f) Pressure flaking. (After I. Tattersall, E. Delson, and J. A. Van Couvering, eds., Encyclopedia of Human Evolution and Prehistory, Garland, 1988)

a larger-brained hominid, often designated as archaic *H. sapiens*. Between 1.7 Ma and 200,000 years ago, these populations spread over much of the Old World, occupying tropical and temperate environments. In most of eastern and southeastern Asia, stone technologies were more similar to Oldowan sites, perhaps because of the extensive use of perishable cutting materials such as split bamboo. Some prehistoric sites of this technological stage also show evidence of the use of fire.

It is important to remember that a direct correspondence does not exist between stone industries and human species. The earliest Oldowan tools are generally and controversially attributed only to early members of the genus *Homo*, but continued to be made by all subsequent human species in various parts of the world. *Homo ergaster* had appeared in Africa by 1.8 Ma, but Acheulean tools are currently thought to have emerged 300,000 years later. Archaic and anatomically modern *H. sapiens* are both associated with Middle Paleolithic technologies.

Middle Paleolithic. Beginning over 100,000 years ago, the large hand axes and cleavers so characteristic of the Acheulean were replaced in many areas by smaller, standardized flake tools such as side scrapers, which were probably used for hide working and woodworking, and unifacial points for projectile points or butchery knives. Hard- and softhammer techniques were common during this industrial stage. Some Early and Middle Paleolithic technologies included prepared-core or Levallois methods that served to produce tool blanks of predictable size and shape.

In Europe and the Middle East, most of these industries are associated with the fossil hominid *H. sapiens neanderthalensis* (Neandertal). In some sites in Africa and Israel, however, Middle Stone Age industries have been found with anatomically modern human fossils of *H. s. sapiens*. Compared with previous hominids, Neandertals appear to have been adapted to a wider range of environments, including the severe cold of Europe during the last glaciation, and hunting may have become more common. *See* NEANDERTALS.

Later Paleolithic. Major changes in lithic technology are observed after 40,000 years ago. Blade blanks have been found on Early Paleolithic sites and were struck from prepared cores during the Middle Paleolithic. However, blades were produced from different core technologies during the Late Paleolithic and were made into a variety of tool forms, such as end scrapers, burins, and awls. In blade technology, blanks at least twice as long as they are wide appeared. Blades can be produced by hardand soft-hammer percussion, but in the finest blade industries a new technique, indirect percussion or punch technique (Fig. 2e), was employed. A punch of antler, horn, or other material was set against the edge of a blade core and struck at the opposite end with a hammer. This technique allowed for more precise flaking and control. However, at the present time indirect techniques are clearly indicated only during the postglacial Mesolithic Period.

In many places, bones and antler tools become common for the first time as well, as can be observed in the form of spear points, barbed harpoon heads, needles, spear throwers, and the earliest representational sculpture and engraving. The emphasis on stone tools to make other tools in a range of materials seems to have accelerated markedly during this period, and it is likely that wooden tools were more common as well.

More geographical variation and faster technological changes occurred, which may suggest different cognitive and language skills compared with earlier hominid forms. These later Paleolithic technologies often take one of three forms: a large blade industry (Late Paleolithic); a small blade industry with small geometric microliths (Mesolithic, Epipaleolithic, or Later Stone Age); or a technology based on recurrent projectile point forms, often bifacially worked. A new technique, pressure flaking (Fig. 2*f*), was employed as early as the Late Paleolithic in some areas, in which small flakes or bladelets were removed by pressing them off an implement with a pointed piece of suitable material, such as antler, bone, or horn. The finely executed projectile points that are characteristic of such places as western Europe, the Horn of Africa, and the Americas were often fashioned by this technique.

Later Paleolithic hunters and gatherers appear to have become more adept at predicting the location of resources and exploiting potential foods than were earlier populations. Big-game hunting was characteristic of some areas; the spear thrower (atlatl) and the bow and arrow appear to have been invented during the last 25,000 years, with single stone points or tips made of composite microliths. Harvesting of fresh-water and marine resources became more common, and the concentration on certain plant resources is first documented during this time period. All of these industries subsequent to 30,000 years ago are associated with anatomically modern humans, *H. s. sapiens*.

Food-producing technologies. With the advent of food-producing communities (called Neolithic or New Stone Age in the Old World), beginning around 9000 years ago, new stone implements became common, including ground stone tools (especially axes and adzes) and seed-grinding equipment (querns and grindstones). Ground stone tools could be fashioned by first flaking or pecking and then final grinding and polishing, usually using an abrasive grit or sand with water against another piece of rock. They permitted the use of coarser-grained raw material than was previously employed. Most of these ground stone tools were hafted onto a handle of some sort. Desirable lithic raw materials or tools often became a major commodity of trade networks.

The advent of metalworking technologies, especially in the Old World, gradually brought about a decline in stone artifacts. By 2000 B.C., stone tools had been replaced by metal implements in much of Eurasia. Subsequent migrations of peoples and diffusion of ideas accelerated this replacement. Nicholas Toth; Brooke Blades

Ceramics

Ceramics refers to a class of materials made of clay mixed with a binding liquid, usually water, which are shaped and hardened through exposure to heat. Depending upon the types of clays used, porosity, and temperatures reached in kilns, they are classified as earthenware or terracotta, stoneware, and porcelain. Earthenware ceramics found in archeological contexts are fired at relatively low temperatures, about 1100-1832°F (600-1000°C); stoneware and porcelains, about 2200-2600°F (1200-1400°C). Some early ceramics were sun-baked.

The earliest known earthenwares are figurines discovered at Dolni Vestonice in former Czechoslovakia, dating to approximately 28,000 years ago. Containers of fired clay are found somewhat later, occurring in archeological contexts by 10,000 B.C.E. (Before the Common Era) in the Near East and Japan (the Jomon) and slightly later in other parts of the Old World. In the New World, they appear around 3000 B.C.E. Prehistoric peoples used ceramics for a variety of purposes. In addition to containers and figurines, clay was made into amulets, ornaments, bricks, tiles, house models, storage pits, and mudbrick architecture. Clay objects or tokens may even have been the earliest precursors to writing. Hand-modeled tokens first appear in 8000 B.C.E. in the form of geometrics and other shapes. Each token represented a particular good and was part of the first known abstract numerical system. This system was largely replaced around 3200 B.C.E., when clay tablets impressed with shapes like the older tokens began to be utilized. Other innovations in ceramic technology appeared later. Stoneware and porcelain are known from the fourteenth or fifteenth century B.C.E. in the Far East. However, the technology of stoneware production had developed earlier in the Indus Valley in Pakistan at around 2500 B.C.E., when it was used to produce bangles. The technique apparently was lost until its reinvention.

Manufacture. Prehistoric potters probably used many of the same procedures adopted by presentday potters throughout the world (**Fig. 3**). Several basic steps must be considered in producing ceramic objects: procuring raw materials, refining clays, shaping and finishing objects, applying decoration to them, drying and heating them.

Obtaining raw materials involves gathering or mining from natural deposits or through trade or purchase. Raw materials include clay, the binding liquid, and the inclusion of minerals or organic matter, such as sand, shell, fiber, crushed rock, or reused pottery bits. Their inclusion improves the workability of clays and minimizes cracking during drying or firing. Potters often have preferences for different types of clay, especially as to color. Many prehistoric potters preferred clays rich in iron oxides that imparted an earthy tone to the final products.

Refining clays involves extracting foreign matter, as well as drying, crushing, sieving, or soaking the clay in water in order to settle out heavy particles.

Shaping ceramics can be accomplished using a variety of hand-building methods, such as pinched, coiled, slabbed, or molded techniques. The most rudimentary is the pinched pot that involves stroking a lump of clay into a desired shape. Coiled and slabbed constructions are additive techniques in which clay coils or slabs are successively added to a base. Molds are forms onto which clay is pressed. The mold provides a structure for complex shapes or relief decoration. Finally, the potter's wheel is a rotating device that, when spun, allows the potter to use centrifugal force and hand pressure to build up vessel walls out of lumps of clay. It appears not to have been utilized in the prehistoric New World, but is known by 4000 B.C.E. in the Near East, South Asia, and China.

Finishing objects involves scraping and smoothing to remove excess materials or irregularities. It may



Fig. 3. Potters from the village of Harappa, Pakistan, display their wares in front of a replica of a two-chambered kiln used by ancient potters in the Indus Valley civilization. (*Courtesy of R. P. Wright*)

also include adding details such as handles.

Decorative effects may be applied either before or after drying the ceramic. Materials such as nets or basketry may be impressed onto the wet ceramic to create a textured surface. Techniques employed after the clays are dried include cutting or incising the surface and applying slips or painted designs. Slips are clay-based coatings to which pigments or colorants have been added to enhance the overall color of the surface of the ceramic. They also may be applied as painted designs to parts of the object when the ceramic is leather-hard.

Drying is a critical step in the production process and may occur at various stages of manufacture. The amount of time necessary is dependent upon a variety of factors, especially weather conditions. Cold or damp weather increases the amount of drying time required.

Firing ceramics makes them hard and relatively watertight. The simplest firing method is bonfires in which ceramics are stacked in an open clearing and fuel (bark, straw, or dung, for example) is placed under, over, and around the objects. Firing temperatures attained in bonfires are relatively low (about 1100-1700°F or 600-900°C), and firing time is of short duration. The kiln is an important technological development because it protects objects from drafts and allows heat to be contained. The two most common types are the pit kiln and doublechambered kiln. Pit kilns are dug into the ground and fuel and ceramics are stacked together, sometimes within an enclosing wall. In double-chambered kilns the fuel and ceramics are placed in separate parts of the kiln. The advantages of double-chambered kilns are the longer duration of firing, higher temperatures reached (2400°F or 1300°C), and added protection during cooling. Updraft kilns were developed in the Near East, South Asia, and China some time around the fifth millennium B.C.E. See CLAY, COMMERCIAL; KILN.

Ceramic invention. There is no direct evidence for how ceramic technology was invented. One theory proposes that the techniques involved in cooking were the precursors to pottery production. Processes such as grinding, mixing, kneading, shaping, and firing occur in both cooking and pottery production. In addition, the proximity of clay-lined depressions, ovens, and hearths in household courtyards may have resulted in the accidental firing of mud clay. Other technologies linked to the invention of pottery are basketry, since the technique of coiling was used in both. Mud architecture also has been proposed as a precursor because the building up of vessels in slab construction bears similarities to early mud architecture construction.

Function. Changes in storage and cooking technology were among the many improvements that occurred as a result of the invention of ceramic containers. In contrast to basketry and nonceramic containers in which stone boiling was the major cooking method, the refractory properties of ceramics made it possible to cook food at higher temper-

atures with sustained heat. In the Near East, the appearance of ceramics generally coincides with broad changes in diet and strategies for obtaining food. Several of the earliest domesticated foods, such as barley, wheat, oats, peas, and sorghum, contain toxins that can be broken down through heating at sufficiently high temperatures. Ceramics also were effective containers for use in brewing and fermentation of grains or fruits because of their durability during fermentation and storage. *See* CERAMICS; POTTERY. Rita P. Wright

Metals

Scholarly speculation on the origins, development, and economic consequences of metallurgy can be traced back 200 years, but systematic empirical research did not begin until the 1960s with the emergence of the field of archeometallurgy. Archeometallurgists apply techniques from archeology, anthropology, mining engineering, geology, chemistry, and materials science to the study of past mining and metallurgy. Radiocarbon dating, which first became widely available in the 1950s, is of particular importance to archeometallurgy, since prehistoric mines, furnaces, and slag heaps often cannot be dated by other means.

Origins of metallurgy. The empirical record does not support the popular view that the invention of metallurgy rapidly transformed agriculture and warfare. The earliest known metal artifacts, from Cayönü Tepesi in Anatolia, date to about 7000 B.C. and are ornaments of hammered native copper, not functional tools or weapons. Native copper is found in the oxidized upper zones of many copper deposits together with brightly colored copper minerals such as malachite, azurite, dioptase, and turquoise. All of these were used for decorative effect long before the first use of metals, and it appears that copper was initially valued for its color, luster, and malleability.

Metals are very scarce in archeological assemblages dated between 7000 and 5000 B.C. Toward the end of this interval, the first unequivocal evidence of the smelting of metals from ores appears (as small ornaments of metallic lead) in Anatolia, the Near East, and the Aegean. The earliest appearance of metals in quantity is in the Balkans, where the spectacular gold and native copper burial goods in the Varna cemetery and the substantial copper mines of Ai Bunar and Rudna Glava are all dated around 4000 B.C. It is now accepted that metallurgy in the Balkans developed independently from metallurgy in the Near East.

Between 4000 and 2000 B.C., the archeology of the lands around the Black Sea (the Balkans, the Caucasus, modern Turkey, and southern Ukraine) is remarkable for the extraordinary quantities of gold (geochemically the scarcest of the metals used in antiquity) in burial and votive deposits, of which those of Troy II/III (2600-2100 B.C.) are the most familiar to western readers. In Mesopotamia and the Levant, gold was much less common than silver, which was recovered from smelted lead metal by cupellation from about 3300 B.C. In both the Balkans and the Near East, the smelting of copper from ore developed between 4000 and 3000 B.C. from a crude form of crucible smelting, using relatively pure oxide or carbonate ores, to the use of low shaft furnaces. The latter could attain the temperatures $(1150-1300^{\circ}C; 2100-2370^{\circ}F)$ required to form a fully molten slag from the nonmetallic gangue (quartz, clay minerals, and so on) in ores of lower grade, and could produce much more metal per smelt than crucible smelting. These advances were presumably made possible by innovations in the supply of air to the charcoal fuel, probably a transition from lung-powered blowpipes to bag bellows of animal hide.

By 3000 B.C., copper was widely used for tools and weapons, and its use spread rapidly east and west, reaching both Britain and China between 2500 and 2000 B.C. It is still unclear whether the technology itself was spread by migrant metalworkers, or whether the appearance of traded copper objects stimulated local invention of smelting processes. The only regions of sub-Saharan Africa that had metallurgy before 1000 B.C. were the Nile Valley and the Red Sea coast.

In the New World, metallurgy was independently invented in two regions. Native copper was used from about 3000 B.C. in midwestern and eastern North America for ornaments, burial goods, and display of status, but was not smelted before the arrival of European colonists after A.D. 1500. The earliest known metallurgy in the Andes dates to about 1300 B.C. The smelting and alloying of copper by lung-powered blowpipes spread through much of South America, but did not reach Mexico until about A.D. 400, and appears never to have reached southwestern North America. Prehistoric South American metallurgy is notable for its sophisticated treatment of gold, silver, and copper, both as pure metals and as alloys. *See* GOLD METALLURGY.

Copper alloys. In all regions but China, the earliest widespread use of copper is marked by the use of copper-arsenic or copper-antimony alloys. At concentrations of 2–5 wt % arsenic or antimony, these alloys make harder and more durable cutting edges than pure copper; at higher concentrations, they have attractive colors and produce clear musical tones when struck. It is not yet clear whether these alloys were made by deliberate co-smelting of the parent minerals, or whether they were the unintended consequence of smelting complex ores. Either way, it is clear that metalworkers recognized variation in the smelted product and selected appropriate alloys for specific uses.

After 2000 B.C., in the Near East and the Balkans these alloys were replaced by bronze, the alloy of copper with tin. Tin is rarely found in copper ores, so this alloy was deliberately produced by the addition of tin oxide or tin metal to molten copper. There are no major tin deposits in these regions, so the sources of the tin used are still a matter of controversy. *See* COPPER METALLURGY.

Iron and steel. Iron ores are much more abundant than those of copper or tin but are more dif-

ficult to reduce to metal. Copper smelters around the Mediterranean typically added iron ore to their furnaces to form a fluid slag; and as smelting techniques improved after 2000 B.C., small amounts of iron metal were sometimes produced with the copper. Yet bronze was preferred to iron, even though tin had to be obtained by long-distance trade. It was not until after the collapse of many Mediterranean civilizations and long-distance trade, around 1200 B.C., that iron began to be widely used.

Unlike copper, iron could not be fully molten in these small furnaces, so the product was small solid lumps (blooms) of iron or steel, which were then forged by the blacksmith to shape. Steel (iron with 0.3-1.5 wt % carbon) was made as early as 1100 B.C., but was not consistently produced until after 500 B.C. in western Eurasia. Iron was first smelted in West and East Africa about 500 B.C., but did not reach southern Africa until about A.D. 400. It was never smelted in the New World before European colonization, though forged iron meteorites have been recovered from Native American sites.

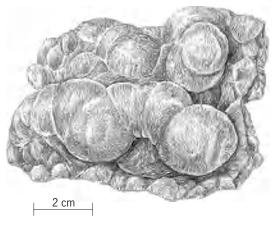
Iron metallurgy in China took a very different turn with the invention about 600 B.C. of the blast furnace to produce liquid cast iron (1.5-4 wt % carbon). The earliest known blast furnaces in Europe are in Sweden and Germany and date to about A.D. 1100. While the former may be a case of diffusion from China, the latter developed from the local bloomery iron smelting tradition. *See* IRON METALLURGY; METALLURGY. David Killick

Bibliography. D. E. Arnold, *Ceramic Theory and Cultural Process*, 1985; P. T. Craddock, *Early Metal Mining and Production*, 1995; J. Gowlett, *Ascent to Civilization*, 1984; M. Inizan, H. Roche, and J. Tixier, *Technology of Knapped Stone*, 1992; P. M. Rice, *Pottery Analysis: A Sourcebook*, 1987; O. S. Rye, *Pottery Technology: Principles and Reconstruction*, 1981: A. O. Shepard, *Ceramics for the Archaeologist*, 1976; I. Tattersall et al. (eds.), *Encyclopedia of Human Evolution and Prebistory*, vol. 2, 2000; R. F. Tylecote, *The Early History of Metallurgy in Europe*, 1987; R. J. Wenke, *Patterns in Prebistory: Humankind's First Three Million Years*, 3d ed., 1989.

Prehnite

A mineral, $Ca_2(Al,Fe^{3+})(OH)_2[Si_3AlO_{10}]$, with Al in parentheses in octahedral and Al in brackets in tetrahedral coordination by oxygens. The $[Si_3AlO_{10}]$ tetrahedral sheet reflects its perfect platy cleavage. Owing to difficulty in securing suitable single crystals, the atomic arrangement of the mineral had not been known until recently. Prehnite crystallizes in the orthorhombic system, but various space groups have been assigned to it owing to Al-Si ordering, such as Pncm, P2cm, and monoclinic (pseudoorthorhombic) P2/n.

The mineral usually occurs as stalactitic aggregates



Prehnite in reniform aggregates, Paterson, New Jersey. (Specimen from Department of Geology, Bryn Mawr College)

or as curved crystals (see **illus.**), has a vitreous luster, and is yellowish green to pale green in color. Hardness is $6-6^{1}/_{2}$ on Mohs scale; specific gravity 2.8–2.9. Common occurrences include vesicular basalts such as the Keweenaw basalts in the Upper Peninsula of Michigan, and the Watchung basalts in New Jersey. Associated minerals include zeolites, calcite, drusy quartz, datolite, and pectolite. *See* SILICATE MINERALS. Paul B. Moore

Prenatal diagnosis

The identification of disease before the birth of a fetus. It often implies genetic diagnosis, but identification of anatomical defects as well as assessment of fetal functions and maturity are also considered. Some of the relatively common diseases that can be diagnosed prenatally are Tay-Sachs disease, cystic fibrosis, Duchenne's muscular dystrophy, hemophilia A, congenital adrenal hyperplasia, thalassemia, and sickle cell anemia.

Ultrasound imaging. Ultrasound is a method of tissue imaging that is based on analysis of reflected high-frequency sound waves directed at the tissue. This method has revolutionized obstetrics perhaps more than any other technology. Ultrasonic data have vastly improved the understanding of normal growth and development, thus permitting earlier and more accurate diagnosis of fetal disease. Fetal movements can also be observed, allowing assessment of functional well-being. Such information has enabled physicians to evaluate the fetus itself rather than relying on indirect signs that can be detected by examining the mother. This change, in turn, has made it practical to consider the fetus as a patient for whom prompt newborn treatment can be planned. The fetus can even be considered a candidate for surgical therapy or medical therapy before birth.

The benefits of ultrasound imaging of the fetus for patient care include the identification of the number of fetuses and their position in the womb; a more accurate determination of fetal (gestational) age, so that fetal growth retardation and postterm (overdue) gestation are diagnosed more accurately; the ability to diagnose most fetal anomalies with obvious structural or functional implications; and the improved assessment of associated structures, such as the placenta, umbilical cord, and amniotic fluid. *See* MEDICAL IMAGING; MEDICAL ULTRASONIC TO-MOGRAPHY.

Assessment of fetal well-being. A traditional way of monitoring the baby's tolerance of labor was to listen to the mother's abdomen with a stethoscope. Modern technology has made possible the direct measurement of fetal heart rate, using techniques such as electrocardiography. Monitoring during labor consists of analyzing fetal heart rate changes occurring in response to uterine contractions. Certain abnormal patterns in the fetal heart rate may suggest possible asphyxia (inadequate delivery of oxygen to tissues and inadequate removal of carbon dioxide).

Basic surveillance of fetal status before labor is often accomplished in part by having pregnant women count the number of fetal movements over a given time. In addition, nonstress testing is performed by checking for fetal heart rate accelerations following fetal movements. Contraction-stress testing looks at fetal heart rate changes following uterine contractions, which can occur either spontaneously or in response to the hormone oxytocin.

Another measure of fetal well-being is the biophysical profile. It is an index of chronic (long-standing) asphyxia and combines nonstress testing with ultrasonic measurement of fetal body tone and movements, fetal breathing measurements, and amniotic fluid volume. Such tests may indicate a high likelihood of fetal asphyxia and may prompt the clinician to use more direct methods of analysis such as a sample of fetal blood obtained from the fetal skin during labor or from the umbilical cord.

Assessment of fetal lung maturity. Respiratory distress syndrome of the newborn is the major cause of illness of preterm infants. Many cases result from an inadequacy of surfactant, a fatty substance that is normally produced by mature fetal lungs at term and that permits lung expansion after birth. Unfortunately, gestational age is not always a good predictor of lung maturity. However, chemical analysis of amniotic fluid for surfactant produced by the fetus is possible. Certain values for the surfactant phospholipids strongly suggest functional lung maturity, thus making respiratory distress syndrome very unlikely. If for medical reasons the fetus must be delivered before the lungs have reached maturity, steroid hormones may be used to induce surfactant production. See INFANT RESPIRATORY DISTRESS SYNDROME.

Genetic diagnosis. Obstetricians typically review the health history of the pregnant woman, the fatherto-be, and their families in order to identify any possible heritable disorders in the families. Certain risks related to the patient's age, race, and geographic origin may be noted. Based on these assessments, genetic testing of the unborn child may be discussed or recommended. Using cells from the fetus, a prenatal genetic diagnosis can be made for at least 10% of those disorders that are known or assumed to result from a gene mutation. Such diagnosis is based on deoxyribonucleic acid (DNA) analysis or on detection of an abnormal enzyme or other protein produced by the defective gene. Expert counseling must be available to the parents to ensure complete understanding of the strengths, limitations, and results of any testing procedure on the fetus. *See* HUMAN GENETICS.

Amniocentesis. Once the feasibility of withdrawing a small amount of amniotic fluid was demonstrated for the management of Rh incompatibility, it was realized that genetic diagnosis could be based on analysis either of fetal biochemical products in the fluid or of living fetal cells contained therein while posing little or no risk to mother or fetus. In amniocentesis, 20–30 cm³ of amniotic fluid is aspirated with a long needle and syringe. This procedure is performed at about 15–17 weeks' gestation. If multiple fetuses are present, the gestational sacs can be sampled separately. *See* RH INCOMPATIBILITY.

Chorionic villus sampling. At 9–11 weeks' gestational age, a thin catheter may be inserted under ultrasound guidance through the uterine cervix and into the placenta, where about 10–25 mg of chorionic tissue is aspirated. An alternate method for obtaining the specimen involves the insertion of the catheter via a needle through the abdominal and uterine walls before obtaining the placental specimen. This technique permits a biopsy of the placenta, which is primarily of fetal origin, thus providing living fetal cells for genetic analysis.

Applications. As more precise prenatal diagnosis becomes available, the opportunities for fetal therapy continue to expand. Imaging techniques, such as magnetic resonance imaging, may facilitate visualization of elusive structures, such as the fetal brain, and may even permit assessment of metabolic wellbeing. Refinements in ultrasound techniques will permit more refined fetal surgery. Fetal distress may become classifiable by its cause, permitting rational intrauterine therapy, such as nutritional treatment for growth-retarded fetuses. Placental bloodflow compromise in maternal conditions such as hypertension, and fetal circulatory disturbances accompanying chronic asphyxia, may be detectable with Doppler ultrasound, and noninvasive assessment of fetal lung maturity by ultrasound or other techniques may become possible. Finally, very early prenatal diagnosis may become feasible by biopsying the preimplantation embryo when genetic disease is likely. Gene therapy of abnormal embryos may also become a clinical possibility. See CONGENITAL ANOMALIES; PREGNANCY. **Donald McNellis**

Bibliography. R. K. Creasy and R. Resnik (eds.), *Maternal-Fetal Medicine: Principles and Practice*, 4th ed., 1999; K. C. Due, *Fetal Echocardiography: Principles and Applications*, 1988; R. G. Edwards and P. Hollands, New advances in human embryology: Implications of the preimplantation diagnosis of genetic disease, *Hum. Reprod.*, 3(4):549–556, 1988; S. G. Gabbe, J. F. Niebyl, and J. L. Simpson (eds.), *Obstetrics: Normal and Problem Pregnancies*, 3d ed., 1996; D. Maulik and D. McNellis, *Doppler Ultrasound Measurement of Maternal-Fetal Hemodynamics*, 1987.

Press fit

A force fit that has negative allowance; that is, the bore in the fitted member is smaller than the shaft which is pressed into the bore. Tight fits have slight negative allowance so that light pressure is required to assemble the parts; they are used for gears, pulleys, cranks, and rocker arms. Medium force fits have somewhat greater negative allowance and require considerable pressure for assembly; they are used for fastening locomotive wheels, car wheels, and motor armatures. *See* ALLOWANCE; FORCE FIT; SHRINK FIT. Paul H. Black

Pressure

Pressure on a surface immersed in a fluid is defined as the normal force per unit area. Pressure always tends to compress the object on which it acts. This definition identifies the so-called static pressure *p*.

Another distinction is made between the absolute pressure, which is measured with respect to a perfect vacuum (that is, zero pressure), and the gauge pressure, which is measured with respect to the atmospheric value. In the International System of Units (SI), the pressure unit is 1 Pa (pascal), corresponding to a force of 1 N (1 newton = 1 kg m/s²) acting on a surface area of 1 m². In the English system, the pressure unit is 1 psi (pound per square inch), corresponding to a force of 1 lbf acting on a surface area of 1 in.²; 1 psi \approx 6895 Pa. Other common units of pressure are 1 bar = 100,000 Pa, 1 atmosphere (atm) = 101,325 Pa, 1 in. $H_2O \approx 249$ Pa, and 1 torr $(1 \text{ mm Hg}) \approx 133 \text{ Pa}$. The standard atmospheric pressure at sea level is 101,325 Pa, whereas the pressure in interstellar space approaches zero. When considering mixtures of ideal gases (that is, such that each component does not influence the others), one may define the partial pressure of each component by assuming that it occupies the entire volume separately from the others; then the pressure of the mixture would be the sum of the partial pressures of all components. See DALTON'S LAW; FORCE; GAS; METRIC SYSTEM; PHYSICAL MEASUREMENT; UNITS OF MEASURE-MENT.

In static fluids with uniform density, the pressure difference between two points at different elevations is called the hydrostatic pressure difference; it increases downward and is equal to $\rho g \Delta b$ ($g \approx 9.81 \text{ m/s}^2 \approx 32.2 \text{ ft/s}^2$ is the gravitational acceleration and Δb is the elevation difference). To define a "static" pressure within a moving fluid, one must average the normal stresses in three orthogonal directions. In addition to their static pressure, flowing fluids are characterized by their total pressure $p_o = p + 1/2\rho V^2$ (ρ is the density and V is the flow

velocity). This is the pressure that would be achieved if the fluid were brought to rest, as would be the case if a fluid jet impacted on a wall. The term $\frac{1}{2}\rho V^2$ is referred to as dynamic pressure. *See* PRESSURE MEA-SUREMENT. Stavros Tavoularis

Pressure measurement

The determination of the magnitude of the normal force applied by a fluid to a unit area. Pressure measurements are generally classified as absolute pressure, gauge pressure, or differential pressure. Absolute pressure is the pressure referenced to perfect vacuum, which by definition has zero pressure. Gauge pressure is the difference between an absolute pressure and the pressure of the atmosphere. Differential pressure is the difference between any two pressures, neither of which is atmospheric. *See* PRESSURE.

Pressure measurement is one of the most common activities in the laboratory, whether for educational purposes or for research, because the pressure level has a significant effect on most physical, chemical, and biological processes. Pressure levels of a great variety of fluids, including air, water, steam, and oil, are measured routinely in all power, manufacturing, and process industries as well as in powered vehicles of most types. In the process industries, pressure is measured and controlled to maintain uniformity of product, to guide in safe plant operation, to determine pumping head for fluid transfer, and to measure other variables indirectly, including weight, liquid level, temperature, flow and density of fluids, and hydraulic forces.

Pressure can be measured by a variety of instruments whose operation is based on different principles and according to which they can be classified as liquid-column manometers, deadweight pressure gauges, mechanical pressure gauges, and electrical pressure transducers. Each of these general categories is subdivided into more specific types. With the exception of the liquid-column manometers and the deadweight pressure gauges, whose output can be related to pressure from first principles, all other pressure gauges and transducers have an empirical response, which means that they need to be calibrated versus a standard (usually a manometer or a deadweight pressure gauge). Factory calibration is normally provided, but it may be necessary to check the instrument's response at regular intervals, as it may be affected by temperature changes, humidity, vibrations, and other disturbances.

The term pressure gauge commonly signifies a selfcontained instrument that senses pressure and indicates its value, usually by some mechanical linkage on a graduated scale. The term pressure transducer may also be used to identify some pressure gauges, but usually applies to more complex instruments, which comprise a pressure-sensing element; a device that converts motion or force produced by the sensing element to a change of an electrical, mechanical, or pneumatic parameter; and sometimes components for the conditioning of the output, such as temperature compensation. A transducer may include an indicating or recording device or may generate a pneumatic output or an electric signal that can be used as input to other devices for recording or control purposes. A related term is pressure transmitter. Although pneumatic and mechanical transducers are commonly used, electrical measurement of pressure is often preferred because of a need for long-distance transmission, higher accuracy requirements, more favorable economics, or quicker response. Especially for control applications, pneumatic pressure signal transmission may be desirable over electrical where flammable materials are present. *See* PROCESS CON-TROL.

Liquid-column manometers. A liquid-column manometer consists of a vertically oriented U-shaped tube, which is partly filled with a nonvolatile liquid. Manometric fluids are color-coded and most commonly consist of water, alcohol, or various mineral oils, while the previously common use of mercury is currently restricted by safety regulations.

If one leg is left open to the atmosphere and the other one is connected to a gas (**Fig. 1**), the difference *b* (head) in liquid level between the two legs is a direct measure of gauge pressure p_g in the gas, according to the relationship

$$P_g = \rho g b$$
,

where ρ is the density of the liquid and $g \approx 9.81 \text{ m/s}^2 \approx 32.2 \text{ ft/s}^2$ is the gravitational acceleration. If each leg is connected to a gas, the manometer measures the differential, rather than the gauge, pressure. The level rises in the low-pressure leg and drops in the high-pressure leg.

If the manometer is connected to a liquid (which must be immiscible with the manometric fluid) instead of a gas, the relationship $p_g = \rho g h$ has to be corrected for the pressure corresponding to the column of the experimental fluid. The smaller is the difference between the densities of the manometric and experimental fluids, the higher is the sensitivity of the manometer, which means that a higher column is generated by a given pressure difference. Thus, for measuring air pressure, an alcohol manometer is more sensitive than a water manometer, which in turn is more sensitive than a mercury manometer. Other arrangements have been devised to increase the sensitivity of the manometer or the ease of reading its output. A common type is the inclined

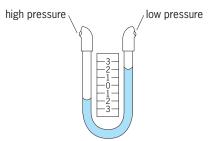


Fig. 1. Liquid-column (U-tube) manometer.

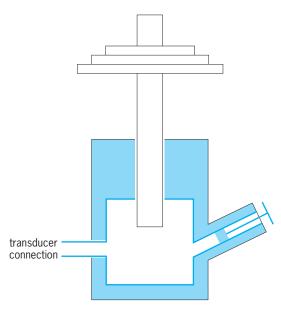


Fig. 2. Deadweight pressure gauge.

manometer, in which the tube is oriented by an angle to the vertical direction, such that the elevation of the liquid surface is read along an inclined scale.

A special type of liquid-column manometer is the mercury barometer. It consists of a single vertical glass tube, whose top part is evacuated and sealed, while its bottom end is immersed in a small tank of mercury, which is open to the atmosphere. The height of the rising column of mercury measures the absolute atmospheric pressure. *See* BAROMETER.

Deadweight pressure gauges. These gauges (**Fig. 2**) are not normally used to monitor pressure, but serve as standards for the calibration of other pressure instrumentation because they are very precise. The pressure is measured directly as the ratio of a precisely known weight required to balance a piston, whose end is immersed in a pressurized hydraulic fluid, and the piston's precisely known area. The pressure gauge or transducer that is to be checked or calibrated is connected to the hydraulic reservoir. Weights corresponding to the desired pressure are placed on the piston, and the hydraulic pressure is adjusted with the use of a pump. The piston and weights are rotated to reduce the effect of friction.

Mechanical pressure gauges. These are in wide use throughout industry due to their low cost, sturdiness, and freedom from the operational limitations of liquid manometers. There are three common elements that serve as sensors of pressure differences: Bourdon, diaphragm, and bellows. These elements may be designed to produce either motion or force under applied pressure. The more common motion type may directly position the pointer of a concentric indicating gauge; position a linkage to operate a recording pen, or a pneumatic relaying system to convert the measurement into a pneumatic signal; or position an electrical transducer to convert to an electrical signal. Depending on its design, it may measure absolute, gauge, or differential pressures. Accuracies vary depending on materials, design, and precision of components.

Bourdon-spring gages (**Fig. 3**), in which pressure acts on a shaped, flattened, elastic tube, are by far the most widely used type of instrument. When the tube is pressurized, it deforms such that its crosssectional shape tends toward a circular one. These gauges are simple, rugged, and inexpensive. Common types have a C-, spiral-, helix-, or twisted-tube element.

In diaphragm-element gauges, pressure applied to one or more contoured diaphragm disks acts against a spring or against the elastic deformation of the diaphragms, producing a measurable motion (**Fig. 4**). Size, number, and thickness of the disks determine the range.

In bellows-element gauges, pressure in or around the bellows moves the end plate of the bellows against a calibrated spring, producing a measurable motion (**Fig. 5**).

Electrical pressure transducers. Such transducers convert the pressure sensed by an element to an electrical signal by a resulting change of an electrical property, including resistance, capacitance, or inductance.

Variable-resistance pressure transducers. Pressure is measured in these transducers by an element that changes its electrical resistance as a function of pressure.

Many types of such transducers use a movable contact, positioned by the pressure-sensing element. One form is a contact sliding along a continuous resistor, which may be straight-wire, wire-wound, or nonmetallic such as carbon. If the cross section of the resistor is constant, the change in resistance is proportional to the motion of the contact. The cross

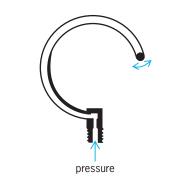


Fig. 3. Bourdon-spring gauge with C-tube element.

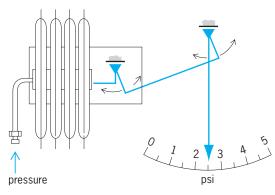


Fig. 4. Diaphragm-element gauge.

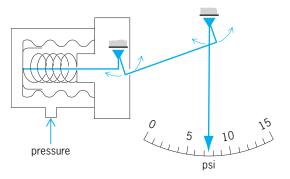


Fig. 5. Bellows-element gauge.

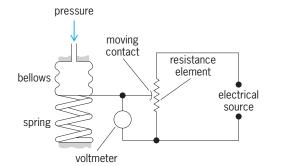


Fig. 6. Variable-resistance pressure transducer.

section may be made nonuniform to give a nonlinear relation between motion and change of resistance. The resistance element may be curved or part of an arc for convenience in measuring angular motion.

Figure 6 shows one type of resistive pressure transducer. A bellows opposed by a precisely designed spring senses the pressure and converts the pressure to a linear motion of the plate between the bellows and the spring. The plate bears a contact which wipes the surface of the precision wire-wound resistor. If a constant potential (AC or DC) is maintained across the resistor and if the resistance of the voltmeter is high with respect to the resistor, the measured voltage is a precise measure of the pressure. *See* ELECTRICAL RESISTANCE; RESISTANCE MEASUREMENT.

Strain-gauge pressure transducers. These might be considered to be resistive transducers but are usually classified separately. Strain-gauge pressure transducers convert a physical displacement into an electrical signal. When a wire is placed in tension, its electrical resistance increases. The change in resistance is a measure of the displacement, hence of the pressure. Its advantages include high resolution and small size. The strain gauge is used in conjunction with a bridge circuit. Often four resistors are connected in a bridge circuit so that an applied pressure leaves one pair of resistors in compression and the other pair in tension, yielding the maximum output change. *See* BRIDGE CIRCUIT; STRAIN GAUGE.

Variable-reluctance pressure transducers. In this type, a change of pressure is converted into change of magnetic reluctance when one part of a magnetic circuit is moved by a pressure-sensing element—Bourdon tube, bellows, or diaphragm. Motion produces a

change of magnetic reluctance (reluctance is the resistance of some material to the transmission of a magnetic field) in a magnetic circuit, which is directly related to pressure. The change of reluctance is usually within one or two coils, wound about the magnetic core in the magnetic circuit. *See* MAG-NETISM; RELUCTANCE.

A representative reluctance-changing device is shown in **Fig.** 7. A Bourdon-type or other pressuresensing device tilts the armature. The reluctances in the magnetic paths A and B are determined chiefly by the lengths of the air gaps between the armature and the core. The inductance and inductive reactance of each winding depend on the reluctance in its magnetic path. *See* INDUCTANCE; REACTANCE.

If the armature is at a neutral symmetrical position, the air gaps are equal, and the inductive reactances X_{LA} and X_{LB} are equal. Change of pressure decreases one air gap and increases the other, thus changing the ratio of the inductive reactances X_{LA} and X_{LB} . These changes can be used in a variety of circuits to produce an electrical signal that is a measure of pressure.

Variable-inductance pressure transducer. A change in inductance and inductive reactance of one or more windings is produced by the movement of a magnetic core that is positioned by a Bourdon tube or other pressure-sensing element. Unlike the action of a reluctance-type transducer, the inductance change is caused by a change in air gap within the winding, rather than in a relatively remote portion of the magnetic circuit.

Figure 8 shows a representative ratio-type inductive device. The pressure-sensing element moves the core in response to changes of pressure. When the core is in a central position, the inductances of the two coils are equal. When a pressure change moves the core, the ratio of the two inductances is changed. Energy is supplied to the coils by the same bridge circuit that measures the ratio of inductances.

Another form of inductive pressure transducer is

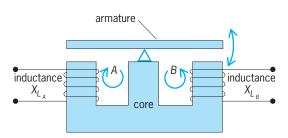


Fig. 7. Variable-reluctance pressure transducer.

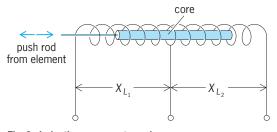


Fig. 8. Inductive pressure transducer.

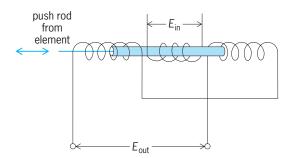


Fig. 9. Linear variable differential transformer.

the linear variable differential transformer (LVDT; **Fig. 9**). When the core is centered, equal voltages are induced in two oppositely wound secondary windings and the output voltage is zero. A change of pressure moves the core, increasing the voltage induced in one secondary and decreasing the voltage induced in the other. The change in output (differential) voltage is thus a measure of the pressure.

Piezoelectric pressure transducers. Some crystals produce an electric potential when placed under stress by a pressure-sensing element. The stress must be carefully oriented with respect to a specific axis of the crystal. *See* PIEZOELECTRICITY.

Suitable crystals include naturally occurring quartz and tourmaline, and synthetic crystals such as Rochelle salts and barium titanate. The natural crystals are more rugged and less subject to drift. Although the synthetic crystals offer much higher voltage output, an amplifier is usually required for both types. Crystal transducers offer a high speed of response and are widely used for dynamic pressure measurements in such applications as ballistics and engine pressures.

Variable-capacitance pressure transducers. Almost invariably, these sense pressure by means of a metallic diaphragm, which is also used as one plate of a capacitor. Any variation in pressure changes the distance between the diaphragm and the other plate or plates, thereby changing the electrical capacitance of the system. The change in capacitance can be used to modify the amplitude of an electrical signal. *See* CA-PACITANCE.

Semiconductor pressure transducers. Technologies, such as photolithography, developed for the manufactur-

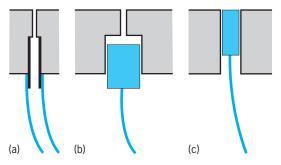


Fig. 10. Pressure taps and types of connections to transducers. (a) Plastic tubing connected to transducer. (b) Transducer in cavity. (d) Transducer flush with wall.

ing of integrated circuits have also been applied to the fabrication of silicon-based pressure transducers. Pressure is measured by its relationship to a piezoelectric potential or to a change in resistance or capacitance affected by the deformation of a semiconductor film. The great advantage of these transducers is their small size: miniature transducers are available with sizes as small as 1 mm (0.04 in.), while much smaller ones, with sizes comparable to 1 μ m, are referred to as MEMS (microelectro-mechanical) transducers. The latter can be manufactured in large, relatively inexpensive arrays, and they are very sensitive to pressure fluctuations. *See* MICRO-ELECTRO-MECHANICAL SYSTEMS (MEMS); MICROLITHOGRAPHY.

Pressure-sensing in pipes. When measuring the pressure of fluids flowing in pipes and channels, it is important not to disturb the fluid by the measuring instrumentation. Most commonly, small orifices [pressure taps, 1 mm (0.04 in.) or less, in diameter] are drilled through the wall, and the flow is connected to a manometer or pressure transducer via plastic tubing (Fig. 10a). The orifice has to be free of burrs and perpendicular to the wall. This method cannot measure accurately pressure transients of relatively fast pressure fluctuations. To improve the response, a small transducer may be embedded in a small cavity near the orifice (Fig. 10b) or flush with the internal wall (Fig. 10c). See PHYSICAL MEASUREMENT; VACUUM MEASUREMENT. Stavros Tavoularis; John H. Zifcak

Bibliography. R. P. Benedict, Fundamentals of Temperature, Pressure and Flow Measurement, 3d ed., 1984; E. O. Doebelin, Measurement Systems, 5th ed., McGraw-Hill, 2004; P. W. Harland (ed.), Pressure Gauge Handbook, Marcel Dekker, 1985; E. Ower and R. C. Pankhurst, The Measurement of Air Flow, 5th ed., Pergamon Press, 1977; S. Tavoularis, Measurement in Fluid Mechanics, Cambridge University Press, 2005; J. G. Webster (ed.), Mechanical Variables Measurement, CRC Press, 2000.

Pressure seal

A seal is used to make pressure-proof the interface (contacting surfaces) between two parts that have frequent or continual relative rotational or translational motion; such seals are known as dynamic seals, as compared with static seals. While the pressure in seals is lower than that in gaskets, the motion hinders their effectiveness so that there are more types of seals than gaskets, each type attempting to serve its environment. The materials are leather, rubber, cotton, and flax, and for piston rings, cast iron. The forms of nonmetallic seals are rectangular, V ring, and O ring. Cartridge seals are available for rolling-contact bearings. Special seals include carbon ring and labyrinth seals for turbines and mechanical seals for pumps. See GASKET; STEAM TURBINE. Paul H. Black

Pressure vessel

A cylindrical or spherical shell structure capable of withstanding internal pressure and external loadings at nozzles, supports, and attachments. The pressure vessel is designed with consideration for all possible modes of failures (such as fatigue, fracture, or collapse) and with provision of rational margins of safety against each type of failure. The Boiler and Pressure Vessel Code has been developed by the American Society of Mechanical Engineers (ASME), and the Code has been widely accepted as a standard for designing pressure vessels.

The pressure vessel may be subdivided into four regions: closure (upper) head, nozzle-shell course, beltline, and lower head regions (see **illus.**).

Heads. There are three basic types of heads: flat heads, conical and toriconical heads, and dished heads (hemispherical, ellipsoidal, or torispherical, a configuration consisting of one spherical part and one toroidal part). The highest elastic stress in an integral flat head occurs at the junction between the cylinder and the head. When the head is thicker than

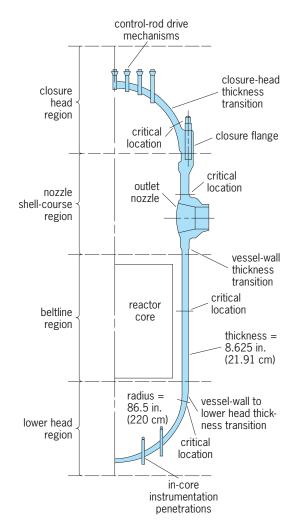


Diagram of a typical pressurized water reactor vessel. (After Pressure Vessel and Piping: Design and Analysis, A Decade of Progress, vol. 1, p. 326, American Society of Mechanical Engineers, 1972)

the cylinder (the usual case), this high stress is always in the cylinder. The highest stress in a cylindrical pressure vessel with a conical head is the axial stress at the junction. This is similar to the situation in a flat head. The conical head, unlike the flat head, also has large hoop stresses at the junction due to the tendency of the base of the cone to deflect inward and the cylinder to deflect outward when internal pressure is applied. The toroidal transition will help to reduce the critical stresses at the junction. The hemispherical head may be considered as a special case of the ellipsoidal head, and the torispherical head may be designed as an approximately ellipsoidal head. As the head becomes flatter and the shell becomes thinner, the possible weakness of the dished head begins to appear at the knuckle and at the junction to the shell, but even so, the stresses are markedly lower than those found in flat or conical heads. An important characteristic of dished heads is that for the flatter shapes and thinner walls there is a tendency to develop relatively high circumferential compression in the knuckle. Since this is accompanied by axial tension, there is a strong tendency toward yielding at this location, which has an important effect on the stability of the head.

Beltline. The required thickness of a cylindrical or spherical vessel under internal pressure loading is based on the criterion that the maximum hoop stress must not exceed the allowable value at operating temperatures. Since the allowable stress never exceeds 25% of the ultimate tensile strength of the material, the implication is that there is a safety factor of 4 against bursting due to excess pressure. Since ductile metals are used for the majority of pressure vessels, and the bursting of a ductile pressure vessel involves large plastic strains, it is recommended that plastic analysis be performed to determine the wall thickness.

Nozzle-shell course. Most pressure vessels are elements of systems and, as such, are subjected to loadings at the supports, nozzles, and other attachments. These loadings may sometimes produce stresses of greater importance than those produced by pressure. The higher stress concentration occurs at the junction between the nozzle and the shell. Finite element method of analysis provides good stress and deformation estimates around the nozzle. *See* BOILER; FINITE ELEMENT METHOD; HEAT EXCHANGER; HIGH-PRESSURE PROCESSES; NUCLEAR REACTOR. Y. S. Shin Bibliography. J. S. Cheung and L. S. Ong, *Pressure*

Pressurized blast furnace

Vessel and Piping Technology, 1993.

A blast furnace operated under higher than normal pressure. The pressure is obtained by throttling the off-gas line, which permits a greater volume of air to be passed through the furnace at lower velocity and results in an increasing smelting rate. *See* IRON METALLURGY; PYROMETALLURGY, NONFERROUS.

The process advanced markedly after its introduction in the United States in 1943, and is now

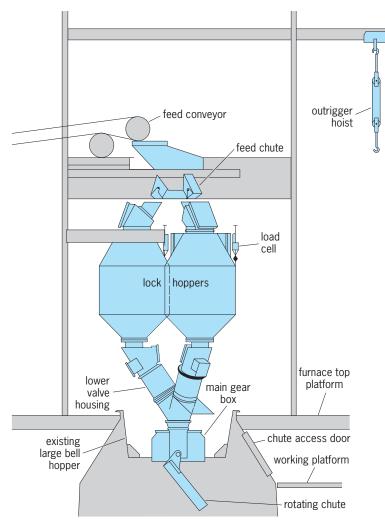


Diagram of a pressurized blast furnace with a Wurth top. (After R. A. Powell, The blast furnace top of the future, Iron Steel Eng., 50(1):86–90, January 1973)

utilized on all modern blast furnaces in the world. High-top-pressure operation has made possible the design of mammoth blast furnaces having over 140,000 ft³ (4000 m³) inner volume and hearth diameters of over 45 ft (14 m); they produce over 11,000 tons/day (10,000 metric tons/day) compared with 1540 tons/day (1400 metric tons/day) in 1943. The process permits large increases in the weight of high-temperature air blown into the bottom of the furnace at lower gas velocities, thus increasing the rate of smelting and decreasing the rate of coke consumption, and also permitting smoother operation with less flue dust production through decreased pressure drop between bottom and top pressures.

Many improvements have also been made in the mechanical features of the flow diagram in the **illustration.** Novel charging mechanisms, such as the Wurth top (lockhoppers with rotating shute), have essentially eliminated wear maintenance problems and improved the distribution of ore, coke, and limestone charged into the furnace. Also, at top pressures of 43 lb/in.² (3 kg/cm²), the throttling valve has been replaced by an expansion turbine to recover power from the top gases. Instrumentation and con-

trols have also been vastly improved. See FURNACE CONSTRUCTION. Bruce S. Old

Bibliography. A. Chatterjee, *Beyond the Blast Furnace*, 1993; W. G. Davenport, *Flash Smelting: Analysis, Control, and Optimization*, 1987; M. Higuchi, M. Ilzuka, and T. Shibuya, High top pressure operation of blast furnaces at Nippon Kokan K. K., *J. Iron Steel Inst.*, September 1973; Iron and Steel Institute of Japan (ed.), *Blast Furnace Phenomena and Modelling*, 1987; B. S. Old, E. L. Pepper, and E. R. Poor, Progress in high pressure operations of blast furnaces, *Iron Steel Eng.*, 25(5):37-43, 1948; O. D. Rice, New blast furnace design boosts iron yields, *Iron Age*, 185(4):96-98, 1960.

Prestressed concrete

Concrete with stresses induced in it before use so as to counteract stresses that will be produced by loads. Prestress is most effective with concrete, which is weak in tension, when the stresses induced are compressive. One way to produce compressive prestress is to place a concrete member between two abutments, with jacks between its ends and the abutments, and to apply pressure with the jacks. The most common way is to stretch steel bars or wires, called tendons, and to anchor them to the concrete; when they try to regain their initial length, the concrete resists and is prestressed. The tendons may be stretched with jacks or by electrical heating.

Prestressed concrete is particularly advantageous for beams. It permits steel to be used at stresses several times larger than those permitted for reinforcing bars. It permits high-strength concrete to be used economically, for in designing a member with reinforced concrete, all concrete below the neutral axis is considered to be in tension and cracked, and therefore ineffective, whereas the full cross section of a prestressed concrete beam is effective in bending. *See* CONCRETE BEAM.

An especially desirable characteristic of prestressed concrete is that as long as the material is maintained in compression it cannot crack. If cracks should appear under small overload, they generally will close when the load is removed. Sometimes concrete is prestressed principally to prevent cracking.

Basic principles. The effect of compressive prestresses is like picking up a pile of books by applying pressure to the end pair. If the pressure is large enough, none of the books will slip out.

If this concept were applied to a concrete beam in actual practice, steel tendons would be tensioned and placed along the centroidal axis of the beam. The resulting prestress would result in a uniform compression at every section (**Fig. 1**). Loads would produce both tensile and compressive stresses at the middle of the span. The prestress would combine with these to increase the compression and cancel out the tension. The whole concrete section would be effective in resisting bending, and there would be no cracks. *See* CONCRETE BEAM.

In practice, however, tendons are rarely placed along the centroidal axis. A smaller prestressing force is required, and therefore less steel for the tendons, if the steel is placed below the centroidal axis of the beam. With the eccentric prestress, stresses at each section of the unloaded beam may vary from tension at the top to compression at the bottom (**Fig. 2**).

When loads are applied to the beam, they produce both tensile and compressive stresses at the middle of the span. At the top of the beam they cause compressive stresses, which are reduced by the tensile

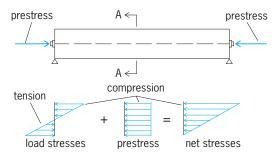


Fig. 1. Diagram showing stresses at section A-A of a concrete beam with uniform prestress.

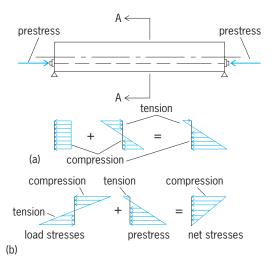


Fig. 2. Stresses at section A-A of a beam with eccentric prestress. (a) In unloaded beam, simple-bending stress component is largely counteracted by uniform compression component. (b) In loaded beam, tension components are counteracted and only compression remains.



Fig. 3. Beam with tendons draped in a vertical curve.

prestress there. Elsewhere, the tensile stresses produced by the loads are counteracted by the compressive prestress.

With this arrangement of the tendons, there is a possibility that near the ends of the beam the tensile prestress may exceed the compressive stresses produced by the loads. The net tension may be undesirable, even though very small. To avoid this condition, the tendons may be draped in a vertical curve (**Fig. 3**). The distribution of prestress at any section of a beam so prestressed is similar to that for straight tendons applying an eccentric prestress except that the stresses decrease from midspan toward the ends, as do the bending stresses due to the loads. The draped arrangement of the tendons also is advantageous in counteracting diagonal tension near the beam ends.

Continuous beams may be prestressed in a similar manner. The tendons may be placed near the bottom of the beams near midspan and near the top over supports.

Tendons. Tendons generally are made of highstrength steel so that they can serve at high working stresses. The reason for this is that losses in stress due to shrinkage and plastic flow of the concrete are relatively high. If the tendons were given a small tension, they might lose nearly all the prestress in a few months. But at a high tension, the loss might be only about 15%, because the increase in stress loss is smaller than the increase in prestress.

Wires are used for prestressing much more frequently than bars because of their greater strength. Wires may be used singly, in pairs, in cables composed of several parallel wires, or in strands. They may be stretched by electric heating, but by far the most common method of tensioning is with jacks. Various devices are used for gripping or anchoring tendons, including swaged fittings on strands, bolts threaded on bars, wedges, and buttonheads on wires.

Prestress losses. One reason for loss of prestress is elastic shortening of concrete and steel due to compression of the concrete. Another reason is creep, or plastic flow, on concrete. It is an inelastic deformation dependent on time that occurs under constant stress.

Concrete shrinks when it dries and chemical changes take place. Shrinkage is dependent on time but not on stresses due to external loading.

Other possible losses in prestress that should be considered include those due to creep of the steel and to friction when the tendons rub against the concrete. *See* CREEP (MATERIALS).

Procedures for estimating the loss of prestress due to shrinkage, creep, friction, and other causes are given in design codes, such as the American Concrete Institute *Building Code Requirements for Reinforced Concrete*, ACI 318, and the American Association of State Highway and Transportation Officials *Standard Specifications for Highway Bridges*.

Pretensioning and posttensioning. Two methods are used in fabricating prestressed beams. In one method, the concrete is bonded to the stretched steel before the prestress is applied. This is called pretensioning. In the other method, posttensioning, the prestress is applied initially through end anchorages and the concrete may or may not be bonded later to the steel.

In pretensioning, the steel is laid through the beam forms and stretched between external abutments. Next, concrete is placed in the forms and allowed to set. When it has gained sufficient strength, the external pull on the tendons is relieved, transferring the prestress to the concrete through bond. The pretensioning method can be used on casting beds several hundred feet long to mass-produce many beams simultane- ously.

In posttensioning, the tendons are prevented initially from bonding to the concrete, usually by encasement in sheaths. The concrete is placed in the beam forms around the sheathed tendons and allowed to set. When it has gained sufficient strength, jacks are used to tension the tendons, and in so doing, the jacks react against the ends of the beam. The tendons then are anchored to the concrete to apply the prestress, and the jacks are released and removed.

Frequently, grout is forced into the sheaths to bond the tendons to the beam concrete. This gives the prestressed beam greater reserve strength and better crack control under overload. Posttensioning appears to be most advantageous for long-span beams and for assembling precast beam components in the field.

Circular prestress. Circular tanks, pipe, or the ring girder of domes may be prestressed, in contrast to the linear prestressing used for beams, by wrapping with steel bars or wires under high tension. Spe-

cial machines have been developed for rapid circular prestressing with wire. *See* REINFORCED CONCRETE; STORAGE TANK. Frederick S. Merritt

Bibliography. P. W. Abeles et al., *Prestressed Concrete Designers Handbook*, 3d ed., 1981; M. Fintel, *Handbook of Concrete Engineering*, 2d ed., 1985; M. K. Hurst, *Prestressed Concrete Design*, 2d ed., 1998; P. Jones, *Concrete and Masonry*, 1983; A. H. Nilson, *Design of Prestressed Concrete*, 2d ed., 1987.

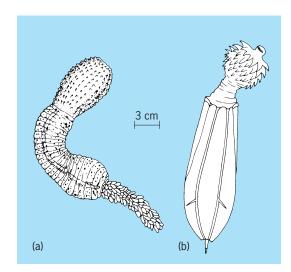
Priapulida

One of the minor groups of wormlike marine animals. Once linked with the Echiurida and the Sipunculida under the term Gephyrea, it has since been recognized as a separate phylum of the animal kingdom with uncertain zoological affinities. *See* ECHIURA; SIPUNCULA.

Priapulida inhabit the colder waters of both hemispheres. They burrow in mud and sand of the sea floor, from the intertidal region to depths of 14,800 ft (4500 m).

The phylum is a small one with only two genera, *Priapulus* and *Halicryptus*. In the former there are seven known species: *P. caudatus* and *P. bicaudatus*, which occur in the northern temperate zone and Arctic waters; *P. australis*, found around the southern tip of South America; *P. tuberculatospinosus* from Antarctic seas; *P. borridus* from the coast of Uruguay; and *P. atlantisi* and *P. profundus* from bathyal and abyssal depths, respectively, off the northeastern coast of the United States. There is but one species of *Halicryptus*, *H. spinulosus*, found in northern waters.

Priapulids are small to medium-sized animals, the largest specimen attaining 6 in. (15 cm) in length. The body of *Priapulus* is made up of three distinct portions: proboscis, trunk, and caudal appendage (**illus**. *a*). Separated by a constriction from the trunk, the bulbous, introversible proboscis usually constitutes the anterior third of the body and is marked by 25 longitudinal ridges of papillae or spines. The



Priapulida. (a) Priapulus adult and (b) larva.

mouth is located at the anterior end of the proboscis and is surrounded by concentric rows of teeth; during feeding the mouth region is everted for the capture of prey. The cylindrical trunk is annulated, but not segmented, and is often covered with irregularly dispersed spines and tubercles. At the posterior end of the trunk there are three openings: the anus and two urogenital apertures. Continuous with the body cavity, the caudal appendages consist of hollow stalks with numerous lateral vesicles. Proposed functions of the caudal appendage are respiration and chemoreception. *Halicryptus* has no caudal appendage and the proboscis is only one-sixth of the body length.

The body wall of priapulids is composed of a cuticle (periodically molted), epidermis, circular and longitudinal muscle layers, and a thin, nucleated membrane which lines the entire body cavity. Originating on the body wall and attached to the mouth region and pharynx are retractor muscles which control the introversion of the mouth and proboscis.

The digestive tract consists of a muscular pharynx, a long, straight intestinal tube, and a short rectum which leads to the terminal anus. Priapulids are predacious, feeding on polychaetes, small crustaceans, and other priapulids which they swallow whole.

The simple nervous system includes a nerve ring encircling the pharynx and a midventral nerve cord situated within the epidermis.

A pair of elongate urogenital organs lies in the body cavity, one on either side of the intestine. Each organ is made up of a central duct into which the gonad opens on one side and the excretory cells (solenocytes) on the other side. The duct carries both excretory products and gametes to the exterior through a terminal urogenital pore.

Sexes are separate. Gametes are shed into the seawater where fertilization takes place. The larva, similar in morphology to the adult, possesses a retractable proboscis armed with spines (illus. *b*). Differing from the adult, the larval trunk is encased in a cuticularized lorica, composed of several plates, which is eventually shed. Mary E. Rice

Bibliography. L. H. Hyman, *The Invertebrates: Acanthocephala, Aschelminthes, and Entoprocta,* vol. 3, 1951; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms,* 2 vols., 1982; H. L. Sanders and R. R. Hessler, *Priapulus atlantisi* and *Priapulus profundus*: Two new species of priapulids from bathyal and abyssal depths of the North Atlantic, *Deep Res.,* 9:125-130, 1962.

Prilling

Solidification of droplets of molten material freefalling against an upward stream of air in a tower. It is a process used extensively in nitrogen fertilizer manufacturing. Melt is dispersed in the top of the tower at a temperature just above the solidification point of the material being processed. The dispersion arrangement, air flow, tower dimensions, and feed material are selected so the droplets approach a spherical shape and solidify before reaching the bottom of the tower.

Prilling has long been a major way of agglomerating ammonium nitrate and urea because of its relative simplicity and economy. It has been estimated that prilling urea is slightly less costly than pan granulation. Prills of ammonium nitrate and urea are smooth, spherical, dust-free, and moderately rugged, but usually are smaller and more fragile than granulated fertilizers.

Tower heights for prilling ammonium nitrate and urea usually range from 72 to 240 ft (22 to 72 m). A process for prilling ammonium nitrate has been developed in which the melt droplets fall through an upward-moving cloud of suitable airborne dust in a tower only 23 ft (7 m) tall, dust conditioning of the prills is accomplished simultaneously, and control of dust emissions can be accomplished economically.

Prills, mostly <0.1 in. (3 mm) in diameter, are too small for effective use in blends of granulated fertilizers, and their production in larger sizes increases cost and lowers product quality. Because of the large volume and high velocity of air through prilling towers, abatement of dust, and especially of ammonium nitrate fume, can present problems and increase production cost. Despite these disadvantages, prilling remains in wide use. There have been reports of economical production of N-P and N-P-K fertilizers in a prilling process said to be cleaner than granulation. However, in production of urea there is a trend toward preference for granulation in new installations because of ease of granule size control and of superior granule hardness. See DRYING; FERTILIZER. E. O. Huffman

Bibliography. G. M. Blouin, Urea Costs, Granulation vs Prilling, TVA Bull. Y92, April 1975; R. W. R. Carter and A. G. Roberts, Proc. No. 110, The Fertiliser Society, October 23, 1969; G. Drake, Proc. No. 136, The Fertiliser Society, October 25, 1973; F. E. Steenwinkel and J. W. Hoogendonk, Proc. No. 109, The Fertiliser Society, April 24, 1969.

Primary vascular system (plant)

The arrangement of conducting elements which serves for two-way transportation of substances between different parts of a plant. The conducting elements are of two principal kinds: xylem, which is mainly responsible for the conduction of water together with dissolved inorganic substances upward from the root to other plant organs; and phloem, which is mainly responsible for the conduction of food materials (assimilates), a flow which may take place in either direction. In the shoot region of the plant, xylem and phloem are usually associated into vascular bundles (**Fig. 1**). In the root, however, they usually alternate with one another on different radii. *See* PHLOEM; XYLEM.

This system of vascular tissue functions while the young shoots and roots are growing in length and

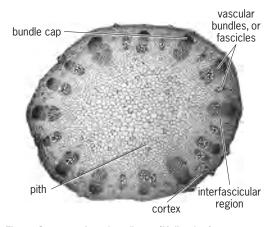


Fig. 1. Cross section of sunflower (Helianthus) stem.

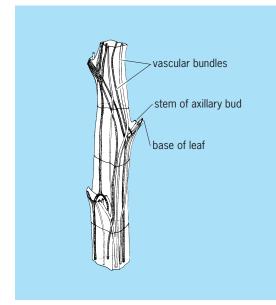


Fig. 2. Primary vascular system of the stem of the potato (Solanum tuberosum). (After A. J. Eames and L. H. MacDaniels, An Introduction to Plant Anatomy, 2d ed., McGraw-Hill, 1947)

while their tissues are undergoing differentiation and maturation. For this reason it is known as the primary vasular system, and is the only vascular tissue found in many plants. However, in most plants that increase in girth after elongation has ceased, much of the additional tissue concerned in this secondary growth is vascular. This secondary vascular tissue is formed by the activity of a cambium, which in nearly all cases develops within the primary vascular bundles.

Course of bundles in the stem. The vascular bundles in the shoot of most higher plants are interconnected to form a complex system (**Fig. 2**). The arrangement of the bundles in this system is subject to much variation in different plants, but in most dicotyledons it consists, as seen in cross section, of a ring of bundles toward the periphery of the stem. The ground tissue of the stem outside this ring is the cortex, and that to the inside of it is the pith. Cortex and pith are connected between the bundles by ground tissue known as medullary rays.

The bundles of the ring usually run up the stem more or less vertically without branching except at the positions where leaves are attached (nodes). Here the bundles frequently branch and become interconnected. Each leaf receives a supply of one or more vascular bundles which are continuous with bundles of the stem. In most monocotyledons the bundles are not arranged in a single ring, but are scattered through much of the ground tissue. This appearance is due to some of the individual bundles not running vertically in the stem, but inclining toward the center of the stem and then at a higher level returning to the stem periphery.

Concept of the stele. The term stele has been used for all tissues inside the cortex, that is, for the primary vascular system together with the pith, if any. Three major types of stele are commonly recognized. (1) The protostele consists of a solid core of xylem surrounded by phloem, but with no pith. In general, protosteles are found in the stems of most primitive vascular plants (lycopods, psilophytes, and some ferns) and in the roots of most vascular plants. (2) In the siphonostele, a cylinder of vascular tissue encloses a pith. The cylinder may be pierced by gaps at the insertions of the leaves. This stele is characteristic of many ferns. (3) In the dictyostele, the vascular cylinder is composed of bundles separated by medullary rays as well as by leaf gaps. This is the type of stele characteristic of seed plants.

Types of bundle. Vascular bundles vary in structure chiefly with reference to the arrangement of xylem and phloem. The most common type of bundle in seed plants is the collateral, in which the phloem appears on one side of the xylem (**Fig. 3**). The phloem may be rather deeply embedded in the xylem in a collateral bundle; then the xylem assumes the shape of a V in cross section (**Fig. 4**). In a bicollateral vascular bundle the phloem occurs on both sides of the xylem (**Fig. 5**). The third main type of vascular bundle is the concentric, which occurs in two forms. In the amphivasal bundle (**Fig. 6**), the xylem surrounds

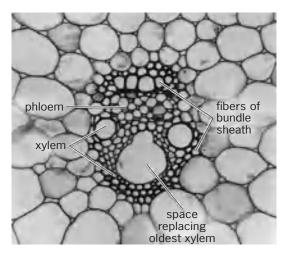


Fig. 3. Cross section of a closed collateral vascular bundle from stem of maize (Zea mays). (After K. Esau, Plant Anatomy, 2d ed., John Wiley and Sons, 1967)

the phloem, whereas in the amphicribral (**Fig.** 7) the phloem surrounds the xylem.

In addition to the conducting elements, vascular bundles contain various amounts of supporting tissue, usually fibers. These may form a complete sheath enclosing the vascular tissues, or they may appear as strands outside the phloem or the xylem, or on both sides. As seen in cross section, such strands are often called bundle caps.

Origin and development. Shoots and roots grow in length by additional tissue being formed at their extremities by growing points, or apical meristems. At first the tissue consists of more or less uniform undifferentiated cells, but gradually the arrangement of tissues found in the mature stem or root is acquired. The future vascular system can first be distinguished as strands of cells, narrower in cross section, known as procambial strands, that run through the general ground tissue. In the shoot, these arise in relation to the primordia of the leaves and develop into the pattern of vascular bundles characteristic of

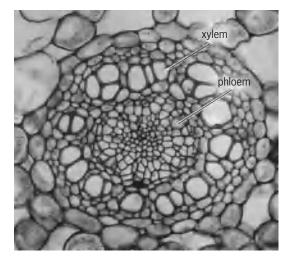


Fig. 6. Cross section of a concentric amphivasal vascular bundle from rhizome of sweet flag (*Acorus calamus*). Here xylem surrounds phloem.

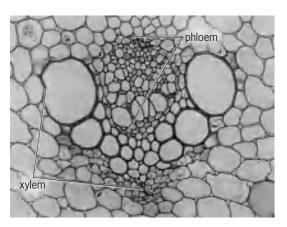


Fig. 4. Cross section of a closed collateral vascular bundle from stem of *Asparagus*.

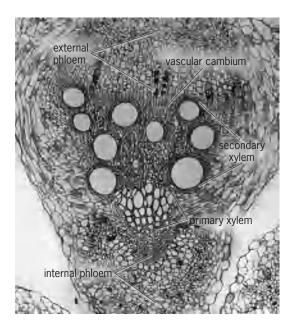


Fig. 5. Cross section of an open bicollateral vascular bundle from stem of squash (*Cucurbita*). (*After K. Esau*, *Plant Anatomy*, 2d ed., John Wiley and Sons, 1967)

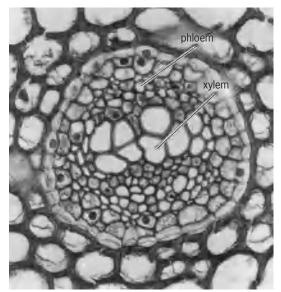


Fig. 7. Cross section of a concentric amphicribal vascular bundle from rhizome of *Polypodium*, a fern. (*After K. Esau*, *Plant Anatomy*, 2d ed., John Wiley and Sons, 1967)

the plant. In the root also, procambial strands arise in positions corresponding to the pattern of xylem and phloem strands found in the mature root.

During differentiation of a procambial strand into a vascular bundle in the shoot of a dicotyledonous plant, the cells lying next to the pith usually mature as xylem elements, and those toward the cortex as phloem. Differentiation then proceeds toward the center of the strand, and in this way a typical collateral bundle is formed. If all the cambial cells develop into xylem or phloem, the resulting bundle is incapable of further growth and is commonly called a closed vascular bundle. Such bundles are characteristic of lower vascular plants, monocotyledons, and those dicotyledons that have no secondary growth. If secondary growth occurs, the last procambial cells become cells of the vascular cambium. These cambial cells are capable of continued growth and division, their products differentiating as further xylem and phloem elements. Such enlargement of the vascular bundle by a cambium is referred to as secondary growth. The sectors of cambium within each vascular bundle usually become linked into a complete cambial cylinder by the development of similar cambial growth in the medullary rays. These linking sectors are known as interfascicular cambium and also produce secondary xylem and phloem. As a result, the secondary vascular tissues form a continuous cylinder and not a system of distinct strands. *See* PLANT GROWTH; PLANT TISSUE SYSTEMS. William B. Philipson

Primates

The mammalian order to which humans belong. Primates are generally arboreal mammals (living in trees) with a geographic distribution largely restricted to the Tropics. Unlike most other mammalian orders, the primates cannot be defined by a diagnostic suite of specializations, but are characterized by a combination of primitive features and progressive trends. These include.

1. Increased dominance of vision over olfaction, with eyes more frontally directed, development of stereoscopic vision, and reduction in the length of the snout.

2. Eye sockets of the skull completely encircled by bone.

3. Loss of an incisor and premolar from each half of the upper and lower jaws with respect to primitive placental mammals.

4. Increased size and complexity of the brain, especially those centers involving vision, memory, and learning.

5. Development of grasping hands and feet, with a tendency to use the hands rather than the snout as the primary exploratory and manipulative organ.

6. Progressive elaboration of the placenta in conjunction with longer gestation period, small litter size (only one or two infants), and precocial young.

7. Increased period of infant dependency and more intensive parenting.

8. A tendency to live in complex, long-lasting social groups.

It has been recognized for a long time that many of these features are adaptations for living in trees. However, it has been proposed more recently that primates may have developed their specializations as a consequence of being visually directed predators, living among the smaller branches of the forest canopy or undergrowth, that captured insects with their hands.

Classification of the primates is as follows:

Order Primates Suborder Strepsirhini Infraorder Lorisiformes Superfamily Lorisoidea Family: Lorisidae (lorises) Galagidae (bushbabies)

Infraorder Lemuriformes Superfamily Lemuroidea Family: Cheirogaleidae (dwarf lemurs) Lepilemuridae (sportive lemur) Lemuridae (true lemurs) Indriidae (sifakas, indri, woolly lemur) Daubentoniidae (aye-aye) Suborder Haplorhini Hyporder Tarsiiformes Superfamily Tarsioidea Family Tarsiidae (tarsiers) Hyporder Anthropoidea Infraorder Platyrrhini Superfamily Ceboidea Family: Callitrichidae (marmosets, tamarins) Cebidae (capuchins, squirrel monkeys, douroucoulis, titis) Atelidae (sakis, uakaris, howler monkeys, spider monkeys, woolly monkeys) Infraorder Catarrhini Superfamily Cercopithecoidea Family Cercopithecidae (Old World monkeys) Superfamily Hominoidea Family: Hylobatidae (gibbons, siamang) Hominidae (orangutan, gorilla, chimpanzees, humans)

Strepsirhines. There are two major groups of primates: the strepsirhines or "lower" primates, and the haplorhines or "higher" primates. Strepsirhines have elongated and forwardly projecting lower front teeth that form a toothcomb, used for grooming the fur and for obtaining resins and gums from trees as a source of food. The digits of the hands and feet bear flattened nails, rather than claws, except for the second toe, which retains a sharp toilet claw for grooming. They also have a moist, naked rhinarium and cleft upper lip (similar to the wet noses of dogs). Most strepsirhines are nocturnal, with large eyes and a special reflective layer (the tapetum lucidum) behind the retina that intensifies images in low light. Compared with haplorhines, the brain size is relatively small and the snout tends to be longer.

The strepsirhines are subdivided into two major groups: the lorisoids, which are found throughout tropical Africa and Asia, and the lemuroids, which are restricted to Madagascar.

Lorisoids. These include the galagids or bushbabies (Galago, Otolemur, Euoticus, and Galagoides) and the lorisids or lorises (Loris, Nycticebus, Perodicticus, Pseudopotto, and Arctocebus). They are small nocturnal primates, in which the largest species, the greater bushbaby, weighs only about 1 kg (2 lb). Their diet consists mainly of a combination of insects, fruits, and gums. Lorisoids are semisolitary, living in small, dispersed social groups. Each male has a territory that overlaps with that of several females, but individuals usually associate only during courtship and mating. At other times,



Fig. 1. Mouse lemur (*Microcebus murinus*). (Photo by David Haring/Duke Lemur Center)

communication is maintained by scent marking or vocalizations. Bushbabies are widespread in sub-Saharan Africa, ranging from tropical rainforests to woodland savanna. They are specialized vertical clingers and leapers, with elongated lower limbs. Lorises progress much more cautiously, employing a slow deliberate mode of quadrupedal locomotion. They are specialized for this type of movement by having powerful grasping hands with a large thumb and highly reduced second finger, a long flexible body, and a tail that is very short or absent.

Lemuroids. The greatest diversity of strepsirhines is found on Madagascar, where more than 30 species are represented, belonging to five different families. The cheirogaleids or dwarf lemurs include eight species of small lemuroids belonging to Microcebus, Mirza, Allocebus, Cheirogaleus, and Phaner (Fig. 1). This group includes the smallest known primate, the pygmy mouse lemur (Microcebus myoxinus), which weighs as little as 25 g. All cheirogaleids move quadrupedally. They are nocturnal, sleeping during the day in nests or holes in trees, often in small groups. Cheirogaleus is unusual in that it hibernates during the dry season, utilizing reserves of fat stored in its tail. Cheirogaleids are omnivorous, feeding mainly on fruits, flowers, gums, and insects. Like lorisoids, semisolitary males have ranges that overlap that of several females. Litters can be as large as four, but one or two infants are most common.

Lepilemuridae contains a single living genus, Lepilemur (the sportive or weasel lemur). They are medium-sized nocturnal lemurs, weighing less than 1 kg (2 lb), and are specialized for vertical clinging and leaping. Their diet consists almost exclusively of leaves and flowers. They are unique among primates in reingesting their feces in order to extract adequate nourishment from their poor-quality diet. True lemurs (Lemur, Eulemur, and Varecia) and bamboo lemurs (Hapalemur) are included in the family Lemuridae. The best-known member of this group is the ringtailed lemur, which is common in captivity. True lemurs are cat-sized primates (about 2-4 kg or 4-9 lb) that move quadrupedally in the trees and on the ground. They are mostly active during the day, and eat a variety of leaves, fruits, and flowers. Lemurs are gregarious and live in large social groups. Bamboo lemurs specialize in eating bamboo

shoots and pith. They are generally smaller than true lemurs, and live in one-male groups.

The indriids include the indri (Indri), the sifaka (Propithecus), and the woolly lemur (Avahi). All are specialized for vertical clinging and leaping. The indri and sifaka are diurnal, and among the largest strepsirhines (3-7 kg or 7-15 lb), while the woolly lemur is nocturnal and somewhat smaller (about 1 kg or 2 lb). The indri is unusual among lemuroids in having a vestigial tail. All indriids subsist primarily on a diet of leaves. The indri and woolly lemur live in groups consisting of an adult pair and their offspring, while sifakas live in small, mixed groups. The aye-aye (Daubentonia madagascariensis) is a rare and specialized lemuroid belonging to its own family, the Daubentoniidae. It has a suite of unique specializations for obtaining insect larvae from dead wood. These include large ears for locating the larvae; continuously growing incisors, like a rodent, for gnawing at the bark of trees; and a long slender wirelike middle finger to extract the larvae from holes. Aye-ayes are the largest of the nocturnal lemuroids (2-3 kg or 4-7 lb).

Tarsiers. These tiny primates (weighing only about 120 g) from the islands of Southeast Asia all belong to a single genus, *Tarsius* (**Fig. 2**). They are nocturnal with the largest eyes of any primate, and unlike strepsirhines they lack the tapetum lucidum. They have legs that are much longer than the forelimbs, a long almost hairless tail with a tuft at the tip, and fleshy pads on the ends of their fingers—all adaptations for a specialized lifestyle as vertical clingers and leapers. Unlike strepsirhines, which only have a toilet claw on their second toe, tarsiers are unique among primates in having toilet claws on both their second and third toes. Their diet consists primarily



Fig. 2. Philippine tarsier (Tarsius syrichta). (Photo © 2006 William Flaxington)

of insects. In the past, tarsiers have been grouped together with the strepsirhines as prosimians, because they retain many primitive features lost in higher primates. However, tarsiers share a number of distinctive specializations with anthropoids (for example, the development of a bony partition that partially separates the eyeball from the chewing muscles behind, lack of a toothcomb, a dry hair-covered nose and undivided upper lip, a short snout with reduced facial whiskers, and the structure of the placenta) that suggest that they are more closely related to each other than either is to the strepsirhines. For this reason, tarsiers and anthropoids are classified together as haplorhines.

Anthropoids. This group contains the platyrrhines or New World monkeys and the catarrhines or Old World monkeys, apes, and humans. Anthropoids are distinguished from strepsirhines and tarsiers in having a larger brain, relatively small eyes (all anthropoids are diurnal, active by day, except for the nocturnal douroucouli from South America), eye sockets almost completely enclosed by a bony septum, the two halves of the lower jaw fused in the midline rather than separated by a cartilage, small and immobile ears, the hands and feet bearing nails with no toilet claws (except for the callitrichids that have secondarily evolved claws on all fingers and toes), a single-chambered uterus rather than two-horned, and a more advanced placenta.

New World monkeys. The platyrrhines from South and Central America are a diverse group of primates comprising more than 50 species and 16 genera. Primatologists have had a difficult time establishing a classification of platyrrhines that reflects their evolutionary interrelationships, and no consensus has been reached. There is agreement, however, that three distinct clusters can be defined: the callitrichids, the pitheciines, and the atelines. The last two groups appear to be closely related and are commonly included together in the family Atelidae. The relationships of the remaining platyrrhines are uncertain, and they are often placed together for convenience in the Cebidae.

All platyrrhines are arboreal, and they are widely distributed throughout tropical forests extending from Mexico to northern Argentina. They are small to medium-sized primates ranging from 100 g to 15 kg (0.2 to 33 lb). Platyrrhines exhibit a variety of quadrupedal locomotor types ranging from squirrellike scrambling to leaping and forelimb suspension. Atelines and capuchin monkeys are unique among primates in having a specialized prehensile tail that can grasp around branches for extra support.

The smallest New World monkeys are the callitrichids (100-700 g). This group includes the marmosets (*Cebuella* and *Callitbrix*), tamarins (*Saguinus* and *Leontopithecus*), and Goeldi's monkey (*Callimico*). The callitrichids are distinguished from other platyrrhines in having claws on all of their digits, associated with their squirrellike ability to cling to and scamper up large tree trunks, and in the absence (except for *Callimico*) of the third molar. Their diet consists mainly of gums, fruits, and insects. Marmosets have large and specialized incisors for gouging trees to obtain exudates. Callitrichids generally live in small, extended family groups containing a single adult female and one or more adult males. They typically produce twins (except for *Callimico*) that are cared for primarily by the males.

Cebids include the capuchin monkey (Cebus), which is a medium-sized platyrrhine, weighing about 3 kg (7 lb). Their teeth and dextrous hands are specialized for opening and processing hard seeds and fruits, which form an important part of their diet. Capuchins are highly intelligent primates, and toolusing behaviors have been reported. They have a semiprehensile tail which, unlike the larger atelines, is fully hair-covered. Capuchins live in small groups with several adult males and females. The squirrel monkey (Saimiri) is smaller than the capuchin and more lightly built. Squirrel monkeys travel quadrupedally through the forest, often leaping between branches and trees. Their diet consists mainly of fruits, supplemented by insects. They live in large social groups, sometimes of up to 200 individuals, dominated by a hierarchy of adult females. The cebids also include the douroucoulis (Aotus) and titis (Callicebus), which are small (about 1-2 kg or 2-4 lb), short-faced, long-tailed monkeys that live in monogamous family groups. They are agile quadrupeds, adept at leaping. Their diet consists of fruits, leaves, and insects. The douroucouli is exceptional in being the only nocturnal anthropoid. Like many strepsirhines, douroucoulis have relatively large eyes, but lack the tapetum lucidum and retain cones in the retina for color vision (typical of anthropoids).

The pitheciines are medium-sized platyrrhines (2-3 kg or 4-7 lb) that include the sakis (Pithecia and Chiropotes) and uakaris (Cacajao). They are striking monkeys with highly distinctive coats and facial hair patterns. Pitheciines have a specialized dentition associated with predation on hard seeds and nuts. They often live in large social groups, although Pithecia generally travels in family units. The atelines are the largest platyrrhines (5-15 kg or 11-33 lb), and are characterized as a group by having a long prehensile tail that bears a naked grasping pad on its underside. Howler monkeys (Alouatta) are heavy-set, slow-climbing arboreal quadrupeds that subsist mainly on fruits and leaves. Most groups contain one adult male and several females with their offspring. Howlers advertise their presence to neighboring groups by emitting loud roars. They have an enormously inflated hyoid bone below the chin that acts as a resonating chamber to enhance these vocalizations. The spider monkeys and woolly spider monkeys (Ateles and Brachyteles) have long and slender limbs (Fig. 3). The thumb is reduced to a small tubercle. They are much more agile and acrobatic than howler monkeys. During feeding on ripe fruits and young leaves, they hang below branches using different combinations of their limbs and prehensile tail.

Catarrhines. The catarrhines include all anthropoid primates from Africa, Asia, and Europe. There are two main groups: the cercopithecids or Old World monkeys, and the hominoids or apes



Fig. 3. Spider monkey (*Ateles geoffroyi*) from Panama. (*Photo by Roy Fontaine/WPRC AV Archives*)

and humans. Catarrhines are distinguished from platyrrhines by a reduction in the number of premolars from three to two in each half of the upper and lower jaw, and the development of a tubelike (rather than ringlike) tympanic bone that supports the eardrum.

Cercopithecoidea. Old World monkeys are widely distributed throughout sub-Saharan Africa and tropical Asia, and also occur in the extreme southwestern tip of the Arabian Peninsula, northwest Africa, Gibraltar (their only European record), and East Asia. They are a highly successful group comprising more than 80 species. They are distinguished from other anthropoids in having bilophodont molar teeth that bear a pair of transverse crests. They also have naked, roughened sitting pads on their rumps, called ischial callosities—a feature that they share with hylobatids. In addition, most Old World monkeys are highly sexually dimorphic, with males considerably larger than females.

Cercopithecids are subdivided into two subfamilies: the Cercopithecinae and Colobinae. Both groups have developed distinctive dietary specializations. The cercopithecines are predominantly fruit eaters, and have cheek pouches for storing food, broad incisors, and low cusps on their molars. In contrast, colobines have complex sacculated stomachs for the digestion of leaves, narrow incisors, and high molar cusps. Colobines are also distinguished from cercopithecines in having a shorter face, a broader distance between the eye sockets, longer hindlimbs for leaping, and thumbs that are highly reduced.

Cercopithecines are exclusively African, except for the highly diverse and successful macaques (Macaca), which are widely distributed from Morocco and Gibraltar to Southeast Asia. The 19 species of macaques (Fig. 4) occupy a wide range of habitats from temperate woodland to tropical forest, and they exhibit different degrees of preference for traveling on arboreal or terrestrial substrates. They are medium-sized (3-10 kg or 7-22 lb) quadrupedal monkeys with omnivorous diets. Macaques are closely related to the mangabeys (Cercocebus and Lophocebus) and baboons (Papio, Mandrillus, and Theropithecus) of Africa, which are grouped together in the tribe Papionini. The mangabeys are large, forest-dwelling monkeys with slender limbs and long tails. Baboons are the largest Old World monkeys (with males exceeding 30 kg or 66 lb). The savanna baboons (Papio) are widespread throughout sub-Saharan Africa, and they even occur in the Arabian Peninsula. They prefer woodlands and savannas, where they travel mainly on the ground, although they do require trees or rocky cliffs for sleeping in at night. Baboons, like most papionins, live in large, complex social groups with up to 80 individuals, dominated by a hierarchy of adult males. The hamadryas baboon from the dry savannas of Ethiopia and Arabia is unusual in living in small one-male groups. Mandrills and drills (Mandrillus) live in the dense tropical rainforests of western Central Africa. Adult males are distinctive in having long muzzles with brightly colored faces and rumps. The gelada



Fig. 4. Long-tailed macaque (Macaca fascicularis). (Photo by Roy Fontaine/Monkey Jungle)



Fig. 5. Guenon species, De Brazza's monkey (Cercopithecus neglectus). (Photo by Aaron Logan/LIGHTmatter Photography)

baboon (*Theropithecus*) is restricted to the treeless highlands of Ethiopia. It is a ground-living quadruped that specializes in feeding on grasses. The hands and feet have short digits as a specialization for terrestrial locomotion, while the thumbs are long and dextrous to facilitate the manipulation and harvesting of small food objects. Geladas, like hamadryas baboons, live in one-male groups with up to 20 individuals.

The guenons (Cercopithecus) and their allies (Miopithecus, Allenopithecus, and Erythrocebus) are mostly forest monkeys. They are a diverse and widespread group from sub-Saharan Africa represented by more than 20 species. Guenons (Fig. 5) are mediumsized (3-7 kg or 7-15 lb), colorful, arboreal monkeys that live in large gregarious social groups. The vervet (Cercopithecus aethiops) is unusual in spending a good deal of time on the ground (Fig. 6). The diet of guenons consists of a combination of fruits, insects, and young leaves. The closely related talapoin (Miopithecus), from western Central Africa, is the smallest Old World monkey (weighing in at just over 1 kg or 2 lb). The patas (Erythrocebus) is a graceful, slender-limbed, fast-running monkey that lives on the grassy plains, savannas, and dry woodlands of equatorial Africa.

There are two major groups of colobines or leaf



Fig. 6. Vervet monkey, or green monkey (*Cercopithecus* aethiops). (*Photo by Gary M. Stolz/U.S. Fish and Wildlife* Service)

monkeys: the African colobus monkeys and the Asian langurs. Colobus monkeys are subdivided into three genera, mainly based on their coat color: the blackand-white colobus (*Colobus*), the red colobus (*Piliocolobus*), and the olive colobus (*Procolobus*). The black-and-white colobus and red colobus are large monkeys (8–10 kg or 18–22 lb), while the olive colobus is smaller (3–4 kg or 7–9 lb). Their diet consists of leaves, with some fruits, although the black-and-white colobus specializes in eating hard seeds. Most African colobines live in large multimale groups, but black-and-white colobus monkeys live in small groups with a single adult male.

The colobines from Asia are more diverse than those from Africa, with over 20 species belonging to six genera. *Presbytis* and *Trachypithecus* are found throughout the forests of South and Southeast Asia. The most familiar and most adaptable representative is the large, semiterrestrial Hanuman langur from India and Sri Lanka, which lives in a variety of habitats ranging from forest to semidesert, and even occurs in urban centers. The other members of this group are arboreal, forest-dwelling monkeys ranging in size from 5 to 12 kg (11 to 26 lb). They mainly eat leaves, but fruits and seeds form an important part of their diets.

The other Asian colobines are commonly referred to as odd-nosed monkeys because of the peculiar shape of the nose. The best-known species is the proboscis monkey (Nasalis) from the mangrove swamps of Borneo. These large monkeys are remarkable leapers and excellent swimmers. Adult males, which are more than twice as large as females, have an enormous pendulous nose, while females have a petite upturned nose. The related simakobu (Simias) is found only on the Mentawai Islands, just off the coast of Sumatra. It is a distinctive colobine with a very short tail and relatively long forelimbs. The douc langur (Pygathrix) and snub-nosed monkeys (Rhinopithecus) from China and Indochina are large, handsome monkeys with flattened or upturned noses. The Golden snub-nosed monkey (Rhinopithecus roxellana) from the mountain forests of southern China is the largest living colobine, with males weighing almost 40 kg (88 lb).

Hominoidea. This is the superfamily to which apes and humans belong. Hominoids are distinguished from cercopithecoids in having primitive nonbilophodont molars, larger brains, longer arms than legs (except in humans), a broader chest, a shorter and less flexible lower back, and no tail. Many of these specializations relate to a more upright posture in apes, associated with a greater emphasis on vertical climbing and forelimb suspension.

Hominoids can be classified into two families: the Hylobatidae, which includes the gibbons and siamang, and the Hominidae, which includes the great apes and humans. The gibbons and siamang (*Hylobates*) are the smallest of the hominoids (4–11 kg or 9–24 lb), and for this reason they are sometimes referred to as the lesser apes. The nine or so species are common throughout the tropical forests of Asia. They are remarkable in having the longest arms of any primates, which are 30–50% longer than their

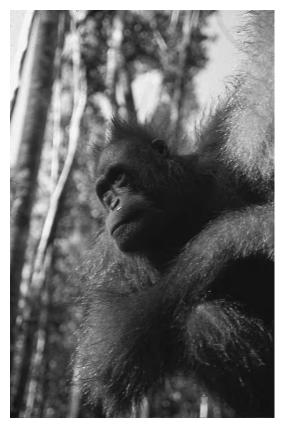


Fig. 7. Orangutan (Pongo pygmaeus pygmaeus) from Borneo.

legs. This is related to their highly specialized mode of locomotion, called brachiation, in which they swing below branches using only their forelimbs. Gibbons are fruit eaters, while the larger siamang incorporates a higher proportion of leaves in its diet. Hylobatids live in monogamous family groups in which males and females are similar in size (unlike most catarrhines). Both sexes defend a territory, mainly through complex vocalizations or songs, which serve to advertise their presence to neighboring pairs. As in Old World monkeys, hylobatids sit upright on branches to sleep, resting on their ischial callosities, rather than sleeping in nests like great apes.

The great apes include the orangutan (Pongo) from Asia and the gorilla (Gorilla) and chimpanzees (Pan) from Africa. These were formerly included together in their own family, the Pongidae, to distinguish them from humans, who were placed in the Hominidae. However, recent anatomical, molecular, and behavioral evidence has confirmed that humans are closely related to the great apes, especially to the African apes, and for this reason most scientists now classify them together in a single family, the Hominidae. The orangutan is restricted to the tropical rainforests of Borneo and northern Sumatra (Fig. 7). They are large, arboreal primates that climb cautiously through the trees using all four limbs for support. Orangutans subsist mainly on fruits. As is typical of other great apes, they lack ischial callosities and build nests to sleep in at night. Adult male orangutans are about twice as heavy as females, and the sides of their faces are adorned with huge fatty flaps. Orangutans are the least sociable of the hominoids and tend to lead solitary lives. Males aggressively defend their territory from other males, which encompasses the ranges of several female consorts.

The gorilla, the largest of the hominoids, has a disjunct distribution in tropical Africa. Because of their great size, gorillas are almost entirely terrestrial, although females and young individuals frequently climb trees. Nests are often built on the ground. Gorillas move quadrupedally, and like chimpanzees, the hands are specialized for knuckle walking in which the weight of the animal is borne on the upper surface of the middle joints of the fingers. Mountain gorillas eat a variety of leaves, stems, and roots, while lowland gorillas eat a greater proportion of fruits. Groups consist of a dominant male, called a silverback, as well as several adult females, subadults, and infants.

There are two species of chimpanzees, the common chimpanzee (Pan troglodytes) and the bonobo or pygmy chimpanzee (Pan panicus). The common chimpanzee is widely distributed in the forests and woodlands stretching across equatorial Africa, while the pygmy chimpanzee is restricted to the tropical rainforests of the Congo. Both species nest and feed in trees, but they mostly travel on the ground. Common chimpanzees have eclectic diets, including meat, which they obtain by hunting small to mediumsized mammals. Tool-using behaviors are common, and more than a dozen simple tool types have been identified. Chimpanzees are gregarious and sociable, and they live in large multimale communities that divide into smaller subgroups for foraging. See APES; FOSSIL APES; FOSSIL HUMANS; FOSSIL PRIMATES; MAM-Terry Harrison MALIA: MONKEY.

Bibliography. J. G. Fleagle, *Primate Adaptation* and Evolution, 2d ed., 1998; R. A. Mittermeier et al., *Lemurs of Madagascar*, 1994; J. R. Napier and P. H. Napier, *Natural History of the Primates*, 1985; R. M. Nowak, *Walker's Mammals of the World*, Johns Hopkins University Press, 1999; A. Richard, *Primates in Nature*, 1985; B. B. Smuts et al. (eds.), *Primate Societies*, 1987; R. Tuttle, *Apes of the World*, 1987.

Prime mover

The component of a power plant that transforms energy from the thermal or the pressure form to the mechanical form. Mechanical energy may be in the form of a rotating or a reciprocating shaft, or a jet for thrust or propulsion. The prime mover is frequently called an engine or turbine and is represented by such machines as waterwheels, hydraulic turbines, steam engines, steam turbines, windmills, gas turbines, internal combustion engines, and jet engines. These prime movers operate by either of two principles (**Fig. 1**): (1) balanced expansion, positive displacement, intermittent flow of a working fluid into and out of a piston and cylinder mechanism so that by pressure difference on the opposite sides of the piston, or its equivalent, there is relative motion

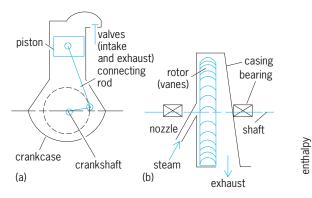


Fig. 1. Representative prime movers. (a) Single-acting four-cycle, automotive-type internal combustion engine. (b) Single-stage, impulse-type steam turbine.

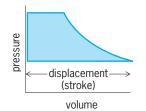


Fig. 2. Pressure-volume diagram (indicator card) for ideal, no-clearance, fluid-displacement types of prime mover.

of the machine parts; or (2) free continuous flow through a nozzle where fluid acceleration in a jet (and vane) mechanism gives relative motion to the machine parts by impulse, reaction, or both. *See* IM-PULSE TURBINE; INTERNAL COMBUSTION ENGINE; RE-ACTION TURBINE; STEAM ENGINE.

Displacement prime mover. Power output of a fluiddisplacement prime mover is conveniently determined by pressure-volume measurement recorded on an indicator card (**Fig. 2**). The area of the indicator card divided by its length is the mean effective pressure (mep) in pounds per square inch, and horsepower of the prime mover is given by Eq. (1), where

Horsepower =
$$\frac{\text{mep} \times Lan}{33,000}$$
 (1)

L is stroke in feet, *a* is piston area in square inches, and *n* is number of cycles completed per minute. Actual mep is smaller than the theoretical mep and may be related to the theoretical value by diagram factor or engine efficiency (**Table 1**). *See* MEAN EFFECTIVE PRESSURE.

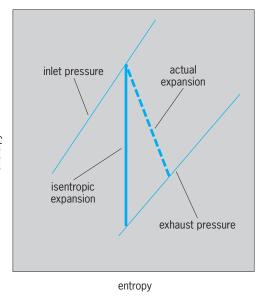


Fig. 3. Enthalpy-entropy (Mollier) chart of performance of steam- or gas-turbine type of prime mover.

Acceleration prime mover. Performance of fluid acceleration (hydraulic) prime movers is given by Eq. (2), where Q is water flow rate in cubic feet per

Horsepower =
$$\frac{QH}{8.8} \times \text{efficiency}$$
 (2)

second, and H is head in feet. For heat-power prime movers of the fluid acceleration type, actual properties of the thermodynamic fluid, as given in tables and graphs, especially the Mollier chart, permit the rapid evaluation of the work or power output from the general energy equation which resolves to the form of Eq. (3), where b is the enthalpy in Btu/lb, and the

$$\Delta W$$
, Btu/lb of fluid = $b_{\text{inlet}} - b_{\text{exhaust}}$ (3)

inlet and exhaust conditions can be connected by an isentropic expansion for ideal conditions, or modified for irreversibility to a lesser difference by engine efficiency (**Fig. 3** and **Table 2**). Fluid consumption follows from Eq. (4) or Eq. (5).

Fluid consumption, $lb/hph = 2545/\Delta W$ (4)

$$lb/kWh = 3413/\Delta W \tag{5}$$

In the fluid acceleration type of prime mover, jet velocities experienced in the nozzles can be found in feet per second, for nonexpansive fluids,

Туре	Size, hp	rpm	Stroke, in.	Bore: stroke ratio	Piston speed, ft/min	Brake mep, Ib/in. ²	Diagram factor, or engine efficiency	
Steam engine	25-500	100-300	6-24	0.8-1.2	400-600	50-100	0.6-0.8	
Automobile engine	10-300	2000-4000	3-5	0.9-1.1	1000-2000	50-100	0.4-0.6	
Aircraft engine	100-3000	2500-3500	4-7	0.8-1.1	1500-3000	100-300	0.4-0.6	
Diesel, low-speed	100-5000	100-300	10-24	0.8-1.0	500-1000	40-80	0.4-0.7	
Diesel, high-speed	25-1000	1500-2000	3-6	0.8-1.0	800-1500	50-100	0.4-0.6	

Туре	Rating, kW	Number of stages	Head, ft; or pressure, lb/in. ²	Temperature, °F (°C)	Exhaust pressure, in. Hg abs	rpm	Tip speed, ft/s	Efficiency
Pelton waterwheel Francis hydraulic	1000–200,000	1	500-5000 ft	Ambient	atm	100-1200	100-250	0.75–0.85
turbine Propeller (and Kaplan) hydraulic	1000–200,000	1	50–1000 ft	Ambient	atm†	72–360	50-200	0.8–0.9
turbine	5000-200,000	1	20-100 ft	Ambient	atm†	72-180	70-150	0.8-0.9
Small condensing steam turbine	100-5000	1–12	100-400 lb/in. ²	400–700 (200–370)	1–5	1800-10,000	200-800	0.5-0.8
Large condensing steam turbine	100,000-1,000,000	20-50	1400-4000 lb/in. ²	900–1100 (480–590)	1–3	1800-3600	500-1500	0.8-0.9
Gas turbine	500-20,000	10–20	70-100 lb/in. ²	1200–1500 (650–820)	atm	3600-10,000	500-1500	0.8-0.9

by Eq. (6). For expansive fluids, they may be found by Eq. (7),where *H* and ΔW are as given above and *C*

Jet velocity =
$$C\sqrt{2gH} = 8.02 C\sqrt{H}$$
 (6)

Jet velocity =
$$C\sqrt{2g\Delta W} = 223.7 \ C\sqrt{\Delta W}$$
 (7)

is the velocity coefficient, seldom less than 0.95 and usually from 0.98 to 0.99.

Selected representative performance values of some prime movers are presented in Tables 1 and 2. *See* GAS TURBINE; HYDRAULIC TURBINE; POWER PLANT; STEAM; STEAM TURBINE; TURBINE.

Theodore Baumeister

Bibliography. E. A. Avallone and T. Baumeister III (eds.), *Marks' Standard Handbook for Mechanical Engineers*, 10th ed., 1996.

Primitive gut

The tubular structure in embryos which differentiates into the alimentary canal. The method by which the primitive gut arises depends chiefly on the yolk content of the egg.

Two-layered blastoderms. Eggs with small or moderate amounts of yolk usually develop into spherical blastulae which invaginate at the vegetative pole to form double-walled gastrulae. The invaginated sac extends in length to become the primitive gut. In some groups, such as the echinoderms, chaetognaths, chordates, and amphibians, its external opening, the blastopore, persists as the anus. The mouth forms as a new opening by contact of the opposite end of the gut with the skin, which then perforates. In other groups, such as the annelids, arthropods, and molluscs, the blastopore becomes the mouth, and a new anal opening is formed.

Three-layered blastoderms. Animals such as fish, reptiles, and birds, having more yolk than can be cleaved, form flattened gastrulae consisting of three-layered blastoderms surmounting the yolk. Mammals also belong in this group, although the yolk has been lost secondarily in all except the monotremes. The

head is formed by a folding of the blastoderm upon itself. The endodermal layer within the head fold becomes the pharynx. This foregut is extended by an anterior growth of the whole head and by the union of lateral endodermal folds at its posterior boundary. In most forms, the hindgut arises by a similar folding in the opposite direction, the tail fold, at the posterior end of the blastoderm. The shark is an exception in which the hindgut is formed by the union of two lateral endodermal ridges beneath the notochord, much as the medullary plate folds into neural tube above the notochord. The midgut remains open until the endodermal layer of the blastoderm surrounds the yolk. The communication of the midgut with the yolk sac gradually narrows to an umbilicus as the foregut, hindgut, and lateral body folds press in from all sides, but the midgut floor is brought into position only with the absorption of the yolk sac. Mouth and anus form by contact of the pharynx and hindgut with the ectoderm at their respective ends of the body. Biochemical incompatibilities of ectoderm and endoderm cause such areas of contact to perforate and form the body orifices.

Teleost fish are an exception to the above account in that the true endoderm does not spread beneath the whole blastoderm, but is confined to the embryonic area. Along the axis of the body beneath the notochord, the endoderm condenses into a solid rod which then hollows out secondarily to form the primitive gut. *See* CLEAVAGE (DEVELOPMENTAL BIOL-OGY); GASTRULATION; OVUM. Howard L. Hamilton

Primulales

An order of flowering plants, division Magnoliophyta (Angiospermae), in the subclass Dilleniidae of the class Magnoliopsida (dicotyledons). The order consists of three families: the Myrsinaceae, with about 1000 species; the Primulaceae, with about 1000 species; and the Theophrastaceae, with a little more than 100 species. These are plants with sympetalous flowers; that is, the petals are fused by their margins to form a corolla with a basal tube and terminal lobes. The functional stamens are opposite the corolla lobes, and there is a compound ovary that has a single style and two to numerous ovules which usually have two integuments and are on a free-central or basal placenta. The Myrsinaceae and Theophrastaceae are chiefly tropical and subtropical woody plants, but the Primulaceae are mostly herbaceous and are best developed in north temperate regions. Primrose (*Primula*) and cyclamen are familiar members of the Primulaceae. *See* DILLENIIDAE; MAG-NOLIOPSIDA; PLANT KINGDOM.

Arthur Cronquist; T. M. Barkley

Printed circuit board

The printed circuit board (PCB) has been in use for over a century, but it is still one of the most important parts of electronic devices, as the circuitry links all components together into an economical and efficiently integrated system. Once called printed wiring since they were first used to replace hand wiring in early telephone switchboards and radios, printed circuit boards are now ubiquitous. The printed circuit board consists of patterned arrays of conductors, such as copper, that are held onto and within dielectric materials, such as epoxy. But many other materials are used with dozens of processes to produce a myriad of circuit types that enable thousands of products.

History. The earliest documentation of circuitry is a 1902 patent by inventor Albert Hanson, who wanted to help solve the "density dilemma" of the telephone exchanges. The telephone switchboard required countless telephone lines to be brought into proximity to allow an operator to manually connect the correct parties together. Hanson recognized that individual wires were not the best answer and conceived of an invention where cut-out strips of metal were glued to a flexible nonconductive material, such as paper or cloth. This flexible wiring array would allow conductors to be placed close together in an orderly and more easily manufactured system. Hanson thus invented a circuit that increased density (more signals per area), which could reduce cost by mass production. Hanson's designs included conductors on both sides and multilayers. Figure 1 shows the Hanson circuits from the patent figures.

Other early circuit inventors, including Thomas Edison, suggested using materials that could be painted on. Ideas moved from painting to printing, and the printed circuit concept was well on its way by the 1920s. World War II drove printed circuit technology to a new level, as electronics engineers sought robust and highly reliable assemblies for bomb and torpedo proximity detonators. The result was ceramic thick-film (CTF) circuitry, where ceramic-metal (cermet) inks were screen-printed onto alumina and fired in a high-temperature oven to fuse the conductor pattern onto substrate. This basic thick-film method, while improved, is still used today to fabricate ceramic printed circuits. Later, chemi-

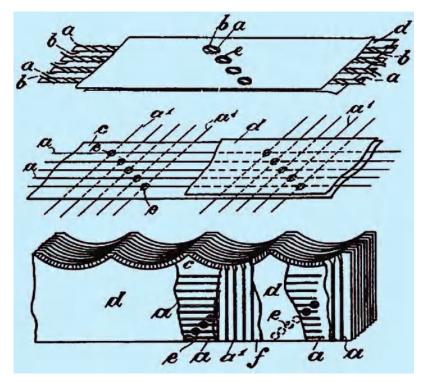


Fig. 1. First documented printed circuit; British patent 4,681, filed in 1903 by Albert Parker Hanson.

cal etching (an adaptation of photoengraving) gained favor over direct printing methods. Most modern circuits are not really printed.

Importance. The printed circuit board is at the center of modern electronics products, and its value cannot be overstated. Every computer, Internet server, digital camera, cellular phone, television, pacemaker, and satellite has one or more printed circuit boards. Printed circuit boards have become specialized during a century of progress, and products range from tiny circuits for watches and hearing aids to extremely large ones for telecommunications and space stations. Today's circuits are efficiently mass-produced using technology from many disciplines, especially chemistry, physics, and mechanical engineering.

Materials. The printed circuit board uses two basic materials: dielectrics and electrical conductors.

Dielectrics. The dielectric must isolate the conductor array, while serving as the mechanical support platform. Although thousands of dielectrics exist, only a few are used by the printed circuit board industry. Dielectric materials can be divided into inorganic, including glass and ceramic, and organic, a much larger group of polymer-based compositions. *See* DIELECTRIC MATERIALS; POLYMER.

1. *Ceramics*. Most ceramic materials have excellent electrical insulation resistance and mechanical characteristics. Several fabrication methods have been developed to produce ceramic substrates and circuitry. Ceramics are distinguished by very high temperature performance that exceeds any organic substrate. Alumina (Al_2O_3), a strong hard material, is the most common ceramic circuit substrate used. Aluminum nitride (AlN) is also used, it has good thermal and mechanical properties, and exceptional thermal conductivity for heat dissipation. *See* CERAM-ICS; INSULATION RESISTANCE TESTING.

2. *Glass*. Circuits are fabricated onto glass by the direct deposition of metals or other conductors. Glass is used for thin-film circuits, optical products, and flat-panel displays. Once the glass is metallized, electronic components can be attached by soldering and other methods. Glass circuits are much less common than ceramic and organic types. *See* SOLDERING.

3. Epoxy-glass. Thermoset (cross-linked) resins have long been used to make printed circuit board dielectrics, and nearly all plastics are good electrical insulators. Early plastics, such as Bakelite, amino resins, and phenolics, were used on some of the first printed circuit boards, but epoxy resins have become the de facto standard. Pure epoxy resin does not have the right thermal and mechanical properties for good printed circuit boards, but reinforcing it with inorganic fillers provides an adequate material. Electronic-grade glass weave (E-glass) is combined with epoxy resin and cured to produce a composite with higher strength and lower thermomechanical expansion than the polymer alone. Brominated epoxy is often added to reduce flammability (to pass Underwriters Laboratories ratings); this product is generally referred to as FR4. Different grades and types of epoxy-glass laminates are used for the many printed circuit board applications. Thickness can range from a few thousandths of an inch to very thick boards for large printed circuit boards that support heavy components. See COMPOS-ITE MATERIAL; GLASS; POLYETHER RESINS; POLYMERIC COMPOSITE.

4. *Polymer films*. Unreinforced polymer film is used to make flexible printed circuits (FPC)—a special, but extremely valuable subset of organic printed circuits. Flexible printed circuits must bend and are therefore made from thin polymer films without reinforcing glass. Polyimide (PI), one of the highest-temperature organic polymers, is inherently flame-retardant, and is the most common flexible substrate. It was introduced in the 1960s as DuPont's Kapton[®]. Today many suppliers offer polyimide circuit products. The next most important film is polyester, which is lower in cost but temperature-limited. Polyester generally is not used when soldering is required.

5. *Other materials*. Polyimide rigid composites are used for high-reliability rigid printed circuit boards, such as military assemblies. Although properties are excellent, this is one of the most costly printed circuit board dielectrics. Many specialty or lower-volume substrates are offered to meet special needs, such as fluoropolymers for very high frequency applications.

Conductors. All printed circuits use electrical conductors to carry power, ground, and electronic signals. Copper is by far the most common conductor and the industry standard. But alloys, other metals, metal composites, and even nonmetals are

used as conductors. *See* CONDUCTOR (ELECTRICITY); COPPER.

1. *Copper foil.* Two basic types of copper are used, common electrodeposited (ED) and special rolled/annealed (RA), which has high fatigue resistance making it useful for flexible printed circuits. Copper thickness can have a wide range, but 3-37 micrometers is common. Thicker copper is used for high-current circuitry. Very thin copper cannot carry much current and is used with build-up methods discussed in "Manufacturing process" below.

2. *Cermet (ceramic/metal composite)*. Ceramic circuits often use conductive inks made from metals, such as silver/palladium mixtures, and a smaller amount of ceramic binder. These ceramic/metal mixtures are referred to as cermets. The cermet is printed onto dielectric and fired at high temperatures (800–1200°C; 1500–2200°F) to fuse the ink. *See* CERMET.

3. *Thin film*. Circuit conductors can be placed on glass, ceramic, and other nonporous dielectrics by vacuum deposition. Metal can be deposited as atoms and atom clusters by energetic processes in a vacuum chamber. Thin films of copper, silver, tungsten, gold, and many other metals can become applied circuitry as thin as 1 μ m.

4. *Polymer thick film*. Polymer thick film (PTF) is the organic analog of cermet and uses a polymer binder to hold the metal particles instead of ceramic. Silver-filled ink is the most common, but carbon inks can also be used alone or mixed with silver ink. Thermoset epoxies, as well as solvent-soluble thermoplastics, are used to make printable inks, which are dried or polymerized to durable circuit conductors. Polymer thick film is a low-temperature system that can be used on nearly any dielectric, including paper.

Environmental regulations. Recent regulations, especially Reduction of Hazardous Substances (RoHS), restrict certain metals such as lead and cadmium. In the future, regulations could restrict halogens, including bromine. Printed circuit boards can no longer use the once popular tin-lead solder finish, and brominated epoxy may need to be replaced in epoxy laminates in the future.

Designing. The simplest design strategy is to involve printed circuit board designers in the product concept from the very beginning and make sure that the design team understands the technology. For example, a rigid board designer may not be qualified to design flexible printed circuits.

Computer-aided design. Computer-aided design (CAD) has been used for printed circuit boards for several decades. It would be nearly impossible to design and layout a complex printed circuit board without it. A printed circuit board can have dozens of conductor layers interconnected by vertical pathways (vias) numbering in the tens of thousands. CAD is an efficient routing tool (best wiring paths), and is used to check for design-rule compliance. CAD software is used to generate artwork for photoimaged circuits, as well as for "direct write" systems which are used

mostly for prototyping. *See* COMPUTER-AIDED DESIGN AND MANUFACTURING.

Computer simulations. Many modern circuits are more than wires. For example, high-frequency circuits have become an integral part of the electronics and circuit tuning may be required. Simulation software can virtually run a circuit design and measure electrical characteristics to modify and adjust the printed circuit board before it is built.

Tool design. The printed circuit board industry uses several hard tools such as dies for cutting out panels, but the trend is to move toward a data-driven "no tool" environment. Tools include artwork, a photographic film with clear and opaque (black) areas to define the exposure pattern for the photoimaging process. For tool generation, CAD files are typically electronically transmitted to laser photo-plotters that print the artwork. Metal cutting/blanking tools are also generated from CAD files.

Prototyping, testing, and qualifying. Designers recommend building prototype printed circuit boards, except for relatively standard products and simple modifications of existing parts. However, errors are made even with advanced CAD and simulation systems. A small run of printed circuit boards assembled with components, followed by testing, is a time-proven method. Even a "correct" printed circuit board may experience production and assembly problems that can be eliminated by adjustments to the design and artwork. The prototype, or a short production run on the factory line, can be used to generate parts that can be tested per the relevant specifications in order to pass or qualify the product.

Types (material-based classification). Printed circuit boards can be divided into two categories, organic and inorganic, but three common types are recognized: organic rigid (rigid board), organic flexible (flex circuit or FPC), and ceramic.

Ceramic/inorganic. These are specialty circuit boards that are very robust, extremely reliable, and heat-tolerant, but they are more expensive than other types (**Fig. 2**).

1. *Thick film.* The original ceramic circuits were produced by screen-printing cermet ink onto the substrate and firing it in a high-temperature oven. This is still a popular method. Although silverpalladium cermet ink is commonly used, copper and gold inks are available. Copper, since it can oxidize and become poorly conductive, may require an inert atmosphere, such as nitrogen, in the oven, which somewhat offsets the material savings.

2. *Thin film*. Thin-film circuits can use ceramic, glass, or other nonporous durable dielectric materials. Metal, including silver, platinum, palladium, gold, copper, titanium, and others, can be deposited through a mask to define a conductor pattern. Alternatively, the entire surface can be metallized and undesired metal etched away.

3. *Copper layer*. Copper may be bonded to the substrate surface by plating or vacuum deposition, followed by etching to generate the conductor array.

Organic. Circuits made from carbon-containing materials, especially plastics, were the earliest printed

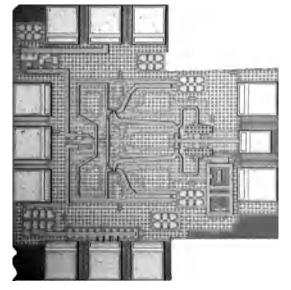


Fig. 2. Ceramic circuit for high-frequency radio.

circuit board substrates and are the dominant materials today. They can be divided into rigid and flexible printed circuit boards.

1. *Rigid*. Typified by epoxy-glass FR4, rigid circuits are the largest class of printed circuit boards (**Fig. 3**). Rigid circuits are used in nearly every application unless special properties, such as flexibility or harsh environment capability, are required. Copper foil is applied to a lay-up of epoxy resin and glass weave, followed by heating in a press to polymerize the epoxy and bond the foil. Copper may be bonded to one side (single-sided) or both sides of the (double-sided) dielectric. The conductor patterns are generally formed by chemically etching.

2. *Flexible*. Flexible circuits comprise a thin (25-to 125- μ m) dielectric plastic film with copper on one or both sides. Processing may be similar to methods used for rigid printed circuit boards, but with special handling to accommodate the flexibility. Continuous-roll (roll-to-roll) processes can be used to increase productivity and throughput. Flex is an enabling technology used for products

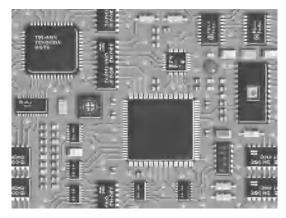


Fig. 3. Common printed circuit board made from etched copper on epoxy-glass substrate, shown with assembled components.

that combine motion and electronics, such as disk drives.

3. *Rigid-flex*. Rigid and flexible circuits can be combined by lamination to form printed circuit boards that are used in military, avionics, and some commercial products. Rigid-flex provides the maximum volumetric density, making it valuable for avionics and other complex systems. Designs typically bond the rigid substrate only to selective areas of the flex so that the final printed circuit board can be efficiently folded into a compact volume (**Fig. 4**). Some of the most complex and expensive printed circuit boards are rigid-flex designs.

4. *Molded*. The idea of molding a circuit from plastic resin has been tried with a modest level of success since the 1950s. Plastic resin can be molded into the required three-dimensional (3D) shape, followed by the addition of appropriate conductor patterns by several proven methods. The resulting printed circuit board can serve as the circuit board, interconnect, and housing. Three-dimensional molded printed circuit boards are produced today but mostly for specialty applications.

Special types. There are numerous specialty printed circuit boards for high- and low-volume products. For example, calculators use a specialty printed circuit board in which cost is paramount. Low-cost calculator circuits with integrated keypads can be made from thin polyester film printed with conductive carbon ink to avoid the cost of metal and chemical etching.

Thermal control. Two basic methods are used for increasing thermal dissipation in printed circuit boards used for power modules and other high-heatproducing products. Ceramics, such as high thermal aluminum nitride, can be used. Organics typically have poor thermal conductivity but can be modified with thermally conductive inner layers or cores. Printed circuit boards with cores are more difficult and costly to produce, and are used only where excessive thermal energy must be controlled.

Manufacturing processes. Dozens of circuit processes have been developed, but the basic processes deal with conductor pattern formation, or image placement. The pattern-generating method is often used to classify the printed circuit board production

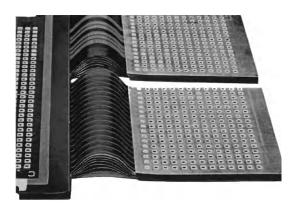


Fig. 4. Rigid-flex circuit commonly used for military and avionics.

and can determine board performance. Conductor patterns must be accurately manufactured, but vertical (z-axis) interconnects must also be fabricated for printed circuit boards with more than a one conductor layer.

Material preparation. The starting point for printed circuit board manufacturing is the substrate, and there are two classes: dielectric only, and laminates and clads with conductive material, especially metal foil, already mated to the dielectric. A clad is directly bonded metal with no adhesive. The first step is to prepare the substrate, laminate, or clad by cutting to the desired size for the printed circuit board line, if necessary. Secondary steps include cleaning, oxide removal, and surface preparation for the next process step.

Metallization. A laminate or clad does not require metallization since the foil layer provides the necessary metal for the conductor pattern. But some printed circuit board processes start with dielectric only and add metal during fabrication. Metal can be added by a mass process such as vacuum metallization or by a selective method that produces the conductor pattern directly. Metal plating is the most common process for creating patterned conductor arrays. Printing of conductive cermet or polymer thick film (PTF) inks can also be viewed as metallization, but many use the term only for mass deposition of metal over the entire substrate surface. *See* ELECTROPLATING OF METALS.

Through-hole formation. Double-sided and multilayer printed circuit boards generally require vertical connectors between layers. A hole must be formed through the substrate as the first step. Drilling has been the standard method for rigid (organic) printed circuit boards, while thin flexible circuitry has mainly relied on mechanical punching. Laser drilling, or machining, is increasingly popular as equipment costs drop and power efficiency increases. While ceramic substrates can be laser-drilled and cut, it is more common to punch holes in "green" ceramic, a pliable and workable material that has not been fired in an oven (kiln).

Conductor patterning (image placement). The conductive pattern-generating process determines the circuit density (conductors per unit width) of the printed circuit board and can significantly influence the cost. While dozens of patterning processes have been proposed and evaluated, printed circuit board manufacturing uses only a few.

1. *Fully additive*. In the additive process, the conductors are added to the dielectric. Many of the early circuit processes were additive, especially conductive-ink printing, and conductors were placed exactly where required. Additive copper methods fell out of favor early, but have become important, especially for high-density fine-line circuitry. Conductors can be printed for ceramic circuits, making this a true printed circuit method. Cermet ink is screen-printed with a high-precision screen of woven metal (or polymer) selected to give the desired ink thickness and feature resolution. The printing screen is coated with an ink-blocking emulsion,

having an open pattern corresponding to the intended conductor image. The pattern in the emulsion is generated by a photoimaging method using artwork produced by a laser plotter from the CAD files for the particular printed circuit board. Once printed, the ink is finished by heating to drive off volatile organics, and then fired at temperatures near 1000° C (1800° F) to fuse and form the ink into a durable metallic conductor. In the co-fire process, cermet ink is printed onto green (uncured) ceramic, and the substrate and ink are fired concurrently. The co-fire process also allows holes to be punched in the substrate prior to firing, so that the cermet ink can be printed into the holes to connect the top and bottom conductors.

Polymer thick film uses the same screen-printing methods, including through-hole printing; the primary difference is in the materials. It uses polymer binders and substrates to yield substantially lower cost, but with lower performance.

Metal, especially copper, can be selectively plated onto a substrate that has been sensitized with a plating catalyst. While the catalyst can be applied selectively, a more desirable method is to apply a plating resist, or mask, over the catalyzed surface to restrict and contain the plated metal. A plating resist also prevents the copper from plating outward from the catalyzed area. While fully additive copper can generate very high density circuits, the process is rather slow and expensive.

2. Subtractive etching. The subtractive patterning process begins with a copper foil laminate and removes, or selectively subtracts, the metal, leaving the desired conductor pattern. The most common method is to apply an etch resist and chemically etch away the exposed copper. The resist can be printed directly onto the copper laminate, then hardened with heat or ultraviolet (UV) energy, and finally submerged in an etching solution. Once the exposed metal is removed, the resist is stripped off the conductor pattern using another chemical solution. While the printing resist is low in cost, the circuit pattern resolution is limited to the capability of the printing method, usually screen printing. Although carefully controlled screen printing can resolve features down to 50 μ m, this print-and-etch process is used for high-volume production at 150to 200- μ m center-to-center spacing (pitch) for better yields. Print-and-etch is one of the lowest-cost methods and is used for making a large number of low- to moderate-pitch printed circuit boards.

The subtractive etching process is also used with photoresists, where the initial pattern is produced in the resist using UV radiation. The resist is applied to copper laminate as a dry film or as a liquid, which is printed, sprayed, or otherwise applied, followed by drying. An artwork film of the conductor pattern can be placed against the resist and exposed to highenergy UV. The exposed resist areas become insoluble (negative-acting) or soluble (positive-acting) depending on the resist chemistry. The resist is then developed to remove undesired material, thus allowing etchant to contact and remove metal to produce the conductor pattern. The resist is stripped after etching. Photoetching is a popular printed circuit board process capable of fine-line geometries approaching 25 μ m, but 50- to 100- μ m minimum trace widths are more typical.

3. Semiadditive. The semiadditive process combines additive and subtractive in an efficient process capable of very high circuit density (less than 10- μ m conductor widths). The method is increasingly popular for flexible printed circuits, which often require very high density. Metal, usually thin (2- to $5-\mu m$) copper known as seed or bussing, is deposited over the entire surface of the dielectric by a vacuum deposition method such as sputtering. A plating resist is applied over the thin metal, and then imaged and developed to form a pattern. Metal, usually copper, is electrolytically plated using the thin layer of copper as the electrical buss. Once the desired thickness of copper is plated, the resist is stripped away. The remaining thin copper is easily etched away without damaging the relatively thick plated-up conductors. The method is more common with flexible circuits, since thin copper can be applied in a cost-effective and continuous-roll fashion. This is more difficult for rigid printed circuit board substrate. See SPUTTER-ING.

Connecting conductor layers. A double-sided circuit has a top and bottom conductor layer produced by one of the above processes. Most double-sided circuit designs require connections between the two layers at specific locations. The most common vertical connection method is the plated through-hole (PTH) [Fig. 5]. First, holes are formed through the substrate by drilling, punching, or lasing. The holes are catalyzed and then plated with a thin layer of electroless copper (no current needed). The electroless copper allows more efficient electrolytic copper plating to complete the process. The resulting hollow conductor looks like an eyelet but is called a barrel. More recently, smaller holes have been used that will plate shut to form solid through-hole posts, or solid vias. See ELECTROLESS PLATING.

Multilayer fabrication. There are two strategies for increasing circuit density: more lines per width using finer geometries, and additional conductor layers. Since circuit-patterning density is near its practical limit, multilayer printed circuit board fabrication is used to increase density. The two common multilayer (ML) methods are lamination of double-sided circuit

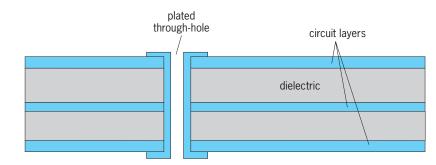


Fig. 5. Cross section of a plated through-hole for connecting circuit layers.

stacks and build-up alternate layers of conductors and dielectrics.

1. *Conventional lamination*. The conventional common multilayer process starts with double-sided circuits described earlier, but without any holes. The layers are carefully aligned, followed by lamination in a heated press that may also have vacuum capabilities to help eliminate air voids between layers. The laminated stack-up is then drilled to create through-holes, which are plated with copper similar to double-sided printed circuit boards, to connect the desired layers. This time-tested method is preferred. Rigid printed circuit boards, flex, and rigid-flex all use this process.

2. Co-fired and co-lamination multilayers. A commonly used process for ceramic circuits is to form the top and bottom conductor patterns, and the vertical through-hole connections, before combining the individual double-sided circuits. Circuit layers can be printed on green ceramic while simultaneously screen-printing into the through-holes. The patterned green layers are stacked, registered, pressed together, and then fired to cure the ceramic substrate, fuse the ink, and bond the layers into one interconnected multilayer circuit.

A somewhat analogous process has been applied to flexible circuitry, called co-lamination. Doublesided circuits are manufactured, including plated through-holes. The circuit layers are aligned together, and an interposer layer is inserted between each circuit pair. The interposer is a special laminating adhesive with conductive material that interconnects the adjacent circuit layers. The stack-up is heated in a press, whereby the adhesive bonds the layers together and creates the vertical connections between the layers. Co-lamination and co-firing offer higher density as well as assembly efficiency. While co-fired multilayer ceramic printed circuit boards are popular, the organic analog, co-lamination, has not yet caught on.

3. *Build-up multilayer*. The build-up process can use ceramic thick film, polymer thick film, or rigid printed circuit boards. In all cases, the layers are formed sequentially, one at a time. Ceramic and polymer thick film circuits both start by forming the first layer of patterned conductors. Next, dielectric is applied over the conductors by screen printing, or a coating method. Through-holes can be formed by screen printing or by using a photoimageable dielectric. Once the dielectric layer is finished, a second circuit layer is applied by screen printing, or some other means. The steps can be repeated until the desired number of layers is formed. Ceramic buildup has been replaced largely by co-firing, but the method is still used with polymer thick film.

Build-up can also be used with organic copper printed circuit boards. The first circuit layer is conventionally produced, and then a dielectric layer is applied and photoimaged to enable through-hole formation. Metal is next applied to the dielectric surface and into the holes by sputtering or plating, followed by patterning. The steps can be repeated to add more layers but is often limited to just the top layer, which can be a high-density pattern. The process can also start with a conventional multilayer circuit, using the build-up just for the high-density top and bottom layers.

Surface finishes. While copper is an excellent electrical conductor with good mechanical properties, it is also chemically active and will react with oxygen, carbon dioxide (slowly), and other airborne materials in the atmosphere. Some bare-copper printed circuit boards are sold, but most exposed copper is protected with a surface finish. The simplest approach is treatment with a solution containing an organic solderability preservative (OSP), which provides moderate protection. Although the organic solderability preservative is the simplest and lowest-cost method, protection will diminish over time and the preservative is easily abraded. A more durable finish is provided by plating metal over the copper. Gold has long been one of the most popular circuit finishes because it is inert and is suitable for wire bonding, which allows integrated circuit chips (die) to be directly connected to the printed circuit board. A barrier layer of nickel is first applied to the copper conductors to prevent the gold from diffusing into the copper. More recently, immersion silver has become a popular finish that is easily applied by dipping the printed circuit board into a silver plating bath. The immersion plating process is self-limiting and stops when no more copper is available at the surface. Modern immersion silver baths contain inhibitors to prevent tarnishing and general corrosion. Silver, gold, and organic solderability preservatives are the most common lead-free finish for solder assembly, but gold must be used for wire bonding and durable edge connectors.

Final testing. Although modern printed circuit board processes minimize defects, dozens of problems can arise during the many process steps. Contamination as small as a particle can ruin a circuit by producing a defect in the photoresist pattern. Essentially all printed circuit board makers electrically test finished circuits, and many use automatic optical inspection (AOI). The methods, especially automatic optical inspection, can also detect trends and thus help adjust the process before out-of-spec printed circuit boards are produced.

Depanelization (singulation). Manufacturing efficiency requires that the processes use their full area capacity, and this often means that there are many circuit boards per panel. Once a printed circuit board panel is completed and tested, the individual circuit boards are cut out or otherwise singulated; depanelize is a more common term in the rigid circuit industry. Rigid circuits can be sawed, punched out, cut with a water jet, or laser-cut. Thinner flexible circuits can be laser-cut or die-cut with a blanking tool. The printed circuit boards may also be partially precut to allow easy separation at this stage. Ceramic circuits can be precut, sawed, or laser-scribed to allow mechanical separation (snapstrate) later.

Basic applications. Wherever there is electronics, there will be printed circuit boards, but circuitry is also used for electrical-only interconnects including

light-emitting diode strip lighting. The range of circuit applications, and the corresponding designs, is amazing when one considers that a circuit can fit inside an "invisible" hearing aid, or fly over the Earth as a 100-ft-long (33-m) flexible power bus for the solar arms in the *International Space Station*. While highvolume products, such as personal computers, use standard printed circuit boards, tens of thousands of other applications require unique circuitry with special properties. Electronics connects the world and the printed circuit board connects electronics.

Interconnect only. Printed circuits, especially flexible types, may be used to connect electronic subsystems together as flex cables with no components. Flex circuits for inkjet printers are a good example of interconnect-only, where electrical power and signals are conveyed to a moving assembly. Stationary circuits, such as the behind-the-dash instrument cluster circuits, may also be connect-only products. Most flex cables are single-sided, but higher density can require double-sided cables or several flexes bonded together. However, the introduction of surface mount technology (SMT) provided the ideal component form factor for flex, and components may be easily added to cables.

Component assembly platform. From the very beginning, the printed circuit has been used as a platform for vacuum tubes, coils, and other heavy components. Large and heavy components is one of the reasons that the industry moved from thin flexible substrates to thick rigid boards to better support the weight. The solid-state electronics revolution, catalyzed by the invention of the transistor, still required an assembly platform for attaching transistors, preferably by soldering. Although lighter in weight, packaged transistors still used a somewhat rigid platform, and the domination of rigid printed circuit boards continued. The next electronic device revolution, solid-state integration in the form of the integrated circuit (IC), increased the demands for component platforms since the integrated circuit enabled much more powerful and complex products such as the personal computer. Today, most integrated circuits and passive components are packaged for surface-mount assembly, which enables much higher densities, full automation, and smaller lighter products. See ELECTRONIC PACKAGING; INTEGRATED CIRCUITS.

Chip carrier. Although most integrated circuit chips are generally provided in a package format that allows easy solder assembly to printed circuit boards, chips can be directly attached to printed circuit boards. Unpackaged chips (bare die) may be wirebonded, connected using a tape carrier package (tape automated bonding; TAB), or directly attached using flip-chip technology. Some refer to bare-die attachment as chip-on-board (COB). Flexible circuits were the first printed circuit boards used as chip carriers, beginning in the mid-1960s. Small rigid printed circuit boards, both ceramic and organic, became popular in the early 1990s when packaging designs went from perimeter connections to the area array to increase the number of connections but keep size

small. The plastic ball grid array (PBGA), a small printed circuit board with solder balls on the bottom, has been increasingly popular. Today, printedcircuit technology is used to make organic and ceramic packages, including the newer chip-scale package (CSP), which is only slightly larger than the integrated circuit chips.

End uses. Since it is not practical to list all the printed circuit board uses, the most important and more interesting ones are covered.

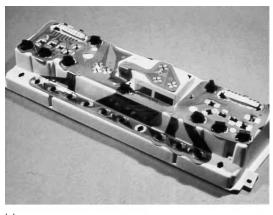
Aerospace. Aircraft and space vehicles generally have sophisticated electronics for sensing, control, guidance, and communications. Some may have very complex military electronics for defense and offense. Space and weight are typically limited for most of the craft, and reliability requirements must be extraordinary even in the harsh environments encountered in flight. Some of the most complex and unusual printed circuit boards are used in aerospace and often combine flex, rigid, and even ceramic types into a fully integrated system that can fit into a restricted volume.

Automotive. The automotive industry began using printed circuits about a half-century ago, and all types are used today. Rigid printed circuit boards are used for entertainment, safety, security, and engine control systems, although more heat-tolerant ceramics are often preferred for "under-the-hood" applications where temperatures can exceed 150° C (302° F). Flexible circuits can be used for the same applications as rigid, but flex uniquely solves instrument connections, where the circuit must be shaped to fit around the back of the instrument cluster (**Fig. 6**). Most flexible cluster circuits are connection-only types made from low-cost polyester film that is not readily soldered, but the trend is to attach components.

Computers and peripherals. Computers have always used printed circuit boards. The early mainframes developed by IBM in the 1960s used high-density, high-layer-count ceramic multilayer to achieve extraordinary reliability and thermal management. Personal computers use mostly organic printed circuit boards, but flex is used where there is motion and as the interconnect to flat-panel displays (FPDs). Disk drives use very thin and tiny flex circuits that allow the read-write head to move over the spinning media disk. Remarkably, these flexible disk-drive circuits may be flexed a billion times without damage. Computers remain one of the largest printed circuit board markets.

Consumer electronics. Consumer electronics is a very large and diversified end-use area for mostly organic printed circuit boards, but ceramics may be used for radio-frequency applications and some packages. Cameras, players, educational toys, and hundreds of other products use printed circuit boards. Even rechargeable smart batteries use circuitry. Rigid printed circuit boards are the most common, but small products often use flex circuitry to reduce the size and weight of the final product.

Industrial. Thousands of industrial machines use printed circuit boards since nearly all have



(a)

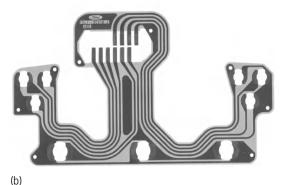


Fig. 6. Flexible circuits. (a) Assembled automobile instrument cluster. (b) Individual flexible circuit.

electronics or electrical devices that need sensing, monitoring, and control.

Medical. Medical electronics is a large and growing sector. Medicine has advanced substantially by the adoption of electronics, and this trend is accelerating. Medical electronics circuitry applications can be as simple as a home-use blood pressure meter or as complex as a CAT scan machine, requiring the power of a supercomputer. All types of printed circuit boards are used. Sensors and electrodes can use flexible circuits, including those made with polymer thick film technology. Many control panels also use flex in the form of membrane switches, which permit sealing of the unit for safety. Wearable and implantable medical products with specialized printed circuit boards include hearing aids and pacemakers.

Military. Military electronics remain an important market for printed circuit boards, with stringent specifications (MIL-STD-883F) for harshenvironment reliability. Avionics is one important area of military electronics. All military vehicles, from large ships to humvees, add more electronics of higher sophistication are adopted and advanced. Portable electronics is also an important sector for communications, detection, sensing, and control. Portable military electronics is increasing as the military strive to be a more flexible and nimble force.

Telecommunications. Communications is now the driver for electronics, a position once held by com-

puters. Communications continues to advance at an accelerated pace, and each product needs one or more printed circuit boards. All types of circuits are used, and each can play a unique role. Ceramic circuit modules are ideal for higher radio-frequency applications. Rigid printed circuit boards continue to be the standard for nearly all products, but may be assisted by flex when a small-form factor requires a thinner substrate, conformability, or motion. Cell phones are the highest-volume telecom market, but newer personal products, such as Apple's iPod, are now a large outlet for rigid and flexible printed circuit boards.

Special applications. The above applications cover a very wide range, but there are others worth noting. The *International Space Station's* solar array power bus circuit may be the world's longest circuit and the most space-resistant. These circuits, exceeding 30 m (100 ft) in length, are made of very robust polyimide, coated with silica to prevent damage by the energetic solar wind. Tiny flexible circuits are being tested in experimental site-restoration devices that will eventually be connected to the retina or optic nerve. In a typical office, school, or library, there are hundreds of printed circuit boards, mostly hidden inside products. While often unseen, printed circuit boards enable the world of technology, a requisite for modern existence.

Future directions. Circuits may switch to nonmetals in the future. Nanotechnology has been able to produce carbon nanotubes, which can be used for transistors and even circuitry. Nanoimprinting and ink-jetting of materials is advancing. While metal conductors have been used for 100 years, a change is almost certain during the coming century. Ken Gilleo

Bibliography. R. Cadenhead and D. DeCoursey, The history of microelectronics, Part One, *Int. J. Microelectr.*, 8(3):14–30, September 1985; C. F. Coombs, *Coombs' Printed Circuits Handbook*, 2001; A. Hanson, British Patent 4,681, filed, 1903; K. Gilleo, *Handbook of Flexible Circuits*, 1992; K. Gilleo, The history of the printed circuit, *Print. Circ. Fab.*, 22(1):18–21, January 1999; R. S. Khandpur, *Printed Circuit Boards*, 2005.

Printing

The processes for reproducing text and images using ink on paper. These include all the processes from typesetting through finishing of the printed product. Prepress operations for conventional plate processes consist of layout and design, typesetting, and platemaking. Traditionally, these were manual laborintensive operations, but now most are digital, using desktop publishing systems.

There are five general printing categories: (1) relief printing, including letterpress and flexography; (2) planographic printing, including offset lithography, screenless lithography, collotype, and waterless printing; (3) intaglio, including gravure and engraving; (4) stencil and screen printing; and (5) digital printing, including electrostatic, inkjet, and laser printing.

Each printing process has specific requirements for its printing plates or image carriers. The printing press unit has a cylinder for mounting the plate and an inking system to feed ink to the plate, as well as a means for feeding paper or other substrates into the printing units and a delivery device for collecting the printed product. Digital printers differ, depending on the process; for example, electrophotographic printers are more like high-speed copiers.

After the sheets are printed, most are put through some postpress finishing operations to make a functional product. Postpress operations include folding, binding, embossing, die cutting, foil stamping, and so on.

Evolution of Modern Printing

The modern printing industry is the result of six major inventions and developments. (1) About 1550, Johannes Gutenberg invented the process of duplicating movable type and printing from it using special inks and a press designed from a wine press. The printing industry remained essentially unchanged for over 350 years. (2) In 1799, Alois Senefelder invented lithography, which provided a practical and convenient means of producing illustrative materials. Previously, woodcuts and copperplate engravings, both requiring highly skilled artists and craftspersons, were the only means of producing illustrations in printing. (3) About the same time, the industrial revolution introduced mechanization to the printing press, first by steam power and later by electricity. The design of multicolor sheet-fed, web-fed (rolls of paper), and offset presses followed. (4) Later in the nineteenth century came the invention of photography, first by the silver halide process of wet plates, followed by dry plates and film, the glass screen for halftone photography which made it possible to produce illustrations photographically, and color films in the twentieth century; and then, photomechanical processes for making printing plates were developed. These techniques stimulated the invention of gravure, collotype, and other processes such as photoengraving for letterpress, albumen and other bichromated coatings, and photopolymers. (5) The Linotype machine, invented by Ottmar Mergenthaler in 1886, mechanized typesetting; teletypesetting followed in the 1920s and phototypesetting around 1950. (6) The introduction of electronics and computers in the early 1960s revolutionized the printing industry and many of its processes and techniques.

To put these developments into perspective, it was 440 years from the time that Gutenberg invented the process of duplicating movable type to Mergenthaler's invention of the Linotype machine. However, it was only 68 years from the introduction of the Linotype machine to the invention of phototypesetting by the Harris Corporation in 1954. In the graphic arts, the trend of accelerated technological change continued. Fourteen years later, in 1968, satellite data were transmitted simultaneously to multiple sites to produce The Wall Street Journal. Computer graphics on cathode-ray tubes took hold in 1974, first for engineering, architectural rendering, automobile design, and then for graphic arts applications. In 1977, the Mitsushita Corporation in Japan produced the first prototype of a color monitor that printed out the content on its display in full color. Mitsushita built an inkjet printing system into a television set to demonstrate the potential of integrated media. This was the forerunner of a desktop publishing system. In 1979, the first practical combination of multiple graphic arts technologies from Scitex in Israel demonstrated how nearly all prepress functions (page layout, retouching, typesetting, graphic arts photography, and imagesetting) could be integrated in one system, called the Response 300. In 1980, Gannett demonstrated that satellites could transmit color data to produce the news paper USA Today. See COMPUTER GRAPHICS; MICROCOMPUTER.

By 1980, technological transitions in the printing industry were occurring so rapidly that it became impossible to point to individual developments. From 1980 to 1985, deregulation of telecommunication in the United States opened the airwaves for broad commercial applications of communication media. From 1985 to 1990, desktop publishing technology evolved and matured. Multimedia applications grew and matured between 1990 and 1995. And ondemand digital color printing was the focus during 1995 to 2000.

The use of electronics and computers has rendered obsolete many techniques and led to the creation of new systems, equipment, and processes that have increased the speed precision of production, including scheduling and control; have simplified the operation of complicated systems in prepress, press, and postpress; and have improved the consistency and quality of the output of the printing processes.

Digital imaging and printing has resulted in the development of three specialized printing markets: ondemand printing, variable-data printing, and shortrun color printing. On-demand printing produces documents at the time they are needed and usually within 24 hours. Variable-data printing allows for customized and personalized text or graphics. Short-run color printing allows for full-color printing in runs from 100 to 5000 impressions, which previously had been too expensive to print by conventional processes.

Composition

This is the assembly of all type, art, illustrations, and other elements into the final page format or layout for the press. This was traditionally a manual operation, requiring paste-up of page elements as paper prints or stripping of films on a flat, but now it is done digitally by desktop publishing.

Typesetting. Typography is concerned with the visual impact, effectiveness, and readability of the printed piece. Typesetting is the basic method for setting type, including cast metal or hot type (set either by hand or machine) and phototypesetting.

These methods have given way to electronic imagesetting. The term "typesetting" is independent of its output technology. *See* TYPE (PRINTING).

Metal composition. Hand typesetting, similar to Gutenberg's method, is done with individual metal characters assembled into lines. The individual characters or letters are placed in a composing stick, which is held in one hand, while the letters are selected from a type case with the other hand until a full line is set. Blank spaces are inserted between words to justify the line, that is, fill it out to the desired width. This is repeated line by line until all the copy has been set. For line spacing, or leading, metal strips or slugs called leads are inserted between the lines. The process of hand composition is very slow, about two characters per second, and is used mainly for fine art printing.

Machine composition, which is now obsolete, was a major advance in typesetting, as it increased speed to about five characters per second. Machine-set type was produced on any of several machines. Linotype and Intertype machines cast one line of type at a time. They cast a one-piece line or slug of type to a predetermined length. On both machines, at the touch of a key, a matrix, which is a mold for each character, was released from a magazine or storage case. When all the matrices or characters for one line were assembled, the line was automatically justified by spacebands or wedges, which mechanically expand the space between words. The line moved into the casting mechanism, where molten metal was forced into the character molds. Some line-casting machines set type semiautomatically from perforated tape. These tape-controlled systems sped the typesetting operation to about 10 characters per second, allowing one operator to handle several linecasters.

In all these methods, the lines of type were locked up in a chase, which is a heavy rectangular steel frame, on a large table known as a stone. Empty spaces were filled up with "furniture" of wood or metal blocks, and quoins (steel wedge-shaped devices) were placed on the sides of the form and tightened slightly until all the parts were made level, after which they were tightened to hold the form securely in place.

Phototypesetting. Metal typesetting gave way to phototypesetting, which reduced the number of production steps leading to the completed printing plate. Phototypesetters varied in electronic sophistication, automation, imaging speed, and typographic versatility. They produced composition from simple to full-page multicolumn formats combining text and images to be output on photographic paper, film, or polymer-based plates.

All photographic typesetters consisted of three elements: a master character image, a light source, and light-sensitive material. Phototypesetting systems have evolved through three generations of changes. First-generation units were adapted from linecasting. In these machines, the master character image was carried on the matrix, which contained a film negative of the character that was then photographed instead of cast. Second-generation units used electromechanical principles. Third-generation units were electronic, using cathode-ray-tube (CRT) or laser technology. Also known as digitized typesetting, this process produced characters as a series of minute dots or lines, from which the resultant image was transferred to photographic material. Because third-generation typesetters employed a beam of electrons, they were capable of setting type at extremely high speeds (500–1000 characters/s).

All second-generation phototypesetters had master character sets that included all of the characters and symbols for a particular type style that were stored as disks, film strips, grids, or film segments mounted on a drum. In most phototypesetters, the master character image was enlarged by individual lenses or by a zoom lens to achieve the final type size.

In using a second-generation phototypesetter, characters were selected by referencing certain marks on the photomatrix. When the proper mark was reached, a burst of light from a xenon flash lamp exposed the character through the master font image. The beam of light was directed to the correct optical path by mirrors or prisms; the character was enlarged or reduced by lenses and then positioned properly on the photosensitive output material. Because newspapers were pioneers in the use of phototypesetting, its speed has been measured in newspaper lines per minute (lpm). A standard line is 8 point in type size, 11 picas in width, with 30 characters. Most second-generation phototypesetters averaged 50 lpm, with a high of 150 lpm (25-75 characters/s).

Third-generation phototypesetters were capable of greatly increased speeds (over 1000 lpm) and created complete pages of text, headlines, and lineart graphics in position. Characters were stored as digital information that was converted into dot images. In operation, patterns of the required characters were called from memory and imaged in the correct size at precise locations on the page or format. The finished composition was then exposed from the tube to the photographic paper or film.

Although character images appear solid with accurately contoured shapes after imaging and printing, they are actually patterns of characters consisting of adjoining sweeps of scan lines, similar to the images created on television screens. The line density or number of scan lines to the inch (resolution) determines the typographic quality of the imaged characters and the imaging speed of any particular digitized system. The more scan lines to the inch, the higher the resolution, or sharpness, of the characters and the slower the imaging speed.

Electronic imagesetting. Typesetting now is often done using an electronic imagesetter to produce the films for platemaking. In imagesetting, photographic paper or film is exposed using raster or vector techniques. Imagesetting uses electronic files instead of conventional camera technology. It yields higher quality and less expensive printing.

Imagesetter users are under constant pressure to produce quickly and inexpensively, not only because electronic media are raising customers' expectations on throughput times, but also because print runs are becoming shorter.

Some imagesetters use multiple lasers and a 30,000–40,000 rpm spinning mirror that traverses the inside of a drum. The lasers sit next to each other and use the same optical path. Multiple lasers increase speed. To keep costs down, some imagesetters use a single beam that is split, with the beams being deflected. A five-faceted spinning disk with a holographic deflection surface is used to direct the laser beam on to the film surface. Each rotation of the deflector produces five scan lines, making it up to five times faster than a conventional mirror-based system at the same rotation speed.

Platesetters are high-resolution imaging devices that use a laser to expose a photosensitive plate. This process is also known as computer-to-plate (CTP). It produces better quality than the traditional film-based process. CTP can cost less and allows for quicker turnaround times, but requires investment in equipment, training, and changing workflows.

Desktop publishing. Desktop prepress has evolved as a process for preparing pages for reproduction on printing presses, copying machines, or electronic printers. The multipurpose nature of the personal computer and the increasing pagination ability of word-processing programs have advanced desktop publishing. The essential part of this approach is the use of typographic characters and the integration of boxes, shapes, line art, low-resolution pictures, and other graphic elements into page form.

It is more appropriate to define this application as page processing, since its approach to putting pages and documents together is applied at all levels of workstation and system configurations. The use of high-level graphics-oriented workstations with multitasking functions permits increased productivity.

A further step in the evolution of imagesetting, incorporating the advances that have been made in digitization and scanning, is the integration of text with pictures, illustrations, diagrams, and so on. This has brought typesetting head to head with developments in page composition systems. Once a page is in a digitized form, it can be used in many ways. It can remain in storage, or it can be transmitted to remote locations where it can be output to an imagesetter, a photographic film, a platemaker, an inkjet printer, or an electronic reproduction system.

Page makeup. Most programs for page makeup can position type so that manual assembly of the page elements is reduced. Heads and text can be output according to their page position. To accomplish this, commands must be given to the output device in coded form (by menu or keystroke sequences) that translates the positioning information into machine instructions. The term "computer-assisted makeup" refers to workstations that show positions of image elements and actual or simulated typeface, point size, line length, leading, and so forth as they will appear in final pages. Passive computer-assisted makeup devices are sometimes known as soft typesetters or preview screens, since they show only how the typesetter will set the information according to the commands. Interactive computer-assisted makeup devices allow the operator to make changes in the typographic format on the screen and to see instantly how the changes will appear on the page.

Almost all modern systems include the ability to view type, graphics, and pictures on-screen as a simulated page with all the items in position. This can be in a preview mode, where coded text is translated into a screen image, or interactive, which involves typographic context at all times. The interactive integration of text and graphics has no single descriptive name; hence, the term "publishing" is often used as a synonym.

Page description language. A page description language is a computer language or software that describes an entire page, including text, graphics, lines, and halftones, as a series of codes and allows for viewing or output on any device capable of decoding the language.

PostScript is the most popular page description language. Invented by Adobe Systems, PostScript consists of software that forms the desired image on an output device such as a laser printer or an imagesetter. PostScript is commonly used for text and line art. In the latter case, it is referred to as encapsulated PostScript (EPS).

PostScript's advantage is that the code is converted to images in an interpreter in the output device, not the computer. Therefore, the images are readable or printable regardless of the computer platform being used.

Essentially, PostScript allows for device independence, or the ability to generate virtually identical output on devices made by different manufacturers, so long as they can interpret PostScript commands. The original version of PostScript is known as PostScript Level 1. It was one of the most important elements in the invention of desktop publishing. The latest version of PostScript, called PostScript 3, provides support for digital printing, image setting, and color separation and correction.

Raster image processing (ripping). Imagesetter raster image processors (RIPs) are software-based running on standard, fast, and easily replaceable workstations. A RIP is a microcomputer that uses a laser to create images out of lines of dots (rasters). It translates commands in a page description language, such as PostScript, or in other formats, such as PDF, TIFF, EPS, and JPEG into a precise image of a page. Everything a RIP produces is in pixels. The RIP then sends the rasterized data to a recording device (imagesetter) for output on paper, film, or plates.

To create a halftone dot, the RIP divides the area to be imaged into halftone cells. Each halftone dot then is created out of the pixels available within each halftone cell. The speed of a RIP generally determines how fast a printer can produce the first copy of a page.

Preflighting. Preflight is the process of checking an electronic document before it is moved into the

printing process. The process makes sure that all the elements are correct and within the specifications for printing. Since nearly all work submitted to printing companies is in electronic form, files sent from a remote site or delivered on a disk must be verified for consistency, presence of graphics and font files, and proper selection of colors. Preflight checks include color mode, cropping, bleeds, photo sizes, resolution, retouching, masking, gray balance, proper spelling, page order, type size and style, trapping, dot gain, flattening of layers, screen lines per inch, files saved in proper format, and if the file has the most recent changes.

Process Photography

Process, or graphic arts, photography comprises the photographic techniques used in the traditional graphic arts processes. Until the introduction of electronic scanners, color electronic prepress systems, and desktop publishing systems, all the conventional printing processes used photography for the production of negatives or positives to make the printing plates by a photomechanical process. Today, process photography is obsolete, as film is rarely used in the printing processes.

Copy preparation. The original or copy is the starting point in the printing processes. It can be in many forms, including text and/or pictorial matter such as line drawings, diagrams, black-and-white and color photographs, color transparencies, paintings, other art subjects, and subjects created on computers. In some cases, the original image can be produced directly on the plate. In all other cases, the original is converted for reproduction by photomechanical or digital means.

In printing that uses traditional photographic techniques, originals are photographed to convert them into the proper positive or negative films for the plate processes by which they will be reproduced. The copy must be clean, sharp, and in focus. If the final printed job consists of type and pictorial matter and everything is at the correct size, the typeset copy and pictures can be pasted-up in position before photography. On complicated jobs, such as color advertising, magazines, and books, in which the copy usually comes from a number of sources, the separate pieces of copy are photographed at the correct size and assembled in the film form. This completed form is known as a flat, which is ready for platemaking. Digital copy preparation, where all copy is assembled on a monitor, has rendered these manual and film techniques obsolete. Hence, copy can also be prepared electronically. Computers can be used to generate graphics, and special software and design stations have been developed to produce layouts, design artwork, and compose pages on a monitor and to output fully composed pages directly on a plate or printing press cylinder.

Films. Photomechanical processes use three types of photographic emulsions. High-contrast orthochromatic or panchromatic films are used for line and halftone photography. Low-contrast orthochromatic commercial types of films are used for single-color

gravure and for other continuous-tone photography. Low-contrast panchromatic films are used for colorseparation photography. Special films are made for scanners and imagesetters, and special daylight films are available to make contact films for image assembly. Rapid-access processing using continuous-tone developers is used to speed development of laserexposed films. The favored film substrate is polyester film, which has good mechanical properties and dimensional stability. *See* EMULSION; PHOTOGRAPHIC MATERIALS.

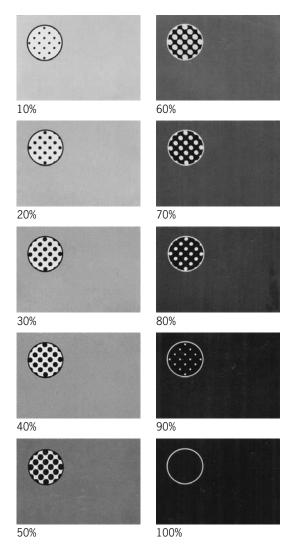
Halftone photography. Halftones are an essential part of the print production process. They are optical illusions that are necessary to compensate for the inability of most printing processes to reproduce photographs and art subjects.

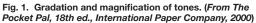
Printing processes, such as lithography, letterpress, and screen printing, cannot print continuoustone reproductions such as photographs and art subjects that have many intermediate tones between the highlights and shadows. These binary printing processes can print only a solid impression of the ink on the press or no ink at all. For these processes to reproduce photographs and art subjects, the appearance of continuous tones must be simulated by the optical illusions produced by images using the halftone principle.

The human eye, at a normal reading distance of 10-16 in. (25-40 cm), is not capable of distinguishing two points separated by about 0.004 in. (0.0016 cm). This is equivalent to a linear screen ruling of about 125 lines per inch (lpi) and 125 blank spaces with the same width as the lines. When two such rulings are superimposed at right angles, a cross-line screen of 125 lpi is produced.

Halftone screens. When a halftone screen is positioned in front of a photographic film and an image is photographed through the screen, the dots formed on the film are proportional in size to the amount of light reflected from the corresponding area on the original. These dots are amplitude-modulated; that is, they have equal spacing between the centers of the dots and vary in size or area. Screens with finer rulings than 125 lpi (such as 133, 150, up to 500 lpi) are not resolved by the normal unaided eye, reproduce finer detail, and produce higher-quality reproductions. Screens with lower rulings (such as 60, 80, 100, and 120 lpi) are visible to the naked eye, show less detail, are easier to print, and are used for newspaper and less critical printing.

Tone reproduction. Tone reproduction (Fig. 1) and contrast are two important conditions that determine the quality of the reproduction. By using a stepped gray scale, good tone reproduction in halftone photography is achieved when the darkest area (shadow) of the subject prints as a solid and the lightest area (highlight) as a white, with no dots in either area. The intermediate tones of the gray scale will have varying sizes of dots, from about 3% dot area in the light or highlight end to about 95% dot area in the dark or shadow end of the scale. High contrast is achieved when two or three steps of the gray scale in the shadow end print solid or when several





steps in the highlight end print white, with a corresponding increase in density difference between the other steps of the scale. Low contrast is achieved when the solid contains 80–90% dots or the white end of the scale has 10–20% dots, with a corresponding decrease in density in the rest of the gray scale. A number of special techniques, such as highlighting or dropout, posterization, and flashing, can be used to change the contrast of the reproduction in local areas.

Digital halftones. Most halftones now are produced digitally in raster image processors or in imagesetters. Digital halftone algorithms have been developed to simulate photographic halftone reproduction with screen rulings, screen angles, and amplitude modulation. These have the same limitations as photographic halftones, with the additional problem that 15 and 75° angles have irrational tangents because the tangent cannot be expressed as a ratio at two integers (horizontal and vertical spots), which results in very long computation times. Some algorithms avoid this problem by using rational tangent angles of 18.4 and 71.6°, but the use of these angles requires varying the screen rulings for each angle to avoid moiré patterns. Conventional halftone screening is called amplitude-modulation (AM) screening. Amplitude means size, so AM screening breaks up an image into dots of varying sizes to simulate the original image. The use of frequencymodulated (FM; also called stochastic) screens, in which all the dots are the same size but vary in spacing, eliminates many of the problems of amplitudemodulated screens as they are independent of screen rulings and screen angles. *See* MOIRÉ PATTERN.

Amplitude-modulated screening. During the 1980s, many advances were made in computer hardware and software that allowed for increased electronic image alterations such as retouching, color correction, and gray component replacement. At the same time, advances were made in computer storage and raster image processors for more efficient image data processing. Along with the ability to electronically scan, digitize, store, retrieve, and manipulate images, there were advances in digital screening. However, because color electronic prepress systems remained linked to high-end electronic color scanners, digital screening centered on the simulation of the traditional amplitude-modulated (AM) halftone dot. See COMPUTER STORAGE TECHNOLOGY; IMAGE PROCESS-ING.

Although it was recognized that dot shape played a significant role in image detail, the electronic digital imaging techniques of the mid-1980s still centered on a fixed screen ruling (lines per inch or lines per centimeter), and the amplitude (size) of the dots was modulated (increased or decreased) relative to the gray levels being reproduced. This type of digital screening is known as a deterministic or order dither, which uses an error distribution process to assign halftone dots to a bitmap in a coherent area (Fig. 2). That is, each halftone dot is usually produced from four scanned pixels. The pixel is divided into a matrix of single recording dots. For example, a typical matrix of 6×6 single dots would produce a halftone dot consisting of $4 \times 6 \times 6$ or 144 single dots.

The number of scanned pixels and the selected recording dot matrix depend on the imagesetter

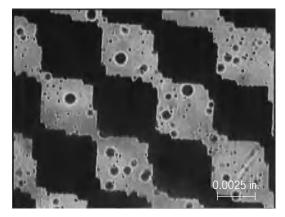


Fig. 2. Photomicrograph showing a typical AM deterministic digital halftone, 50% tone at 150 lines per inch.

used, but 6×6 , 8×8 , and 12×12 matrixes are common. The recording dots form a coherent area on a bitmap that consists of several contiguous single recording dots to produce an amplitudemodulated halftone dot. In general, the coherent areas are selected to produce symmetrical dot shapes, such as square, round, or elliptical, but they can also be arranged to produce irrational dot shapes. Thus, a dot arrangement in the form of a bitmap can be defined for each tone value and stored in memory.

Although the ordered dithering process and digital AM screening have proven to be an excellent method for producing halftones, there are limitations. There is a physical and mathematical relationship between the number of gray levels, the dots per inch, and the screening ruling, such that there is always a trade-off between resolution and the number of reproducible gray values. Also, the algorithms for conventional screen angles of 15 and 75° can never be exactly calculated, and require long and burdensome computational times because of their irrational tangents.

Stochastic screening and frequency-modulated screening. Stochastic and frequency-modulated (FM) screening allows outputting high-quality color images at lower resolutions by using a different approach to screening technology. Unlike amplitude-modulated (AM) screening, which varies the size of the dot to simulate different shades of color, stochastic screening keeps the dot size the same and varies the number of dots used. Stochastic screening uses randomly placed dots instead of the traditional halftone dots aligned along visible screen angles (Fig. 3). Stochastic and FM screening have become synonymous, although FM screening is not as random as the word "stochastic" implies. In effect, most stochastic screening products available are variations of FM screening.

The concept of stochastic screening is that dots placed randomly will not cause moiré patterns. Completely random positioning does not work well, however. It may prevent interference patterns, but it can cause graininess and other visual artifacts. FM screening products apply advanced algorithms to deter-

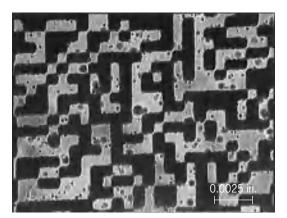


Fig. 3. Photomicrograph showing a typical FM stochastic digital halftone, 50% tone at 2400 dots per inch.

mine the optimum placement of dots to appear random, yet not cause visual artifacts.

In FM screening, the concepts of screen angle and frequency no longer apply. Because the dots are randomly placed, there is no direction to the dots. The variable spacing of the dots means there is no fixed spacing and therefore no screen frequency.

Applications for FM screening span the industry, including newspapers, magazines, and commercial printers. Commercial printers who do high-end color work, such as brochures, art reproductions, and annual reports, are interested in FM screening as a way to achieve the 600-lpi-equivalent. Commercial printers with imagesetting operations are good candidates for FM screening because they can control the entire output production cycle. With the ability to generate custom FM screens for any combination of paper, ink, and imagesetter parameters, printers can set up a completely controlled environment for their full range of work.

Magazine printers with imagesetting operations can use FM screening to produce high-quality color. Once the dot gain calibration has been performed in collaboration with the printer for the press, paper, and ink, the magazine can use these settings for every production run.

Newspapers can also use FM screening to good advantage. Newspaper industry tests with frequencymodulated screening show higher-quality color and image detail than with conventional screening because of the relative coarseness of the conventional screen used with newsprint. Newspapers can output pages more quickly at lower resolutions, and the color images will be of much higher quality than if conventional halftone screening were used.

Color Reproduction

The basic principle of color reproduction in modern color printing of pictures lies in the use of only three primary colors to create the illusion of seeing in the reproduction the hundreds or more colors the eye sees in the original. Color printing with inks is the controlled proportioning of red, green, and blue light, and this is accomplished in an indirect way called the subtractive process. Printing is done with yellow, magenta, and cyan inks because these colors ideally are complementary to blue, green, and red respectively, and each can individually shut off the reflectance of red, green, and blue light from the paper surface.

Until 1893, colors were commonly printed separately with ink on paper for commercial printing, book illustrations, and art picture reproductions. As many as 40 colors were sometimes printed to facsimile-match an oil painting. While some copperor steel-plate engravings were used, plus many woodblocks, 95% of all color printing up to 1900 was done from hand-drawn images on stone. On a litho stone, an artist would draw (crayon) or paint (tusche) a separate printing image for every important color in the original. The opaque mineral ink pigments used were similar to those used in the original oil painting. **Primary colors.** In modern color printing, only three printing ink colors are needed. This came about by bringing together the invention of photography, the halftone screen, coal tar, and synthetic dyes to make transparent color inks and color separation filters. All of these technologies were coupled with a better understanding of color vision.

The human retina has photoreceptors called cones that are separately stimulated by long red wavelengths, medium green wavelengths, and short blue wavelengths of the light spectrum. The brain interprets thousands of hues, tints, shades, and intermediates from different proportions and amounts of the three-color signals sent through the optic nerves. When the three printing ink colors are controlled in the same proportion and amounts as in the original, the brain accepts these as matching. Even with imperfect photographic plates, James Maxwell in the 1860s demonstrated this principle by using three lantern slide projectors whose lenses were separately covered with red, green, and blue filters. He simulated the color appearance of objects he had photographed through red, green, and blue filters. It took almost 40 more years to improve the color sensitivity of photographic plates, invent the halftone screen, and produce inks transparent enough to make this a practical reality in printing.

Practically all commercial color printing also uses a fourth color, black. Technically, black is not a color but the absence of color. In its use, black creates no new hues beyond those in the basic three-color printing but adds depth, contrast, and stability to the colors of the image. When quality transparent inks are printed on low-absorbency, high-gloss, coated papers, the three colors overprint to an intense black, making a separate printing of black ink unnecessary. However, black ink is printed for type, as well as to give more intense blacks and contrast when lower-quality inks and papers are used. *See* COLOR.

Color separation. Originally, all input copy for color printing consisted of large oil paintings, watercolors, pastels, or dye-transfer or photographic color prints. These were reflection copy, and huge process cameras with four arc lamps were usually used to reduce the lens image size. However, most modern input copy consists of small color transparencies. Smaller enlargers are used instead of huge reducing process cameras. Larger color transparencies, groups, and page layouts are also same-size color-separated by contact in vacuum frames. Even this revolution of downsizing color separation input and equipment is being superseded by the electronic age of scanners, in which no full-size picture image is color-separated at once. Copy input of color transparencies, photographic color prints, or reflective hand-drawn art is mounted on a rotating transparent cylinder and color-separated spot by spot in lines only a few thousandths of an inch wide and in only a few millionths of a second per spot. Faster flatbed charge-coupled scanners can make four-color separations in a minute or less. See CAMERA; CHARGE-COUPLED DEVICES.

The first step in any color reproduction system is known as color separation. The photographic image is exposed separately through red, green, and blue filters. Each filtered separation image must also be exposed through a halftone screen to create differentsize dots that control the hues and tone values. When the filtered color separation and halftone screen exposures are made simultaneously to the same photographic film, it is known as the direct screen process. For many years, the indirect process was most commonly used. In this, the filtered color-separation negatives are normal continuous-tone photographic images, and the halftone dot screening is introduced in a second photographic step with a high-contrast film to create positive dot images. These positive dots could be more conveniently color- and tonecorrected in size by hand dot-etching steps. Improved and better-controlled photographic masking steps minimized the need for hand dot etching so that the direct screen method became more popular

Primary color inks. The actual final press plate images can be exposed through either halftone negatives or halftone positives, but the printing-plate image for the press will be positive. Ideally, this image should be transferred to the printing paper with a process ink that is complementary to the color of the filter that made each color separation image. Yellow ink is used to print from the blue-filtered image because yellow ink absorbs the blue light; magenta ink (not red) is used to print the green-filtered image; and cyan ink (not blue) is used to print from the redfiltered image. This is the first step in the subtractive process of color printing. Each ink, printed as 100% color, acts like a transparent color filter and subtracts one of the additive primary colors of light from the spectrum of the white light reflecting from the unprinted white paper surface. Each of the three colored process inks used in printing absorbs approximately one-third of the white light spectrum. However, being imperfect, they each also absorb more than their proper third of the spectrum. This prevents the reproduction of some of the brightest reds, oranges, greens, purples, and blues. The color from printing that the eye sees is what is left after the ink's subtraction; the secondary and intermediate colors mostly required in reproduction are always less pure and bright than those available from the white paper.

Good yellow inks are the most efficient (about 95%) of the three colors. They usually reflect 100% of the red light, but are typically warmer in hue than they should be theoretically because they also absorb some green light. Cyan ink is the major handicap to reproducing bright pure greens since its hue is bluer than correct because it absorbs considerable amounts of green light. Magenta ink is typically the most inefficient because it absorbs substantial amounts of the blue light that it should reflect 100%. Printers usually refer to this primary ink as red or process red because most commercial magentas have a hue error of more than 50%. The theoretically correct magenta ink would have

a more purple hue, reflecting red and blue equally. Therefore, commercial magentas make acceptable reds, but their blues and purples are less pure and bright.

The three primary colors seldom print alone in most pictures, and their performance in twoand three-color mixtures becomes pictorially more important to the reproduction. If a four-color-ink printed picture is examined with a 10-power or stronger magnifying glass, no more than eight actual colors will be found: yellow, magenta, and cyan appear as separate dots; two-dot overlaps make red, green, and blue; three-dot overlaps make black; and then there are some areas of unprinted white paper between dots. The illusion of hundreds of different colors seen visually without the magnifying glass is the additive color mixture averaging of these eight different areas.

Neutral gray. Perfect process ink would require equal halftone dot sizes in the three colors to reproduce neutral gray. Because the yellow, magenta, and cyan hues of process inks are not equally pure, an important production control is to create a specific inequality of the color dot sizes in a gray scale. This three-color inequality will vary with the use of different sets of process inks, but the cyan dot must always be larger than the other two.

Variables. The theory of color reproduction by printing inks is well understood, but in practical and economic terms process inks and paper are less than the optimum quality. And while the separate inks individually may transfer well to the paper, they do not always all print fully on each other (a problem known as trapping). Another press variable may be that different pressures, ink consistencies, and paper absorptiveness result in larger or different dot sizes. Thus, not all printed color reproduction is automatically of optimum production quality.

Proofing

For most high-quality printers, a key part of the workflow has been the contract proof. The proof is a document presented to a client for approval prior to printing. Printers wishing to provide high value to their customers now have an expanding range of proofing options.

Contract proofs have changed dramatically. Yet as the developments of laser scanning, desktop publishing, and computer-to-plate systems have improved the quality of the printing process, trends in contract proofing is often driven more by cost than by quality.

Virtual proofs. Virtual proofs, sometimes called soft proofs, are images on a monitor, as opposed to hard proofs on paper, film, or some other material. A virtual proof is used to check the pages ready for transmission to a printing plant as well as to detect errors (in the RIP, lack of fonts, colors). In virtual proofing, the cyan, magenta, yellow, and black bitmaps are converted into red, green, and blue on the monitor. Virtual proofing does not eliminate the need for a hard-copy proof, but it does allow the correction cycle to occur faster and reduces the number of hard proofs that are needed.

Virtual proofing systems are increasingly used for comparisons at the press. However, the process demands a lot from the monitor and the lighting environment. All parts of the system have to be thoroughly calibrated. Virtual proofing provides press operators with a general idea of what the printed page should look like. One problem is that virtual proofing is not user-friendly for customers or production operators for making changes or corrections, and for transporting proofs between operators and to customers. There is also the issue of two or more monitors on which the proof is viewed being calibrated differently and, hence, displaying images that appear different in color. However, advances in color management tools, improved electronic file transfer methods, and upgraded computer monitor technology are prompting printers to pursue virtual proofing methods.

Calibrating the monitor requires determining its color gamut and adjusting it to match what the printing press will produce. This must be consistent among all parties viewing the proof. For example, when proofs are transmitted to a magazine's art director, advertiser, and printer, all parties must simultaneously review the page files and indicate any corrections to the soft proofs with accuracy in minutes.

Virtual-proofing methods can range from PDF files via e-mail or FTP sites, to systems that allow users to review content and layout, to more complex colormanaged systems that seek to emulate hard-copy proofs. While users may adopt soft proofing as a workflow tool to speed production, whether they embrace it as a replacement for the hard-copy contract proof depends on the comfort level of printers and print buyers.

While a large amount of current online soft proofing simply involves posting PDF and JPEG files on password-protected areas of a Web site, many e-commerce vendors and even some makers of prepress-workflow systems have added online virtual proofing capabilities to their products. Online virtual proofing is especially valuable for projects with tight deadlines.

However, the process is not for all situations and particularly not for those involving color-critical projects because calibrating displays can be difficult. The process can be valuable for applications that warrant it, such as repeat jobs, projects involving Pantone or spot colors, or clients who trust the production vendor to produce accurate color.

Remote hard-copy and remote proofing. Traditional analog and digital halftone proofing systems have a high acquisition cost, making them inaccessible to many smaller printers and impractical for low-volume remote-proofing applications. Lower-cost inkjet proofing systems can deliver prints with six colors at high print resolution. These printers are powerful proofing tools, capable of output that simulates traditional analog proofs. They have color profiling capabilities and the ability to print on special media.

Remote proofing involves sending files to clients (or vice versa) for output on a proofing device, or soft proofing via a computer monitor. Either or both can be integrated into a printer's workflow depending on client needs and expectations.

The availability of lower-cost alternatives, which simulate the output of analog and digital proofing systems, lowers the price of proofing and places hard-copy color proofing at more points in the workflow. To ensure consistent print quality between systems, some manufacturers have integrated densitometers that automatically make color measurements and adjustments to keep color values within specifications.

Electrophotographic color copiers and printers are also being used as proofing devices. They may be more expensive to acquire than inkjet proofing systems. However, they usually have a lower colorpage cost. The lower page cost and a larger paper capacity make electrophotographic units more suitable than inkjet for printing short-run color work. Hence, a color copier can be a dedicated proofing system and handle other graphic and office printing.

A challenge to printers is getting their customers to adopt remote proofing. The long-standing, stable accounts that generate high dollar volume are more receptive to remote proofing than newer customers or customers with infrequent print work. Wide acceptance of remote hard-copy proofing also depends on the development of inexpensive systems that are easier to implement and maintain.

Not all remote proofing workflows rely on the transmission of job files to a client for output on a remote proofing system. Some printers install colorcalibrated digital proofing systems at customer locations and let them produce their own proofs, which are submitted along with the job files for print production.

Also aiding remote proofing are inkjets with improved color gamut, light-fastness, and stability. In addition, the cost of consumables on these inkjet printers is lower than that for high-end digital proofing systems.

Remote proofing in general has presented challenges for printers with critical-color clients, but emerging new technologies and products are addressing this issue. For color-managed remote proofing to succeed, there has to be a relationship between the designer, the print buyer, and the printer so that all involved agree on paper and standard ink sets, and all of the components that affect color. One of the most important considerations is that the lighting under which the proof is viewed must be controlled.

Image Assembly and Imposition

Rarely used any longer, the image assembly, or stripping, operation is a skilled manual procedure for constructing and assembling negative or positive films in such a precise manner as to provide the platemaker with complete film assemblies for producing printing press plates. Image film assembly was a simple manual procedure involving the positioning of camera films on a predrawn support flat (**Fig. 4**). The most important skill was the planning and drawing of the layout arrangement, known as an imposition. Computerized systems are now used for the layout and assembly of images.

Machines have been developed that automatically control image placement, practically eliminating hand assembly methods. In these step-and-repeat machines, computer programs control films and their positions on the plate, and optical imposition cameras that photograph pages onto a film mounted on a movable back whose movement is in the x, y plane. Advances in the technology have allowed plates to be exposed digitally in platesetters by completely imposed computer programs.

Platemaking

Platemaking is the final prepress process. Without good plates, even a sophisticated printing system cannot produce optimum results. Therefore, the methods by which plates are made are crucial to the outcome of the printing process.

Platemaking processes. Printing plates or cylinders are made in one or a combination of seven different ways: (1) manually by using hand tools, engravers, knives, and so forth, to produce plates for the relief, intaglio, and stencil processes; greasy crayon or tusche can be used for making hand-drawn lithographic plates; (2) mechanically by engraving and geometric lathes, ruling machines, pantographs, and so on, for the relief and intaglio processes; (3) electrochemically by electrodeposition of metals to produce longer-wearing images; (4) electromechanically by electronic engravers to make relief and intaglio image carriers; (5) electrostatically by using xerographic principles to make lithographic plates; (6) photomechanically by combining the photographic process to produce images and a mechanical process to make plates; (7) digitally by laser exposure or other radiant energy from direct digital information in storage or prepress systems. The photomechanical method is the most widely used, while digital platemaking is increasing and will dominate in the future. See DIGITAL CONTROL.

Each of the printing processes has different platemaking or image-forming characteristics and requirements. Letterpress image carriers (plates) consist of engravings of single subjects locked in position with other engravings and cast-metal type into an imposition form for printing on a flatbed or platen press. Plates for rotary presses are duplicate plates made from curved molds of the original imposition forms. These can be of cast type metal called stereotypes, electroplated metal called electrotypes, and cast plastic or rubber plates. Rotary letterpresses have been developed that use photopolymer-coated wraparound metal plates exclusively. Flexographic plates can be cast rubber plates made from engraving forms, but flexography uses mainly photopolymer-coated wraparound metal plates. Lithographic plates are wraparound metal (usually aluminum) plates with light-sensitive

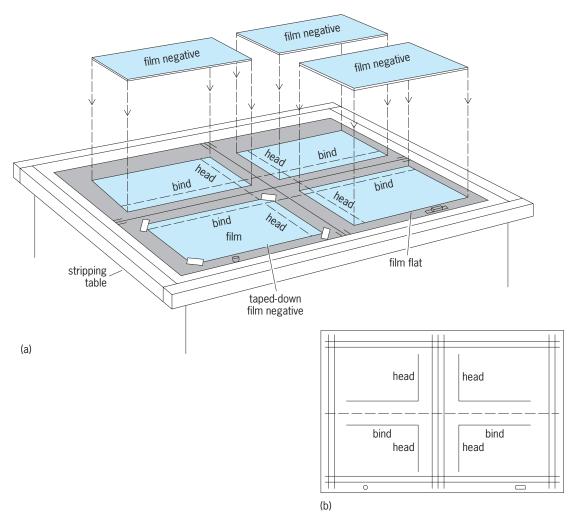


Fig. 4. Image film assembly. (a) Assembly setup. (b) Page position layout.

coatings. Most gravure uses copper-plated steel cylinders on which intaglio cells are etched by electromechanical engraving machines. Screen-printing stencils are made on polyester, nylon, or stainless steel fine-mesh screens, using band-cut stencils, transfer films, or light-sensitive coatings. Electrolytically produced rotary screens with perforated images and internal ink squeegees are also used for screenprinting.

Most of these processes use photomechanical methods that consist of photographic films exposed on light-sensitive coatings either directly coated on the printing surface or transferred to it. Since most photomechanical coatings are of slow speed, highenergy pulsed xenon or metal halide light sources are used for exposures, which are usually controlled by light-integrating devices. Step-and-repeat machines are used for exposing large plates for lithography, flexography, and wraparound letterpress where multiple exposures, as for label printing, are required. Automatic processing machines are used for making lithographic plates. The important feature of light-sensitive coatings is that on exposure to light, physical characteristics, such as solubility in water or other chemicals, change so that the coatings

can be developed to produce images that serve as the printing images or as resists for producing the printing images. The present trend is toward computer-to-plate systems, whereby plates are laserimaged directly from a computer. Some of these systems use thermal exposing as opposed to light exposure.

Photoengraving. This is the photomechanical process used to produce relief plates for letterpress, letterset, and flexographic printing. There are two general types of engravings: line and halftone. Zinc, magnesium, and copper are the metals commonly used for photoengraving. In the United States, zinc and magnesium have been used almost exclusively for line engravings, and copper for halftones and fine-line engravings.

Photopolymer plates. Polymeric materials are used to produce relief images by chemical means. Several types of photopolymers are used, such as poly(methacrylates), nylon, polyamide, urethanes, and photosensitive synthetic rubbers. The preferred combination has been nylon-type photopolymers on steel plates used on the press with magnetic plate cylinders and ultraviolet curable ink. *See* POLY-MER.

Letterpress declined in use because hard relief plates (metal and most photopolymers) require very precise preparations and adjustments (makeready) prior to printing, especially to compensate for the variable pressures exerted by the different-size image elements. To transfer ink, more pressure per unit area is exerted on small highlight dots than on large shadow dots or solid areas. To achieve equivalent ink transfer, considerable pressure adjustment is required.

Flexographic platemaking. Flexography uses flexible or resilient relief plates that alleviate the differential pressure/makeready problems of the inflexible relief plates used in letterpress. As a result, flexography has become the dominant relief printing process and has replaced letterpress in many newspaper, label, packaging, and corrugated-box printing applications. It uses rubber duplicate plates made from relief engravings for printing corrugated boxes and some packaging, and special resilient photopolymer relief plates for newspapers, pressure-sensitive and other labels, and milk cartons and other packaging materials. Seamless rubber-coated cylinders for printing gift-wrapping papers and wallpapers are made digitally by laser-engraving the rubber.

Gravure platemaking and cylinder making. Gravure is printed principally on web presses from etched cylinders of copper-plated steel and is generally called rotogravure. A small amount of sheet-fed gravure is printed mostly from thin copper plates wrapped around the plate cylinders of special singlecolor and multicolor presses. Sheet-fed gravure is capable of very high quality pictorial reproduction and is used for proofing and short runs of art and photographic reproductions, as well as annual reports.

Gravure differs from other printing processes in that it can print varying amounts and densities of ink to simulate continuous-tone imaging, as in photographic prints. The intaglio cells, cut or engraved and etched to various sizes or depths below the surface of a plate or cylinder, vary in volume and transfer different amounts of ink corresponding to the tonal values of the image areas being printed. Two processes for producing the cells represent two different types of cells. Conventional gravure, which is used for making most of the plates for sheet-fed gravure, produces cells all with the same areas but variable depths. Electromechanical engraving, the process used for almost all rotogravure printing, produces cells with variable areas and variable depths.

Conventional gravure. In this method, a sensitized carbon tissue (paper coated on one side with a pigmented gelatin layer sensitized by bichromates) or transfer film is contact-printed through a continuoustone positive film and given a second exposure by contact with a gravure screen consisting of a grid of 150 or 175 transparent lines and opaque square dots to the inch with a ratio of line to dot width of 1:3. The exposed tissue or film is moistened and squeegeed into contact with the clean copper surface of the plate or cylinder. The image is developed in water until a gelatin relief resist is produced. The image is etched with ferric chloride solutions, which etch the copper to different depths corresponding to the thickness of the gelatin resist in the different tone areas. The fully exposed areas such as the clear lines in the screen remain unetched and provide lands or bridges of common height to support the doctor blade during printing. The image areas consist of cells, all with the same area and variable depths depending on the tone values. This process is used mainly for high-quality single-color and processcolor printing. It suffers from short life on the press, as the shallow highlight dots wear prematurely, causing undesirable changes in tone reproduction.

Electromechanically engraved cylinders. Electromechanical engraving systems consist of three parts (Fig. 5): a scanning unit with a rotating drum on which photographic prints are mounted in position for printing and scanned by one or more reading heads; a computer that processes the image density data recorded by the scanning heads and converts the data into electrical impulses; and an engraving unit that uses the electrical impulses to drive diamond styli on one or more heads that engrave gravure cells in the rotating copper-plated steel printing cylinder. The depth of the cells varies with the strength of the electrical impulses, which correspond to the densities of the same areas on the originals. The electromechanical engraving system produces 4000 cells per second. The cells vary in area, depth, and volume, corresponding to the tone values they represent (Fig. 6). After engraving and before mounting on the press, the cylinder is polished to remove all burrs and chrome-plated to resist wear from the doctor blade on long runs.

Halftone gravure. The use of continuous-tone images for electromechanical engraving required high skills for retouching and correcting the films, prints, and engraved cylinders, and the production of expensive press proofs. The shift to the use of the same halftone images as produced for lithography has resulted in many improvements in cylinder making. Color separation and corrections have been simplified, and off-press and offset press color proofs can be used. Reading heads are used to defocus the halftone images on electromechanical engraving machines, and the hues of gravure color inks for magazine printing have been revised to match the inks formulated to the *Specifications for Web Offset*

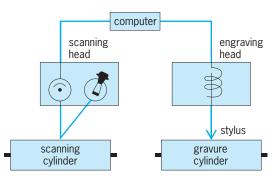


Fig. 5. Electromechanical engraving system.

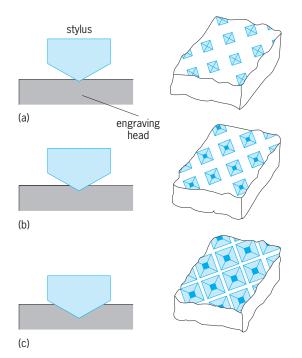


Fig. 6. Electromechanical engraving. Depth of cells and shapes and sizes as engraved by the stylus corresponding to tone values: (a) highlights, (b) middle tones, and (c) shadows.

Publications, which in turn match most off-press and offset press color proofs. There are also cost savings, since the same films can be used to produce the proofs and images for printing by lithography or gravure.

Filmless gravure. Gravure was the first process to use computer-to-plate (cylinder) digital technology to eliminate the use of films for making the printing cylinder. Interfaces have been developed between prepress color systems and electromechanical engraving systems, so that the digital data are fed directly to the electromechanical engraving computer to produce the electrical impulses to drive the engraving styli, thus making the scanning cylinder, prints, and reading heads unnecessary. The use of filmless gravure also eliminates the need for halftone gravure.

Photopolymer gravure. Photopolymer plates are also available for gravure. They are similar to letterpress photopolymer plates except that the image or printing areas are intaglio (recessed) instead of relief. Hard polymers, such as nylon, are used to resist wear from the doctor blade, which scrapes the excess ink off the surface of the plate. Such plates make gravure competitive with lithography and flexography for run lengths under 100,000, especially in packaging.

Lithographic platemaking. The main metal used for platemaking in lithography is aluminum with a mechanically, chemically, or electrolytically roughened surface. The image areas of lithographic plates are ink-receptive (nonwettable by water), and nonimage areas are water-receptive (nonwettable by ink). The wider the difference in ink and water receptivity of the two areas, the easier it is to control printing and the longer the plate will run on the press. The chemistry of the production of the images is related to the changes that occur in the photosensitive coatings on exposure to light and the development process.

Chemistry is also involved in making the nonimage areas water-receptive and desensitized to ink. One way to improve water wettability of the nonprinting areas is by roughening the plate surface. Another method employs surface treatments such as anodizing or treating with silicates or phosphates. Wettability is also improved and maintained by the use of hydrophilic (water-receptive) materials such as gum arabic or synthetic gums, which have good adhesion to metal surfaces. Gums are used in combination with phosphoric acid and salts such as diammonium monohydrogen phosphate and zinc or magnesium nitrate. Known as a densitizing etch, the mixture improves the water wettability of the metal surface to which it is applied.

Surface plates. The light-exposed coating on these plates becomes the image that accepts ink on the press. There are two types of surface plates: additive and subtractive plates. For additive plates, an inkreceptive lacquer or ink is applied to the image during development. Subtractive plates are precoated with an ink-receptive lacquer that is removed from the nonimage areas during development. Until the introduction of presensitized plates in 1950 and "wipeon" plates in 1957, the most popular surface plates were made with bichromated albumin or casein. These have become obsolete because of pollution problems and difficulties in printing, with image areas blinding (refusing to take ink) and nonimage areas scumming (wetting with ink) because of poor densensitization.

Additive plates are easy to make but are not as consistent and durable as subtractive plates. Diazo presensitized and wipe-on plates are simple to make. They print well because their surface treatments prior to coating are very water receptive, allowing printing with a minimum of water and reducing the tendencies for the images to blind and the nonimage areas to scum. Presensitized plates are precoated by the manufacturer; wipe-on plates are coated in-plant with diazo resin, usually by applying the coating with a simple roller. After exposure with a negative film, presensitized and wipe-on plates are handled in the same way. An emulsion developer, consisting of a pigmented lacquer, acidified gum solution, and a suitable emulsifier, is rubbed over the whole plate. The gum dissolves the coating in the unexposed (nonimage) areas and makes them waterreceptive, while the lacquer deposits on the exposed (image) areas, making them visible and ink-receptive. The plate is washed and gummed, and it is ready for use.

Additive plates are subject to abrasive wear on the press, so they are not suitable for runs much over 50,000 on sheet-fed presses or about 100,000 on newspaper and other web presses. The prelacquered subtractive presensitized plates are capable of runs almost double this.

Subtractive plates are coated by the manufacturer with a light-sensitive layer. After exposure, development removes the coating from the nonprinting areas, leaving the exposed coating in the image areas. Subtractive plates are generally more consistent and capable of longer press runs.

Positive plates are made from coatings that use naphthoquinone diazides. After exposure, alkaline developing solutions remove the exposed areas and leave the unexposed ones as the printing areas. These image areas are inked to protect them from light degradation, and the plate is gummed to protect the nonimage areas.

Photopolymer plates are the longest-running of lithographic surface plates. Some use polymers that can be developed in aqueous solutions, thus eliminating problems of water pollution. They are all subtractive plates for which processing is very simple. All are organic compounds that are very inert and abrasion-resistant. Light exposure changes the molecular size and molecular weight of some, and cross-links others to produce very durable images that are capable of runs in excess of 250,000 impressions. An important feature of some photopolymer plates is that run life on the press can be further increased by baking the plates in an oven at $450-550^{\circ}F(220-270^{\circ}C)$ for about 4–5 min. This treatment increases life on web presses to over a million impressions.

Bimetal plates. Unlike surface plates, the coating on bimetal plates serves as a stencil for etching the unexposed areas. The plate consists of copper plated on aluminum or stainless steel, on top of which a photosensitive coating is applied. After the coating is exposed and developed, the plate is etched with a ferric chloride solution to remove copper in the unexposed areas. The final printing plate consists of copper in the image areas and aluminum or stainless steel in the nonimage areas. These very rugged plates can take considerable abuse on the press from excessive pressure and papers with gritty surfaces, and are capable of runs of 1–5 million impressions.

Bimetal plates became popular because enlargement in the size of the halftone dot, or dot gain, can be controlled by etching rather than exposure. Photopolymer plates made from positives are preferred because the images have less dot gain than those on plates made from negatives. However, because the dot gain correction is erratic, many weboffset publication printers turned to bimetal plates, where the dot gain correction is proportional and controllable. Disposal of etching wastes is a problem, however.

Waterless plates. These plates differ from surface plates in that they use silicone rubber in the nonprinting areas and the image is printed from ink on the metal base. Silicone rubber has very low surface energy, preventing it from being wet easily with ink. Therefore, plates with silicone rubber in the nonimage areas can be printed without dampening solution, or water, eliminating the need for an ink-water balance that is the source of many printing problems, including long setup time and excessive paper waste. The process is sensitive to increases in temperature, which can cause printing in the nonimage area, called toning. Presses printing waterless plates need special inks and a temperature-control inking system. With these controls, waterless plates have shorter setup time, higher press speeds, and exceptional print quality in screen rulings up to 500 lpi. *See* SILICONE RESINS.

Digital plates. These plates are used in computerto-plate digital printing systems. They are made directly from digital data from high-end color electronic prepress systems or imagesetters interfaced with color desktop publishing systems, and they are used for printing on conventional lithographic presses. There are three types: high-speed lightsensitive plates, ablation plates, and heat-sensitive polymer plates.

High-speed light-sensitive plates. These plates are exposed by lasers in imagesetters. There are four types: silver halide, electrophotographic, high-speed dye-sensitized photopolymer, and hybrid.

Silver halide plates on polyester bases are used for short-run single-color and spot-color printing. They consist of silver halide emulsions processed in special developers. Silver halide plates on metal bases for process color printing were introduced in 1990.

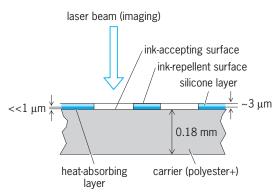
For newspapers that receive satellite transmissions, electrophotographic plates are coated with organic photoconductors, charged by corona discharge, and imaged by digital-data-driven visible-light lasers and liquid toners.

High-speed dye-sensitized photopolymers are used to produce long-run lithographic plates. The photopolymers used for conventional plates are mainly sensitive to ultraviolet radiation. As is done with photographic emulsions to produce orthochromatic, panchromatic, and infrared-sensitive films, special dyes are added to the photopolymers to extend their light sensitivity into the visible spectrum, thus increasing their exposure speed. A number of such plates are available for use. Most have light sensitivities for exposure by a medium-power 30–50-milliwatt (blue) argon ion (488-nanometer) laser.

Hybrid plates use two photosensitive coatings on a metal plate. The top coating is a silver halide emulsion whose light sensitivity can be varied to handle a full range of speeds, and a full gamut of spectral responses from ultraviolet to blue, green, red, and infrared lasers. The bottom coating is a photopolymer with good performance characteristics on the press. The top coating provides a wide range of light sensitivities to accomplish imaging for contact, projection, camera, or digital computer-to-plate imaging, with controlled dot gain and image contrast. The bottom coating provides ease of printing characteristics that result in high productivity and consistent quality such as plate life, durability, ease of makeready, and maintenance of inkwater balance, ink transfer, register, and other running controls on the press. See LASER; ULTRAVIOLET RADIATION.

The top layer (silver halide) is exposed by a laser in the platesetter, and then the plate is developed to form an opaque image. Next, the plate's photopolymer coating is exposed to ultraviolet light through the clear (top) image areas and then developed. Plate life can range from 150,000 impressions to over 1 million if the coating is baked after processing.

Ablation plates. These are lithographic printing plates with slightly depressed (ablated) wells in the image areas. They were introduced as waterless plates with silicone rubber in the nonimage areas, and the printing wells were produced by spark discharge, driven by digital data. Because ablation is a dry process (no liquid development), four-color plates can be imaged simultaneously directly on the press, with an additional feature that the pixel distribution for each color is counted as the plates are being exposed and the data are used to automatically set the ink fountains. The spark discharge system has been replaced by laser diode technology with considerable improvement in printed image quality (**Fig. 7**). Laser diode technology can also be used to make





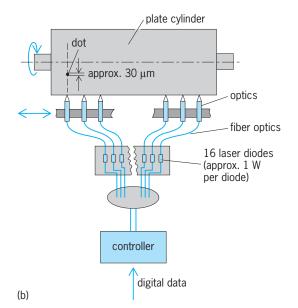


Fig. 7. On-press plate imaging for waterless offset printing with thermal laser ablation technology. (a) Printing plate (thermal plate) structure for waterless offset and thermal ablation process. (b) Principle of the imaging head as used for plate imaging in the system.

plates that can run in a conventional lithographic ink-water balance mode.

Heat-sensitive polymer plates. A thermally sensitive polymer responds to heat but not light, so it can be handled in ordinary room light. It is imaged in a proprietary platesetter, fitted with a laser diode (heat source). The plate is processed with an aqueous solution. Run lengths of over 1 million impressions, without baking, have been achieved, as well as reproductions of fine image details and resolutions of 4.8 μ m, the equivalent of a 1% dot on a 600-lpi halftone screen.

Printing Processes

Two categories of printing presses are sheet-fed presses and web presses. Sheet-fed presses print on individual sheets of paper, as opposed to continuous rolls of paper used on web presses. Some types of sheet-fed and web presses can print on both sides of the paper at the same time; this is called perfecting. Sheet-fed presses range in size from small presses that print on letter-size paper to presses that handle large sheets of paper. In addition, sheet-fed presses can print on cardstock, plastics, metal, and other substrates. A sheet-fed press may consist of multiple print units, with each printing a different color.

Web presses print on continuous rolls of paper or other substrates. Web presses are used for highvolume printing of magazines and newspapers. They typically are faster than sheet-fed presses. Web presses can consist of several connected units for printing different colors and providing finishing processes such as cutting, folding, and punching.

High-speed commercial web presses use wide rolls of paper for printing magazines and books, and use heat to set the ink (heat-set web). Cold-set web presses are used for printing on newsprint or other paper on which ink is dried by absorption and oxidation.

Offset lithography. Offset lithography is the most popular and widely used process and represents approximately 45% of all printing. In offset lithography, an image is transferred to the paper using inks, specially treated plates, and an offset cylinder. The image and nonimage areas of the printing plate are on the same plane, as opposed to plates with raised or etched type or images.

Lithographic plates' image areas are ink-receptive and water-repellent, and the nonimage areas are water-receptive and ink-repellent. The term "offset" refers to the image being not printed directly on the paper but transferred (offset) from the plate to another surface that makes contact with the paper. An important function of the lithographic printing press is to provide a means for transferring ink and water to the plate and maintaining the proper ink-water balance during printing.

In offset lithography, an image is produced on a thin aluminum plate, although other metals, plastics, and paper can be used. Rollers apply ink (usually oilbased) and fountain solution (mostly water) to the plate. Since oil and water do not mix, the oil-based

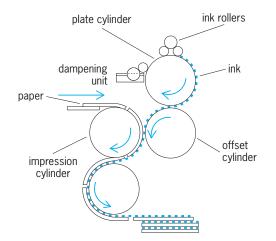


Fig. 8. Lithographic printing process.

ink adheres only to the image areas. The inked image is transferred to a rubber blanket mounted on a cylinder and then transferred to the paper as it passes between the blanket and an impression cylinder beneath the paper (**Fig. 8**).

Lithographic printing press. Lithographic presses are rotary presses; that is, the plate is mounted on a cylinder that rotates during printing. Because the plate is comparatively thin, from 0.005 in. (0.125 mm) for a small press to 0.02 in. (0.5 mm) for a large press, uniformity is very consistent, which reduces makeready time.

For sheet-fed presses, the plate cylinder is provided with the means to rotate it independent of its gear drive, making possible the desired front-toback positioning of the image on the sheet (**Fig. 9**). Systems have been developed that accurately position the image on the plate to minimize or eliminate the need to change the position of the plate cylinder. Some plants install pins in the plate cylinder that correspond to the pin system used to coordinate image positioning in platemaking operations. Newer electronic printing presses automatically position the plates.

Dampening system. When the lithographic press is running, the plate cylinder carries the plate under two different roller systems. The first system that contacts the plate is the dampening system. In a conventional dampening system, the dampener is separate from the inking system and uses cloth-covered form rollers and a ductor roller that alternately contacts the fountain (dampener reservoir) roller and the dampening oscillator roller. Some dampening systems feed the solution to the first ink form roller, which transfers it to the plate. Others spray the solution on the plate. The surface of the plate is slightly dampened, but the image areas, being hydrophobic (water-repellent), resist wetting. The dampening (fountain) solution is about 98% water, with the other 2% made up of a variety of chemicals, which in most systems make the solution acidic. Alkaline dampening solutions are also used, primarily in lithographic newspaper printing. Automatic devices are used for mixing, supplying, and replenishing fountain solutions, as well as measuring their pH and conductivity and controlling their temperature.

Inking system. After dampening, the plate passes under the ink form rollers. The nonimage areas of the plate do not accept the ink if properly dampened, but the image areas do. Lithographic ink is thixotropic, as well as very viscous. These attributes necessitate a complex inking system, consisting of drum, oscillator, riders, and fountain rollers made of hard materials such as steel or hard rubber (**Fig. 10**). With the exception of rider and fountain rollers, the steel rollers are usually covered with copper, ebonite, or nylon to resist the tendency of acids in the dampening solution to desensitize them. Another reason for this complex inking is the fact that some of the

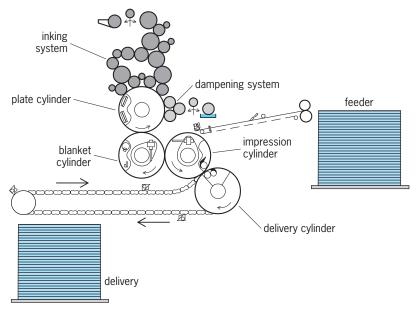


Fig. 9. Single-color sheetfed press. (After L. DeJidas and T. Destree, Sheetfed Offset Press Operating, 2d ed., Graphic Arts Technical Foundation, 1994)

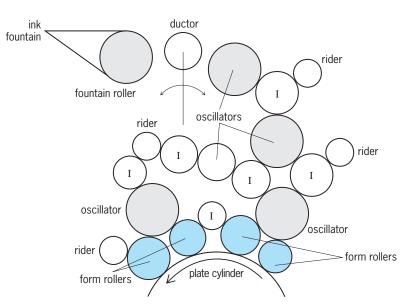


Fig. 10. Typical lithographic press inking system. I = intermediate roller. (After L. DeJidas and T. Destree, Sheetfed Offset Press Operating, 2d ed., Graphic Arts Technical Foundation, 1994)

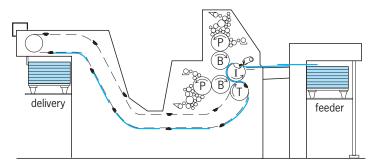


Fig. 11. Two-color press that has a common impression cylinder. P = plate cylinder. B = blanket cylinder. I = impression cylinder. T = transfer cylinder. (*After L. DeJidas and T. Destree, Sheetfed Offset Press Operating, 2d ed., Graphic Arts Technical Foundation, 1994*)

dampening solution is picked up by the ink, and the higher the surface area of the rollers, or the more rollers, the faster the dampening solution is evaporated from the inking system to prevent emulsification. Ink fountains have keys for adjusting the flow of ink to the ink train. Setting the ink keys manually is a tedious and time-consuming operation. Electronic plate scanners are available for setting ink fountains automatically. There are also means to calculate the ink demands of digital plates to set the ink fountains automatically.

Waterless systems. Special presses are built for the printing of waterless plates. They have no dampening system, but temperature-controlled inking rollers. Most systems use refrigerated water circulation in the ink distributor or vibrator rollers. Others use special fans to blow cool air over the inking rollers. Inks for waterless printing have a control temperature range of about 10° F (5.6°C) between good ink transfer and toning, or scumming. Since the blanket is dry, waterless plates require the use of higher grades of paper to avoid the accumulation of lint and other paper debris on the blanket during printing.

Blanket cylinder. After inking, the plate contacts the blanket cylinder. The cylinder is equipped with blanket mounting and tightening devices, just as the plate cylinder has similar devices to mount and tighten the plate. A principal difference is that the plate cylinder can be rotated independently of its driving gear, while the blanket cylinder cannot. The position of the plate cylinder in the frame of the press is fixed, but the blanket cylinder of all but some small-sized presses is so mounted that its position with respect to the impression cylinder can be changed. This feature makes it possible to change the clearance between blanket surface and impression-cylinder surface to accommodate printing on substrates with a wide range of thicknesses. For example, sheetfed presses can be adjusted to print on onionskin or board.

Impression cylinder. After receiving the ink image from the blanket cylinder, the blanket contacts the paper, or substrate, that is carried on a third, impression cylinder. At the lead edge of the cylinder is a device to receive and firmly hold the front edge of the sheet to be printed during the cycle when the blan-

ket contacts the paper. The device is known as the gripper bar, a shaft across the length of the cylinder on which is mounted a series of fingers known as grippers. As the sheet is drawn through the printing nip (the area where blanket and substrate contact each other), the leading edge of the cylinder reaches a position where the impression cylinder grippers release the sheet, while other sets of grippers mounted on traveling chains (one on each side of the press) carry the printed sheet to a delivery pile and release it there.

Lithographic press configurations. The basic lithographic offset press is sheet-fed. It can be fed from a roll by attaching a paper-sheeting device at the feeder. But one of the most important features of the lithographic offset sheet-fed press is the ease with which a multicolor press can be assembled. Presses with up to 10 units are built to meet special needs. If a plate, a blanket, and an impression cylinder in combination are considered as a printing unit, a two-color sheetfed press comprises two printing units with feeding and delivery components. A six-color sheet-fed press comprises six such units.

One deviation from this approach to multicolor sheet-fed presses is the common impression cylinder design. In sheet-fed presses, two plate cylinders and two blanket cylinders are assembled as couples. Each couple comprises a plate and blanket cylinder. Two such couples, mounted with an impression cylinder between them, results in a two-color common impression cylinder unit (**Fig. 11**). Two such units assembled in tandem, with provisions for sheet transfer from one impression cylinder to the other, become a four-color press. The rotary sheet-fed press has lent itself to the development of presses that can be run as either perfecting (printing both sides of the sheet) or as multicolor, with a very short time required for the changeover of modes.

In-line (tandem) web press. To meet some specialty needs, such as forms and package printing, printing units are assembled in tandem (**Fig. 12**) and fed from

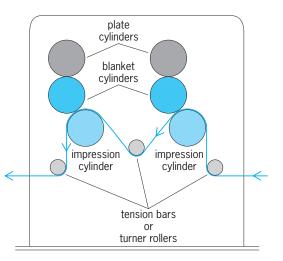


Fig. 12. Unit-type press with printing units assembled in tandem. (After R. Blair and C. Shapiro, The Lithographers Manual, Graphic Arts Technical Foundation, 1980)

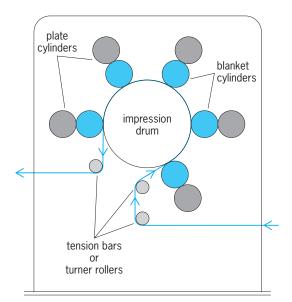


Fig. 13. Common impression drum-type press. (After R. Blair and C. Shapiro, The Lithographers Manual, Graphic Arts Technical Foundation, 1980)

a roll of paper (web-fed). Such a press can print lithographic and indirect (offset) letterpress plates on the same run. Presses have been built with three to five printing units in tandem around a common impression cylinder (**Fig. 13**). This configuration has been discontinued for publications presses, but it is used in newspaper presses for process color printing.

Blanket-to-blanket perfecting. The greatest growth in lithography has resulted from a unique capacity of the offset method of printing. That capacity is simultaneous perfecting, or printing, on both sides of a substrate by two opposing blankets (Fig. 14). A printing unit in this configuration has two plate/blanket cylinders assembled, such that the blanket cylinders run against each other. Each blanket cylinder transfers the ink for its image and, at the same time, acts as the impression cylinder for the opposing blanket cylinder. These presses are known as blanket-to-blanket presses. The configuration lends itself to a variety of assemblies, depending on the type of product being printed. Up to 12 units have been assembled into a single press. In most commercial printing applications, the web is fed straight through. Many of the presses are equipped so that more than one web may be printed simultaneously. On an eight-unit press, for example, two webs can be run, with each being printed in four colors on both sides. This configuration is used on most magazine and catalog presses.

Ink drying on web lithographic offset presses. After printing, the specially formulated (heat-set) inks are dried in high-temperature ovens. Drying is virtually instantaneous because of the rapid and complete evaporation of the ink solvent. Before the web is passed into a finishing operation (usually folding), the heated web is passed over chilling rollers to set the ink, which is in a molten state when the web leaves the dryer. In newspaper printing by lithography, ink drying results from the absorption of the ink vehicle into the paper.

Relief printing. In relief printing, the image to be transferred to paper (or other substrate) is raised above the surface of the printing plate. Ink is applied to the raised surface and then rolled or stamped onto the substrate (**Fig. 15**). The process is similar to using an inkpad and stamp. Letterpress and flexography are forms of relief printing. Although the image to be printed is raised on the printing plate, it does not necessarily create relief or raised lettering, as in embossing.

Letterpress. Letterpress is printing from a raised surface. It is the oldest of the present printing processes and has been replaced mostly by offset and other printing processes. Letterpress printing is still used for specialized and limited-edition printing, as well as for embossing, die cutting, and foil stamping. It represents less than 5% of all printing. The image is

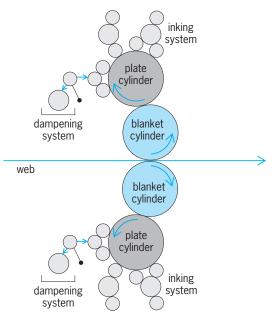


Fig. 14. Typical blanket-to-blanket printing unit of the horizontal type. (After GATF Staff, Web Offset Press Operating, 4th ed., Graphic Arts Technical Foundation, 1996)

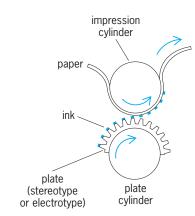


Fig. 15. Relief printing process.

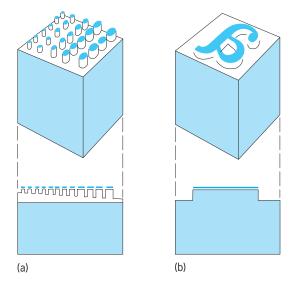


Fig. 16. Printing-image carriers for letterpress. (a) Dots of a halftone. (b) Type face. (After V. Strauss, The Printing Industry, Printing Industries of America, 1967)

wrong reading, or reversed, on the plate and is transferred directly from the plate to the substrate, where the image appears as right-reading (**Fig. 16**). Letterpress plates are typically made of zinc, copper, or a photopolymer.

Three are three major letterpress-printing processes. The platen press (**Fig. 17***a*) has a flat platen and bed. The flatbed cylinder press (Fig. 17*b*) has a flat bed, with an impression cylinder in place of a platen. The rotary press (Fig. 17*c*) has two cylinders, with the flatbed being replaced by a plate cylinder.

The most common size of a platen press is 10×15 in. (25×38 cm; maximum sheet). The platen press configuration, widely used in die cutting, creasing, and embossing of packaging material, can be considered as an offshoot of letterpress printing. For this purpose, platen presses are available in sizes up to about 64 in. (163 cm).

The flatbed cylinder letterpress was the predominant printing press until the rapid growth of lithography began. This configuration has become virtually extinct except for its use in the packaging field as a cutter/creaser and embosser.

Before 1950, the rotary letterpress was the means that identified printing and set the standard for printing quality. It was responsible for the growth of magazine, catalog, directory, book, and commercial

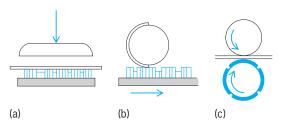


Fig. 17. Letterpress printing presses. (a) Platen press. (b) Flatbed-cylinder press. (c) Rotary press. (*After V. Strauss, The Printing Industries, Printing Industries of America,* 1967)

printing in the United States. The great commonimpression cylinder presses of the 1930s and 1940s almost completely disappeared from the scene as lithography, gravure, and flexography took over the printing markets.

In 1970, rotary letterpress started a comeback with the development of photopolymer plates, ultraviolet-cured ink systems, and the design of high-speed rotary presses with interstation curing. The presses became popular in label printing, where the narrow-web (up to 25 in. or 63.5 cm) rotary press was used extensively for the production of pressure-sensitive labels and packaging for pharmaceuticals, cosmetics, and toiletries.

The belt press is a type of letterpress (**Fig. 18**). It uses two printing belts to which flexible, shallowrelief photopolymer plates produced from photographic negatives are secured with double-sided tape supplied in wide, accurately sized sheets (stickyback). The plates are page size and imposed on the belts in the long-grain direction, on a plate mounter, to permit collating of completed books at the delivery end of the press. The use of belts, instead of the fixed-circumference used in conventional presses, permits a virtually limitless assortment of page sizes and page counts to be printed without time-consuming, costly plate changes. The press is ideally suited for the production of mass-market paperback books.

Flexography. Flexography is a growing process and frequently is used for printing on packaging materials. Flexography uses plates that have a slightly raised image made of rubber or plastic, including photopolymers. The inked plates are rotated on a cylinder that transfers the image to the substrate. Flexography is a high-speed printing process for printing on many types of absorbent and nonabsorbent materials, including plastic, foil, acetate film, brown paper, toilet tissue, bags, pressure-sensitive labels, corrugated board, paper and plastic bags, milk cartons, disposable cups, candy bar wrappers, envelopes, labels, and newsprint. It is also used for in-line numbering, in-line folding, dual-sided laminating, booklet placing, pattern gluing, dual-web registration, in-line UV coating on both sides of the web, promotional labels, booklet labels, instant redemption coupons, and encapsulated game pieces. It is also popular for long-run printing requiring one or two flat colors, and where sharpness and the highest print quality are not required. However, there is a growing trend toward high-quality flexography.

Flexography is a form of rotary-web letterpress. The print area or image consists of a raised surface, or relief, which can be inked and pressed onto the substrate. The rubber, plastic, or other flexible plate material is mounted to the printing cylinder with double-faced adhesive or slid on as a sleeve. Nonimage areas are below the printing surface and do not reproduce. The plates are sometimes backed with thin brass or metal sheets and attached to the cylinder with fastening straps for close register. The inked plates are rotated on a cylinder that transfers the image to the substrate.



Fig. 18. Belt press system. (Cameron)

Flexography inks are fast-evaporating solvent types, water-based types, or ultraviolet-curing types. The presses have a simple ink-distribution system in which a roller rotates in a fountain or trough of ink. The ink is transferred from the fountain roller to the printing plate by an engraved (anilox) roller with cuplike cells that meter the amount of ink transferred. The addition of a reverse-angle doctor blade to the transfer roller is sometimes used to meter the ink. The anilox method has been used on letterpress and lithographic presses and is known as keyless inking.

There are three types of flexographic presses: stack, common impression cylinder, and in-line (**Fig. 19**). In the stack-type press, the color sections are arranged one above the other or one on both sides of the main press frame. In the common impression cylinder press, a central impression cylinder press, also known as the drum, supports all of the color sections around a single impression cylinder mounted in the main press frame. This prevents register shift from color to color. The in-line unit press has separate color units horizontal one to the other, driven by a common line shaft. The in-line press can print on both sides of a web.

The growth of flexography is largely due to rapid improvement of plates, anilox rollers, and inking systems. Approximately 70% of all packaging is printed by flexography. The simplicity of the flexographic printing process, consistency, and relative low-cost make it increasingly popular among print buyers. Growth in corrugated cartons and flexible packaging as well as a trend to move away from glue-applied labels to pressure-sensitive labels should keep flexography in a growth mode for many years.

Flexographic printing has considerable impact on the environment based on its use of rubber and photopolymer plates, solvent-based inks, and hydrocarbon solvents. As a result, water-based inks are increasingly used by flexographic printers, especially when printing on paper substrates such as corrugated containers and newsprint.

Chrome-plated rolls have been replaced by laserengraved ceramic surfaces. The doctor blades now wear instead of the roller. Laser-engraved ceramic rollers wear better with the doctor blade in place.

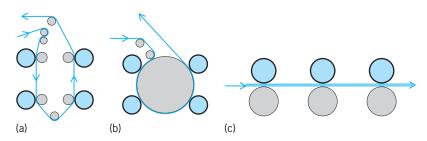


Fig. 19. Flexographic rotary presses. (a) Stack type. (b) Common impression cylinder type. (c) In-line press. (After V. Strauss, The Printing Industry, Printing Industries of America, 1967)

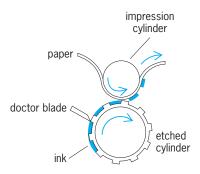


Fig. 20. Gravure (intaglio) printing process.

Several new drying and curing technologies have been developed that reduce energy costs and work well with new ink formulations. Infrared drying systems require moderate capital cost but have high operating costs. They are often used in conjunction with convection air dryers to provide the heating load to the coated web. Radio-frequency systems are an efficient alternative for drying water- and solventbased coatings. There are high capital costs and moderate operating costs. Supplemental convection air dryers are sometimes used in conjunction with radio frequency.

Intaglio printing. In the various intaglio printing methods, the image to be printed is recessed into the surface of the printing plate. The incised image may be etched with chemicals or engraved with tools. Intaglio printing is used for printing paper currency and for fine-art printmaking. The image areas are filled with ink, and excess ink is wiped from the plates. Heavy pressure is applied to transfer the ink from the plates to the paper. Other forms of printing that use the intaglio process include engraving and gravure.

Gravure. Gravure is the opposite of letterpress. It involves printing from a recessed surface (the image is beneath the surface of the plate or cylinder), and the typical plate is a large copper cylinder (**Fig. 20**). The image is made of microscopic wells that are etched or engraved and filled with ink. The plate is rotated on a cylinder, transferring the image to the paper or other material. Gravure printing is often used for high-volume printing of packaging, wallpaper, and gift-wrap, using fast-drying inks. Although less common, gravure printing may also be used for printing magazines, greeting cards, and high-volume advertising pieces. Gravure is also popular for long-run magazines, packaging, and printing on foils, plastics, and cellophane.

Publication gravure printing is done by relatively few companies, which have large printing plants. Packaging and specialty gravure printing is done by a considerably larger number of companies, ranging from large integrated packaging companies with many gravure press units to small captive operations with only one or two press units.

Gravure is a high-cost and time-consuming platemaking process. It is economical for very long runs and for high-quality printing. The advantages of gravure are the long life of the recessed image and the uniformity of ink, resulting from the fact that the quantity of ink printed is determined by cell volume only. Fast-drying solvent inks, which make possible drying between each impression, and the ability to print well on a wide variety of surfaces, are other advantages.

U.S. Environmental Protection Agency (EPA) emission-control regulations require all printing processes to reduce emissions of volatile organic compounds. Gravure has the advantage that the solvent used in gravure inks can be recovered by carbon adsorption and recycled. In addition, nonpolluting water-based gravure inks have been developed and are widely used in packaging and specialty gravure.

The gravure image carrier is a copper-plated cylinder or copper plate on which the image is in the form of cells or cups engraved in the surface. If chrome-plated, the gravure cylinder can run for millions of impressions. A typical gravure cell is 35 μ m deep and 125 μ m square, with 22,500 cells per square inch (150-line screen). The walls between cells are known as the bridge or doctor blade support, and the intersections between the walls are called the corner posts. The gravure cylinder is flooded with a liquid gravure ink. The excess ink is removed with a high-quality steel doctor blade 0.006–0.010 in. (0.15–0.25 mm) thick. The ink is transferred by pressing the paper against the cylinder as it turns, using a rubber-covered impression roll.

The basic printing unit consists of an engraved gravure cylinder at the bottom with a rubber-covered impression roll at the top (**Fig. 21**). The gravure cylinder is geared to, and driven by, a motor-driven shaft, which drives the other gravure or converting units in the line. The impression roll is friction-driven through the web by the gravure cylinder. For the wide presses, used in publication printing, the impression roll includes a deflection-compensating

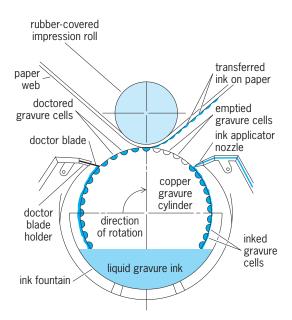


Fig. 21. Gravure printing unit.

mechanism to minimize bending. Mechanical impression screws or hydraulic or pneumatically operated impression pistons apply pressure to the end journals of the impression roll, pressing the rubber against the gravure cylinder.

The gravure cylinder runs in an ink fountain or trough. Spray nozzles may also be used to apply ink at a point close to the exit side of the nip. The ink fountain on high-speed presses using volatile inks is totally enclosed except for a small area between the doctor blade and the nip and on the exit side of the nip. Enclosure of the ink fountain reduces solvent evaporation and permits the use of fast-drying solvents.

The doctor blade holder is provided with an oscillating mechanism to permit adjustment of the blade angle, as well as the horizontal and vertical position of the doctor blade. Doctor blade angles, measuring from a normal to the gravure cylinder, range from 15° on publication presses to 45° on some packaging presses.

The side frames and structural elements of a gravure press unit are usually constructed of heavy cast-iron members to provide the mass and rigidity needed for vibration-free, high-speed running. Gravure press speeds range from 400-800 ft/min (2-4 m/s) in packaging and specialty printing, 1000-2500 ft/min (5-12 m/s) in magazine and supplement publication printing, and up to 4000 ft/min (20 m/s) or more on some printing operations running in line with paper machines.

Once the various press settings are made and the ink viscosity and drying conditions are adjusted for proper ink transfer and printed film characteristics, the gravure printing or coating operations is relatively simple to control. Maintenance of ink viscosity is the main requirement. This is done by periodically adding solvent to the ink to make up for evaporation. The makeup solvent is blended to maintain the primary solvent/resin mixture in the original ink formulation.

Gravure ink transfer is accomplished when the paper is pressed into contact with the liquid ink in the gravure cell. The characteristic hollow dot of the gravure indicates that ink transfer occurs around the edges of the cell where the ink first contacts the paper, and capillary action draws a portion of the remainder of the ink to the paper. In 1967, an electrostatic ink transfer process was introduced in the gravure industry. In this process, the paper web is charged by application of a dc voltage to the top of the paper through a slightly conductive rubber coating on the impression roll. The field lines established from the paper to the grounded gravure cylinder concentrate at the edges of the cell and exert an upward pull on the inductively charged ink, causing the meniscus to be distorted and bulged into contact with the paper. Dramatic improvement in print smoothness resulted.

Digital technology has transformed gravure printing from a craft to a high-technology industry. Virtually all publication and a growing percentage of packaging and specialty gravure cylinders are electronically engraved on electromechanical engraving machines. Predominant is a machine that uses an electromechanically driven diamond stylus (at 4000 cells per second) to engrave pyramid-shaped gravure cells in the copper cylinder. With an eightengraving-head machine, a large publication gravure cylinder can be completely engraved in about 1 h, plus setup time.

Electronic page makeup systems interfaced to electronic engravers permit filmless production of gravure cylinders. As an interim step before switching to entirely filmless gravure prepress systems, the industry began electronic engraving from halftones instead of the traditional continuous-tone material. This technique, also known as halftone gravure conversion, permits gravure to handle much work formerly intended for offset lithography and reduces gravure prepress cost.

Engraving. Used for fine invitations and sometimes business cards and letterhead, engraving creates a raised surface on the front of the paper. A metal die with an impression cut into the surface is filled with an opaque engraving ink, then pressed against paper, creating raised areas of the paper coated with ink.

Screen printing. Screen printing is the simplest of the major printing processes and is popular for shortrun printing. It requires a very thick ink film and is normally a slow manual process, though automated sheet-fed and web presses are available.

Screen printing can accommodate a wide variety of substrates that differ in shape, size, and material, including paper, plastics, textiles, metal, glass, ceramic, leather, and wood. Objects having cylindrical and irregular shapes, such as containers, barrels, glassware, and bottles, are frequently decorated by means of screen printing.

Screen printing is a versatile printing process. Based on simple stenciling techniques, modern screen printing uses precisely woven mesh to support a stencil, where the image is defined by open areas. The open areas allow ink to pass through the top of the screen to the substrate below.

Method. In the basic screen printing method, an industrial woven fabric (**Fig. 22**), generally polyester, nylon, or stainless steel, is stretched to a predetermined tension and affixed to a wood or aluminum frame. Stretching can be accomplished by holding the fabric in a clamping system driven with either mechanical or pneumatic motion to exert the necessary force to pull the fabric to the desired tension level. The fabric is then attached to the frame with a suitable adhesive. Self-tensioning frames can also be used, in which the fabric is clamped into the frame itself and elements in the frame sides are expanded or rolled to create the stretching force.

Fabrics range in mesh count, or fineness, from 25 to over 500 woven threads per inch and have various thread diameters. The woven structure of a screen contains mesh openings that allow ink to pass onto the substrate in the areas not blocked by the stencil during the transfer process. Higher mesh counts are used for fine-detail images, while lower counts are

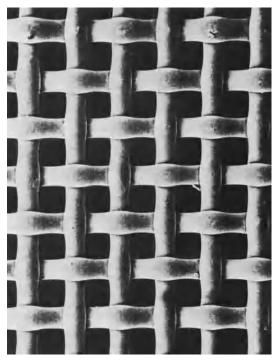


Fig. 22. Enlargement of meshed-screen fabric for screen printing. (From A. Akermann, Influence of mesh on ink density in halftone printing, Technical Guidebook of the Screen Printing Industry, vol. B, Screen Printing and Graphic Imaging Association, 1987)

used to deposit larger amounts of ink, as required by textile applications.

There are several types and classifications of stencil material. One of the simplest is a knife-cut stencil, in which the image is hand-cut from the material and the stencil is then adhered to the mesh. The most popular stencil type is a photosensitive emulsion (direct emulsion) that is coated on a screen and allowed to dry. The emulsion is then exposed through a film positive using a high-intensity ultraviolet light source. The unexposed areas of the emulsion are water-soluble and washed out, while the light-exposed areas become insoluble and remain as the stencil. A capillary-film stencil is precoated on a polyester sheet, applied to a wet screen, and then allowed to dry. It is processed the same way as the direct method. An indirect stencil has a gelatin composition and is exposed and processed independently from the screen and then adhered.

The fully prepared printing screen is mounted onto the press, leaving a slight gap between the bottom of the stencil and the substrate being printed. A polyurethane squeegee is positioned to permit a regular sweeping motion, and ink is introduced to the screen (**Fig. 23**). Ink transfer is accomplished when pressure from the squeegee brings the screen and substrate into intimate contact. The forward stroke moving the squeegee then draws the ink across the screen while creating a shearing force that transfers the ink through the open areas of the stencil onto the substrate. Following ink transfer, the screen separates from the substrate because of screen tension, leaving behind the printed image. In most screen-printing applications, a color must be printed and dried before the next one can be applied. Registration is necessary for multicolor prints, and this is attained by using automated or manual techniques. Textile applications are screenprinted by using direct-contact printing (screen and substrate are completely touching), and colors are printed wet on wet before the ink is cured.

Presses. Various press types are available for screenprinting. The most common is the flatbed press, where the squeegee is drawn back and forth over a stationary screen while the substrate is held firmly in place with a vacuum bed. Textile applications use flatbed presses.

On cylinder presses, the screen is driven back and forth under a stationary squeegee as the substrate is simultaneously moved under the printing point. Similar presses are used to print cylindrical objects, such as cans, by rotating the object under the moving screen.

Most screen-printing presses can handle only piece goods, but web presses are also available for printing roll stock and textiles. Rotary presses are web presses that use a cylindrical screen with a squeegee inside. This press configuration boasts much faster printing speeds than both cylinder and flatbed presses, and is used primarily to print textiles and wall coverings.

Screenless printing. Screenless printing has the potential of coming closer to printing facsimile reproductions of originals than any of the processes using screens or halftones. It has no undesirable moiré patterns caused by overlapping screens, and the colors are purer and more saturated, since they are not degraded by light absorption in the paper between the dots of halftone patterns. Processes involving screenless printing are considered essentially specialty processes because of the difficulty of controlling them. These are collotype and screenless lithography. Digital means also are used to simulate the effects of screenless printing.

Collotype. This type of printing is known in different countries by different names, for example,

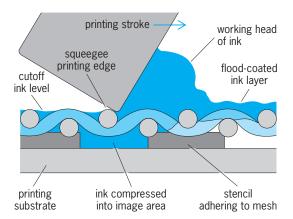


Fig. 23. Basic factors in the screen-printing action. (After E. J. Kyle, Principles of quality control, Technical Guidebook of the Screen Printing Industry, vol. C, Screen Printing and Graphic Imaging Association, March 1975)

photogelatin, phototype, heliotype, heliogravure, albertype, and artotype. The process uses no screen, but the prints closely simulate in gradation true photographic prints.

Originally, collotype printing was a hand process in which production might amount to no more than 100 impressions per day. With a good machine, the modern process may average up to 2000 impressions per day. The plate, usually aluminum, is coated with a layer of light-sensitive (bichromated) gelatin, over which is laid an unscreened (continuous-tone) photographic negative of the copy to be reproduced. When exposed to light, the gelatin is hardened in varying degrees in proportion to the amount of light received from the dark, medium, and light portions of the negative. The plate is developed in water to remove unexposed bichromate and immersed in a glycerin or glycol solution. The result is a printing image that varies not only in thickness (in thousands of an inch) but also in moisture content and consequent ability to repel ink or hold it in varying amounts corresponding to exposure of the gelatin. The harder areas accept less moisture and swell less, thus accepting more ink.

The plate is kept moist by maintaining the pressroom relative humidity from 30 to 95%. When the ink roller passes over the plate, which is dampened by the moisture in the air, the greasy ink is picked up in inverse proportion to the amount of moisture in the partially hardened portions of the gelatin.

Technically no two prints are exactly alike in collotype, because the tone reproduction varies with the moisture content of the gelatin. Each impression removes moisture, so the relative humidity must be gradually increased to compensate for the moisture loss. Eventually, the tone reproduction becomes flat, and the plate must then be soaked again in glycerin or glycol solution and the relative humidity reduced to start a new cycle of printing.

Collotype is a short-run process. Runs of only a few hundred prints are common, with 5000 as the practical maximum. The process has been used extensively for art reproductions and for a wide range of advertising work in both black-and-white and color, including counter and theater displays, posters, blowups (enlargements up to 42×62 in. or 107×157 cm) or advertisement and pictorial copy, and facsimiles of photographs.

Screenless lithography. This is a means of printing lithographic plates made from continuous-tone positives or negatives on an offset press. Best results have been obtained with special photopolymer plates exposed from positives. Unlike collotype, the plates are run conventionally on the press with regular inks and fountain solutions. The random grain pattern of the images, corresponding to the grain of the plate, is almost indistinguishable from the gelatin characteristic of collotype printing, and screenless lithography has been mistaken for collotype.

Plate coating variations, which affect the consistency of printing results, are a disadvantage that has kept the use of screenless lithography to a minimum. The tone reproduction of plate coatings depends on the roughness of the plate surface (grain) and the thickness of the plate coating. Since each is dependent on the other, and neither can be accurately controlled in manufacture, plate results are inconsistent and unpredictable.

Another approach in simulating the effects of continuous-tone reproduction is to use special screens with random grain elements of about 500–600/in. (200–240/cm). An additional method is the production of images on electronic color scanners with very fine elements in random patterns. Inkjet and dye-sublimation color proofing systems also produce screenless images.

Digital printing. First introduced in the 1990s, digital color printing presses have had a major impact on the printing industry sooner than most believed they would. Computer-to-film technology, which bypasses the laborious tasks of graphic arts photography and image assembly, was a major step toward automating the printing processes in the late 1980s. Computer-to-plate technology, an extension of this technology, gained in popularity in the 1990s. However, computer-to-press technology provides the promise of completely integrated printing systems. Direct-to-press color systems were introduced throughout the 1990s, such that by 2000 direct-to-press color-printing systems were competing with commercial printing presses.

On the hardware side, systems that handle a greater variety of substrates, including very lightweight paper, have been developed. The concept of a universal copier/printer also developed. These devices output color and monochrome pages at competitive costs with dedicated color and monochrome printers.

Adding value by integrating all services from frontend to printing to finishing is the key to digital printing, as is flexibility and process improvement. A modern printing operation uses the World Wide Web as a common business tool for receiving information and for executing related business functions. Innovations in raster image processor (RIP), server, and workflow technology are making it easier to integrate digital devices into almost any printer's array of output services.

Within the range of applications that they were designed for, digital presses compare with offset lithography in terms of print quality. While digital systems complement conventional offset lithographic equipment, they do not displace offset as digital presses cannot yet deposit ink or coatings, print complex items such as magazines and die-cut packaging, or turn litho-sized papers into products at thousands of sheets per hour or thousands of feet per minute.

For the most part, digital presses use toners (dry and wet) in place of traditional printing inks. The exceptions are those lithographic presses that use digital technology for imaging plates and then print in the traditional offset manner.

Digital printing press systems have many commonalties. They are all on-demand printing systems, which address the 50- to 20,000-impression market. In addition, they all provide perfecting (two-sided printing) capabilities, require short or no makeready time, and require no warehousing of finished products because all are produced one at a time and shipped to customers upon completion. Systems using dry toners face some print quality deficiencies related to gloss. And those using wet inks provide images approaching offset lithography in quality. Most of the systems provide variable imaging capabilities.

Color copiers. An offshoot of digital on-demand printing presses is digital high-speed color copiers, which are starting to resemble low-end variable-imaging digital printing presses. This technology addresses the need for up to 50 copies but can easily be costeffective for much larger quantities.

Wide-format digital printers. Wide-format digital printers provide individuals and companies with access to inexpensive large-sized prints in quantities of one to approximately 100. A growing number systems are available that can produce full-color digital prints in sizes ranging from 36 in. (91 cm) to billboardsize sheets. There are two components to these systems: a large format printer and a raster image processor.

Large-format systems print on a variety of substrates, including paper and Mylar, using inkjet, electrostatic, or thermal wax transfer technology. This market is growing rapidly and is finding its niche in quick printing and on-demand printing. Users of this technology include advertising agencies, screenprinters, service bureaus, and in-plant printing departments of many businesses. The applications are numerous and include items such as art-on-demand, backlit signage, transportation advertising displays, engineering drawings, maps, murals, posters, window graphics, and billboards.

Dye sublimination. Dye-sublimation printers use solid dyes and heat to produce photo-quality images. Cyan, magenta, yellow, and black dyes are coated in page-sized panels on a roll of transparent film. Heating elements in the print head vaporize the dyes onto a specially coated paper. The intensity of the color is controlled by precise variations in temperature.

Dye-sublimation systems print color in continuous tones one color at a time, instead of dots of ink. The color is absorbed into the paper rather than sitting on the surface. Therefore, the output is more photorealistic, more durable, and less vulnerable to fading than other ink technologies. Dye-sublimation printers are favored by some graphic designers for highend proofing and for applications that require the best possible color.

Inkjet. Typically, an inkjet printer forms images by spraying tiny droplets of liquid ink onto paper. The small size and precision placement of the ink dots produce near-photo-quality images.

Two primary inkjet technologies are thermal bubble, or bubble jet, and piezoelectric. The former uses heat and the latter uses a crystal and an electric charge to apply the ink. Inkjet printers are inexpensive and produce good color but can be slow, especially for short-run color printing. The best results are normally achieved when printing to specially coated inkjet papers. Inkjet printing is popular for variable-data printing, including addressing, coding, computer letters, sweepstakes forms, personalized direct mail advertising, and imprinting of forms and invoices. All these applications involve variable information, which is the basic advantage of digital printing. An important application of inkjet printing is color proofing.

Inkjet printing is also classified as continuous drop (a stream of ink) and drop-on-demand (ink is squirted only when needed). In inkjet printing, the ink is ejected through a nozzle under pressure and transformed into uniform drops by vibration of a piezoelectric crystal; the drops are charged electrostatically and deflected by electrostatic charges.

There are four basic types of inkjet printing systems. One uses a computer-activated single nozzle that oscillates across a substrate much like an electron beam scans a television tube to produce a picture. The second method uses a bank of nozzles, each of which is digitally controlled by a computer program. The third is bubble-jet, which uses heat to produce bubbles that form the ink droplets. The fourth uses phase-change technology, in which color sticks that are solid at room temperature undergo two phase changes during printing: one when the stick melts to form liquid drops, and the other when the ink droplets hit the paper and solidify.

Modifications of inkjet systems are used for shortrun, large-format poster or billboard printing, where runs are very short (1–100 impressions) and speed is not an urgent requirement. The large-format market also includes posters, outdoor advertising, transportation advertising, and in-store displays in sizes up to 20×60 ft (6×18 m) in about 5 h. The substrate is a sheet of vinyl that is wrapped around a huge cylinder 17.5 ft (5.25 m) in diameter (58-ft or 17-m circumference) for imaging. The images are painted with 16 paint jets, called dot shooters, which discharge as many as 31 million dots to create the final image. Single-sheet billboards require derricks for installation, and most are backlit for night viewing. *See* INKJET PRINTING.

Laser. Static electricity is the principle behind laser printers. A revolving drum or cylinder builds up an electrical charge, and a tiny laser beam pointed at the drum discharges the surface in the pattern of the letters and images to be printed, creating a surface with positive and negative areas. The surface is then coated with toner—a fine powder that is positively charged, so it clings only to the negatively charged areas—which is then transferred onto the paper. The paper then passes through heated rollers, which fuse it to the paper. Color lasers make multiple passes, corresponding to the toner colors.

Speed and lower operation costs make laser printers attractive. Laser printers have their own command language that translates the digital information in a file into the bitmap images printed on the page. Adobe PostScript capabilities of many laser printers make them popular with graphic designers and desktop publishers who often use EPS images and PostScript files.

Electrostatic and electrophotographic. Electrostatic printing is similar to xerographic or electrophotographic printing. It uses toner particles, and the image is formed from an analog original. There is no print drum required. Toner particles are attracted directly to the paper, and no optical system is required. The entire copier glass is exposed at once, and the electrostatic charge is directly deposited on the paper. Liquid or dry toners can be used. Liquid toners are fused to paper by hot air. Dry toners are fused to paper by pressure.

Electrophotographic printing is similar to highspeed copier technology. It operates on the principle of xerography. However, electrophotographic printers operate from digital data and use lasers. Images are created by a moving laser beam across an imaging drum. The image is developed with dry or liquid toners and works well for single-color printing. Highquality color printing is possible, but slow relative to other printing processes.

Electrophotographic systems have drums coated with photocondcutors that are charged overall with a corona charge. The coating is exposed to an image by light that discharges the charge in the nonimage areas. The image is formed using powder or liquid toners in the charged areas, and then transferred to a substrate and fused, or fixed, by heat, solvent vapor, or some other method (**Fig. 24**).

One example is the two-unit web-fed electrophotographic press, which uses cylinders coated with organic photoconductors and liquid toners to produce the image and has many special in-line finishing features and provisions for printing one color on both sides of the paper or two colors on one side. The press prints variable information at speeds up to 300 ft/min (90 m/min), and is capable of changing up to 25% of the information from impression to impression.

Two types of full-color electrophotographic digital presses are used. One uses laser imaging, a special liquid toner, and essentially a single-color offset press with a plate, special heated blanket, and a cold impression cylinder; paper feed and pile delivery system; a means for turning prints over and printing the reverse side (perfecting); and some finishing options. Printing has 800-dpi resolution. Single-color prints can be produced at A3 size (11.5 \times 16.5 in.; 297-420 mm) at a press speed of 4000 sheets/h. Since color printing is done by consecutive impressions of the same sheet on the impression cylinder, the output for four-color (one-side prints) is 1000 sheets/h. Duplex printing (two sides) reduces the output to 500 sheets/h, 8 sheets/min, or 32 pages/min (two sides). The system is capable of printing six colors.

The other electrophotographic digital press is a web press with eight in-line units (four on each side of the paper), so it can print both sides of the paper in one pass. Exposure of the photoconductor plates on each cylinder of the printing units is done with light-emitting diodes assembled into arrays of

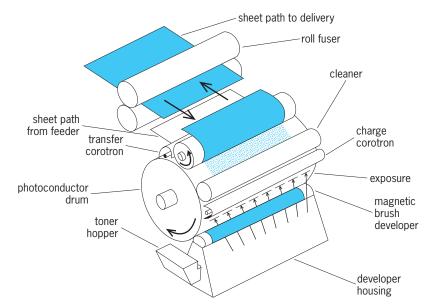


Fig. 24. Typical digital press print unit.

7400 diodes, spaced at a density of 600/in. (236/cm), corresponding to a spatial resolution of 600 dpi. The system has variable spot intensity, with each spot having 64 gray levels, by applying different amounts of a special dry, dual-component toner made up of small particles. The image is composed of single-line screens with continuously variable widths and placed at different angles, according to the amount of color needed. *See* LIGHT-EMITTING DIODE.

There are two digital electrophotographic printing systems for larger-print, poster-type images. One consists of a computer front-end digital printing system that incorporates a wide array of digital color scanners and a 44-in.-wide (340-cm) and a 54-in.wide (137-cm) electrostatic color plotter modified by a special toner control unit. Imaging is done with a special frequency-modulated (FM) screening algorithm. The system is used for indoor and outdoor posters and paneled billboards. The toners are specially treated to resist fading from exposure to ultraviolet radiation. The other system is a composite electrophotographic system, consisting of existing components with extensive software and hardware modifications. It uses specially treated toners and substrates, as well as absorbing overlaminants to resist ultraviolet radiation. Prints 34 in. (87 cm) wide and up to 26 ft (8 m) long can be produced. Larger images can be panelized. Prints are used for fleet and transit advertising, industry expositions and exhibits, visual merchandising and displays, murals, outdoor exhibits and events, banners, and prototyping. See ORGANIC CONDUCTOR; PHOTO-COPYING PROCESSES.

Thermal. Falling somewhere between dyesublimation and solid ink technologies are thermal wax printers. Thermal wax printers produce vibrant color but require very smooth or specially coated paper or transparencies for best output.

There are two thermal printing processes: direct thermal and wax transfer. In direct thermal printing,

a specially coated paper, when heated by a print head, reacts by turning color. In the wax transfer process the same chemical principle is used; however, a special coated ribbon is used which releases pigmented material. Characters are printed one at a time by a print head that moves back and forth.

Printing Inks

Changes in the technology of printing ink have been driven by environmental regulations, workplace safety, and increasing press speeds. In the United States, federal and state regulations have dominated the technology of all inks since the 1980s. Regulations have spread to include ink formulation, labeling, shipping, solvent type and content, emissions of volatile organic compounds, heavy-metal content of inks, and solid- and liquid-waste effluents from printers and ink manufacturers. All of these factors affect ink technology and development. *See* HAZ-ARDOUS WASTE.

Paste inks. In lithographic printing, a dominant consideration has been the increasing speed of running web and sheet-fed presses. Web offset presses are capable of printing at speeds of 3000 ft/min (900 m/min). For heat-set paste inks, the drying rate is an important consideration, since the print must dry rapidly at high speeds.

Sheet-fed inks. Sheet-fed lithographic presses have also increased in printing speed and run as fast as 18,000 impressions/h. The ink formulations need to be adjusted so that the drying (setting) times are fast enough to avoid the transfer of ink from a printed sheet to the back of the next sheet (set-off) in the stack of output sheets. Most of these formulations have extremely quick setting characteristics and become tack-free within a few seconds after printing.

Ultraviolet curing. Ultraviolet curing is favored, particularly for clear overprint varnishes and in applications that require instant drying and processing of the dried prints. Typical applications include packaging with in-line die cutting, packages for cosmetics and liquors, and compact-disk packaging. The growth of ultraviolet overprint varnishes has been mostly in the area of book and magazine covers printed with either ultraviolet-set or conventional heat-set or sheet-fed inks. The driving forces in these applications are protection, glossy appearance, and enhanced color brightness.

Electron-beam curing. High-energy, linear cathode electron-beam generators are used to cure inks and coatings in Europe and the United States, particularly for the manufacture of aseptic paper/foil packaging for beverages such as citrus drinks, milk, and wine. Food products in such packages have a shelf life of several months without refrigeration. Typically, these inks are printed in a web press and overcoated with a clear electron-beam-curable varnish.

Packaging inks. Air-pollution regulations have shifted the emphasis toward water-based liquid inks. While the use of these inks and varnishes in flexog-raphy or gravure packaging is not new, most applica-

tions have been for paper or board printing. *See* AIR POLLUTION.

Compared to flexography, packaging gravure is used for a wider variety of substrates, resulting in more adhesion problems. As a result, binders are used that require solvents that are difficult to recover and separate in carbon-bed absorption units. Thus, many packaging gravure printers use solvent incineration or recovery instead of water-based inks. Most of the water-based inks in this market are limited to paper and board substrates.

More than 50% of the flexographic packaging inks are water-based. Some plastic film applications (especially bread bags) have resisted the use of waterbased flexography inks because of the nature of the additives in the films, which impede adhesion and wetting. In addition, water-based inks are slowerdrying, limiting the printing speeds.

Another packaging area involves radiation curing and a high-speed web offset, wide-web paperboard press. This press uses offset lithography paste inks formulated to be cured with ultraviolet radiation or electron-beam energy to produce finished cartons at speeds of up to 1400 ft/min (420 m/min).

Publication gravure. The major type of ink used throughout the world for this high-speed printing process (up to 3000 ft/min or 900 m/min) is a toluene-based liquid ink using a metallized rosin binder. The toluene vapors emitted during the drying of these inks are captured by carbon absorption beds. The beds are regenerated, and the recovered toluene is reused in ink manufacture. While this system is usually better than 95% effective, there are problems in corrosion and decreased recovery efficiency. In addition, health and safety regulations restrict the amount of toluene vapor allowable in the pressroom atmosphere (around 50 parts per million).

Screen-printing inks. There are four categories: aqueous, evaporative/oxidizing, catalytic, and ultraviolet. In aqueous inks, heat is required to evaporate the primary vehicle, water, leaving the pigments and binders behind to form the dry printed image. Evaporative/oxidizing inks contain solvents or oil vehicles. As with aqueous inks, the vehicle heating is used be evaporate the vehicle or oxidize the oil. Catalytic systems incorporate reactive resins in a chemical process by which the ink is transformed from a liquid to a solid. Ultraviolet-curable inks contain no water, oil, or solvents, and change from a liquid to a solid when exposed to specific wavelengths of ultraviolet light This type of ink can be applied to a wide range of plastics, metals, and paper. See DRYING OIL.

Screen-printing inks are available in hundreds of formulations to address different applications. Many specialized screen-printing inks are used to achieve specific effects and functions. Included are thermochromatic, photochromic, fragrant, fluorescent, scratch-off, conductive, ceramic, glitter, and metallic inks.

While other printing processes are primarily restricted to direct and indirect printing on paper, plastic, and metal sheets, screen-printing is capable of imaging virtually any planar or threedimensional substrate. Another distinctive characteristic of screen-printing is its ability to lay down a heavier ink deposit than other processes, resulting in bright vibrant color reproduction and exceptional lightfastness for outdoor applications. *See* INK.

Printing Papers

Printing papers are designed primarily for printing one or more colors on each of the two sides. Secondary requirements are that the paper meets the needs of operations subsequent to printing, including folding and binding, and the needs involved in end use. To fulfill the diversified demands of modern printing, printing papers are manufactured in many different finishes (coated and uncoated), degrees of whiteness, colors, basis weights, grades, and types.

Various fibrous and nonfibrous materials, in combination with different degrees of fiber refining and adjustment in paper machine settings, are used to obtain the balance of paper properties required for the desired print quality, press performance, and demands of a specific printing process. Generally, the properties are maximized in accordance with other secondary properties for folding, binding, strength, and cost, and include dimensional stability to maintain printing register, opacity to minimize show-through of the printed image, the degree of smoothness and compressibility demanded by the printing process, and a surface optimized for printability.

Often, a balance must be struck in the fiber and finishing conditions to optimize strength and stiffness to enable the press to feed the sheet mechanically without difficulty and surface properties for printability. Paper coated with minerals such as kaolin and calcium carbonate, and binders such as styrene-butadiene latex and starch can provide a superior printing surface without compromising sheet strength.

Letterpress. For letterpress printing, microsmoothness (without pits or voids) of the paper surface closely relates to the completeness of image transfer from the printing plate to the paper, and to the extent to which halftone dots are a true reproduction of those on the plate. A level paper surface uniformly absorbent to printing ink is required for consistent density, color, or gloss of the printed image. Compressibility of the paper must be adequate to assist its surface in contacting all areas of a relief printing plate.

Flexography. Paper requirements for flexographic printing are less demanding than for letterpress or planographic printing. Flexography uses comparatively low-tack inks that dry by low-temperature evaporation or absorption into the paper. Consequently, the stresses applied to the paper's surface are less.

Modern photopolymer plates combined with special press designs made flexographic a printing technology capable of process color at the quality level of low-end offset lithography. In order to achieve this quality, flexographic paper must have a smoother and flatter surface than for offset. Flexography is largely water-based, and thus the paper and board used must resist surface deterioration or dimensional change when wetted.

Offset lithography. Offset lithography places more stringent demands on paper than other printing processes.

Sheet-fed. Paper fibers must be well bonded, with sufficient water resistance to prevent them from softening and picking (rupturing at the surface) under the stresses applied to paper as it is repeatedly separated from the inked rubber blanket during multicolor printing. Excessive moisture pickup from repeated contact with water of the lithographic dampening system could lead to intolerable fiber expansion, changes in paper dimensions, and printing misregister. Paper surfaces must be exceptionally clean and have no loosely bonded material that can be lifted off and transferred to the offset rubber blanket, thereby introducing objectionable defects in the prints. Additionally, the paper must not release materials such as fillers or coatings that accumulate on the offset blanket or react with the fountain solution to change its chemistry.

Flatness of the paper sheet with straight undistorted sheet edges is a critical requirement. In addition, it is generally necessary that the paper's moisture content be in balance with the relative humidity of the pressroom to maintain satisfactory register for multicolor printing.

Web. Requirements are basically the same as those for sheet-fed offset lithography, except the moisture requirements of the printing paper is generally less than for sheet-fed offset printing and need not be in balance with pressroom relative humidity. Heat-set web offset papers usually have porous coatings for rapid escape of water vapor in order to prevent blistering during ink drying. A coated web offset paper must be sufficiently strong to fold without cracking after loss of moisture.

To maintain satisfactory printing register and runnability, rolls must unwind with even tension, remain flat across the web, and have minimum paper stretch during printing.

Gravure. Gravure printing requires a flat paper surface that has micro-smoothness for uninterrupted ink transfer from the inked cells to the paper. The compressibility of the paper aids in contacting the surfaces of the small gravure cells. High paper-surface strength is not essential. Paper and board for packaging are often gravure-printed with water-based inks, which must have appropriate characteristics for wetting and ink spread. Dimensional stability is important for color registration, as the inks are dried with hot air between color stations. Paper rolls must feed into the press with even tension and flatness across their wide webs for satisfactory runnability.

Uncoated papers. Uncoated printing papers are manufactured from bleached chemical wood pulps, a mixture of chemical and mechanical wood pulps,

a combination of chemical wood pulps, recycled fibers, or cotton or linen fibers depending upon type, grade, and availability. The hardwood and softwood species that produce short- and long-fiber chemical wood pulps, respectively, are commonly used in various proportions in the papermaking furnish (the mixture of fibrous and nonfibrous materials) to achieve the desired balance of properties for a printing paper. Other components of the papermaking furnish that determine the characteristics of an uncoated printing paper are fillers, sizing, dyes, and starch.

Alkaline papers. The uncoated fine-paper industry largely converted from acid to alkaline papermaking to take advantage of high brightness filler such as precipitated calcium carbonate, and to provide greater permanence. Alkalinity is provided by substituting a synthetic material for the rosin-alum size used in acid papers. Even wood-containing papers are being produced under neutral conditions to avoid strength loss from acid hydrolysis. However, permanence is not considered crucial for daily newspapers, so newsprint remains an acid paper grade.

Newsprint. Newsprint is used chiefly for the printing of newspaper by web offset, some flexography, and letterpress. Its minor uses are advertising supplements, government publications, comic books, and commercial printing. In the United States, the basis weight of newsprint is defined as the weight in pounds per 3000 ft² (ream) of paper. In almost all other countries, including Canada, the basis weight is defined in grams per square meter of paper. Newsprint ranges in basis weight from 25-32 lb per ream (40-52 g/m²). Newsprint can be manufactured by using one or a mixture of the following: thermomechanical pulp, stone-ground wood, chemithermomechanical pulp, recycled newspapers and magazines, and reinforcing chemical pulp. Mechanical fibers that originate from ground wood or thermomechanical pulps have lower brightness and yellow more readily with aging than chemical wood fibers, but they are used for their lower cost and high opacitv.

Mechanical pulp-containing papers. These printing papers are made with varying percentages of mechanical and chemical wood pulps, depending upon the grade. They resemble newsprint, but are somewhat brighter and have a smoother finish produced by supercalendering or hot soft calendering. Their basis weights normally range 20-45 lb per ream (30-68 g/m²). Typical uses are for the printing of periodicals, catalogs, directories, advertising circulars, newspaper supplements and magazine sections, and paperbacks.

Wood-free papers. This group includes offset, book, bond, envelope, and text papers.

Offset lithography provides the versatility of printing papers with all types of finishes. Finishes include antique, eggshell, vellum, wove, smooth, machine, and English. Offset papers may have distinctive patterned finishes obtained by embossing. Their brightness range extends from approximately 75 to 90 or more, depending upon their quality and grade. Offset papers can be manufactured with extra opacity and in various colors to fulfill the varying needs of the printing industry. Standard basis weights range 40–120 lb per ream (59–178 g/m²).

Uncoated printing papers for book manufacturing are offset papers that are made to bulk a specified number of pages per inch and to assure that the total thickness of the finished book conforms to the predesigned thickness of its cover or casing. Other requirements for book manufacturing are different white shades and finishes that relate to the esthetic qualities of a book and other needs peculiar to book binding.

Text papers are high-quality printing papers made solely from chemical pulp, in many different finishes or textured surfaces, and in white and colors. Some text papers are watermarked and have a deckled edge. Common basis weights are 60, 70, and 80 lb per ream (89, 104, and 118 g/m²) plus matching cover weights. Text papers are printed mostly by sheet-fed offset lithography.

Lightweight printing papers are characterized by their extraordinarily high opacity for their basis weights, which range approximately 18–40 lb per ream (25–60 g/m²). Their fiber composition may consist of chemical pulp, recycled cotton, linen fiber, or some combination of these. Lightweight papers are printed mostly by offset lithography. They are used for the bulk reproduction of bibles, encyclopedias, reference books, directories, and manuals; and for lowering first-class mailing costs of printed matter such as financial reports, prospectuses, and proxies.

Coated papers. Mineral coatings such as kaolin, calcium carbonate, or other pigments, and coating binders such as starch or various synthetic binders such as styrene butadiene or acrylic latex are applied to the base paper predominantly by blade coating as well as by roll and air-knife coating either on or off the paper machine. Coating formulations and base papers differ, depending upon the printing process, the quality level, and end-use considerations.

Grade levels of coated papers from the highest to the lowest quality are sometimes designated in six grades: premium and numbers 1, 2, 3, 4, and 5. The premium and number 1, 2, and 3 grades have chemical-pulp base stocks. Grade 4 may or may not contain mechanical fiber. The number 5 grade is made with about 50% mechanical fiber. The higher grades of coated papers are characterized by higher gloss, whiteness, and superior print quality.

Premium and grades 1, 2, 3, and 4 are used for sheet-fed offset lithography. Grades 3, 4, and 5 are used mostly for web offset and gravure printing. Basis weights of coated paper for sheet-fed lithography extend 60–100 lb per ream (90–150 g/m²) with matching cover weights.

The number 4 and 5 coated grades made with mechanical fiber are used primarily for publication and periodical printing by rotogravure, or web offset, in a basis weight range of 28–50 lb per ream (40–75 g/m²). Catalog and commercial printing

represents the remaining usage of these lightweight coated papers.

Dull-finish enamel. Dull-finish enamels have a lower gloss than their high-gloss counterparts and are used where the printing quality of an enamel surface with low paper gloss is desired.

Cast-coated paper. Mirrorlike gloss-coated papers having an exceptionally smooth level surface and uniform ink absorptivity are made by a unique method of casting the coating on its base paper against a heated polished metal cylinder. Binders such as casein are used to provide the right adhesion, release, and porosity.

Matte-coated paper. Matte-coated papers provide better lithographic reproduction than uncoated papers while still providing a gloss-free background for printing. Ground calcium carbonate is a common coating pigment.

Coated-one-side papers. Coated-one-side paper is used for the printing of flexible packaging, labels, and book jackets. This type of paper is designed for offset lithography, flexography, or rotogravure and for other postprinting requirements such as lacquering, bronzing, embossing, die cutting, and labeling.

Ink-jet printing. This requires a paper surface that has the capacity to dry the impinging drops of ink by absorption without wicking or spreading of the ink for maximum printing density and sharpness. Uncoated papers for single-color ink jet are generally well sized, and papers for high-quality color ink jet are generally coated with a transparent layer such as silica or gelatin.

Recycled papers. Since paper represents a major part of municipal solid waste, much effort and promotion have been devoted to recycling the paper waste and using recycled papers for printing. In the United States, recycled newsprint has been used for years, but the use of recycled papers for business and commercial printing started in the 1980s. Three cost constraints affect the use of recycled fibers: sorting the waste paper, deinking it, and transporting it to mill and deinking sites far from populated areas. *See* PAPER.

Synthetic paper. Synthetic paper was developed in the 1970s when paper resources were scarce and petroleum was inexpensive. It was also a time when the capacity to produce paper in the United States had decreased because of papermaking machine shutdowns by the Occupational Safety and Health Administration (OSHA) and the EPA, resulting from the discovery of unsafe working conditions and air and water pollution. It was also thought that diminishing wood resources could drive up costs of paper substrates.

Synthetic papers are substrates made of materials other than the natural fibers of trees and the resulting pulp. Most synthetic paper is produced from polypropylene or polystyrene, which may contain minerals and other additives. In addition, any film produced and/or treated to improve its printability, as well as any process that improves the printability of an opaque film substrate, is considered synthetic paper. Synthetic paper has the look, feel, and processing capabilities of traditional printing paper but is stronger and more durable. Synthetic papers have a higher tensile strength and support higher processing speeds, as well as improved moisture, grease, and chemical resistance.

Ink selection for synthetic paper is important. Most synthetic papers perform best with thin ink films and low mineral oil content, because absorption can be minimal even with auxiliary drying. Oxidizing inks are often recommended for synthetic papers because they dry through evaporation and do not totally rely on absorption. The absorptive nature of the surface of the substrate will determine ink/water balance and ink film thickness when printing on synthetic paper. Clay coatings allow ink absorption, which aids in printing.

Thermal transfer, direct thermal, inkjet, and in some cases laser printing are markets for synthetic paper. Digital printing for short runs is an application for synthetic and tear-resistant papers. Inkjet printing is the most popular, while laser applications are growing.

The dry toner fusion process is popular, but excessive heat causes dimensional instability such as shrinkage, cracking, curling, and sometimes delaminating. Newer copy machines and laser printers have addressed these problems with adjustabletemperature fusion rollers and tear-resistant papers.

In terms of the environmental friendliness, synthetic papers' raw materials are nonrenewable (petrochemicals and minerals), but they can be recycled or burned as an energy source. Since production of synthetic paper is costly, manufacturers waste little in processing.

Synthetic paper is experiencing a high growth in many printing segments. In the past, applications for synthetic paper were for outdoor posters and billboards, maps, and manuals for special purposes. Synthetic papers now are being used for products they are replacing, such as cotton, metal, and foil for tags and labels; laminates of all types; artists' materials; and paper and board. Tag and label applications are growing rapidly, with clothiers and industrial end-users accepting synthetic papers' reliability in wear and print performance. Products often include thickness from 0.003 to 0.01 in. (0.08 to 0.25 mm), or basis weights from 40 to 200 lb (18 to 90 kg). Many products are diversifying into cream shades from high-brightness stocks, and most contain antistatic agents to avoid double feeding

Another large growth area for synthetic paper is lamination, with identity cards, menus, and tickets being target markets. However, the largest applications are still outdoor banners, maps, and manuals. Traditional paper products compete heavily in these markets because of the large volumes available, somewhat short lifecycles, and lower costs. Synthetic paper improvements in print performance and converting (such as die cutting, embossing, and folding) offer graphic arts producers a high-end alternative and a high profit potential.

Finishing Processes

Finishing processes include folding, stitching, book casing, gluing, perfect binding, scoring, perforating, creasing, sheeting, foil stamping, collating, trimming, die cutting, embossing, laminating, punching, numbering, shrink-wrapping, bundling and banding, in-line finishing, demographic finishing, segmentation, and fulfillment, and mailing and distribution. Finishing processes continue to be enhanced by automation and computerization.

Demographic finishing is a process that personalizes a publication or other print medium for individuals or geographic regions. It typically involves inline inkjet printing from a subscription database and bar codes to address publications or direct-mail advertising, eliminating the need for mailing labels. This technology is widespread in the periodical and catalog industry.

Inline finishing involves binding and finishing operations that are integrated with printing presses and take place as continuous operations, including trimming, slitting, die cutting, embossing, gathering, folding, stitching, numbering, punching, and addressing.

There are significant amounts of automation from production, to bundling and wrapping, to delivering jobs, to loading docks. This involves using bar coding and computer-controlled finishing systems. There is automation for setting up finishing equipment and for removing jobs from press and moving them to the bindery.

The ability to feed precollated sets from a digital press or copier and add a preprinted cover to the set is now possible. The unit then finishes the product by stitching, folding, and trimming. Customer databases allow the operator to retrieve, update, delete, and maintain job requirements. *See* DATABASE MANAGEMENT SYSTEM.

There are robotic palletizers, electronic overhead conveying systems, and automatic, programmable stitching/trimming machines. Some presses have inline UV coating units to enable users to print, UVcoat, slit, and rewind or cut digitally printed pieces in a single pass at full press speed.

Digital presses are driving finish-on-demand developments as markets for digital printing and on-demand finishing accelerates. There is also ondemand print-to-mail. Print-to-mail uses stored or real-time data to create personalized communications with recipients. The Internet is driving this. The most sophisticated use is for data streaming from Web sites to drive personalized communications. Servers residing in the mailing facility capture input information from the site. This drives digital printers, which create customized letters and marketing pieces. They are loaded into an inserting machine, which uses a digital camera-based vision system to verify that the personalized printed pieces match each other. The camera system can also use verified bar code information to instruct an inline inkjet printer to address the outer envelope with the matching address and additional personalization. See INTERNET; WORLD WIDE WEB.

Bar code readers enable systems to recall preprogrammed jobs and provide fully automatic slitting, cutting, and creasing. A creasing unit ensures that color-printed documents do not crack when folded.

The major digital print vendors have software codes for describing the attributes of a print job, from file handling and correction, through printing, collating, and finishing. Nearly all vendors of the on-demand printing and finishing equipment are implementing CIP4, the open interface standard for all digital-printing hardware components. Harvey R. Levenson

Bibliography. Bridg's CTP Handbook for the Graphic Arts, 2000; M. H. Bruno, Pocket Pal, 19th ed., 2004; M. H. Bruno, Principles of Color Proofing, 1986; G. G. Field, Color and Its Reproduction, 3d ed., 2004; Foundation of Flexographic Technical Foundation, Flexography: Principles and Practices, 6 vols., 5th ed., 2000; Gravure Education Foundation and Gravure Association of America, Gravure: Process and Technology, 2d ed., 2003; IDEAlliance SWOP, SWOP: Specification for Web Offset Publications, 10th ed., 2005; H. R. Levenson, Complete Dictionary of Graphic Arts and Desktop Publishing Terminology, 1995; H. R. Levenson, Understanding Graphic Communication: Selected Readings, 2000; F. J. Romano, Digital Basics 3.0, 2001; F. J. Romano, Digital Printing: Mastering On-Demand and Variable Data Printing for Profit, 2000; F. J. Romano and R. M. Romano, The GATF Encyclopedia of Graphic Communications, 1998; R. Sasso, Publishing Timeline 2000, 2000; L. Wilson, What the Printer Should Know about Paper, 3d ed., 1998.

Prion disease

Transmissible spongiform encephalopathies in both humans and animals. Scrapie is the most common form seen in animals, while in humans the most prevalent form is Creutzfeldt-Jakob disease. This group of disorders is characterized at a neuropathological level by vacuolation of the brain's gray matter (spongiform change). They were initially considered to be examples of slow virus infections. Experimental work has consistently failed to demonstrate detectable nucleic acids-both ribonucleic acid (RNA) and deoxyribonucleic acid (DNA)-as constituting part of the infectious agent. Contemporary understanding suggests that the infectious particles are composed predominantly, or perhaps even solely, of protein, and from this concept was derived the acronym prion (proteinaceous infectious particles). Also of interest is the apparent paradox of how these disorders can be simultaneously infectious and yet inherited in an autosomal dominant fashion (from a gene on a chromosome other than a sex chromosome).

Disorders. Scrapie, which occurs naturally in sheep and goats, was the first of the spongiform encephalopathies to be described. Over the years, an increasing range of animal species have been recognized as occasional natural hosts of this type of disease. Examples of these disorders include

transmissible mink encephalopathy, chronic wasting disease of mule deer and elk, and an epidemic form in cattle in the United Kingdom called bovine spongiform encephalopathy. The last was believed to have been introduced by the accidental contamination of the food chain by meat and bone meal protein supplement derived from sheep presumably harboring scrapie. As a reaction, various measures, including a feed ban, were put in place, with recent surveys indicating the bovine spongiform encephalopathy epidemic is on the decline. The British surveillance unit and its detailed case-control study set up to monitor the incidence of Creutzfeldt-Jakob disease has shown no increase in the number of cases of this disorder, and no geographical clustering. However, the recent occurrence of Creutzfeldt-Jakob disease in persons under the age of 42 has raised the possibility that the bovine disease may have spread to humans. At this stage, the overall risk of transmission appears small. So far, animal models have indicated that only central nervous system tissue has been shown to transmit the disease after oral ingestion-a diverse range of other organs, including udder, skeletal muscle, lymph nodes, liver, and buffy coat of blood (that is, white blood cells) proving noninfectious.

The currently recognized spectrum of human disorders encompasses kuru, Creutzfeldt-Jakob disease, Gerstmann-Sträussler-Scheinker disease, and fatal familial insomnia. All, including familial cases, have been shown to be transmissible to animals and hence potentially infectious; all are invariably fatal with no effective treatments currently available.

Kuru has been almost eradicated since the cessation of ritualistic cannibalism, but occurred endemically among the Fore and some neighboring people of the Eastern Highlands of Papua New Guinea. The disorder was believed to be transmitted by the consumption of infected central nervous system tissues. In its typical form, the symptoms of kuru included a relentlessly progressive loss of muscular coordination (ataxia) associated with an unusual tremor; the duration of the entire illness spanned 1–2 years.

Creutzfeldt-Jakob disease is a rapidly progressive deterioration of intellectual abilities (dementia), which proceeds to death with a median duration of illness of around 4 months. As the disease evolves, it is typically accompanied by a range of additional clinical features which include ataxia, muscular rigidity, and spontaneous irregular limb jerks (myoclonus). For 90% of affected individuals Creutzfeldt-Jakob disease develops sporadically, without any clues as to causation. Approximately 10% of individuals have a genetic basis and occur within families. The final important epidemiological subgroup of Creutzfeldt-Jakob disease comprises those cases arising after accidental person-to-person transmission (iatrogenic).

Gerstmann-Sträussler-Scheinker disease is relatively uncommon, and presents as a chronic progressive familial ataxia with somewhat distinctive neuropathological features. Fatal familial insomnia is the rarest of the disorders, and is characterized by rapidly progressive and untreatable insomnia, ataxia, and myoclonus, with pathological examination revealing severe and selective degeneration of a restricted part of the brain (the thalamus). See BRAIN.

Molecular biology. Although not completely resolved, the prevailing view is that the transmissible agent of these disorders is an abnormal conformational form of the constitutively expressed prion protein molecule. The prion protein molecule is normally produced by a wide variety of cells throughout the body, but especially the central nervous system. Although initial investigations minimized its probable biological importance, this host-encoded glycoprotein probably plays a role in receptor-mediated signal transduction, and disordered function may be capable of explaining some of the clinical features observed in Creutzfeldt-Jakob disease.

It has been suggested that in sporadic cases of Creutzfeldt-Jakob disease, the change in the prion protein molecule occurs spontaneously as a random or chance event. However, once in this abnormal shape, the prion protein molecule becomes relatively resistant to enzymes usually capable of degrading proteins (proteases), and also serves as a template or nucleating factor for the conversion of other prion protein molecules, through self-replicating polymerization. In some of the prion disorders, especially Gerstmann-Sträussler-Scheinker disease, the accumulation of the abnormal prion protein molecules can be detected as deposits of amyloid, which may also represent a mechanism of neurotoxicity inducing the spongiform change. However, the precise sequence of events subserving the translation of this subtle molecular change into a devastating fatal neurodegenerative disorder is unknown.

The study of the familial prion diseases has led to the recognition of a number of causal mutations within the prion protein gene located on chromosome 20. Each mutation is generally associated with a particular form of clinical illness, but considerable variability in clinical and neuropathological expression can occur within affected members of the same family. These mutations are thought to greatly increase the chance of the prion protein molecule adopting the abnormal configuration, leading to self-replicating polymerization. Similar to sporadic Creutzfeldt-Jakob disease, the consequent accumulation of abnormal prion protein molecules is thought to be linked to the eventual disruption of cellular function. If the abnormal form is introduced into an individual, it serves as the template for subsequent autocatalytic polymerization. Such a model attempts to reconcile the paradox of how this group of disorders can be transmissible and at the same time inherited in an apparent autosomal dominant fashion.

Although no causative mutation has been demonstrated in the prion protein gene to explain sporadic examples of Creutzfeldt-Jakob disease, a genetic component may still play a part in the susceptibility to developing this disorder, with normal variations in the gene possibly altering the likelihood of developing this and pituitary hormone-related Creutzfeldt-Jakob disease.

Human-to-human transmission. A variety of mechanisms of human-to-human transmission have been described, and include contaminated instruments such as electroencephalographic depth electrodes, dura mater homografts, corneal transplants, and cadaverically derived human pituitary growth hormone and gonadotrophins. Transmission is due in part to the ineffectiveness of conventional sterilization and disinfection procedures to control the infectivity of transmissible spongiform encephalopathies. Numerically, pituitary hormone-related Creutzfeldt-Jakob disease is the most important form of humanto-human transmission of disease. Although kuru and Creutzfeldt-Jakob disease are experimentally transmissible to nonhuman primates through the oral route, successful human-to-human transmission appears to require more invasive methods of inoculation of the infectious material, usually in the form of tissue transplants or injectable therapies or through the use of contaminated neurosurgical instruments.

Epidemiological evidence suggests that there is no increased risk of contracting Creutzfeldt-Jakob disease from exposure in the form of close personal contact during domestic and occupational activities. Nonetheless, a very small number of physicians, including pathologists and neurosurgeons, and a slightly larger number of allied health personnel (including histologists) have developed Creutzfeldt-Jakob disease. However, there is no clear evidence to suggest that occupational exposure poses a significant risk of contracting Creutzfeldt-Jakob disease, especially if precautionary and preventive guidelines are followed.

Incubation periods in cases involving human-tohuman transmission appear to vary enormously, depending upon the mechanism of inoculation. In the case of pituitary hormone therapy and kuru, which presumably utilize peripheral methods of inoculation, incubation periods vary considerably from a few to more than 30 years, with infected individuals usually exhibiting an ataxic syndrome. However, when the infectious material is given more direct access to the central nervous system, as with intracerebral depth electrodes, the incubation periods are usually much shorter, generally of the order of 12–24 months, and cognitive decline similar to sporadic Creutzfeldt-Jakob disease is the usual manifestation.

Current evidence suggests that transmission of Creutzfeld-Jakob disease from mother to child does not occur. This finding accords with the much larger epidemiological experience seen in the related disorder kuru. The evidence encompasses children born to mothers during both the presymptomatic and clinical phase of their illness. The explanation for the apparent transplacental and postpartum barriers to transmission remain to be elucidated.

Transmissibility. The transmissibility of human spongiform encephalopathies was first reported in 1966 after kuru had been successfully transmitted to three chimpanzees. Within 2 years, the successful transmission of Creutzfeldt-Jakob disease using the same animal model was announced. Using nonhuman primates (chimpanzees, squirrel and spider monkeys), transmission rates range from 90% for

sporadic Creutzfeldt-Jakob disease, through 95% for kuru, and up to 100% for cases of human-to-human transmission of Creutzfeldt-Jakob disease. Two important factors pertaining to transmissibility are the method of inoculation and the dose of infectious material administered. A high does of infectious material administered by direct intracerebral inoculation is clearly the most effective method of transmissibility and generally provides the shortest incubation time, which in the primate model is approximately 10 months. Peripheral methods of inoculation (for example, subcutaneous, intraperitoneal) are less effective routes for transmissibility which can be partly offset by the administration of higher doses of infectious material.

Smaller nonprimate host animals have been successfully utilized in transmissibility experiments. However, transmissibility rates are generally far lower, despite the use of similar inoculation techniques. Familial forms of disease are less transmissible, averaging around 70%. Within the familial group, the transmission rate also appears dependent on the underlying mutation in the prion gene.

There is some evidence of a species barrier to transmissibility. The prion protein gene appears to modulate both scrapic susceptibility and incubation times, and the more dissimilar the prion protein genes, the greater the species resistance to transmissibility.

Central nervous system tissue has been consistently shown to be the most infectious material in animal transmissibility experiments. However, other tissues have been shown to be infectious, although at lower rates of transmissibility, and include lung, liver, kidney, spleen, and lymph nodes. Although cerebrospinal fluid has been shown to be infectious, vaginal secretions, semen, sputum, saliva, nasal mucus, and tears are not. The testis, prostate, skeletal muscle, peripheral never, adipose tissue, adrenal gland, and thyroid gland have failed to transmit the disease; the status of urine and placental tissue is less clear.

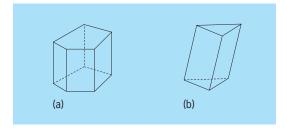
The transmissibility status of blood and blood products is of some concern. White blood cells (buffy coat) and whole blood have been shown to be infectious and transmissible to nonprimate animals when taken from individuals during late stages of Creutzfeldt-Jakob disease. Clearly such findings raise some concern regarding the possible hematogenous transmission of disorders from individuals in the preclinical or clinical phases of their illness, but despite claims to the contrary there is no well-documented and verified example of a Creutzfeldt-Jakob disease being transmitted in this manner in humans.

See MUTATION; NERVOUS SYSTEM DISORDERS; SCRAPIE; VIRUS INFECTION, LATENT, PERSISTENT, SLOW. Colin L. Masters; Steven J. Collins

Bibliography. P. Brown, M. A. Preece and R. G. Will, "Friendly fire" in medicine: Hormones, homografts, and Creutzfeldt-Jakob disease, *Lancet*, 340:24–27, 1992; D. M. Knipe et al. (eds.), *Fields' Virology*, 4th ed., 2001; M. S. Palmer and J. Collinge, Human prion diseases, *Curr. Opin. Neurol. Neurosurg.*, 5:895-901, 1992; S. B. Prusiner, Natural and experimental prion diseases of humans and animals, *Cur. Opin. Neurobiol.*, 2:638-647, 1992.

Prism

A polyhedron of which two faces are congruent polygons in parallel planes, and the other faces are parallelograms (see **illus**.). The bases *B* are the congruent



Prisms configurations. (a) Right. (b) Oblique.

polygons; the lateral faces are the parallelograms; the lateral edges are the edges not lying in the bases; and the perpendicular distance between the bases is the altitude l. Sections parallel to the bases are congruent to the bases. A prism is a right prism if its lateral edges are perpendicular to the bases; an oblique prism otherwise. A prism is called a triangular prism if its bases are pentagons; and a parallelepiped if its bases are parallelograms. The volume of any prism is equal to the area of its base times its altitude (V = Bb). See POLYHEDRON. J. Sutherland Frame

Prismatic astrolabe

A surveying instrument used to make the celestial observations needed in establishing an astronomical position. The instrument (**Fig. 1**) consists of an accurate prism, a small pan of mercury to serve as an

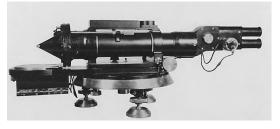


Fig. 1. A prismatic astrolabe, used to make celestial observations. (U.S. Naval Oceanographic Office)

artificial horizon, an observing telescope with two eyepieces of different power, level bubbles and leveling screws, a magnetic compass and azimuth circle, adjusting screws, flashlight-battery power source, light, and a rheostat to control the intensity of illumination.

By using a fixed prism, the instrument measures a fixed altitude, usually 45° . As shown in **Fig. 2**, a light ray *R* from a star enters the upper face of the prism and is refracted so as to pass down the tube of the telescope to the eyepiece. At the same time a parallel ray *R'* from the same star is reflected from the mercury horizon to the lower face of the prism and is also refracted down the telescope tube.

As a rising star increases altitude past that for which the instrument was constructed, the direct image appears to move upward from the bottom of the field of vision to the top. The image reflected by the mercury horizon appears to move downward from top to bottom. Thus, shortly before the rising star reaches the fixed altitude, the rays (*a*) and (*a'*) produce images at the bottom and top, respectively, of the field of view as shown in inset (1). As the star continues to rise, the two images approach each other. At the established altitude the rays (*b*) and (*b'*) both produce images at the center of the field of view, as shown in inset (2). As the star continues to rise, the two images continue their apparent movement, again separating. For a setting body

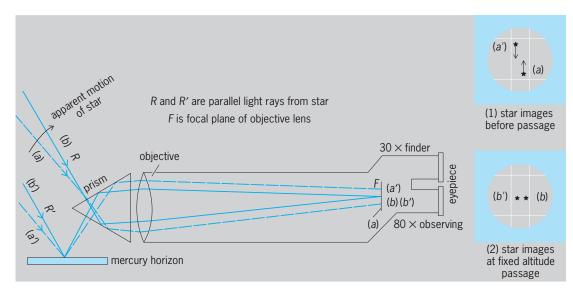


Fig. 2. Optics of a prismatic astrolabe. (U.S. Naval Oceanographic Office)

the apparent motion is in the opposite direction. The prism, when installed, is rotated slightly about the telescope axis so that the two images will be close together on a horizontal line when the fixed altitude is reached, as shown, rather than coincident.

A fixed altitude is used to minimize error that is due to variations from standard atmospheric refraction. Each accurately timed observation provides one line of position. The established position is at the center of the inscribed circle of the figure formed by a number of lines of position. An increase in the number of observations decreases the random error of the position.

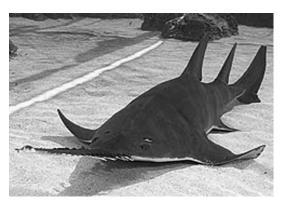
An "astronomical position" determined by a prismatic astrolabe is subject to a small but sometimes serious error due to deflection of the vertical by gravity anomalies in the Earth. Alton B Moody

Pristiformes

An order of batoid fishes comprising one family (Pristidae), two genera, and seven species and commonly known as sawfishes. Sawfishes are superficially similar to sawsharks (order Pristiophoriformes) in having an elongate sharklike body, sharklike locomotion, two distinct dorsal fins, a distinct caudal fin, and a long bladelike rostrum armed on both sides with large teeth but their fundamental structure is that of other batoids. Each rostral tooth is an enlarged dermal denticle homologous to those of the sawsharks and the dorsal spines of other sharks. *See* BATOIDEA; ELASMOBRANCHII; PRISTIOPHORIFORMES.

Sawfishes are found in tropical and subtropical shallow coastal waters and occasionally enter freshwater. The saw is used to grub in the sand and mud for bottom-dwelling invertebrates. Also, the toothed blade is used to slash to and fro in schools of small fish, stunning or killing them. The maximum length of the smallest species is 1.4 m (4.6 ft); the largest is 7.6 m (25 ft), possibly more, placing them among the largest elasmobranchs. Regardless of their size and their possession of a formidable weapon, unprovoked attacks upon bathers are highly unlikely.

All species of the family are listed as critically en-



Sawfish (Pristis microdon). (Photo from Aquarium of the Pacific, Long Beach, California)

dangered on the International Union for the Conservation of Nature and Resources (IUCN) Red List of Threatened Animals. The principal threat to all sawfishes is fisheries, both targeted and bycatch (unintentional capture by fishing gear), because their long tooth-studded saw makes them extraordinarily vulnerable to entanglement in any sort of net gear. *See* ENDANGERED SPECIES. Herbert Boschung

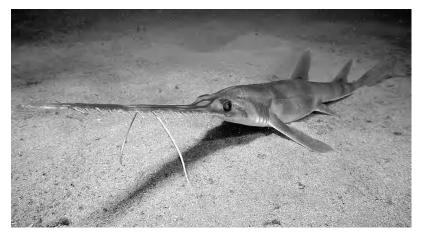
Bibliography. L. J. V. Compagno, Checklist of Chondrichthyes, pp. 503-547, in W. C. Hamlett (ed.), Reproductive Biology and Phylogeny of Chondrichthyes: Sharks, Batoids and Chimaeras, Science Publishers, Enfield, NH, 2005; L. J. V. Compagno, Checklist of living elasmobranches, pp. 471-498, in W. C. Hamlett (ed.), Sharks, Skates, and Rays: The Biology of Elasmobranch Fishes, Johns Hopkins University Press, Baltimore, 1999; L. J. V. Compagno and P. R. Last, Pristidae, Sawfishes, pp. 1410-1417 in K. E. Carpenter and V. Niem (eds.), FAO Identification Guide for Fishery Purposes: The Living Marine Resources of the Western Central Pacific, FAO, Rome, 1999; M. R. de Carvalho and J. D. McEachran, Family Pristidae (sawfishes), pp. 17-21, in R. E. Reis, S. O. Kullander, and C. J. Ferraris, Jr. (eds.), Checklist of the Freshwater Fishes of South and Central America, EDIPUCRS, Porto Alegre, Brasil, 2003; R. Froese and D. Pauly (eds.), FishBase, World Wide Web electronic publication, <www.fishbase.org>, version (05/2005); J. S. Nelson, Fishes of the World, 4th ed., Wiley, New York, 2006.

Pristiophoriformes

An order of sharks comprising one family (Pristiophoridae) and two genera, commonly known as sawsharks. Sawsharks (see **illustration**) are superficially similar to sawfishes, which are rays, but are distinguishable from all other elasmobranchs by the following combination of characters: snout shaped as a long flat blade bearing alternate large and small teeth on each side that are weakly embedded and periodically replaced; one pair of long barbels on the underside of the blade; no visible dorsal fin spines; anal fin absent; large spiracles; and ovoviviparous, with embryos feeding solely on yolk. *See* ELASMOBRANCHII; PRISTIFORMES; SELACHII.

Sawsharks occur on continental and insular shelves and slopes in temperate and tropical waters of the western Atlantic (around the Bahamas, Florida, and Cuba), the southwestern Indian Ocean off South Africa, and the western Pacific from southern Australia to Japan, rarely entering estuaries.

The genus *Pliotrema* (one species) has 6 pairs of gill slits, serrated larger rostral teeth and teeth with prominent transverse ridges on basal ledges, and a maximum length of 170 cm (67 in.). *Pristio-phorus* (four described species) has 5 pairs of gill slits, smooth larger rostral teeth and teeth without prominent transverse ridges on basal ledges, and a maximum length of 80 to 137 cm (32 to 54 in.),



Longnose sawshark (Pristiophorus cirratus). (Photo by Marty Snyderman, www.martysnyderman.com)

depending on the species. Sawsharks cruise the bottom feeding on bony fishes, shrimps, and squids. Herbert Boschung

Bibliography. L. J. V. Compagno, Checklist of Chondrichthyes, pp. 503-547 in W. C. Hamlett (ed.), Reproductive Biology and Phylogeny of Chondrichthyes: Sharks, Batoids and Chimaeras, Science Publishers, Enfield, NH, 2005; L. J. V. Compagno, Checklist of living elasmobranches, pp. 471-498, in W. C. Hamlett (ed.), Sharks, Skates, and Rays: The Biology of Elasmobranch Fishes, Johns Hopkins University Press, Baltimore, 1999; L. J. V. Compagno, FAO species catalogue, in Sharks of the World: An Annotated and Illustrated Catalogue of Shark Species Known to Date-Part 1: Hexanchiformes to Lamniformes, FAO Fish. Synop. (125) Vol. 4, pt. 1, pp. 212-249, 1984; R. Froese and D. Pauly (eds.), FishBase, World Wide Web electronic publication, www.fishbase.org, version (05/2005); J. S. Nelson, Fishes of the World, 4th ed., Wiley, New York, 2006.

Private branch exchange

A customer premises communications system that provides its station subscribers with a wide variety of telephony services and access to public switched telephone network (PSTN) trunk circuits provided by local exchange carrier (LEC) and interexchange carrier (IEC) services. Standard private branch exchange (PBX) system hardware elements include the common equipment (cabinets, shelf carriers, printed circuit boards) and telephone instruments. Peripheral servers may also be included in the system design configuration to support systems management and voice messaging requirements. A PBX system is privately owned and physically located at the user subscriber's premises. Telephony services provided by a PBX range from basic communications operations (provisioning of dial tone and switched connections between peripheral endpoints) to station user-controlled call processing features (such as call transfer, call forwarding, and call waiting) to advanced software applications (such as multisystem networking and call contact centers).

A defining attribute of a PBX system is that trunk circuits are segmented into customer-defined groups for incoming and outgoing call requirements, and allocated to station subscribers as pooled resources. Specific trunk routes and individual circuits are assigned on a call-by-call basis based on the PBX's automatic route selection (ARS) feature as programmed by the system administrator. In contrast, trunk circuits are not pooled in a small-premises communications system designated a key telephone system (KTS). KTS station subscriber telephone instruments have designated key buttons that are used to dial and answer off-premises calls. PBX subscribers have no control over the trunk circuit they use; KTS subscribers can select an available outside-line circuit for placing off-premises calls by depressing a specific key on their telephone instrument. See KEY TELE-PHONE SYSTEM.

Elements. There are several architecture design elements common to all PBXs, regardless of underlying system technology (analog or digital transmission, circuit-switched or packet-switched): common control elements, the switching network, service circuit boards, and port interface circuit cards.

Common control elements include, at minimum, a call processor and memory storage for generic feature software and the customer database. A PBX common controller may be a proprietary carrier housed in a proprietary equipment cabinet, a proprietary server cabinet or server blade, or a customer-provided server conforming to manufacturer-recommended specifications.

The switching network used to connect PBX peripheral endpoints may be a circuit-switching system (center stage switch, local transmission buses) integrated into proprietary common equipment cabinets or an external Ethernet LAN (local-area network) packet-switched network. *See* DATA COMMUNICATIONS; LOCAL-AREA NETWORKS; PACKET SWITCHING.

Service circuits include tone senders (such as dial and busy), dual-tone multiple-frequency (DTMF)

detectors for push-button analog telephones, registers to temporarily store dialed digits, and input/output (I/O) interfaces for peripheral devices such as cathode-ray-tube (CRT) terminals and printers. *See* COMPUTER PERIPHERAL DEVICES; TELEPHONE.

Port interface circuit cards are required to support station equipment [analog, digital, ISDN BRI (Integrated Services Digital Network Basic Rate Interface), IP (Internet Protocol)] and off-premises trunk circuits [GS/LS (Ground Start/Loop Start), DID (Direct Inward Dialing), E&M (Ear and Mouth), T1interface, ISDN BRI/PRI (Primary Rate Interface), Auxiliary, FX (Foreign Exchange)]. Distinct port circuit cards are usually required for each type of station or trunk port. *See* INTEGRATED SERVICES DIGITAL NETWORK (ISDN).

System designs. PBX systems are currently available in a wide range of system designs based on several call processing, switching, and hardware equipment platforms, and individual models can be cost-effectively designed to support relatively small port capacity requirements (less than 20 station subscribers) or extremely large configuration needs (tens of thousands of station subscribers). PBX systems compete with medium-line-size KTS-hybrid systems, local telephone company Centrex services, and third-party-hosted service offerings, but are the dominant system for enterprise customers with port size requirements above 80 stations.

Evolution of digital PBX. Since the 1970s, PBX system technology has evolved through several distinct phases as mechanical analog telephone switching systems were transformed by advances in semiconductor technology and computer software programming. The two major PBX technology breakthroughs during the 1970s were computer-stored program

control and digital circuit switching using time division multiplex (TDM) transmission buses for internal system transport between port circuit terminations. There were competing techniques used to digitize voice, but pulse code modulation (PCM) using 8-kilohertz sampling and 8-bit encoding emerged as the standard, and remains so for circuit-switched communications systems. Another important PBX system innovation was electronic telephone instruments with programmable line-feature buttons. *See* MULTIPLEXING AND MULTIPLE ACCESS; PULSE MODU-LATION; SWITCHING SYSTEMS (COMMUNICATIONS).

Important PBX system design advances during the 1980s included stackable modular cabinets, dispersed processing, and digital telephone instruments with data terminals interfaces. This decade also saw the introduction of many evolutionary PBX feature and function options, including intelligent transparent private networking, integrated automatic call distribution (ACD) and reporting, voice messaging system interface, and open application interface (OAI).

Since the mid-1990s, the most important PBX technology innovation has been the integration of Internet Protocol (IP) communications standards into the core system design. The first IP telephony system designs were characterized by an Ethernet LANconnected call telephony server, IP telephone instruments, and media gateways for analog telephone instruments and PSTN trunk circuit interface requirements. This type of PBX system is now referred to as a client-server IP telephony system. An IP telephony system utilizes Ethernet packet-switched networks for premises transport of call control or voice communications signals, and IP WAN (widearea network) routers for off-premises transmission

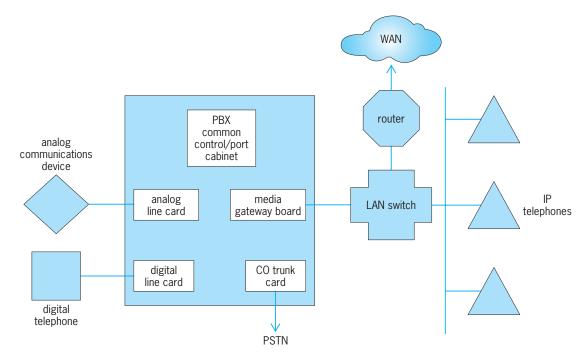


Fig. 1. IP-enabled PBX system design. PBX = private branch exchange, CO = central office, PSTN = public switched telephone network, WAN = wide-area network, LAN = local-area network, IP = Internet Protocol.

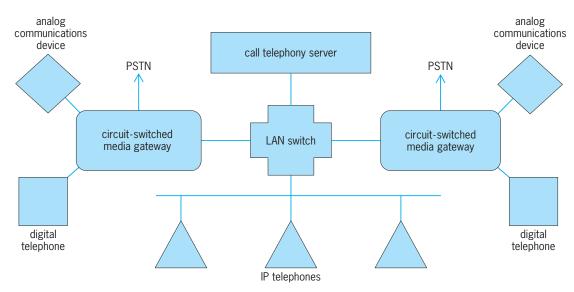


Fig. 2. Converged IP-PBX system design.

requirements. Most of the early IP telephony systems were based on ITU H.323 communications protocol specifications, but the evolving standard that all systems will most likely support is the Internet Engineering Task Force (IETF) Session Initiated Protocol (SIP) specifications, particularly core standards such as RFC (Request for Comments) 3261, RFC 1889, and RFC 2833. *See* INTERNET; WIDE-AREA NETWORKS.

IP telephony PBX system designs. There are three basic IP telephony PBX system designs: IP-enabled, converged, and client-server (sometimes referred to as softswitch).

The following are definitions of each design category:

1. IP-enabled PBXs are communications systems originally based on a circuit-switched network design that are upgraded to support IP telephony requirements using fully integrated media gateway port interface circuit cards to support IP station equipment and/or voice over IP (VoIP) trunk circuits (**Fig. 1**). *See* VOICE OVER IP.

2. Converged IP-PBXs are communications systems based on either a traditional common control cabinet or a LAN-connected call telephony server. The common control cabinet or call telephony server can support distributed circuit-switched common equipment cabinets across a LAN-WAN infrastructure. IP peripheral endpoints can be supported using direct call signaling from the primary call control processor or through integrated media gateway interface boards (**Fig. 2**).

3. Client-server IP-PBXs are communications system based on a LAN-connected call telephony server with direct call control signaling to IP telephone instruments. Media gateway equipment is required to support traditional TDM-PCM endpoints, such as analog telephone instruments or circuit-switched PSTN trunk circuits (**Fig. 3**).

IP-enabled PBXs share the following system design attributes:

1. System architecture based on a traditional circuit-switched system design platform.

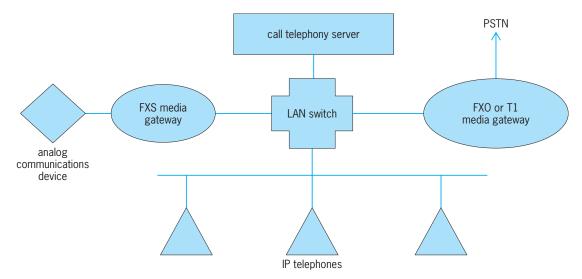


Fig. 3. Client-server IP-PBX system design. FXS = foreign exchange subscriber, FXO = foreign exchange office.

2. A proprietary common control complex housed in a dedicated control cabinet or carrier shelf.

3. Port cabinets or carriers with port interface card slots that physically connect port circuit card terminations to integrated TDM transmission bus backplanes.

4. Integrated media gateway boards that support LAN-connected IP voice terminals.

5. Integrated media gateway boards that support VoIP trunk calls.

Converged IP-PBXs share the following system design attributes:

1. Common control may be a proprietary traditional common control complex housed in a circuitswitched common-equipment cabinet or carrier, a self-standing call telephony server; or a proprietary call processor server blade.

2. The common control supports distributed LAN-WAN-connected TDM-based port cabinets or carriers for traditional circuit-switched analog and digital port interface requirements.

3. LAN-distributed circuit-switched cabinets or carriers require media gateway boards to communicate with each other and provide transmission links between TDM-PCM ports and IP peripheral endpoints.

Call control signaling over the LAN between the system common control and peripheral IP endpoints may require an intermediary media gateway or control signaling board if a traditional common control design is implemented. Server designs usually have direct signaling links to the LAN and all peripheral ports.

Client-server IP-PBXs share the following system design attributes:

1. A call telephony server as a common control complex.

2. Direct call control signaling over LAN-WAN to all IP peripheral endpoints and distributed media gateways.

3. Media gateways used for passing call control signaling and providing transmission channel connections to all non-IP ports.

Media gateways may be embedded in a proprietary call telephony server, or be provisioned as desktop modules, self-standing carriers, or circuit boards housed in Ethernet LAN switches or IP WAN routers.

Generic software features. The PBX system generic software includes hundreds of basic and advanced telephony features and system diagnostic and administration functions. Basic system features include provisioning of dial tone, a uniform numbering and dialing plan, and call routing operations under the control of the customized ARS program. The most widely used and well-known station user features are hold, transfer, forward, conference, call release, and last number redial. Several enhanced applications services can be integrated into the PBX system generic software package, including voice messaging, automatic call distribution (ACD), and networking. IP communications options that can be integrated into the PBX system offering or run on peripheral application servers include presence

management, conferencing (audio, video, and Worldwide Web), and collaboration (file sharing and whiteboarding, the placement of shared files on an on-screen "shared notebook" or "whiteboard").

Voice terminals. There are a variety of voice terminal options available to PBX station subscribers: analog, digital, and IP desktop telephone instruments; personal computer (PC) client softphones; and mobile communications devices (cellular handsets, wireless LAN handsets, and personal digital assistants, or PDAs). Attendant operator positions require specialized console instruments or softphone clients to satisfy high call volume and complex call processing operations. *See* TELEPHONE SERVICE. Allan Sulkin

Bibliography. K. Camp, *IP Telephony Demystified*, McGraw-Hill, 2003; A. Sulkin, IP PBX market chugging along, *Bus. Commun. Rev.*, 36(1):21–29, January 2006; A. Sulkin, IP-telephony enhances bysiness continuity, *Bus. Commun. Rev.*, 36(8):24–26, August 2006; A. Sulkin, *PBX Systems for IP Telephony*, McGraw-Hill, 2002.

Probability

Although probability theory derives its notion and terminology from intuition, a vague statement such as "John will probably come" is as remote from it as the statement "John is forceful and energetic" is remote from mechanics. Probability theory constructs abstract models, mostly of a qualitative nature, and only experience can show whether these reasonably describe laws of nature or life. As always in mathematics, only logical relations and implications enter the theory, and the notion of probability is just as undefinable (and as intuitive) as are the notions of point, line, or mass. An actual assignment of numerical probabilities is frequently unnecessary or impossible. For example, telephone exchanges are based on a theoretical comparison of several possible systems; only the optimal ones are built and the others discarded. Thus a huge industry depends on theoretical models of exchanges which will never exist.

An uncomplicated illustration of the nature of probability models is found in an experiment of Lord Rutherford. To measure radioactive intensity, he proceeded as follows. Observers A1 and A2 counted scintillations on a screen and observed, respectively, N_1 and N_2 scintillations; of these, N_{12} were common to both observers. To estimate the unknown true number X, Rutherford assumed that each scintillation has fixed probabilities p_1 and p_2 to be observed by A_1 and A_2 , and furthermore that the observations are independent in the sense that a scintillation observed by A_1 has still probability p_2 to be observed by A_2 . In reality the likelihood of observing a scintillation varies with growing fatigue and the proximity of the preceding scintillation; also, the observers are affected by common causes and are therefore not independent. Equating probabilities with observed frequencies (another approximation), Rutherford set $N_1 =$ $Xp_1, N_2 = Xp_2, N_{12} = Xp_1p_2$, whence $X = N_1N_2/N_{12}$.

The three equations may be solved for p_1 and p_2 , but these "probabilities" are purely fictitious and, as experience shows, inaccessible to experimental verification. The model is justified by plausibility and success.

Sample space. One speaks of probabilities only in connection with conceptual (not necessarily performable) experiments and must first define the possible outcomes. Thus, by convention, tossing coins results in heads H or tails T; regardless of experimental or philosophical difficulties the age of a person is taken as an exact number and each positive number is taken as a possible age. Throwing two dice results in one of the 36 combinations (1.1), $(1.2), \ldots, (6.6)$. An outcome such as "sum 4" is a compound event which can be further decomposed by enumeration: Sum 4 occurs if the outcome is (1.3), (2.2), or (3.1). Thus it is necessary to distinguish between elementary (indivisible) and compound outcomes or events. Each elementary outcome is called sample point; their aggregate is the sample space. The conceptional experiment is defined by the sample space, and it must be introduced and established at the outset.

For example, the experiment "distributing 3 balls in 3 cells" has 27 possible outcomes (sample points) listed in tabulation (1).

1.
$$\{abc| -|-\}$$
 10. $\{a|bc|-\}$
 19. $\{-|a|bc\}$

 2. $\{-|abc|-\}$
 11. $\{b|ac|-\}$
 20. $\{-|b|ac\}$

 3. $\{-|-|abc\}$
 12. $\{c|ab|-\}$
 21. $\{-|c|ab\}$

 4. $\{ab|c|-\}$
 13. $\{a|-|bc\}$
 22. $\{a|b|c\}$

 5. $\{ac|b|-\}$
 14. $\{b|-|ac\}$
 23. $\{a|c|b\}$
 (1)

 6. $\{bc|a|-\}$
 15. $\{c|-|ab\}$
 24. $\{b|a|c\}$
 7. $\{ab|-|c\}$
 16. $\{-|ab|c\}$
 25. $\{b|c|a\}$

 8. $\{ac|-|b\}$
 17. $\{-|ac|b\}$
 26. $\{c|a|b\}$
 9. $\{bc|-|a\}$
 18. $\{-|bc|a\}$
 26. $\{c|b|a\}$

Note that "n balls in 7 cells" may represent the distribution of n hits among 7 targets, or of n accidents in 7 weekdays, and so on.

Consider next the experiment of placing 3 indistinguishable balls into 3 cells. Whether actual balls are indistinguishable is irrelevant; they are treated as such and, by convention, there is now a space of only 10 sample points. It is listed in tabulation (2).

$$1. \{* * * | -|-\} 4. \{* * | * |-\} 8. \{-| * * |*\}$$

$$2. \{-| * * * |-\} 5. \{* * | - |*\} 9. \{-| * | * *\}$$

$$3. \{-| -| * * *\} 6. \{* | * * |-\} 10. \{* | * |*\}$$

$$7. \{* | -| * *\}$$

$$(2)$$

In playing roulette, each point on a circle represents a possible outcome and the sample space is the interval $0 \le \vartheta < 2\pi$. When one observes the motion of a particle under diffusion, every function x(t) represents a conceivable outcome and the sample space is a complicated function space. **Events.** In examining a bridge hand, one may ask whether it contains an ace or satisfies some other condition. In principle each such event may be described by specifying the sample points which do satisfy the stipulated condition. Thus every compound event is represented by an aggregate of sample points, and in probability theory these terms are synonymous. The standard notations of set theory are used to describe relations among events. *See* SET THEORY.

Given an event *A* one may consider the case that *A* does not occur. This is the negation or complement of *A*, denoted by *A'*; it consists of those sample points that do not belong to *A*. Given two events *A* and *B*, the event *C* that either *A* or *B* or both occur is the union of *A* and *B* and denoted by $C = A \cup B$. In particular $A \cup A'$ is the whole sample space \mathfrak{S} which therefore represents certainty. The event *D*, both *A* and *B* occur, is the intersection of *A* and *B* and written $D = A \cap B$. It consists of the points common to *A* and *B*. If there are no such common points (as in the case of *A* and *A'*), *A* and *B* cannot occur simultaneously and they are called mutually exclusive, written $A \cap B = 0$. The event "*A* but not *B*" is simply $A \cap B'$.

For example, in tabulation (1), the event *A* "one cell multiply occupied" is the aggregate of the points numbered 1-21. The event *B* "first cell not empty" is the aggregate of the points 1, 4-15, and 22-27. Because every point belongs either to *A* or to *B* (or both). $A \cup B = \mathfrak{S}$ is the certain event. Next, $D = A \cap B$ consists of the points 1, 4-14. Finally, *A'* may be described as "no cell empty."

Probabilities in finite spaces. If the sample space \mathfrak{S} contains only *N* points E_1, \ldots, E_N their probabilities may be any numbers such that $P\{E_j\} = 0$ and $P\{E_1\} + \cdots + P\{E_N\} = 1$. The probabilities of all points contained in *A*; thus $P\{\mathfrak{S}\} = 1$. To find $P\{A \cup B\}$ one considers all points belonging to either *A* or *B*, but those belonging to both *A* and *B* are counted only once. Therefore $P\{A \cup B\} = P\{A\} + P\{B\}$ - $P\{A \cap B\}$. In particular, for mutually exclusive events, there is the addition rule $P\{A \cup B\} = P\{A\} + P\{B\}$.

Frequently considerations of symmetry lead one to consider all E_i as equally likely; that is, to set $P(E_i) =$ 1/N. In this case P(A) = n/N where *n* is the number of points in A; for a gambler betting on A, these represent the "favorable cases." For example, in throwing a pair of "perfect" dice, one naturally assumes that the 36 possible outcomes are equally likely. This model does not lose its justification or usefulness by the fact that actual dice do not live up to it. The assumption of perfect randomness in games, card shuffling, industrial quality control, or sampling is rarely realized, and the true usefulness of the model stems from the experience that noticeable departures from the ideal scheme lead to the detection of assignable causes and thus to theoretical or experimental improvements.

How the success of probability theory depends on the disregard of preconceived philosophical ideas and on the readiness to adapt models to unexpected circumstances is illustrated by Bose-Einstein statistics. In the example of tabulation (1), the notion of perfect randomness leads to the assignment of probability 1/27 to each point. In the case of indistinguishable balls, tabulation (2), it has been argued that an experiment is unaffected by failure to distinguish between balls; physically there remain 27 possibilities grouped in 10 distinguishable forms. This argument leads to assigning probability 1/27 to each of the points 1-3, probability 1/9 to each of the points 4-9, and 2/9 to point 10. This reasoning (sound in certain situations) has been accepted as evident in statistical mechanics for the distribution of r particles in n cells (Maxwell-Boltzmann statistics). Surprisingly, it turned out that no physical particles behave this way and it was revolutionary when Bose and Einstein showed that for one type of particle all distinguishable arrangements are equally likely. This model assigns probability 1/10 to each point of tabulation (2). See BOLTZMANN STATISTICS; BOSE-EINSTEIN STATISTICS.

A useful, although vague, intuitive description of probability describes $P{A}$ as the relative frequency of the event *A* if the experiment is repeated many times under identical circumstances. The laws of large numbers render this more precise, but the description often lacks operational meaning. Experiments in agriculture and human sampling cannot be repeated under remotely similar conditions, and in the case of telephone exchanges, useful probability models refer to situations which will never materialize.

Probabilities in infinite spaces. Two examples, unending coin tossing and roulette, illustrate the novel features of this topic.

Unending coin tossing. In the study of limit laws, one must consider potentially infinite sequences of coin tossings. The possible outcome of this experiment is an infinite sequence of heads and tails, and every sequence such as HTHTTH ... represents a simple point. Finitely many tosses are the beginning of an infinite sequence, and the event "first four trials resulted in HTTH" is the aggregate of the infinitely many sequences with the prescribed beginning. Such an event is called an interval of length 4. There are 2^n intervals of length n, and they are mutually exclusive. For reasons of symmetry, one attributes the probability 2^{-n} to each interval of length *n*. Thus the assignment of basic probabilities refers to intervals rather than to points. A point such as HTHT ... is the limit of an infinite sequence of contracting intervals H, HT, HTH, ..., and therefore probability zero must be attributed to each individual point.

The probabilities of other events are similarly defined by limiting procedures. For example, consider the event A that an infinitely prolonged sequence of trials never produces a run of at least two consecutive heads, or two consecutive tails. It is more convenient to enumerate the points of the complementary event A' that two equal symbols do occur in succession. Clearly A' is the union of the infinitely many mu-

tually exclusive intervals *HH*, *TT*; *HTT*, *THH*; *HTHH*, *THTT*, and so on. Here there are 2 intervals of length n = 2, and therefore $P\{A'\} = 2(2^{-2} + 2^{-3} + 2^{-4} + \cdots) = 1$, whence $P\{A'\} = 0$. The indicated result of the experiment is thinkable, but probability zero is attributed to it. A similar, although more complicated, limiting procedure leads to the law of large numbers according to which the event "the frequencies of *H* and *T* in the first *n* trials tend to 1/2 as $n \to \infty$ " has probability one.

Roulette. Here the sample space consists of the angles $0 \leq \vartheta < 2\pi$, and the notion of a perfect roulette assumes equal probabilities for intervals of equal length; thus an interval of length *a* carries probability $a/2\pi$. If the roulette is divided into 32 equal numbered intervals, the event "even number" consists of 16 intervals and has probability 1/2.

The situation encountered here is not peculiar to probability but is common in measure theory. One starts with a collection of basic events, called intervals, and attributes probabilities to them. By simple and natural limiting procedures, probabilities can then be defined for a much wider class \mathfrak{F} of events which are obtainable by applying the operations of set theory to intervals (in finite or infinite numbers). F is the Borel field generated by the intervals. Probability is simply a measure on \mathfrak{F} ; that is, to each event A in \mathfrak{F} , there corresponds a probability $P\{A\} = 0$ which is completely additive. If A is the union of the mutually exclusive events A_1, A_2, \ldots , then $P\{A\} =$ $\Sigma P\{A_i\}$. The probability of the whole space is, of course, unity.

The extension of the addition rule from finitely to infinitely many summands may be defended by considerations of continuity, but ultimately it is justified by its simplicity and its success.

Conditional probability—independence. Suppose that a population of N people includes N_A color-blind persons and N_H females. To the event A "a randomly chosen person is color-blind" can be ascribed probability $P{A} = N_A/N$, and similarly for the event *H* that a person be female one has $P\{B\} = N_H/N$. If N_{AH} is the number of color-blind females, the ratio N_{AH}/N_H may be interpreted as probability that a randomly chosen female be color-blind; here the experiment "random choice in the population" is replaced by a selection from the female subpopulation. In the original experiment, N_{AH}/N is the probability of the simultaneous occurrence of both A and H, so that $N_{AH}/N_H = P\{A \cap H\}/P\{H\}$. Similar situations occur so frequently that it is convenient to introduce Eq. (3) and to call this the conditional probability

$$P\{A|H\} = \frac{P\{A \cap H\}}{P\{H\}}$$
(3)

of the event *A* relative to *H*. This concept is useful whenever it is desired to restrict the consideration to those cases where the event *H* occurs (or where the hypothesis *H* is fulfilled). Thus, in betting on an event *A* the knowledge that *H* occurred would induce one to replace $P\{A\}$ by $P\{A|H\}$. If all sample points are equally likely, $P\{A|H\}$ still represents the

ratio of favorable cases to the total of cases possible when it is known that *H* has occurred.

Despite its simplicity the notion of conditional probability is exceedingly important, and frequently the probabilities in sample space are defined only in terms of conditional probabilities. An example of the concept is provided by a bolt factory in which three machines manufacture, respectively, 25, 35, and 40% of the total. Of their output 5, 4, and 2% are defective bolts. Classification of the bolts according to the number of the machine and the quality (*d* for defective, *c* for conforming) gives the six categories c_1 , c_2 , c_3 , d_1 , d_2 , and d_3 . A random choice of a bolt results in one of these six outcomes, but their probabilities are not given directly. Instead, the data relating to the first machine are given by

$$P\{c_1 \cup d_1\} = 0.25$$
$$P\{d_1 | c_1 \cup d_1\} = 0.25$$

It follows that $P{d_1} = 0.0125$ and similarly for the other points. This example may also serve to illustrate the reasoning following Bayes concerning the probability of causes. Supposing a bolt was found to be defective (hypothesis *H*), what is the probability that it came from the first machine (cause *A*)? Here

$$P{H} = P{d_1} + P{d_2} + P{d_3}$$

= 0.0125 + 0.0140 + 0.0080 = 0.0345
$$P{A \cap H} = P{D_1} = 0.0125$$

and thus the required answer is given by

$$P\{A|H\} = 0.0125/0.0345 = 25/69$$

In tabulation (1) the probability of *H* "ball *a* is in the first cell" equals 1/3, and the probability of *A* "first cell is multiply occupied" = 7/27. Now given that the ball *a* is in the first cell, the conditional probability that this cell is multiply occupied becomes 5/9. The knowledge that *H* has occurred should increase one's readiness to bet on *A*. By contrast, for the event *B* "ball *b* is in the second cell," $P\{B|H\} = 1/3 = P\{B\}$, and so the knowledge that *H* has occurred gives no clue as to *B*. Therefore, *B* is said to be independent of *H* if $P\{B|H\} = P\{B\}$, that is, if

$$P\{B \cap H\} = P\{B\}P\{H\}$$

Clearly, in this case $P\{H|B\} = P\{H\}$ so that *H* is also independent of *B*. Accordingly, two events *B* and *H* are independent of each other if the probability of their simultaneous occurrence follows the multiplication rule $P\{B \cap H\} = P\{B\}P\{H\}$. This notion carries over to systems of more than two events.

Independent trials. The intuitive frequency interpretation of probability is based on the concept of experiments repeated under identical conditions; a theoretical model for this concept can be developed.

Consider an experiment described by a sample space \mathfrak{S} ; for simplicity of language it can be assumed that \mathfrak{S} consists of finitely many sample points E_1, \ldots, E_N . When the same experiment is performed twice in succession, the thinkable outcomes are the N^2 pairs

of sample points (E_1, E_1) , (E_1, E_2) , ..., (E_N, E_N) , and these now constitute the new sample space. It is called the combinatorial product of \mathfrak{S} by itself and denoted by $\mathfrak{S} \times \mathfrak{S}$ with reference to analytic geometry, one speaks of the first and second coordinate of the point (E_i, E_j) . These notions apply equally to infinite sample spaces and to products $\mathfrak{S} \times \mathfrak{S} \times \mathfrak{S}$ \cdots of more than two factors. For example, the cartesian plane of points (x, y) is the product of the real line by itself. In tossing a coin once, \mathfrak{S} contains only the points *H* and *T*; tossing the coin *n* times leads to the *n*-tuple product $\mathfrak{S}: \cdots \times \mathfrak{S}$ whose points have *n* coordinates and are of the form $(HT \cdots T)$.

Probabilities must be assigned to the events in $\mathfrak{S} \times \mathfrak{S}$. The case of dependent trials will be treated in the next section; if the second trial is independent of the first, the probabilities in $\mathfrak{S} \times \mathfrak{S}$ follow the productive rule $P\{E_i, E_i\} = P\{E_i\}P\{E_i\}$.

In the case of *n* tossings of a coin, this rule leads to the probability 2^{-2} for each sample point in agreement with the requirement of equally likely cases. The present approach is more flexible and more general as shown by the Bernoulli trials.

Bernoulli trials. Suppose each trial results in success *S* or failure *F*, and $P{S} = p$, $P{F} = q$ where p + q = 1. (This may be considered as the model of a skew coin.) A succession of *n* independent trials of this kind leads to the sample space of *n*-tuples (*SFFS* \cdots *FS*), and the probability of such a point is the product ($pqqp \cdots qp$) obtained on replacing each *S* by *p* and each *F* by *q*.

This model has obvious applications to repeated observations and to gambling. Independence is an assumption to be verified experimentally. Conceivably a coin could be endowed with memory and avoid runs of more than 17 successive heads. That the sex distribution within families resembles Bernoulli trials is purely a matter of experience. Many gamblers fully accept the independence and yet believe that they can influence fate by using "systems," for example, by skipping the game after each failure, or waiting for a run of 3 successes, and so on. The theorem on systems shows this to be a fallacy; a gambler not endowed with foresight may use any system or random choice of the times when he plays or skips the game; he remains confronted with Bernoulli trials and is exactly in the same situation as if he played at each trial. See DISTRIBUTION (PROBABILITY).

Geometric probabilities. In the interval 0 < x < 1 a point is chosen at random. This interval is the sample space \mathfrak{S} and the probability of each subinterval equals its length. The sample space $\mathfrak{S} \times \mathfrak{S}$ is the unit square of the *x*,*y* plane, and the probability of any figure equals its area. The event "the two successive choices result in a sum <1" is represented by the triangle below the main diagonal and has probability 1/2. The event "the greater of the two choices is < t" is represented by the square 0 < x < t, 0 < y < t and has probability t^2 .

Dependent trials; Markov chains. Many phenomena can be analyzed in terms of dependent trials. In their description the convenient and picturesque terminology of urn models shall be adopted, which should not detract from the general nature of the schemes being presented.

Consider an urn containing *N* balls, of which *r* are red *R* and b = N - r black *B*. Assuming perfect randomness, the probability that a randomly drawn ball be red equals r/N. If the ball is replaced and the procedure repeated, the result is Bernoulli trials with p = r/N. Without replacement, the sample space corresponding to two drawings contains four points *RR*, *RB*, *BR*, and *BB*, to which probabilities are assigned as follows: If the first ball drawn is red (probability r/N), the conditional probabilities of *R* and *B* at the second trial become (r - 1)/(N - 1) and b/(N - 1). By Eq. (3), therefore,

$$P[RR] = r(r - 1)/N(N - 1)$$

$$P[RB] = P[BR] = rb/N(N - 1)$$

$$P[BB] = b(b - 1)/N(N - 1)$$

A more general urn model is obtained by letting the composition of the urn vary from trial to trial. For definiteness consider the following scheme: Each time a ball is drawn, it is replaced, and c balls of the color drawn and d balls of the opposite color are added to the urn. Here c and d are fixed numbers which may be negative. This scheme contains interesting special cases such as the following:

1. When c = d = 0, drawing with replacement occurs, and for c = -1, d = 0, drawing without replacement occurs. In the latter case, the process terminates after *N* drawings.

2. The Polya model of contagion is the special case when c > 0 is fixed and d = 0. Here the drawing of either color increases the probability of the same color at subsequent trials, just as in the contagious disease each occurrence increases the probability of further occurrences. This model represents only a crude first approximation to phenomena of contagion, but it leads to comparatively simple formulas and has been applied with astonishing success to a variety of experiences from sickness insurance to baseball scores.

3. The Ehrenfest model for heat exchange considers two containers, I and II, and N particles distributed in them. A particle is chosen at random and removed from its container into the other. This scheme differs only linguistically from the urn scheme. If the particles in I are called red and those in II black, then each trial changes the color of one ball and gives the special case c = -1 and d = 1.

The probabilities of the various possible outcomes in the general scheme are obtained as above. For example, $P\{RBR\} = r(b + d)(r + c + d)/N(N + c + d)(N + 2c + 2d)$, and so on.

Markov chains represent another important scheme for dependent trials. Suppose that at each trial the possible outcomes are E_1, \ldots, E_N and that whenever E_i occurs the conditional probability of E_j at the next trial is p_{ij} , independently of what happened at the preceding trials. Here, of course, $p_{ij} = 0$ and $p_{ij} + p_{i2} + \cdots + p_{iN} = 1$ for each *i*. The p_{ij} are called transition probabilities. The whole process is now determined if the initial probabilities, π_i , at the first trial are known. For example, $P\{E_a E_b E_c\} = \pi_a p_{ab} p_{bc}$. The probability of the event " E_c at the third trial" is obtained by summation over all a and b, and so on. Markov chains, and their analog with continuous time, represent the simplest type of stochastic process. The Ehrenfest model considered above may be treated as a Markov chain by letting E_i represent the event that container I contains *i* particles. Then $p_{i,i-1} = i/N$, $p_{i,i+1} = (N - i)/N$, and $p_{ij} = 0$ for all other combinations of *ij*. Other examples of Markov chains are the gambler's accumulated fortune, the composition of a deck of cards under random shuffling, and random walks. Important applications are to queueing theory where one encounters also processes with more complicated aftereffects. See QUEUEING THEORY; STOCHASTIC PRO-CESS.

Random variables and their distributions. The theory of probability traces its origin to gambling, and the gambler's gain may still serve as the simplest example of a random variable. With every possible outcome (sample point) there is associated a number, namely, the corresponding gain. In other words, the gain is a function on the sample space, and such functions are called random variables. (In infinite spaces the idea is the same, but a somewhat more cautious definition is in order.) With the same experiment, one may associate many random variables. As an example, consider the sample space of tabulation (1) with probability 1/27 for each point. A typical random variable is the number N of occupied cells; it assumes the value 1 at the three points numbered 1-3; the value 2 at the eighteen points 4-21; and the value 3 at the six points 22-27. One says, therefore, that the probability distribution of N is given by $P{N = 1} = 1/9$, $P{N = 2} = 2/3$, $P{N = 3} =$ 2/9. Another variable is the number X of balls in the first cell. An inspection of tabulation (1) shows that its probability distribution is given by $P{X = 0} =$ $8/27, P\{X = 1\} = 12/27, P\{X = 2\} = 6/27, P\{X = 1\}$ 3 = 1/27. One may also consider the two variables simultaneously and find, for example, that the combination N = 1, X = 0 occurs at two points, whence $P{N = 1, X = 0} = 2/27$. The probabilities of all pairs are given by the joint probability distribution of N and X exhibited in tabulation (4). Adding the

X N	0	1	2	3	Distribu- tion of N	
1 2 3	2/27 6/27 0	0 6/27 6/27	0 6/27 0	1/27 0 0	3/27 18/27 6/27	(4)
Distribu- tion of <i>X</i>	8/27	12/27	6/27	1/27		

entries in the rows and columns gives the distribution of *N* and *X*, respectively, and they are therefore occasionally called marginal distributions. *See* COM-BINATORIAL THEORY.

The "Geometric probabilities" example given earlier may be used to illustrate the case of

continuous random variables. This example considers two consecutive selections of a point in the interval 0 < x < 1. Let *S* be the random variable denoting the sum of the two choices, and L the larger of the two. One sees that for 0 < t < 1 the event $L \le t$ has probability t^2 ; thus, setting $P\{L < t\} = F(t) = t^2$ when $0 \le t \le 1$; for t < 0 and for t > 1 one has trivially F(t)= 0 and F(t) = 1, respectively. This is the distribution function of L. From it can be calculated all probabilities relating to L. Similarly the event $S \leq u$ is represented by the region in the unit square below the line x + y = u; therefore, the distribution function of S, namely, $P{S \le u} = G(u)$, is given by G(u) = 0 for $u \leq 0, G(u) = (1/2)u^2$ for $0 \leq u \leq 1, G(u) = 1 - 1$ $(1/2)(2-u)^2$ for $1 \leq u \leq 2$, and G(u) = 1 for $u \leq 2$. In like manner, the joint distribution function $P\{L \leq L\}$ $t, S \leq u$ = H(t,u) of the pair L,S can be calculated.

Every random variable *X* has a distribution function $F(t) = P\{X \le t\}$. If *X* assumes only finitely many values, then F(t) is a step function. Thus, in tabulation (4), F(t) assumes the values 0, 8/27, 20/27, 26/27, and 1, respectively, in intervals t < 0, $0 \le t < 1$, $1 \le t < 2$, $2 \le t < 3$, and $t \le 3$. In such cases the notion of distribution function is used mainly for uniformity of language. The notion is really convenient when F(t) is not only continuous but also has a derivative f(t) = F(t); then f(t) is called the probability density of *X*. In the above example the variable *L* has a density defined by 2t for 0 < t < 1 and 0 elsewhere; the density of *S* is 0 for u < 0 and u > 2; it equals *u* for 0 < u < 1 and equals 2 - u for 1 < u < 2.

The notion of independence carries over: Two random variables *X* and *Y* are independent if $P{X \le x, Y \le 2} = P{X \le s} \cdot P{Y \le t}$. It is easily seen that for independent variables with distribution functions *F*(*t*) and *G*(*t*) the distribution function of the sum *S* = *X* + *Y* is given by the convolution shown in Eq. (5). In terms of densities, Eq. (5) reads as Eq. (6). In the random choice example, the coordi-

$$P\{S \le u\} = \int_{-\infty}^{+\infty} F(u-s) \, dG(s)$$
$$= \int_{-\infty}^{+\infty} G(u-s) \, dF(s) \tag{5}$$

$$b(u) = \int_{-\infty}^{+\infty} f(u - s)g(s) \, ds$$
$$= \int_{-\infty}^{+\infty} g(u - s)f(s) \, ds \tag{6}$$

nates of the points chosen are independent variables with the rectangular density f(s) = g(s) = 1 for 0 < s < 1. The distribution of their sum *S* which has been calculated above can be found also using Eq. (6).

Expectations. Given a random variable *X* one may interpret its distribution function F(t) as describing the distribution of a unit mass along the real axis such that the interval $a < x \leq b$ carries mass F(b) - F(a). In the case of a discrete variable assuming the values x_1, x_2, \ldots with probabilities p_1, p_2, \ldots the entire mass is concentrated at the points x_i ; if F(x) = f(x) exists, it represents the ordinary mass

density as defined in mechanics. The center of gravity of this mass distribution is called the expectation of *X*; the usual symbol for it is E(X), but physicists and engineers use notations such as $\langle X \rangle$, $\langle X \rangle_{Av}$, or \overline{X} . In the cases mentioned,

$$E(X) = \sum p_i x_i$$
$$E(X) = \int_{-\infty}^{+\infty} x f(x) \, dx$$

In all cases E(X) is given by the Stieltjes integral over x dF(x). (To be precise, one speaks of expectations only when the integral converges absolutely.)

Before discussing the significance of the new concept, a few frequently used definitions are appropriate. Put m = E(X). Then $(X - m)^2$ is, of course, a random variable. In mechanics, its expectation represents the moment of inertia of the mass distribution. In probability, it is called variance of *X*; its positive root is the standard deviation. Clearly,

$$Var(X) = E(X - m)^2 = E(X^2) - m^2$$

The variance is a measure of spread: It is zero only if the entire mass is concentrated at the point m, and it increases as the mass is moved away from m. In the case of two variables X_1 and X_2 with expectations m_1 and m_2 it is necessary to consider not only the two variances $s_i^2 = E[(X_i - m_i)^2]$ but also the covariance $Cov(X_1,X_2) = E[(X - m_1)$ $(X_2 - m_2)] = E(X_1X_2) = m_1m_2$. The covariance divided by s_1s_2 is called the correlation coefficient of X_1 and X_2 . If it vanishes, X_1 and X_2 are called uncorrelated. Every pair of independent variables is uncorrelated, but the converse is not true.

If X_1, X_2, \ldots, X_n are random variables with expectations m_1, \ldots, m_n and variances s^2_1, \ldots, s^2_n , the expectation of their sum $S_n = X_1 + \cdots + X_n$ is always given by $E(S_n) = m_1 + \cdots + m_n$; if all the covariances of X_1 and X_j vanish, then clearly $Var(S_n) = s^2_1 + \cdots + s^2_n$.

When X represents a physical quantity, then $X^* = (X - m)s^{-1}$ represents the same quantity measured from a different origin and in new units. In the physicist's terminology, X^* is the quantity X referred to dimensionless units. In probability, X^* is called the reduced or standardized variable.

It was once assumed that every reasonable random variable has finite expectation and variance. Modern theory refutes this assumption. Many recurrence times in important physical processes have no finite expectations. Even in the simple coin-tossing game, the number of trials up to the time when the gambler's accumulated gain first reaches a positive level has infinite expectation.

Laws of large numbers. To explain the meaning of the expectation and, at the same time, to justify the intuitive frequency interpretation of probability, consider a gambler who at each trial may gain the amounts x_1, x_2, \ldots, x_n with probabilities p_1, p_2, \ldots, p_n . The gains at the first and second trials are independent random variables X_1, X_2 with the indicated distribution and the common expectation $m = + p_i x_i$. The event that an individual gain equals x_i has

probability p_i , and the frequency interpretation of probability leads one to expect that in a large number n of trials this event should happen approximately np_i times. If this is true, the total gain $S_n = X_1 + X_2 + \cdots + X_n$ should be approximately nm; that is, the average gain $(1/n)S_n$, should be close to m. The law of large numbers in its simplest form asserts this to be true. More precisely, for each $\epsilon > 0$ it assures one that

$$P\left\{\left|\frac{1}{n}S_n - m\right| > \epsilon\right\} \to 0 \text{ as } n \to \infty$$

As a special case, one can obtain a frequency interpretation of probability. In fact, consider an event A with $P{A} = p$ and suppose that in a sequence of independent trials a gambler receives a unit amount each time when A occurs. Then the expectation of the individual gain equals p, and S_n is the number of times the event A has occurred in n trials. It follows that

$$\left\{ \left| \frac{1}{n} S_n - p \right| > \epsilon \right\} \to 0$$

That is, the relative frequency of the occurrence of *A* is likely to be close to *p*.

Without this theorem, probability theory would lose its intuitive foundation, but its practical value is minimal because it tells one nothing concerning the manner in which the averages $n^{-1}S_n$ are likely to approach their limit *m*. In the regular case where the X_i have finite variances, the central limit theorem gives much more precise and more useful information; for example, it tells one that for large *n* the difference S - np is about as likely to be positive as negative and is likely to be of the magnitude $n^{1/2}$. When the X_k have no finite variances, the central limit theorem fails and the sums S_n may behave oddly in spite of the law of large numbers. For example, it is possible that $E(X_k) = 0$ but $P\{S_n < 0\} \rightarrow 1$. In gambling language this game is "fair," and yet the gambler is practically certain to sustain an ever-increasing loss

There exist many generalizations of the law of large numbers, and they cover also the case of variables without finite expectation, which play an increasingly important role in modern theory. *See* GAME THEORY; PROBABILITY (PHYSICS); STATISTICS. William Feller

Bibliography. W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. 1, 3d ed., 1968, vol. 2, 2d ed., 1971; H. J. Larson, *Introduction to Probability*, 1995; S. M. Ross, *A First Course in Probability*, 5th ed., 1997.

Probability (physics)

To the physicist the concept of probability is like an iceberg. The part of it which the physicist uses and which is, therefore, in full view for him or her is but a small fraction of what is hidden in other and larger disciplines. The philosophy and mathematics of probability have become increasingly interesting and important. *See* PROBABILITY.

Bernoulli's problem. One of the most basic problems encountered in the application of probability to physics was solved by J. Bernoulli. It concerns the probability of achieving a specified number (*x*) of successes in *n* independent trials when the probability of success in a single trial is known. Denote by *p* the probability of success; let q = 1 - p be the probability of failure. Equation (1) represents

$$C_x^n \equiv \frac{n!}{x!(n-x!)} \tag{1}$$

the well-known binomial coefficient. The probability in question is given by Eq. (2). For example, one

$$w_n(x) = C_x^n p^x q^{n-x} \tag{2}$$

may consider an urn containing a black balls and b white balls; n drawings are to be made from this urn, with replacement of the drawn ball each time. The probability that x white balls shall turn up in these n drawings is required. In this example, Eqs. (3) hold.

$$p = \frac{b}{a+b}$$

$$w_n(x) = C_x^n \left(\frac{b}{a+b}\right)^x \left(\frac{a}{a+b}\right)^{n-x}$$
(3)

Equation (2) defines Bernoulli's distribution; because of the newtonian binomial coefficients which it involves, it is sometimes called Newton's formula. It satisfies relations (4)-(6). Hence, the σ^2 is called

$$\sum_{x=0}^{n} w_n(x) = 1$$
 (4)

$$\bar{x} = \sum_{x} x w_n(x) = np \tag{5}$$

$$\sigma^{2} = \sum_{x} (x - \bar{x})^{2} w_{n}(x) = \bar{x}^{2} - \bar{x}^{2} = npq \qquad (6)$$

the dispersion of the distribution $w_n(x)$. The socalled standard deviation σ equals \sqrt{npq} .

The physicist regards the emission of particles from nuclei, of photons from hot bodies, and of electrons from hot filaments as random phenomena controlled by probability laws. One of the simplest methods for testing this assumption is based upon these formulas. Suppose that a radioactive specimen emits particles for a finite period T, and that their number is x. By repeated observations on the emission of particles in a period of T seconds, one obtains a series of numbers, x_1 , x_2 , x_3 , etc. Now imagine the period T to be subdivided into a very large number, *n*, of intervals, each of length τ , so that $n\tau = T$. Indeed, τ will be assumed to be so small that, at most, one particle is emitted within the interval τ . A single observation made during that infinitesimal interval would therefore yield the result: either no emission

or emission, say, with probabilities p or q. By hypothesis, n such observations yield x emissions. It is seen, therefore, that the problem involves an application of Bernoulli's distribution. While it is difficult to specify the values of p, q, and τ , the result given by Eq. (6) must nevertheless be true. Since q is extremely small, p is very nearly 1. Hence, in view of Eq. (5), $\sigma^2 = npq = nq = \bar{x}$. This relation has been tested experimentally and found to be true in all instances.

Approximations. In many practical applications, the values of *x* and *n* are very large, and a direct use of Newton's formula becomes impossible because the binomial coefficients are difficult to evaluate for large *x* and *n*. The example of particle emission just discussed is a case in point. Under these circumstances, two approximations to Newton's formula are available, one derived by C. F. Gauss, the other one by S. D. Poisson and bearing his name. Gauss's law is the limiting form of Eq. (2), when both *x* and *n* are large, so large in fact that $1/\bar{x}1/\bar{x} = 1/np$ and 1/npq are both negligible. This implies that *p* and *q* are numbers not greatly different from unity. In that case, Eq. (7) holds. This is Gauss's law.

$$\lim_{n \to \infty} w_n(x) = \frac{1}{\sqrt{2\pi px}} \exp\left[\frac{-(x-\bar{x})^2}{2px}\right]$$
$$= \frac{1}{\sqrt{2\pi\sigma}} \exp\left[\frac{-(x-\bar{x})^2}{2\sigma^2}\right] \quad (7)$$

If p is so small that the mean np is of the order of unity in any given application, Bernoulli's distribution is approximated by Poisson's law, Eq. (8). The

$$\lim_{n \to \infty} w_n(x) = \frac{(np)^x e^{-np}}{x!} \tag{8}$$

validity of this formula has also been confirmed in many instances of particle emission.

The density fluctuations which occur in a finite volume of gas present an interesting application of Gauss's law. Suppose that a vessel of volume V contains but a single molecule. One may then select a small element of volume v and imagine successive observations to be made in order to determine whether the molecule is in this element or not. The probability of its being there, which will be denoted by p, is clearly the ratio of the small volume element to the volume of the total container, p = v/V. Now let n observations be made upon the volume v. The probability that, in x of them, the one molecule is found in v is given by Newton's formula. If the vessel contains not one but n molecules, then the probability of finding x of them in v simultaneously is the same as that of finding the one molecule x times in v on *n* successive occasions. Thus, the probability of observing x of the n molecules in v is given by Eq. (9).

$$w_n(x) = C_x^n p^x q^{n-x} \tag{9}$$

Equation (9) may be approximated by Gauss's law. Thus, on substituting p and q, as shown in Eqs. (10) and (11), Eq. (12) results. The expected mean and

$$p = \frac{v}{V} = \frac{1}{k} \tag{10}$$

$$q = 1 - p \tag{11}$$

$$w(x) dx = \frac{k}{\sqrt{2\pi n(k-1)}}$$
$$\times \exp\left[-\frac{(kx-n)^2}{2n(k-1)}\right] dx \quad (12)$$

the dispersion of this distribution must be the same as the corresponding quantities for Bernoulli's distribution. Hence Eqs. (13) and (14) hold. If k is suffi-

$$\bar{x} = np = \frac{n}{k} \tag{13}$$

$$\sigma^2 = npq = \frac{n}{k^2}(k-1) = \bar{x}\left(1 - \frac{\bar{x}}{n}\right) \qquad (14)$$

ciently large, $\sigma^2 = \bar{x}$ approximately.

It is often convenient to introduce a quantity δ known as the relative fluctuation and defined by Eq. (15).

$$\delta = \frac{x - \bar{x}}{\bar{x}} \tag{15}$$

In view of Eq. (13), Eq. (16) holds.

$$\delta = \frac{kx - n}{n} \tag{16}$$

In terms of δ , the law of fluctuations takes on the simple form given by Eq. (17). Here \bar{x} is the mean

$$w(\delta) d\delta = \sqrt{\frac{\tilde{x}}{2\pi q}} \exp\left[\frac{-x}{2q}\delta^2\right] d\delta \qquad (17)$$

number of molecules within v, and q is very nearly equal to 1 if $v \gg V$. Equation (17) is well supported by experiment.

Theory of errors. Perhaps the most important application of probability theory to exact science occurs in the treatment of errors. Measurements are accompanied by errors of two kinds: determinate and random. The former arise from actual mistakes, either on the part of the observer or from faulty instruments; they are not susceptible of mathematical treatment. Random errors, however, because they are numerous, small, and likely to combine in linear fashion, are subject to the laws of probability analysis.

Gauss error law. Let there be *n* measurements of some physical quantity resulting in the numbers X_1 , X_2 , X_3 , ..., X_n . If the true value of the quantity (usually unknown) is denoted by *X*, then the errors are given by Eqs. (18). It can be shown mathematically and

$$x_1 = X_1 - X$$
 $x_2 = X_2 - X, ...$
 $x_n = X_n - X$ (18)

has been confirmed by numerous observations that the relative frequency of occurrence of an error *x* is represented by Bernoulli's distribution which, under the present conditions, takes on the form of Gauss's law, Eq. (19). In this formula the parameter b^2 equals

$$N(x) \, dx = \frac{b}{\sqrt{\pi}} e^{-b^2 x^2} \, dx \tag{19}$$

 $1/(2\sigma^2)$ and is called the index of precision. Clearly, the greater the value of *b* the narrower the distribution given by Eq. (19). The latter is often called the Gauss error law or the normal distribution of errors. In writing it, the assumption has of course been made that the numbers x_1 , x_2 , etc., may be replaced by a continuous distribution.

The probability that a single measurement will contain an error between the limits -a and +a is given by Eq. (20). The integral occurring here is

$$\frac{b}{\sqrt{\pi}} \int_{-a}^{a} e^{-b^2 x^2} \, dx = \frac{2b}{\sqrt{\pi}} \int_{0}^{a} e^{-b^2 x^2} \, dx \qquad (20)$$

called the error function, and is denoted by erf(a); it is tabulated in most textbooks that discuss the theory of errors. The factor in front of the integral has been chosen so that Eq. (21) holds.

$$\int_{-\infty}^{\infty} N(x) \, dx = 1 \tag{21}$$

Returning to the set of measured values X_1, X_2, \ldots, X_n , one may ask: Which is the most probable value (the one most likely to be true) consistent with this set of numbers? The answer is given by the principle of least squares, which affirms that the most probable value is the one for which the sum of the squared errors $x_1^2 + x_2^2 + \cdots + x_n^2$ is a minimum. *See* LEAST-SQUARES METHOD.

From this one may prove directly that the most probable value of X is its arithmetical mean, given by Eq. (22).

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} \tag{22}$$

Kinds of errors. The reliability of a set of measurements, such as the sequence X_1, X_2, \ldots, X_n , is specified by certain measures called errors, but in a slightly different sense from that previously employed. Three kinds of "error" will be described: the average error a, the root-mean-square error m, and the probable error r. All three refer, not to a single measurement as did the quantity x_i , but to the entire distribution in Eq. (19). The average error a is the arithmetical mean of all individual errors without regard to sign, as shown in Eq. (23).

$$a = \frac{\Sigma |x_i|}{n} \tag{23}$$

When the averaging process is carried out with the use of Eq. (19), Eq. (24) is obtained.

$$a = \int_{-\infty}^{\infty} |x| N \, dx = \frac{2b}{\sqrt{\pi}} \int_{0}^{\infty} x e^{-b^2 x^2} \, dx = \frac{1}{b\sqrt{\pi}}$$
(24)

The measure of precision b has already been defined. It is an inverse measure of the width of the

gaussian curve, but it cannot be established without ambiguity from a finite set of measurements X_i . If the most probable value of h is calculated, it turns out to be the root-mean-square error given by Eq. (25).

$$m = \sqrt{\Sigma x_i^2/n} \tag{25}$$

Comparison with Eq. (24) shows that Eq. (26) is

$$a = m\sqrt{2/\pi} \tag{26}$$

valid. The quantity *m* is identical with what was previously called the standard deviation.

The probable error *r* is defined as follows: It marks that value of *x* which divides the area under the curve N(x) between zero and infinity into two equal parts. Hence, *r* is an error such that a given error x_t has an equal chance of being greater or smaller than *r*. Mathematically, the value of *r* is found from Eq. (27).

$$erf(br) = 1/2$$
 (27)

The numerical relations between r, m, and a are given by Eqs. (28).

$$r = 0.6745m = 0.8453a$$

$$m = 1.4826r = 1.2533a$$
 (28)

$$a = 0.7979m = 1.1829r$$

Probability in statistical mechanics. A complex physical system made up of many constituents, for example, molecules, obeys the empirical laws of thermodynamics. Statistical mechanics is that science which attempts to explain these laws by an appeal to the laws of ordinary mechanics. In doing so, it encounters problems of the following sort.

Suppose one wishes to explain why a gas exerts a pressure upon the walls of a container. Pressure is the momentum lost by the molecules per unit time as they strike a unit area of wall. The molecules are very numerous, and their momentum losses vary erratically from instant to instant of collision, and from point to point on the surface. It is necessary, therefore, that some average be taken. Should this average be over the different collisions which a single molecule experiences in time? Or should one take the average at a given time over the whole area under consideration? Clearly, there are many ways of computing the average, each involving a particular collective in the aforementioned sense and each requiring a specification of elementary probabilities.

The science of mechanics does not clearly dictate which collective is the proper one to be used, and there is a considerable latitude of choice. One of these collectives which has proved most successful, the ensemble of J. W. Gibbs, will be discussed briefly.

Phase space. The motion of a single molecule is described in terms of two variables—its position and its momentum at any given time. In the simplest case of motion along a single axis, for example, the *x* axis, two numbers suffice to describe the motion; they are

the position x and the momentum p_x . If a plane is constructed with x and p_x laid off on two perpendicular axes within that plane, the motion of the molecule is represented by a curve in the plane, and the plane is called the phase space of the moving molecule. A molecule whose motion is in a plane requires x, y, p_x , and p_y , that is, four numbers, for its complete description, and these four numbers define a point in a phase space of four dimensions. Similarly, a molecule moving in three-dimensional space has a phase space of six dimensions (axes x, y, z, p_x , p_y , and p_z) and its dynamic behavior is depicted by a curve in this sixdimensional space.

This idea can be generalized and applied to a gas containing N molecules. The phase space of the gas will have 6N dimensions, and the motion of the entire gas corresponds to the trajectory of a single point in this 6N = dimensional space. A single point within it describes the physical condition of the entire gas.

Ensembles. Here Gibbs introduced the notion of an ensemble. He imagined a great number of thermodynamic systems, all similar to the given one. If the latter be a vessel filled with gas, he imagined a very large number of similar vessels, all filled with the same quantity of the same gas. This collection of imaginary vessels is called an ensemble. Each member of the ensemble will have its fate represented by a point moving in 6N-dimensional phase space, and the whole ensemble, when viewed in that space, will appear like a cloud of dust, with each individual dust particle following its own path. The density of this cloud of dust will differ from place to place and will change in time at any given place. From the laws of mechanics it may be shown that this imaginary cloud of dust behaves like an incompressible fluid.

There is a set of conditions, however, under which the cloud will not change its density in time, even though its individual points are in motion. One such condition amounts to the existence of a special density distribution known as the canonical distribution. It has the simple form given by Eq. (29), where *H*

$$D(x_1 \cdots p_n) = \text{ constant} \times \exp\left[\frac{-H(x_1 \cdots p_n)}{kT}\right]$$
(29)

is the energy, T the temperature of the gas, and k is Boltzmann's constant.

This distribution is relevant to the problem discussed above. All dynamical variables, such as the pressure of the preceding example, which need to be averaged in order to correspond to the observables of thermodynamics, are to be averaged over the probability distribution *D*. When this is done, the laws of thermodynamics follow; in that sense, Gibbs' probability distribution provides an explanation of thermodynamics.

The success of Gibbs' theory in classical mechanics is remarkable. In order for it to be applicable, however, to systems which follow the laws of quantum mechanics, certain modifications are necessary. *See* BOLTZMANN STATISTICS; QUANTUM STATISTICS; STATISTICAL MECHANICS; STATISTICS.

Henry Margenau

Probability in quantum theory. Probability enters into both classical and quantum physics, but on very different footings. The toss of a coin provides an example of probability in classical physics. It is possible to learn to toss sufficiently gently, with little spin and the right initial conditions, so as to generally guarantee either outcome, heads or tails. But vigorous tosses with lots of spin produce outcomes beyond practical control. Classically, there is no doubt that the outcome of any particular toss, with its particular initial conditions, is fully deterministic. Rather, what the seeming randomness reflects is that for vigorous tosses the outcome is very sensitive to initial conditions, switching from one result to another with tiny changes in those initial conditions (or tiny changes, too, in external influences such as passing breezes).

In contrast, probability notions enter into quantum mechanics intrinsically, at the very foundations. According to quantum theory, the most that can be known about the state of a system is not enough to determine with certainty the outcome of measurements on that system. That is, repeated experiments on systems in exactly the same quantum state will generally yield a spread of outcomes depending on the particular state and on the dynamical quantity being measured. However, although the outcome of any particular measurement is not predictable, the probability distribution is sharply determined for any given state; and the evolution over time of that state, hence of probability distributions, is fully deterministic. If one of the surprises of quantum mechanics is this probabilistic foundation, another-a feature that gave quantum mechanics its name-has to do with the universe of possible outcomes for measurements of various dynamical quantities. In the case of linear momentum, for example, the possible outcomes range continuously over all magnitudes and directions just as they do classically; so too for position. But for angular momentum, in striking contrast with the classical situation, only certain discrete outcomes are possible; and so too in certain cases for energy measurements. The possible outcomes for certain quantities, that is, are quantized. See ANGULAR MOMENTUM; ENERGY LEVEL (QUANTUM MECHANICS).

But what is most distinctive about quantum mechanics is the rule for composition of probabilities. This can be illustrated with the example of the scattering of a beam of electrons off a small, weakly scattering target centered at some location. Distant detectors record the number of electrons that emerge at various scattering angles, θ , and the angular probability distribution is labeled $P_1(\theta)$. Next, the first target is removed and replaced with another, centered at a new location. The new probability distribution is labeled $P_2(\theta)$. Finally, both targets are set up, each at its previous location (supposing that there is no overlap), and the resulting probability distribution is labeled $P(\theta)$. Classically, an electron detected at any particular angle would have had to scatter off one target or the other, so classically it would be expected that $P = P_1 + P_2$. But this is not what happens quantum-mechanically. Instead, quantum mechanics deals with probability amplitudes, complex quantities whose absolute squares determine the probabilities; and it is the amplitudes to which the simple composition law applies: $A = A_1 + A_2$. This interference of amplitudes is what gives quantum phenomena their wavelike character. *See* INTERFERENCE OF WAVES.

In the case of a single particle moving in a potential, the state of the system, alluded to above, is described by a complex wave function of space (*r*) and time (*t*), $\Psi(r,t)$. The wave function incorporates all that can be known about the state. It evolves in time deterministically, as governed by the Schrödinger equation, a partial differential equation. Quantum theory is concerned with the principles for extracting probabilistic predictions from the wave function and with the application of these principles to various physical problems. To give one elemental example: the probability of finding the particle in an infinitesimal volume element d^3x centered around the position **r** is proportional to $\{\Psi\}^2 d^3x$, where $\{\Psi\}$ is the absolute value of the complex wave function Ψ . The rules for extracting other kinds of probability distributions, for example, predictions concerning the outcome of energy measurements, are equally definite, although mathematically more complicated. See NONRELATIVISTIC QUANTUM THEORY; QUANTUM MECHANICS; SCHRÖDINGER'S WAVE EQUA-Sam Treiman TION.

Bibliography. W. T. Grandy and P. W. Milonni (eds.), *Physics and Probability*, 1993; S. P. Gudder, *Quantum Probability*, 1988; R. W. Hamming, *The Art of Probability for Scientists and Engineers*, 1993; A. Papoulis, *Probability, Random Variables and Stochastic Processes*, 3d ed., 1991; B. P. Roe, *Probability and Statistics in Experimental Physics*, 2d ed., 2001; D. ter Haar, *Elements of Statistical Mechanics*, 3d ed., 1995.

Problem solving (psychology)

Problems represent gaps between where one is and where one wishes to be, or between what one knows and what one wishes to know. Problem solving is the process of closing these gaps by finding missing information, reevaluating what is already known or, in some cases, redefining the problem.

The complexity of human behavior is frequently overlooked, so that everyday activities and experiences are often viewed as uneventful and typical. The regularity of overt human behavior is taken for granted, and people often miss the demands that are made upon cognition by the events and activities in their daily lives. Psychologists, however, investigate the complexity of human behavior and seek to understand its nature and action. Pioneers such as William James realized that behavior was not an isolated response to objects and forces in the environment, but rather represented a systematic adaptation to specific environmental demands and opportunities. One kind of behavior that is created in response to the environment is problem solving. Research into this behavior has attempted to discover the information that is processed and the techniques for processing it.

For many years the main thrust of psychological research on problem solving has focused upon the identification of specific behaviors that enable individuals to reach the goal in a given problem situation. The solution to a problem was thought to consist of an infrequently occurring response to the problem statement. Much of the early work along this line tended to focus on what has come to be known as well-structured problems. These are typically situations with a known beginning, a known end, and a well-defined set of intermediate states. Solving a well-structured problem consists of finding an infrequently used path connecting the initial state of the problem with its end state. One example of such a problem is the game of tictac-toe. Having a well-structured problem does not mean, however, that it will be easy to solve. In a game such as chess the set of possible intermediate states (moves) is enormous, and an exhaustive search through this set in order to find a good move is not feasible. People solve well-structured problems such as chess not by exhaustively searching through the set of possibilities, but rather by heuristically identifying good starting places and productive lines of search.

Two early pioneers in the field of human problem solving were Allen Newell and Herbert Simon. According to their view, people solve problems by first constructing an internal representation, called a problem space. This representation contains the information that must be processed in order for a problem to be solved. The problem space contains knowledge of the initial state and the goal state of the problem as well as possible intermediate states that must be searched in order to link up the beginning and the end of the task. Transitions between one state and another are achieved by selecting and applying operations that process information from memory or the task, in order to reduce the differences between what is given and what is to be achieved.

The structure of a given task helps determine the size and complexity of the problem space that is created to solve it. Characteristics of the human information-processing system, such as limited attention and a fallible memory, constrain the problem space so that the actual set of possibilities considered is usually smaller than the total set available. Simon coined the term satisficing to describe the way in which human problem solvers attempt to limit and redefine a task so that the problem space to be attacked is a simpler version of the one that could be described. While the strategy of simplifying or changing a given problem may seem arbitrary, it is also quite common. The same set of facts and information given to someone from the marketing department in a large corporation will frequently be interpreted as a marketing problem, while when given to a person

in the operations area will appear to be a production problem. The fact that people tend to see new situations in terms of their own past experience enables them to cope with problem complexity. But it also means that behavior is often subject to distortion and bias. Successful consultants and top-level managers recognize this and are constantly on the lookout for evidence of satisficing in their own behavior and in the behavior of others. *See* INFORMATION PROCESS-ING (PSYCHOLOGY).

Arenas of study. One means of discovering how the human problem-solving process is carried out by human beings, regardless of the type of problemsolving task encountered, is based upon observations made through the use of thinking-aloud protocols. This technique requires an individual to talk about whatever comes to mind as a problem is being solved. Analysis of transcriptions of such protocols reveals that for individuals who are relatively inexperienced at a task (novices), the activity of problem solving often consists of general strategies for linking up one stage with another in the search for a solution. One such general strategy is the meansends analysis, which consists of identifying some future point in the task, perhaps the end, and attempting to find a means for reducing the difference between one's current status and that future point. A less powerful though more general strategy of a similar sort is referred to by computer scientists as generate-and-test and by psychologists as trial-anderror behavior. It consists of picking a possible answer, trying it out, and if it does not work, trying another. Means-ends analysis and trial-and-error behavior can require large amounts of time to complete if the problem is complex, and may not lead to a solution at all in a practical amount of time. They have been successful primarily in games and relatively simple problem-solving tasks. The analysis of thinkingaloud protocols reveals that for experts the activity of problem solving involves the use of problem-specific and knowledge-intensive methods and techniques (often referred to collectively as heuristics). These heuristics are acquired through experience and represent the basis for expertise in a specific domain of problem-solving tasks such as chess, physics, or medicine.

While in fairly well-defined problem areas such as chess playing, a solution consists of an attempt to convert an initial state of a task (for example, a chess board before the first move is made) into a final or goal state (checkmate) through a series of actions or problem-solving steps, in many other problem areas either the initial state, the final state, or both may be poorly understood (these are referred to as ill-structured problems). If the necessary problemsolving steps are known ahead of time, problem solving consists of finding the right steps to apply at the right time. If some of the problem-solving steps are not known, then the problem-solving process requires creating or inventing new ways to convert one state of a task into another.

Chess. Following early work using protocol techniques with fairly simple, well-structured problems,

investigators began to branch out to the study of more complex problem-solving situations. One of the earliest domains attacked was skill in playing chess. Enlarging upon early work by Adrian DeGroot, investigators in the United States began to study the differences between grand masters and average chess players. This work has shown that the important characteristics of those especially skilled in a problem-solving domain such as chess is not in a better set of general problem-solving skills, but rather in a vast amount of knowledge that the expert uses to identify important cues and especially good paths to take to a potential solution. Thus in the case of chess, the difference between a grand master and lesser player is not in the number of moves that can be searched in attempting to find a good move, but rather in the set of possible moves considered.

Physics. As in chess, novices (undergraduates) as well as experts (graduate students or professors) have been observed solving textbook problems through the technique of protocol analysis. Study of these protocols has identified two types of problemsolving steps: representing the problem situation, and applying physics principles to the problem representation in order to generate a solution. Representation involves translating the problem statement into a standard form that is accessible to physics principles. For many physics problems, this translation may entail identifying the problem as an instance of some standard or prototype problem (for example, a harmonic oscillator, free fall at constant gravity, or a block on a frictionless plane) that has known solution methods. More generally, translation consists of equating aspects of the problem with idealized objects such as massless rods, frictionless surfaces, and point masses, whose behavior can be represented fairly clearly by a few basic principles. Of course, a given representation can be inappropriate or flatly wrong, and even a good representation may not be tractable (for example, the classic three-body problem in physics).

Once a representation is created, problem solving proceeds by finding physics principles that will generate a solution. In some instances, single principles (such as uniform linear motion) will be sufficient to derive a solution, while in other cases combinations of principles (rotational as well as linear motion) are necessary. In order to be applied efficiently either singly or in combinations, principles of physics need to be selected in such a way that they fit the assumptions and data of a given problem. When several principles are involved, the coordination of their application becomes important. Prototypical problems in physics are those for which single principles suffice, or ones for which the coordination of principles is taught explicitly in introductory textbooks (for example, the simultaneous use of momentum and kinetic energy conservation in an elastic twobody collision). In more advanced texts, procedures are identified for handling additional complexities, and algorithms are derived for dealing with general cases (such as lagrangian and hamiltonian equations in advanced mechanics).

Problem solving sometimes fails because relevant principles are either not known or have not been learned. In such cases, specific heuristics or shortcuts must be employed. In mechanics problems, for example, the goal of problem solving is typically to infer the behavior of a physical system over long spans of time, given only a description of its initial state. One heuristic for accomplishing this goal is to begin with the initial state of the system (for example, a body at rest) and piece out the time course of system behavior by applying interactively the laws that say what happens in the next instant of time (this process is often referred to as simulation). A second heuristic for specifying the behavior of a physical system is to integrate the differential laws of motion to yield an expression that describes the state of the system at any point in the future as a function of its initial state. Because mathematical complexities often make either of the above heuristics difficult to carry out in practice, additional heuristics are sometimes employed. One such heuristic involves describing steady-state solutions of the system (the places where the laws of motion state that nothing changes). Another heuristic involves describing asymptotic behavior of the system (what happens to it in the long run). Yet another heuristic is to specify a property of the system (such as energy or momentum) that is invariant over time, even though there is variation in other aspects of the system. These heuristics (and others like them) are typically learned and applied by individuals in specific problem-solving tasks. Indeed, the power of heuristics as problemsolving tools derives from their development in the successful solution of particular types of problem situations.

The use of fairly elegant heuristics (such as those described above) is often associated with level of expertise; they are used more often by experts than novices. Such an association does not mean, however, that individuals at more advanced stages of training and experience do only heuristic thinking, nor even that this is their preferred method of solving problems. Like other cognitive activities, problem solving requires the expenditure of mental effort. Because individuals at all levels of expertise seem to prefer to expend the least amount of effort that will do the job, experts as well as novices generally solve problems by relating them to their past experience in the form of problem-solving examples (often referred to as prototypes). It is only when truly unfamiliar or difficult problems are encountered that the expert resorts to the more demanding skills of heuristic thinking, which are for the most part unavailable to novices.

Medicine. Many of the features just described for solving problems in chess and physics can also be found in the behavior of individuals attempting to solve the even more ill-structured problems of medicine. Like problem solving in physics, medical problem solving has been studied intensively by using techniques for analyzing thinking-aloud protocols. Whereas in physics equal amounts of effort are often expended on problem representation and solution generation, in medicine one aspect of problem solving deals almost exclusively with finding a good representation (the problem of medical diagnosis).

Solving a diagnostic problem in medicine seems to consist of applying hypotheticodeductive reasoning, in which the basic units are questions, cues, and diagnostic hypotheses. In order to obtain a good representation, diagnosis typically begins with the elicitation of a small number of cues (the patient's signs and symptoms) elicited via questions (usually through the taking of a patient history). These cues suggest a limited number of diagnostic hypotheses which, in turn, suggest new revised hypotheses. Eventually, each hypothesis is evaluated with respect to its relative ability to account for the available cues. Physicians at all levels of skill appear to generate hypotheses in similar quantities, to use those hypotheses to guide collection and interpretation of additional patient data, and to evaluate the relative ability of the hypotheses to account for those data. Differences between experts and novices (and among members of each group) occur in the detailed content of the medical knowledge they possess and in the way it interacts with the specific features of the case they are working on.

At a detailed level, the activity of medical problem solving is guided by specific representations of the problem situation. Images of "what the problem could be" enable forward-searching strategies for the collection of additional information as well as heuristics for generation and evaluation of possible categories within which to interpret this information. Such strategies propose a large number of recurring patterns or prototypical combinations of data (such as combinations of pieces in chess, or combinations of patient signs and symptoms in medicine) that are familiar and can be recognized. The organization of these prototypes is hierarchical, with the more general patterns at the top and the most specific ones at the bottom. Such patterns are stored in memory in the form of templates to be matched against external data, so that a particular configuration of data can be interpreted as an instance of a given pattern or prototype. For the expert, access to a prototype occurs at multiple levels in the stored hierarchy (top, middle, bottom) and is a function of other patterns (pattern-to-pattern links) and data (data-to-pattern links). Unfamiliar experiences are interpreted by building new prototypes. This occurs either through a form of data-driven processing in which commonalities are abstracted from recurring patterns or cues, or through a more conceptually driven form of processing in which higher-level rules are used to derive patterns that can then be matched with unfamiliar data.

General skills. Although much of the work on problem solving today proceeds on a field-by-field basis, some investigators continue to search for a set of general problem-solving skills. Such work suggests that if a more generalized set of skills for solving problems exists, they have much in common with the basic skills of learning. Part of what it means to learn, is to be able to solve the problem of linking up a given situation with some specific goal or outcome.

Attention as a resource. Attention is a psychological process that plays a major role in the development of efficient and skilled performance in many complex problem-solving tasks. One definition of attention is simply a mechanism with a limited capacity for processing information. In this sense, attention provides a bottleneck through which information about the environment, and the interpretation of that information, must pass. Attention can also be thought of as an active process for allocating cognitive processing resources to components of a task. According to this view, problem-solving performance is a function of the demand for processing resources by the mental operations that are applied to achieve a given problem solution. Complex problems in which large amounts of information or knowledge must be processed require efficiency in order to overcome basic human processing limitations. Therefore the performance that is observed in less experienced individuals is often a function of their attempt to cope with limited resources and an overabundance of problemsolving information.

An important characteristic of problem-solving skill is the manner in which experts circumvent their processing limitations. Skilled performance results when individuals are able to focus upon a small set of cues and relevant task information. Skilled performance also results from automatic control over the type of processing resources used. An expert who has developed skill in solving a specific class of problems typically reduces the processing demands of the task by changing an information-processing step that requires more of the available resources into one which requires less but is just as effective (for example, rather than figuring out a complex problem-solving step each time it occurs, the expert simply remembers it). Efficient information processing is an important component of expert skill that is developed through extended practice in specific problem-solving activities.

An example of cognitive efficiency in problem solving is provided by the skill of learning to read. The beginning reader must give attention to discovering and differentiating many visual features such as the lines and angles that compose the letters printed on a page. After repeated attention to such features, the reader begins to code patterns of features as a single unit (a letter). With enough exposure and practice, the reader eventually learns to recode letters into higher-order units such as words and phrases. As these higher-order units are developed, the lowerorder units no longer require conscious attentional processing-they are automatically coded. Such automatic processing allows the skilled reader to allocate attentional capacity to the more complex task of comprehension and deriving meaning from the printed material. See COGNITION.

Garden path problems. The fact that experts achieve their expertise in part through an efficient allocation of problem-solving resources is a two-edged sword. Research has identified classes of problems

in which errors regularly occur, even among fairly expert problem solvers, due to the misapplication of attention. The usual desirable strategy of beginning a task with fairly general concepts and then working toward more specific levels of detail can fail when the structure of a problem is such that the initial choice of a representation either does not apply or else fails to emphasize the kind of knowledge required for a successful solution. Some types of problem situations have the additional property that the dominant interpretation they elicit appears to be correct when in fact it is not. Problems of this type are referred to as garden path problems since they induce performance that has the false appearance of success.

One famous example of a garden path problem is from physics and is attributed to the work of Harvey Cohn. In Cohn's original problem, termed Milko, a milk bottle is filled with a mixture of milk and cream and left to stand. After a time, the milk and cream separate and the cream rises to the top without any change in volume in the container. Cohn asked simply whether the pressure at the base of the bottle is the same or different after the separation of the two fluids. Expert scientists and engineers as well as more novice problem solvers often fail to correctly solve the Milko problem.

Difficulties with Milko appear to be due primarily to an improper choice of a problem representation as well as an incorrect level of detail in reasoning. Pressure change is thought to be associated with change in either the height or amount of a fluid above the base of its container. And since neither of these has changed in the Milko problem, individuals infer that there has been no change in pressure (the garden path answer). What problem solvers fail to consider, however, is that the situation deals with two variables-a nonhomogeneous fluid and a container with slanted sides. If one of these variables is present by itself, there is no pressure change, something which is described in many introductory physics books as Pascal's principle. If both variables are present, however, the pressure at the base of the container is less following the separation of the two fluids. It is interesting to note that if individuals working on the Milko problem are given a second problem in which the shape is made explicit by drawing a flask with slanted sides, the more expert individuals will often recognize their error in the first problem and proceed to do both problems correctly. For some experts and most nonexperts, however, the illusion that the pressure does not change is so great that a third problem is necessary, one in which explicit values are given for the variables of shape and fluid density so that an exact calculation can be made. Faced with calculations showing that the pressure in the separated state is indeed less, experts not only recover and adopt the correct problem-solving approach but also frequently generalize this recovery both forward and backward. They immediately infer that answers to previous problems were wrong, and additional problems of the same type are correctly solved. Novices, by contrast, even when given an opportunity to make calculations, typically do not generalize either to the previous problems or to new ones, and will frequently abandon the correct approach when given another problem.

Garden path problems, due mainly to the allocation of attention to one thing while overlooking another, exist in all fields. In medicine, for example, the possible diseases for a given set of signs and symptoms often have information in common. Because data that trigger the correct disease also trigger others, the wrong hypothesis may be chosen. In fact, in some areas of medicine, research has shown that the early cues in a case inevitably trigger the wrong answer so that the expert must avoid overcommitting processing resources to an early, incorrect response in favor of a later, correct alternative.

Computer applications. Scientists working in the field of artificial intelligence have developed computer programs that are capable of solving problems long thought attackable only by human experts. Computer programs exist that are capable of solving problems in medical diagnosis, oil exploration, and engineering design. Having computer programs that can solve complex, ill-structured problems has made it possible to turn understanding of expert problem solving into a useful resource. Computer programs (usually called expert systems) are currently being developed to give advice to problem solvers working in hospitals, industry, and government.

In addition to potential for providing resources for problem solving and training, programming computers to solve problems has served to highlight some interesting facts about the human problem-solving process. Among the most important of these is the difference that exists between problem solving that proceeds on the basis of recognizing situations as instances of ones that have been dealt with in the past (the use of prototypes), and problem solving that proceeds by reasoning through a solution in terms of basic facts and principles from beginning to end. The former permits great efficiency and economy of thought. Through experience, individuals build up a vocabulary of problem-solving prototypes that represent recurring problem situations. It is these prototypes that form the basis for the ability to recognize new problems as instances of ones that have been solved in the past. On the other hand, research also has shown that these prototypes often serve as a source of misinterpretation, bias, and error.

As mentioned at the beginning of this article and as illustrated with the Milko problem in physics, a housing problem may be interpreted by one individual as landlord exploitation, by another as tenant neglect, and by a third as poverty or racial discrimination. When used to solve problems, prototypes have both descriptive and normative functions. These functions are guided by the metaphors contained in the prototypes. Such metaphors allow one to describe and understand complex problem-solving situations. Many social problems, for instance, can be seen in terms of a health metaphor. Families are sick, housing is decayed, or neighborhoods are unhealthy. A health metaphor may suggest different solutions than a need metaphor. Because prototypes can lead to alternative interpretations of a given problem-solving situation, problem solvers need to be aware of them. By programming computers to solve problems, researchers are learning how to achieve a more balanced and meaningful formulation of a given problem-solving task. *See* ARTIFICIAL INTELLIGENCE.

Instruction. Not surprisingly, teaching problemsolving skill has been a topic of interest to psychologists and educators alike. The biggest payoff for instruction from current research in problem solving has been simply learning more about what task-specific knowledge and strategies need to be learned and where garden paths and other pitfalls may lie as the skills of problem solving are acquired and practiced. *See* LEARNING MECHANISMS; MEMORY. Paul E. Johnson

Bibliography. A. DeGroot, *Thought and Choice in Chess*, 2d ed., 1978; A. S. Elstein, L. S. Shulman, and S. A. Sprafka, *Medical Problem Solving*, 1978; J. G. Greeno, Trends in the theory of knowledge for problem solving in D. T. Tuma and F. Reif (eds.), *Problem Solving in Education*, 1980; J. R. Hays, *The Complete Problem Solver*, 1989; P. E. Johnson, What kind of expert should a system be?, *J. Med. Phil.*, 8:77-97, 1983; J. Larkin et al., Expert and novice performance in solving physics problems, *Science*, 208:1335-1342, June 1980; A. Newell and H. Simon, *Human Problem Solving*, 1972.

Proboscidea

An order of placental mammals containing the largest and most powerful living terrestrial mammals—the elephants. Proboscideans once had a wide distribution, reaching every continent except Australia and Antarctica. Although early forms lacked both significant tusks and a proboscis, they are still formally considered proboscideans. *See* ELEPHANT; MAMMALIA.

Classification. The African elephant (*Loxodonta africana*; **Fig. 1**) historically occurred throughout Africa from the Mediterranean Sea to the Cape of Good Hope, except in parts of the Sahara and some



Fig. 1. African bush elephant (Loxodonta africana). (Photo by Gary M. Stolz/U.S. Fish and Wildlife Service)

other desert regions. Most researchers agree that there are two kinds of African elephant: a smaller, darker forest elephant found in the tropical rainforests of West and Central Africa and a larger, paler bush or savannah elephant in the remainder of the current range. Some feel that these two forms should be regarded as separate subspecies—*L. a. cyclotis* for the forest elephant and *L. a. africana* for the bush elephant; others feel that the forest elephant should be a separate species. In addition, some researchers have identified a pygmy elephant (*L. pumilio*) in the dense lowland jungles from Sierra Leone to Zaire. However, these small forms may just be unusually small adult *L. africana*.

The Asiatic or Indian elephant (*Elephas maximus*; **Fig. 2**) historically occurred from Syria and Iraq east across Asia to Indochina, the Malay Peninsula, China, Sri Lanka, Sumatra, and possibly Java.

A third member of this order is the woolly mammoth (*Mammuthus primigenius*). It once occurred throughout most of Europe and the northern halves of North America and Asia. Mammoths survived up to about 3700 years ago on Wrangel Island in the Arctic Ocean off extreme northeastern Siberia. Intact carcasses have been recovered from permanently frozen ground in Alaska and Siberia. Mammoths had a dense covering of hair and small ears (both adaptations to prevent heat loss), and a trunk that ended in two fingerlike projections at its tip.

Characteristics and description. Proboscideans have a proboscis (trunk), enlarged incisors that form paired tusks, the largest ears of any mammal, and pillarlike legs whose structure prevents the elephant from jumping even a short distance. There are five toes on each forefoot and either three or four toes on each hindfoot. The nose and upper lip are drawn into the long trunk, which is used as a hand. The trunk measures about 6 ft (2 m) in length and weighs about 300 lb (140 kg). It is a boneless



Fig. 2. Asian elephant (*Elephas maximus indicus*) in Kerala region, South India. (*Photo by Nidish K. Viswanathan*)

mass of flesh and muscles that bends easily. With its trunk the elephant can deftly pick up a peanut or a 600-lb (275-kg) log. The elephant smells, drinks, and feeds itself with its trunk. It depends more on smell than on any other sense. The trunk is used to carry food and water to its mouth. Water is sucked up by the trunk and squirted into the mouth. Elephants feel with the sensitive tips of their trunks. They can tell the shape of an object, and whether it is rough or smooth, hot or cold. They also use their trunk to caress their mate and their young. In a fight, the trunk is used to grasp an enemy. The tusks are modifed incisors that are made of ivory. About two-thirds of each tusk sticks out of the upper jaw; the rest is in the skull. The tusks, which continue to grow throughout the life of the animal, are used to dig for food and to fight. The molars, or grinding teeth, are each at least 1 ft (30 cm) in length and weigh about 8.5 lb (4 kg) each. The elephant has only four of these teeth at any one time. One tooth is on each side of each jaw. New molars form in the back of the mouth and push forward as the old teeth in front wear down and drop out. During its lifetime, an elephant grows six sets of four molars each. The last set comes in when the animal is about 40 years old.

Elephants are covered with a thick gray hide that is leathery and tough. Because their skin is 2.5 cm (1 in.) thick, elephants are often referred to as pachyderms. In spite of the thick skin, an elephant is very sensitive to cold. Even a slight frost will give it a severe case of cramps. The short stiff hairs on the body may be spaced as much as 1 in. (2.5 cm) apart, and many wear off, resulting in the skin of some animals appearing almost hairless. The enormous ears [3-4 ft (0.9-1.2 m) wide and up to 5 ft (1.5 m) long] serve as fans and are kept constantly in motion during the heat of the day to circulate the air. Elephants have small eyes and poor eyesight, long eyelashes [over 5 in. (13 cm) long], and excellent hearing. They are strictly herbivorous and feed on grasses, leaves, small branches, bark, berries, mangoes, coconuts, corn, and sugarcane. A wild bull may consume 500 to 600 lb (230 to 270 kg) of food per day.

Midway between the elephant's eye and ear lies the temporal gland. It gives off a dark oily material that has a strong smell and stains the elephant's face. The temporal gland becomes active when the elephant is in a condition called musth. At these times, the animal is either dangerously violent or very quiet. Musth occurs in all male elephants and in some females. Some scientists believe it takes place during mating peroids. *See* GLAND.

Differences between African and Asian elephants. How do African and Asian elephants differ? African elephants are taller and heavier. An adult male (bull) is about 11.5 ft (3.5 m) tall at its shoulder, and weighs from 12,000 to 14,000 lb (5440 to 6350 kg). The largest known African bull stood 13 ft 2 in. (4.0 m) tall. An adult female (cow) weighs from 8000 to 10,000 lb (3600 to 4500 kg). An Asian bull stands about 9 ft (2.7 m) tall and weighs from 7000 to

12,000 lb (3200 to 5400 kg). An Asian cow is about 7.5 ft (2.3 m) tall and weighs from 5000 to 10,000 lb (2300 to 4500 kg). The trunks, tusks, and ears of African elephants are much larger than those of Asian elephants. The trunk of an African elephant has two fingerlike knobs on the tip, whereas only one knob is present on the trunk of an Asian elephant. African elephants have slightly hollowed backs; the backs of Asian elephants are arched. African elephants have a more rounded, sloping forehead, whereas the Asiatic elephant has a bulging, protuberant forehead. Most African elephants are dark gray; most Asian elephants are light gray. Both male and female African elephants possess well-developed tusks. Those of a bull grow from 6 to 8 ft (1.8 to 2.4 m) and weigh up to 80 lb (36 kg). The tusks of an Asian bull may grow 4 to 5 ft (1.2 to 1.5 m) long, although the cow's tusks are always small. African elephants are fiercer and more difficult to train.

Behavior. Wild elephants live in herds of 10 to 50 or more. The leader of the herd is usually a female in this matriarchal society. They may travel long distances in search of food or water or to escape pests. They usually travel in single file at a speed of about 6 mi (10 km) per hour. When going down steep slopes, they slide on their bellies, with the back legs stretched out behind and the front legs extended forward. They are good swimmers and often bathe in lakes and rivers and wallow in muddy water.

Reproduction and development. Elephants have the longest gestation period of any animal—20 to 22 months. Pregnancy normally results in one offspring (calf), although occasionally there may be twins. At birth, an elephant weighs about 200 lb (91 kg) and stands about 3 ft (91 cm) tall. The calf nurses by curling its trunk over its head so its mouth can reach the mother's breasts. The calf sheds its milk tusks at 5 or 6 months after birth, but it continues to nurse for another 2 years and remains under its mother's care for 2 years longer. Elephants reach sexual maturity at about 14 to 15 years of age, and attain their adult size when about 20 years old. Most elephants live for about 60 years.

Elephants have few enemies. Tigers in Asia kill calves. Humans are the main enemy through poaching and habitat destruction. Donald W. Linzey

Bibliography. D. Macdonald, *The Encyclopedia of Mammals*, Andromeda Oxford, 2001; R. M. Nowak, *Walker's Mammals of the World*, 6th ed., Johns Hopkins University Press, 1999.

Procellariiformes

A large order of strictly marine birds found far offshore except when breeding. The procellariiforms, or tube-nosed swimmers, are most closely related to their descendant group, the penguins. *See* SPHENIS-CIFORMES.

Classification. The Procellaritiformes comprise four families: Diomedeidae (albatrosses; 13 species; southern and Pacific oceans); Procellariidae (shearwaters, petrels, and fulmars; 66 species; worldwide); Hydrobatidae (storm petrels; 21 species; worldwide); and Pelecanoididae (diving petrels; 4 species; subantarctic oceans). Some taxonomists have enlarged the Ciconiiformes to include the Procellariiformes, as in the Sibley-Ahlquist taxonomy (a radical new approach to taxonomy based on DNA hybridization studies that differs greatly from the traditional approach). *See* CICONIIFORMES.

Fossil record. The fossil record of procellariiforms is rather incomplete. Most interesting is that a record of fossil albatrosses is known from North America and Europe from the middle Miocene into the lower Pleistocene, even though albatrosses are not now found in the Atlantic Ocean north of the cold waters of the South Atlantic. Their disappearance from most of the Atlantic is, therefore, a recent Quaternary event, and likely related to the lack of suitable nesting islands in the Atlantic Ocean; a fossil albatross colony was recently described from the Quaternary of Bermuda. Few shearwater (Procellariidae) and storm petrel (Hydrobatidae) fossils are known, and little is known of the evolutionary history and past distributions of these groups. One diving petrel fossil dates from the Pliocene of South Africa, where the group is unknown today.

Characteristics. The tube-nosed swimmers are characterized by having their nostrils enclosed in a tube, which is paired in albatrosses; dense plumage; webbed feet, and long wings. They range in size from the sparrow-sized storm petrels to large albatrosses (see illustration), which have the greatest wingspan of all living birds, up to 12 ft (3.7 m). The procellariiforms are excellent fliers, as evidenced by the migratory-actually nomadic wanderingflights of many thousands of miles. Albatrosses and shearwaters are dynamic soarers, when their wings are held rigidly away from the body to use the updrafts along wave crests to power their flights. Their flapping flight is slower and more awkward. Their plumage, which varies from white through shades of gray to black, is generally somber but sometimes has a bold pattern. Procellariiforms swim well, and the pelecanoidids and some shearwaters can dive



Black-footed albatross (Phoebastria nigripes). (Photo from Olympic Coast National Marine Sanctuary/National Oceanic & Atmospheric Administration)

under water, using their wings for propulsion. They are highly pelagic (oceanic) and feed on fish, squids, and crustaceans, which are concentrated into oils and waxy esters in their stomach for transport and feeding to their young, which can digest waxes. Procellariiforms have a well-developed olfactory sense to locate food and apparently to locate their nesting burrows at night.

The larger species mature slowly; some albatrosses begin breeding only after reaching 6 to 8 years of age. Tight pair bonds are formed during courtship, which can take elaborate forms, as is seen in albatrosses. Their nests can be solitary or part of a colony and may be built on the ground as is done by albatrosses, or in burrows as in many shearwaters and storm petrels. The downy young stay in the nest and are cared for by both parents. Development of the young is slow, but the fully grown youngster is very fat, often weighing more than the adults. The breeding season is precarious for procellariiforms, and most species breed on isolated islands. A reduction in numbers has usually resulted from the introduction of mammals such as rats and mongooses which prey on the eggs and young. After the young have learned to fly, they and the adults leave the breeding area and live at sea, usually far from land, until the next breeding season. Some shearwaters and even the small storm petrels follow erratic routes that take them from their subantarctic nesting islands to subarctic oceans. Wandering albatrosses (Diomedea exulans) may fly around Antarctica between breeding seasons.

The diving petrels are confined to the colder waters of the southern oceans, and are generally found year-round in the regions of their nesting grounds. They are convergent in structure and habits with the auks (Alcidae) of the Northern Hemisphere. *See* CHARADRIIFORMES.

Economic importance. Aside from being a welcome sight at sea, the procellariiforms have little contact with humans. Several species of tropical albatrosses, however, have been hunted almost to extinction for their feathers and for oil from their fat. A number of species of shearwaters and petrels are seriously endangered or have become extinct because of the introduction of predatory rodents and mongooses onto their breeding islands. *See* AVES; ENDANGERED SPECIES. Walter J. Bock

Bibliography. M. Brooke, *Albatrosses and Petrels across the World*, Oxford University Press, 2003; J. Warham, *The Behaviour, Population Biology and Physiology of the Petrels*, Academic Press, 1996; J. Warham, *The Petrels: Their Ecology and Breeding Systems*, Academic Press, 1990.

Process control

A field of engineering dealing with ways and means by which conditions of processes are brought to and maintained at desired values, and undesirable conditions are avoided as much as possible. In general, a process is understood to mean any system where material and energy streams are made to interact and to transform each other. Examples are the generation of steam in a boiler; the separation of crude oil by fractional distillation into gas, gasoline, kerosine, gas-oil, and residue; the sintering of iron ore particles into pellets; the polymerization of propylene molecules for the manufacture of polypropylene; the incineration of waste; and the roasting of green coffee beans.

In a wider sense, process control also encompasses determining the desired values. In the narrow sense, it merely refers to keeping process conditions as close as possible to these values, which generally is realized by automation. Modern instrumentation for process control can hardly be thought of without process computers and visual display units.

Process control in the wider sense includes a number of basic functions, which can be arranged as follows:

> Scheduling Mode setting Mode correction and quality control Regulatory control Sequence control Coping with off-normal conditions

Generally, not only do these basic functions determine the performance of conventional process control systems, but they are also predominantly embedded in more advanced integrated intelligent process control systems.

From Conventional to Integrated Intelligent Control

In conventional process control systems, the basic functions are often organized into a hierarchy. At the lower levels, data acquisition is performed directly from the dynamic technical process. Also, basic regulatory control and sequence control are carried out mainly on these levels. Mode correction and quality control as well as coping with off-normal conditions such as fault diagnostics belong to the intermediate levels of the hierarchy. The higher levels include mode setting and scheduling. This bottom-up view of process control is more data-oriented whereas the top-down view emphasizes the goal-oriented perspective.

Advanced process control systems combine the data and the goal orientations by means of a network of data and knowledge flows for achieving integrated intelligent control. The hierarchies of processes, functions, and tasks become less stringent. More overlapping and duplicating functionalities are introduced. The advanced technologies with extended functionalities include expert systems for control and diagnostics, fuzzy process control, and artificial neural networks.

It is important to ensure an integration between conventional control systems and these advanced technologies for the purpose of reaching the overall goals of efficient and safe systems performance as well as sustained and satisfying human operator performance. Thus, the functionalities of well-designed human-machine interaction have gained predominance in advanced process control. Another extension beyond the conventional control hierarchy is the integration of the advanced process control systems into overall plant control and business management. This organizational extension goes beyond planning and scheduling and relates to the next higher levels of the hierarchy. Thereby, integrated intelligent process control systems are embedded into plantwide control and communication. These developments are changing process control toward human-centered automation. *See* COMPUTER-INTEGRATED MANUFACTURING.

Basic Functions

The basic functions listed above will now be discussed.

Scheduling. At the highest level of the basic control hierarchy, the schedule for process operation is determined, that is, the sequence of the required modes of operation. This basic function indicates, for example, whether product A will be made before B, or vice versa, and when the plant will be shut down for general maintenance. In some cases, scheduling is very important, for example, for continuous distillation units processing a variety of crude oils into different series of intermediate products; in other cases it is practically nonexistent, for example, for the manufacturing of standard solvents. The primary goal is to obtain a feasible schedule, that is, a schedule which does not overflow or empty storage vessels for feedstocks and (intermediate) products. Only if a feasible schedule is possible can one try to find an optimum schedule.

Scheduling can be realized by computer algorithms or by human search-and-try, possibly with the aid of computer simulation. *See* OPERATIONS RE-SEARCH; SIMULATION.

Mode setting. At the next level in the hierarchy, the desired values are determined for each individual mode of operation. When little is known about the process operation, the mode is characterized by a recipe, obtained from experience. In cases where knowledge is sufficient to set up a mathematical model, optimal desired values can be computed by a suitable optimizing algorithm. Optimization corresponds to maximization of an objective function, which, for continuous processes, might have the form of the following equation:

$$J = \sum_{i} p_i P_i - \sum_{j} f_j F_j - \sum_{k} u_k U_k - \sum_{l} w_l W_l$$

Here P_i = production rate of product *i*; F_j = consumption rate of feedstock *j*; U_k = use rate of utility *k* (electricity, fuel, cooling water, and so forth); W_l = additional wear rate of process apparatus *l*; and p_i , f_j , u_k , w_l = corresponding unit values and costs. See NONLINEAR PROGRAMMING; OPTIMIZATION.

For integrated production processes, where the product of one process is the feed for another, it is rather difficult to find representative figures for these unit values and costs.

Maximization of the above equation can be interpreted as a compromise between maximum yield of

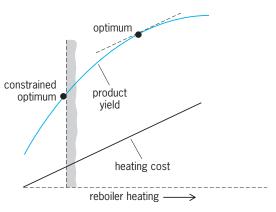


Fig. 1. Optimum operation of distillation column.

valuable product, minimum consumption of raw material, minimum use of energy, and minimum wear of process apparatus.

Figure 1 shows a simple example. It is the compromise between product yield and reboiler heating for a simple distillation column. More heat costs money but increases the rate of valuable product as well. The optimum is at the point where the marginal value and cost are equal, unless this point is above the maximum capacity of the column. Then, of course, the constrained optimum is at maximum capacity.

The optimal values need not always be the same; they can vary because of changes in feed properties, process load, fouling of process equipment, ambient temperature, and so forth. Here an on-line computer (directly coupled to process sensors) is a suitable device for the computations. Expert systems with their extended knowledge-based functionalities may be applied. The results can be shown to the process operators (advisory control), or they can be automatically transmitted to the automatic control system (supervisory control).

Other examples of constrained optima are controlling the catalyst temperature in a chemical reactor near the sintering point for maximizing product yield, and setting the condenser "wide open" in order to minimize heat requirements in distillation. *See* DISTILLATION.

Special cases are mode switches. For instance, a polyethylene reactor has to be switched from producing a lower molecular weight to producing a higher one. The purpose is to minimize the quantity of intermediate material, that is, to move as rapidly as possible from one to the other mode of operation. Here the optimum desired values depend on the time; they follow a path which can be found by dynamic optimization techniques. *See* CALCULUS OF VARIATIONS.

Analogous results are obtained for batch processes where the sequence of operation steps is done in the same process apparatus. For instance, a batch chemical reactor is started with maximal heating, followed, at a critical moment of time, by a switch to maximum cooling in order to reach the desired reaction temperature without overshoot in minimum time. Some continuous processes appear to work more efficiently when they are subjected to cyclic variations than when under equilibrium conditions. Evidently, the optimum desired values here are a periodic function of time.

Mode correction and quality control. The next level in the hierarchy is quality control. In some cases, product quality can be completely specified by concentration, for example, less than 1% impurities. Alternatively, some physical property is a good measure of quality, for example, density or average molecular weight. In still other cases, a performance test has to be done, which bears some resemblance to the way in which the customer uses the product. A straightforward example is the octane number for gasoline, which is determined in a standardized internal combustion engine.

Traditionally, quality measurements are done in a control laboratory by skilled analysts. This incurs a rather long delay, as samples have to be brought to and routed inside the laboratory. A faster response time is achieved by local semiautomatic instruments which can be handled by process operators. Still better results can be obtained with completely automatic quality analyzers, which take their own samples. However, these instruments are available only for the more common and straightforward applications. Moreover, they are rather expensive and require much highly skilled maintenance. Consequently, in practice, quality control often has to be done with sporadic and late quality data. In between, process operators try to estimate product quality changes from the values of more easily measured variables. They can be helped by the on-line computer, provided some type of estimation model is available (Fig. 2).

In discrete-parts manufacturing, statistical techniques are commonly applied to supervise product quality. Applications in the process industries are being generalized to statistical process control as a tool to assist the operator in correcting mode settings. However, superior results can be expected from estimation techniques (Fig. 2). *See* QUALITY CONTROL.

Regulatory control. Most process variables can be measured with inexpensive and reliable sensors. Common ones are temperature, pressure, level (liquid or solid), flow rate, and rotational speed. It is customary to maintain important process variables of these types near desired values by automatic control.

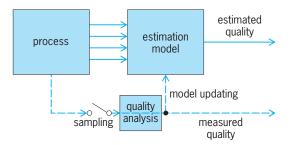


Fig. 2. Estimation of product quality by on-line computer.

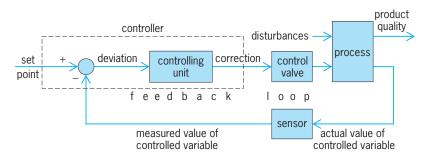


Fig. 3. Block diagram illustrating feedback loop around process.

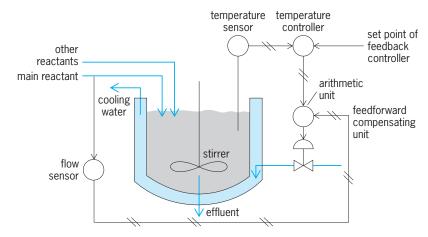


Fig. 4. Feedback and feedforward control on continuous stirred chemical reactor.

Feedback control. The most commonly used method is feedback. In an automatic feedback control system (Fig. 3), a process variable such as the temperature in a continuous stirred chemical reactor (Fig. 4) is to be maintained at a desired value; hence it is called the controlled variable. It is measured by a sensor, for example, a thermocouple. The measured value is compared with the set point, which can be manually adjusted. The difference between the set point and the measured value, called the deviation, is the input of the controlling unit. Here a correction is generated that should eliminate, or at least reduce, any difference between set point and measured value. The correction influences an input of the process, the manipulating variable (in Fig. 4 the rate of cooling water supplied to the reactor jacket) via a control valve. See THERMOCOUPLE.

The process operator still is the master. This worker is relieved from the task of continuously introducing corrections, but still retains the overall supervision. When, for instance, the product quality deviates from the desired value, the process operator will adjust the set point in order to reduce quality deviations.

When a reliable automatic quality analyzer is available, the operator's supervisory action can be partly replaced by a master control loop. The result is a so-called cascade control system, with the accompanying reduction in the operator's work load.

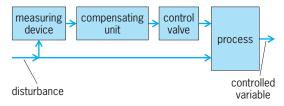


Fig. 5. Block diagram of typical feedforward control.

Feedforward control. Feedforward control is another principle, which is often applied to supplement feedback control (**Fig. 5**).

A process variable that is a source of disturbances (in Fig. 4 the feed rate of the main reactant) is measured, and the measured value is made to influence the control valve via a compensating unit. The compensating unit is adjusted in such a way that the direct influence of the disturbance on the controlled variable is balanced by the indirect influence (through the measuring device, compensating unit, and control valve). In Fig. 4 an increase in main reactant rate will gradually increase the reaction rate in the reactor, hence generate more heat of reaction. A gradual opening of the control valve introduces more cooling, which takes care of removal of the additional heat of reaction; hence the temperature remains constant.

With feedforward control, perfect compensation of disturbance effects is possible in principle. (With feedback control this cannot be achieved because the controlling unit can only generate a correction when there is a deviation.) Perfect compensation, however, requires detailed knowledge of the process dynamics (that is, the way in which processes respond to disturbances and corrections as time progresses). Any significant mismatch between the direct and the indirect influences causes rather severe deviations (100-80 instead of 100-100 makes 20 instead of 0). This does not hold for feedback control, where even large changes in process dynamics have only little effect on performance (a change of 20% does little harm). Another disadvantage of feedforward control is that the designer must provide compensating units for all important sources of disturbances. In view of these difficulties, it is usually more practical to use a combination of feedforward and feedback control, whereby the former helps the latter in reducing the effects of the most important disturbances.

Features of automatic regulatory control. Automatic regulatory control has the following features:

1. It relieves the process operator of tedious timeconsuming control tasks, enabling a few people to supervise and control large production plants.

2. Compared to manual regulatory control, performance is more consistent and reliable. This has led to plants with smaller residence times and less buffering, thus decreasing costs and improving safety (smaller amounts of materials under hazardous conditions).

3. Via a judicious choice of controlled variables,

product quality variations are reduced, which simplifies quality control.

4. Reduction of the effects of disturbances decreases the chance that the plant runs into off-normal or unsafe conditions.

Sequence control. What regulatory control does for continuous process operation, sequence control does for batch-type operation. In its simplest form, sequence control resembles the actions of the program in a washing machine: at predetermined moments of time, valves are opened or closed, motors are started or stopped, and so on.

In more sophisticated versions, the timing is not fixed, but depends on process conditions: in a chemical reactor, catalyst injection is started after the reactor contents have reached a certain temperature, which may take a shorter or longer time. Further, logical conditions are included too: a centrifugal pump is started only if the suction valve is open and the discharge valve is closed.

Automatic sequence control can be realized with programmable logic controllers (PLCs; a digital device which performs logical operations). It enables one process operator to supervise a series of batch processes, relieving the worker of repetitive tasks. Moreover, the accurate reproduction of mode recipes enhances constancy of product quality. *See* PROGRAMMABLE CONTROLLERS.

Automatic sequence control is also being applied in continuous processes for operations like washing of filters, regeneration of catalysts, and switching of heat exchangers. In some processes, for example, electric power generation, automatic sequence control also takes care of startup and shutdown.

Coping with off-normals and faults. This level is of vital importance for safety and plant availability. It is devoted to the detection, diagnosis (identification, interpretation), and compensation of off-normal process conditions and faults in process equipment and instrumentation. Diagnostic functionalities constitute the most important and most difficult phase of fault management.

Well-protected plants have five barriers against hazardous situations:

1. Automatic regulatory and sequence control systems reduce deviations from desired values, hence decrease the probability of undesired values.

2. The operator observes trends in important process variables in order to detect and identify the start of off-normal conditions.

3. If the previous barriers have been ineffective, an alarm signal is generated, usually in auditive and visual form, to alert the operator. The choice of the alarm limit is a compromise between allowing the process operator time to take action and causing too many false alarms.

4. If the operator does not take action, an automatic protection system shuts down the plant or brings it to a safe condition (standby).

5. If the automatic protection system fails to act, the process operator is the last line of defense. The first three barriers are intended to keep the

process going, while the last two barriers resort to standby or shutdown, with accompanying loss of production.

Advanced Technologies

The extended functions of advanced process control that were mentioned above will now be discussed.

Expert systems. In integrated intelligent process control systems, increasing demands and needs of end users are handled by control-system modules with overlapping and extended functionalities. These functionalities are well suited for integration with the above levels of scheduling, mode setting, and coping with off-normal conditions. The knowledge and experience of technical experts are incorporated into the process control system, enabling it to function at a higher intelligent level. Supervisory control and diagnostic functionalities are particularly suited for implementation with real-time expert systems. The expert knowledge is represented explicitly in the knowledge base, while an inference mechanism is responsible for problem solving under a wide range of circumstances. The inference or reasoning mechanism leads to better problemsolving results when the knowledge representations in the knowledge base are more adequately capturing a multitude of process conditions and, correspondingly, a rich repertoire of well-suited human operator behaviors.

Several forms of deep and shallow knowledge may be available, such as model-based knowledge or embedded simulations, rule-based knowledge, and frame-based knowledge with the inheritance of object attributes and values between different levels of an object hierarchy. Thus, procedural as well as declarative knowledge can be represented in the knowledge bases. Several knowledge bases are generally necessary, covering areas such as process plant components, process states (particularly macrostates), failure cases and failure symptoms, fault diagnosis strategies, and control laws. The knowledge to be incorporated into the knowledge bases has to be elicited from different classes of technical experts, such as operators, engineers, and maintenance personnel. Knowledge elicitation and acquisition is an empirical social interaction process between experts and knowledge engineers. It takes much time, needs much experience to avoid pitfalls, and is therefore sometimes called the bottleneck of knowledge engineering and expert systems design. See OBJECT-ORIENTED PROGRAMMING.

The special features of expert systems for industrial applications of dynamic processes include their real-time behavior as well as their online or embedded functionalities within the whole technical system. Process variables are viewed mainly as symbolic expressions in linguistic terms. For example, a temperature is denoted as "high" if the actual numerical value is anywhere between two relevant boundaries. The real numerical values are not of direct interest during the problem-solving process. The coexistence of particular linguistic values of several process variables may define characteristic macrostates of the process control system. Since the variables change their values asynchronically, the real-time aspects of process control require a buffering and regular updating of process variables at the instances of time stamps between successive macrostates. This ensures piecewise-constant conditions for the performance of the reasoning process.

The application of expert systems in supervisory control allows the knowledge-based guidance of the process along desired trajectories between different steady states. Different kinds of regulatory control mechanisms can be influenced by the supervising heuristic expert control. In fault management situations, the diagnostic process can be handled in several steps or phases. After the actual state of conditions is identified, the two steps of hypothesizeand-test of possibly faulty system components are performed. An example is the diagnosis of technical faults in a preheater system of a power plant. Learning mechanisms for improving control through experience are also feasible. *See* EXPERT SYSTEMS.

Fuzzy process control. The methodologies of fuzzy set and systems theory are now applied in process control. Several possibilities for such applications exist. On a higher intelligent level, problem solving and decisions for scheduling and diagnostics are sometimes fuzzy, rather than crisp and precise, in the overall control of complex industrial processes. The process variables themselves may be transformed into fuzzy linguistic variables, based on the degree of membership with which each numerical value of a measured process variable belongs to one or several overlapping linguistic terms. This transformation from precise measurements to fuzzy linguistic variables is called fuzzification.

The inference mechanism of a rule-based expert system may also be described with fuzzy rules and relations. Satisficing rather than optimizing a problem solution is often enough, and sometimes is the only achievable possibility in a complex practical context. Heuristic knowledge for supervisory control and diagnostics can often be formalized only with fuzzy rule-based reasoning. In such cases, evaluation is not restricted to a single path through the rule base, and, indeed, several rules are activated and several inferencing paths contribute to the end result of a problem solution. This end result is expressed in terms of membership functions of linguistic terms. A transformation is then needed into linguistic expressions that are understandable by the human user of the decision outcome in the application context. All definitions of linguistic terms, all membership functions, and all fuzzy rules and relations have to be derived from human expert knowledge by means of empirical studies.

On a lower level nearer to the technical process, regulatory control can be performed with fuzzy logic controllers. As in the case of fuzzy expert systems, precise quantities are changed into linguistic variables by fuzzification (**Fig. 6**). The inference mechanism is now the fuzzy controller, which consists of

linguistic IF-THEN rules such as:

IF temperature is "positive big" or "positive medium" and

IF temperature rate is "negative very small"

THEN coding input change is "positive medium".

A fuzzy decision combines several conclusions of the THEN parts whenever a number of rules are applied in parallel for a fuzzy-logic-based control policy. Most industrial processes cannot interpret fuzzy command inputs. Therefore, the fuzzy controller and decision output need to be transformed into precise control quantities by the defuzzification operation. *See* FUZZY SETS AND SYSTEMS.

Artificial neural networks. Another knowledgebased technology of generally high potential is available with artificial neural networks. Whereas expert systems contain more of the "know-how" type of knowledge such as in heuristics and principles, artificial neural networks are more data-driven and thus represent the "know-what" type of knowledge. Large sets of data such as multi-inputs and corresponding multioutputs of dynamic technical processes can be observed as vectors or patterns of data and can

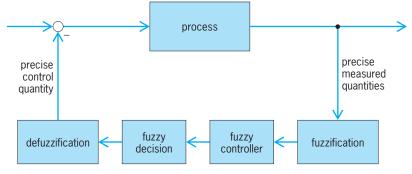


Fig. 6. Closed-loop fuzzy process control.

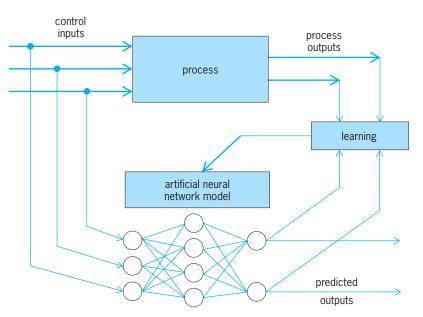


Fig. 7. Neural network for process modeling and control.

appropriately be matched by the general patternmatching capabilities of artificial neural network (ANN) models.

In no way do the parameters of these models relate to any type of explicit knowledge, as in expert systems. The knowledge of artificial neural network models is implicitly contained in the combination of the large number of parameters, the so-called weights, in the overall web-type network.

Figure 7 illustrates the most common use of artificial neural networks in process modeling and control. The process input data are the parallel control inputs at certain moments in time, and the process output data are the corresponding parallel process variables, such as temperatures and flow rates. Empirical models based on process input-output data can be developed, particularly for nonlinear and poorly understood processes such as those found in many chemical and biotechnological plants. Artificial neural network models are powerful empirical models with good approximation capabilities. Typically, neural networks with three layers of artificial neurons are applied: the input layer, the hidden layer as a kind of intermediary, and the output layer.

The weights of neurons in the hidden layer and in the output layer can be tuned by means of a learning or training algorithm. The input-output data of the process are used as training data, and the output data of the artificial neural network model are compared with the process outputs (Fig. 7). The back-propagation algorithm is the most commonly employed learning algorithm in process control applications. It is a gradient descent optimization technique developed for suitable use in multilayered iterative updating of the weights. The knowledge about the process behavior is implicitly represented in the combination of all these weights. However, an application-related interpretation of the parameters is not feasible.

Once the artificial neural network model has been fine-tuned, it can be used to predict outputs in regulatory process control. Various control structures with one or two artificial neural network models have been applied in various combinations with control theoretical algorithms to realize a large number of process control applications. Process monitoring and fault diagnosis can also be supported. Several distinct faults in different process components, such as a leakage in a condenser, can be predicted by different artificial neural network models, each with a typical match of its input-ouput patterns as a representation for the corresponding fault. Practical installations of artificial neural network models for process modeling, control, monitoring, and diagnosis can be found in chemical plants, robots, steel plants, and other applications. Further, neural networks have been combined with fuzzy logic to form neuro-fuzzy networks which achieve better interpretations and robustness than artificial neural network models alone. This is accomplished by a clearer modeling structure, for example, by partitioning the process operation into several fuzzy operating regions. See NEURAL NETWORK.

Process Control Components

The hardware and software implementations of process control systems are accomplished with the integration of a large number of components.

Sensor and actuator systems. The sensor systems and the actuator systems make up the process-near instrumentation. Pneumatic, mechanical, electronic, and mixed types of instrumentation have been realized for practical process control applications. All relevant outputs of a particular technical system, the process variables, have to be measured by suitable sensors. Product and process properties have to be continually measured in real time. Examples of the large number of possible process and product variables to be measured are pressures, temperatures, flow rates, rotational speeds, mass flows and levels, energy consumptions, viscosities, and concentrations. Different physical principles are applied to realize a variety of sensor technologies, including optical, acoustic, inductive, capacitive, radiometric, hydrostatic, chromatographic, and spectrometric. See TRANSDUCER.

The actuator systems transform control signals, mainly electrical outputs of a digital control computer, into the appropriate physical forms of control inputs for the process behavior which has to be influenced. Servo drives such as motors are used to operate the final actuator elements. These can be valves, pumps, heating devices, and all kinds of switches and contacts. *See* SERVOMECHANISM; VALVE.

Distributed control systems. A hierarchical and decentralized organization of all process monitoring and control functions is realized by means of distributed control systems. The required functions include controlling and supervising the process, planning and diagnosing, improving safety and reliability, and displaying information to human users as well as recording and evaluating.

The most common type of information transmission in distributed control systems is the serial data bus. All sensors and control valves are coupled via microcomputers to a bidirectional high-speed transmission line, for example, a coaxial cable or an optical fiber (**Fig. 8**). The microcomputers convert signals from sensors into binary numbers, and put these on the transmission line as a sequence of short (voltage or light) pulses, for example, each of 1 microsecond duration. At the other side the pulses are picked up and stored in a digital register for introduction into a computer. An analogous series of steps takes place between computer and control valves. *See* COAXIAL CABLE; OPTICAL FIBERS.

As all data are transmitted along the same transmission, sender and receiver have to find each other by a communication procedure, coded in pulse trains, the so-called protocol. By way of example, assume that the computer always acts as bus master, both when requesting and when supplying data. Here the computer starts a transmission with an initialization code, which alerts all sensors and control valves, followed by an address code, which activates a particular device. This device responds with a ready code, to which the computer reacts with a request data code. The device then sends the information, if necessary ending with an end data code, and returns to the listening state (to be ready for the next initialization code). More complicated protocols are required when the bus mastership can transfer from one device (or computer) to another. *See* DATA COMMUNI-CATIONS; LOCAL-AREA NETWORKS.

The main feature of this type of data transmission is the saving in cabling and installation costs. With traditional transmission, thousands of pneumatic tubes or electric wire pairs have to be installed and correctly connected. With serial transmission, only one cable (or optical fiber) runs along all devices in the plant, and insertion of address codes replaces tedious testing of wire (tube) connections.

However, a single serial transmission is too risky: when it fails, the complete instrumentation system breaks down. It is therefore desirable to achieve redundancy by installing at least two independent serial transmissions, which follow individual paths to all devices. If one fails, the other can take over.

The data load on the transmission and communication system can be decreased by incorporation of local data processing (Fig. 8). A number of sensors and control valves, pertaining to a small part of the production processes, are connected to a local computer, which takes care of basic information and control functions. Further decentralization can be achieved by so-called smart sensors, which are provided with microcomputers for linearization, calibration, and fault identification. *See* MICROCOM-PUTER; MICROPROCESSOR.

Figure 9 gives a schematic example of a modern distributed control system. It contains many computers for a wide variety of functions, both in local and in central data processing. The latter part is interconnected by a parallel bus, which allows a higher transmission rate than a serial bus. The highest level in this hierarchy of decentralized functionalities is accomplished by the production computer. On the intermediate levels, the medium and lower basic

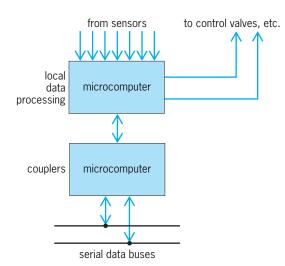


Fig. 8. Local computer system.

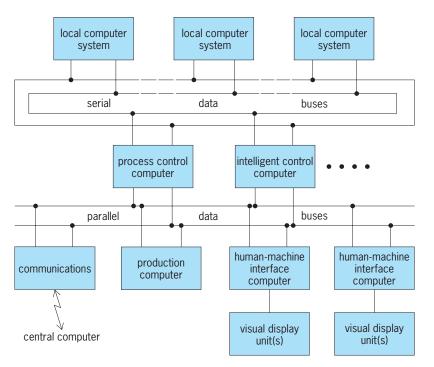


Fig. 9. Schematic example of modern distributed control system.

functions are realized with the process control computer (in the narrower sense), the extended functions by means of the intelligent control computer, and the safety as well as the recording functions with the same or, optionally, additional dedicated computers. The implementation of distributed control systems with a hierarchical decentralized organization and with supervisory control functionalities supports the integration between conventional and integrated intelligent control systems. Further, modern distributed control systems have provisions for coupling to office computer networks and even to distant computers.

A very important feature of distributed control systems is system software for configuration of regulatory and sequence control, as well as of higher-level and intelligent control, and for display of process information. In this way, flexibility is combined with efficiency. *See* DISTRIBUTED SYSTEMS (CONTROL SYS-TEMS).

Information systems. It is often difficult to make a clear distinction between process control systems and process information systems. As automation never is complete (even space satellites are supervised from Earth), there is interaction between process operators and automation systems, which, among others, requires an operator information system. Moreover, individual computers for data processing in various departments of the factory (or the enterprise) are coupled into an integrated information and control system. In this way, sales, design, internal transport, storage, packaging, expedition, accounting, maintenance, and management send and receive information to and from production. For instance, customer orders are entered into the system

in the sales department, and they appear at the production scheduling section, where the sequence of operations is determined for production processes, based on information from storage.

The practical realization of information systems comprises computer networks and management, communication structures and protocols, knowledge and data mining technologies, and qualityassurance architectures and tools. In information systems it is very important to select the type and form of information in accordance with the needs of the user. A plant manager requires quite different reports than a process operator. Systematic user participation during the phases of design and development of process control and information systems becomes a strong requirement.

Human-Machine Interaction

Computerized instrumentation has revolutionized the interaction with plant personnel, in particular the process operators. Traditionally, the central control room was provided with long panels or consoles, on which alarm lights, indicators, and recorders were mounted. Costs were rather high, and surveyability was poor. In computerized instrumentation, visual display units are the main components for information presentation in the human-machine interfaces (Fig. 9). They can provide information in a concise and flexible way, adapted to human needs and capabilities. Many visual display units contain cathode-ray tubes, similar to those used in television sets (**Fig. 10**). *See* CATHODE-RAY TUBE.

A computerized visual display unit offers the designer greater freedom in structuring process information. Process flow schemes can be displayed which indicate the state of process equipment (on/off), alarms, and values of process variables (information for supervisory control).

Historical information can be shown in sectors of the screen. Operators can be offered a choice of display structures and even facilities to generate formats to their individual preferences.

Process flow schemes are visualized for the human operators as topological views on the process in a component-oriented fashion. User-oriented and goaloriented human-machine interfaces can additionally supply functional knowledge. Much more information processing than previously is required in these human-machine interfaces. Goals to be achieved can be related to functions of the process, which again relate to the components of the process. Multimodelbased approaches use fuzzy logic and qualitative modeling adapted to human needs. Different abstraction levels and different degrees of detail can be visualized. The process visualizations include representations of the fulfillment degrees of goals, the violation degrees of conditions, flow functions of mass and energy balances, and ecological gestalt graphs for the relations between process states and control inputs. Some of these process visualizations are fairly abstract and, therefore, the user needs to be able to navigate between them and the more traditional



Fig. 10. Central control room of the ESSO refinery, Rotterdam, Netherlands. (RO Studio, Vlaardingen, Netherlands)

topological views within one human-machine interface.

Intensive participation of prospective end users as well as usability testing is necessary in order to design more advanced human-machine interfaces with supportive knowledge-based functionalities. This participative design approach can be extended further toward the consideration of users' needs early in the development of the whole process control system, ultimately leading to the ideal of human-centered automation.

Other developments in human-machine interfaces for process control include large projection screens and multimedia presentations, in most cases in addition to the process visualizations with visual display units. The main objective of the multimedia approaches for process control is to make the best use of human sensory modalities in addition to vision. This utilization will avoid the visual overload of human operators. Also, it can bring the operator back closer to the environment where audio signals, such as noises from process components, and vibrations can be exploited for transmitting important information. *See* MULTIMEDIA TECHNOLOGY.

Instrumentation system designers can obtain criteria and results about the human-machine interaction from human-factors engineering. *See* HUMAN-COMPUTER INTERACTION; HUMAN-FACTORS ENGI-NEERING; HUMAN-MACHINE SYSTEMS.

Plant-Wide Control and Communication

Modern distributed control systems offer an infrastructure that meets most needs for process control in the wide sense. New developments toward human-centered automation can be realized by combining basic and intelligent control and communication technologies with human-machine and cognitive systems engineering. Communication with the central office and the outside world can support integration of various production activities in the industrial enterprise. Of particular interest are interaction between processing and packaging (or other discrete production activites); integration with logistics control; integrated plant-wide, factory-wide, or enterprise-wide information systems; and combined systems for process model development and recipe improvement. The main problems encountered in these activities are not of a technical nature, but are related to reorganization and to training and changing the attitudes of all involved people. See AUTOMA-TION; CONTROL SYSTEMS. Gunnar Johannsen

Bibliography. F. Jovic, Expert Systems in Process Control, Kluwer (Chapman & Hall), Dordrecht, 1992; T. E. Marlin, Process Control, 2d ed., McGraw-Hill, New York, 2000; M. Polke (ed.), Process Control Engineering, VCH Verlagsgesellschaft, Weinheim, 1994; W. F. Ramirez, Process Control and Identification, Academic Press, Boston, 1994; M. Rao and H. Qiu, Process Control Engineering, Gordon & Breach Science, Newark, NJ, 1993; J. E. Rijnsdorp, Integrated Process Control and Automation, Elsevier Science, Amsterdam, 1991; T. J. Ross, Fuzzy Logic with Engineering Applications, McGraw-Hill, New York, 1995; F. G. Shinskey, Process Control Systems, 4th ed., McGraw-Hill, New York, 1996; S. G. Tzafestas (ed.), Methods and Applications of Intelligent Control, Kluwer, Dordrecht, 1997.

Process engineering

A branch of engineering in which a process effects chemical and mechanical transformations of matter, conducted continuously or repeatedly on a substantial scale. Process engineering constitutes the specification, optimization, realization, and adjustment of the process applied to manufacture of bulk products or discrete products. Bulk products are those which are homogeneous throughout and uniform in properties, are in gaseous, liquid, or solid form, and are made in separate batches or continuously. Examples of bulk product processes include petroleum refining, municipal water purification, the manufacture of penicillin by fermentation or synthesis, the forming of paper from wood pulp, the separation and crystallization of various salts from brine, the production of liquid oxygen and nitrogen from air, the electrolytic beneficiation of aluminum, and the manufacture of paint, whiskey, plastic resin, and so on. Discrete products are those which are separate and individual, although they may be identical or very nearly so. Examples of discrete product processes include the casting, molding, forging, shaping, forming, joining, and surface finishing of the component piece parts of end products or of the end products themselves. Processes are chemical when one or more essential steps involve chemical reaction. Almost no chemical process occurs without many accompanying mechanical steps such as pumping and conveying, size reduction of particles, classification of particles and their separation from fluid streams, evaporation and distillation with attendant boiling and condensation, absorption, extraction, membrane separations, and mixing.

Bulk Products

Process engineering begins with a need for a product and some sense of the required scale of production. The end product is always defined; frequently the desired raw materials can also be stipulated at this stage, although sometimes it is very important to consider wholly different sources (for example, motor fuel from coal, petroleum, or renewable resources) and the degree of vertical integration desired-for example, should the process begin with natural products or purchased intermediates. Then one or more process flow sheets are drawn, depicting the series of steps through which the raw materials must pass in the course of transformation into product. On the flow sheet, all lines represent movement of material from one process step to another, and all lines must show the intended direction of flow. Each processing step is represented by an enclosed area (circle, rectangle, and so on) on whose edges the lines terminate (unless only heat exchange is intended, in which case at least one line goes through the area unbroken to indicate thermal contact without mixing). Each process flow sheet envisions a definite sequence of operations, including the selection of separation processes and choice of chemical reactions. The sequence may contain subsequences that are serial or parallel or enclosed by a recycle stream.

Batch versus continuous operation. In designing almost any process, an early choice between batch and continuous operation must be made. Small amounts of material can be taken through a series of process steps over time, often using the same apparatus (for example, a thermally jacketed stirred tank) for several or even all steps. Commercial production in batches is feasible and often preferred when small amounts are to be manufactured, when a sequence of many similar reaction steps is envisioned, or when it is feasible for multiple products to be manufactured in the same equipment. However, batch operations are labor-intensive and more difficult to control, analyze, and design (if done with the same care given to continuous processes). Continuous operation is easier to analyze, design, and control, but the plant is expensive, unique to the product, and difficult to operate effectively at other than design rate.

Material and energy balances. Once the process flow sheet is drawn, it is possible-for either batch or continuous processes-to generate material and energy balances. These are drawn for the overall process and for each step on the basis of conservation and rate laws as well as certain assumptions, some of which are only temporary. For example, in a mechanical step, such as dewatering of a fiber slurry during paper manufacture, a dilute stream containing 1% dry fiber and 99% water may be presumed to enter a separator at the rate of 22,000 lb/min (10,000 kg/min), to be separated into two product streams, one containing 99% of the fibers fed at a concentration of 5% fibers and the other, a recycle water stream. The proportioning of the fibers between the streams and the composition of the one outlet stream are assumptions that may well be adjusted at later stages. In this example, the conservation laws require that water and fiber be separately conserved since no reaction is envisioned, and it is immediately possible to calculate that the concentrated fiber stream emerges at the rate of 4360 lb/min (1980 kg/min), while the balance of the 22,000 lb/min (10,000 kg/min) leaves as a stream containing about 0.0125% fiber. In a complex process, many hundreds of such steps must be assessed and matched to each other. [In the case of the process step just considered, accepting the 22,000 lb/min (10,000 kg/min) stream constitutes matching this step to a predecessor; and two subsequent steps must be matched to the two exit streams just considered.] When the assessment and matching of individual steps are complete, the overall result must be considered. Is the production rate correct? Is the yield of product satisfactory, or has too much material been turned into by-products or lost in waste streams? Has too much water or heat been used? Is the product pure enough? Frequently one or more of these targets are not met. Then assumptions must be revised and the computations repeated. Sometimes the flow sheet itself must be changed: additional steps added, a new recycle stream provided,

one kind of operation replaced by another. Because of the complexity of this process, it is often executed with the aid of computers. Once a flow sheet has been fixed and consistent heat and material balances reckoned, the duty of each process unit is established, and their individual design can be commenced.

Process units and reactors. For many years chemical engineers have recognized that the design of process units was much more dependent upon the kind of operation under consideration (for example, heat transfer, or fluid flow, or distillation) and much less dependent on what was being processed or where in the sequence of process steps a particular unit occurred. Thus evolved the concept of unit operations and a set of principles and rules for design of process units according to their type of unit operation rather than according to the process itself. Notable among unit operations that have been extensively developed are the membrane processes (reverse osmosis, dialysis, electrodialysis, and electroosmosis), techniques for separating solids from fluids, and methods of controlling transport across phase boundaries such as freeze drying.

Reactor design. In completing the engineering of a process, many features peculiar to it must be considered, such as the selection of pipes and tanks strong enough to withstand the temperatures and pressures of operation and the choice of materials that will not be corroded by the materials being processed. Because the chemical reaction is the most critical step in most chemical processes, reactor engineering is a very important component of process engineering. Chemical reactions may be carried out either with a single phase (homogeneously) or near phase boundaries and interfaces (heterogeneously) with or without a catalyst. Process reactors may be stirred tanks containing one fluid, one fluid dispersed in another, or a suspension. Some reactors are fluidized beds: fine particles which may be either reactant or catalyst kept suspended and moving by a flowing gas or liquid. Tubular reactors are used frequently, filled only with fluid, or packed, or equipped with complex so-called internals. Many reactors are jacketed or otherwise equipped for removing heat. Some reactions are conducted in flames or in illuminated chambers. However, in spite of this great geometric variety, general principles of reactor engineering have been developed, based on expressions for the rates of chemical reactions and the transport of molecules and heat as well as the laws of conservation for mass and energy.

Integration and optimization. The widespread availability of computers has led both to higher integration of processes and to much more nearly optimal design. Integration refers to the increased use of recycle streams to recover heat or material that would otherwise be discarded. It is motivated both by possible economies and by its ability to decrease the burdens imposed on the environment by a process. Integration makes design more difficult, usually results in a higher capital cost for the plant, and may make it more difficult to control. Modern petroleum

refineries tend to be particularly highly integrated. Optimization is a relative term. In practice it denotes computation of the best split of streams and duties among units and selection of the best operating conditions for each unit to optimize some overall quantity (often plant profitability) subject to constraints (such as environmental restrictions, plant size, usage of certain utilities). *See* OPTIMIZATION.

Extreme conditions. Extremes of high temperature and high and low pressure have been extended, so that reactions can be carried out in plasmas at temperatures of $17,500^{\circ}$ F (10,000 K), while pressures may go as high as several hundred thousands of pounds per square inch (several gigapascals) and as low as 10^{-10} torr (10^{-8} pascal). Extension of extremes of feasible operation has been facilitated by development of new materials of construction.

Economic considerations. Process engineering is an activity supported almost entirely by industry, and is thus greatly influenced by economic considerations, much augmented however by regulatory considerations designed to minimize environmental impacts, protect workers, guarantee the safety and efficacy of products, and carry out governmental policies for conservation of resources, especially energy. Still, economics figures heavily in process design: in choosing to make a product, in specifying the degree of vertical integration, in fixing the scale of operations, and in balancing capital costs against operating costs. Other economically significant decisions include estimating the desirable operating lifetime of the process and plant and assessing the degree of risk to be taken in using attractive but not fully proved new technology in the process realization. Thus process engineering utilizes not only principles from the chemical and physical sciences but also detailed assessments of the state of commerce and the economy. See DIALYSIS; DISTIL-LATION; EVAPORATION; EXTRACTION; ION-SELECTIVE MEMBRANES AND ELECTRODES; MECHANICAL CLASSIFI-CATION; MECHANICAL SEPARATION TECHNIQUES; MIX-Edward F. Leonard ING

Discrete Products

The process engineer must consider all available processes capable of changing the shape of a raw material to the geometrical configuration specified on an engineering drawing. Thus, the process engineer must consider not only those processes within his or her own plant but also those in vendor or subcontractor plants. Usually, several ways exist to produce any part, one being best for a given set of conditions. The process engineer has the responsibility of determining which is best. The principal constraints to be considered in the selection of a basic process in order to bring raw material more closely to the specifications of a functional design are (1) type of raw material; (2) size of the raw material that the equipment involved can handle; (3) geometric configurations that equipment characteristic of the process is capable of imparting to the raw material; (4) tolerance and surface finish capabilities of the equipment; (5) quantity of finished parts needed and their

Process	Applicable geometry*	Intricacy of geometry	Applicable materials [†]	Minimum lot for which process is economical	Secondary operation cost	Decision equation for primary process	Expected rate of production per hour	Expected time in weeks to get into production
Die casting	1,2,3,4,5, 6,7,8,9	Х	t	3000	1	$D_t = 1.623N + 1515$	200–500 injections	12
Investment casting	1,2,3,4,5, 6,7,8,9	X	s, t	100	5	$D_t = 2.370N + 320$ $D_s = 1.000N + 320$	10–20 molds	5
Permanent mold	1,2,3,4,5, 6,7,8	Y	s, t	500	9	$D_t = 0.770N + 465$ $D_s = 0.335N + 465$	20–30 molds	10
Plaster mold	1,2,3,4,5, 6,7,8,9	X	t	100	5	$D_t = 2.160N + 320$	15–28 molds	5
Shell mold	1,2,3,4,5, 6,7,8,9	X	s, t	300	6	$D_t = 1.965N + 320$ $D_s = 0.595N + 320$	35–45 molds	5

delivery requirements; and (6) economics of the process. Similar constraints apply in the selection of the most favorable secondary operations. *See* PRO-DUCTION ENGINEERING.

Choice of process. To identify the most favorable basic processes and secondary processes to be used to produce a given functional design, the process engineer may develop a selector guide to assist in the decision-making process.

Selector guide sheets are usually developed for a given size classification and are shown in part in **Tables 1–3**. In using these guide sheets, an analyst refers to the correct sheets for the size of the part under study. There may be several size classifications to handle all products being made by a given company. Next the analyst determines the geometry that best describes the component. Table 1 indicates nine geometric configurations needed to be considered in connection with the five casting processes listed. These nine classifications may be defined and illustrated to facilitate the classification of the geometry of the product being studied. For example:

Class 1 geometry—Solid or partly hollow rounds involving one or more diameters along one axis. The depth of hollow not to be more than twothirds the diameter of the end containing the hollow. Class 2 geometry—Hollow or partly hollow rounds involving one or more outside diameters and one or more inside diameters along one axis. The depth of hollow is more than two-thirds the diameter of the end containing the hollow.

Class 3 geometry—Solid or partly hollow shapes other than rounds such as square, triangular, octagonal or irregular, including one or more crosssectional areas along one axis. The depth of hollow is not more than two-thirds the distance of the major diameter.

Class 4 geometry—Hollow or partly hollow shapes other than rounds involving one or more crosssectional areas along one axis. The depth of hollow is more than two-thirds the major diameter of the hollow.

Class 5 geometry-Bowl-shaped concentric.

Class 6 geometry—Dish-shaped nonconcentric.

Class 7 geometry—Flats with or without configuration.

Class 8 geometry—Flanged.

Class 9 geometry—Complex characterized by unsymmetrical shapes.

The next parameter that is considered is the material used. Table 1 illustrates only two material considerations, ferrous (*s*) and nonferrous (*t*). The required

Process	Raw material form	Raw material cost	Tooling cost	Equipment cost	Setup cost	Production cost	Scrap cost	Expected tolerance, mm	Expected surface finish, micrometers
Die casting	Pig or liquid	2	10	4	4	1	3	±.08 except .18 across parting line	.191
Investment casting	Pig	2	4	6	6	8	4	±.13	.318
Permanent mold	Pig	2	6	6	3	4	4	±.38	.381
Plaster mold	Pig	2	3	6	6	7	3	±.13	.254
Shell mold	Pig	2	4	6	6	4	3	±.13 except ±.22 across parting line	.318

TABLE 3. Part of gu	uide sheet for qua	lity consideration	IS	
Process	Expected reliability	Directional flow lines	Tool marks	Remarks
Die casting	A for Zn B for Al C for brass	No	Yes	Ejection pins may cause tool marks; surface finish depends on surface finish for die; casting of ferrous allows generally not economically feasible
Investment casting	В	No	No	Best for complicated design; no parting line; frozen mercury results in better finish than wax or plastic
Permanent mold	С	No	No	Lightest wall thickness 1/8 in. (3 mm); limited complexity because of rigid molds; not suitable for class 9 geometry
Plaster mold Shell mold	A B	No No	No No	Little finishing required

number of parts needed is estimated. The constraint associated with quantity is the minimum number of parts required before it would become economical to provide tooling for the process under study. Decision equations may be developed for typical manufacturing environments of a company. The final step for the analyst is to solve the decision equations for processes still under consideration; most will already have been eliminated by this time because of the constraints of geometry, material, and quantity. The process that provides the smallest value based on the solution of the decision equation will be considered favorable for use. If Table 1 were given in full, it would be seen that two decision equations are provided for most of the processes. The Ds equation is used in conjunction with ferrous materials, and the D_t equation in conjunction with nonferrous materials.

To do a thorough job of process planning, use of the digital computer is recommended. Manual process selection may not give consideration to all possible alternatives. The computer can handle the mathematical manipulation of solving the various decision equations and the final selection problem. The equation used to develop the computer program based on the parameters shown in Table 1 is shown below,

$$D_t = A_{ijk} \times N + B_{ij}$$

where *A* is the coefficient of *N*, the number of units to be processed, *i* is the process to be considered, *j* is the size classification of the part being process planned, and *k* is the material specified. B_{ij} is the fixed element. *See* DECISION THEORY.

Sequence of secondary operations. In determining the sequence of operations to be performed, the process engineer considers the logical process order and the geometrical and dimensional control that can be maintained with each processing operation.

In the logical process order, basic operations are performed first and final inspection near the end of the process, just prior to a product's packaging for shipment. Holes that require reaming would necessarily be drilled before being reamed and, similarly, threaded inside diameters would necessarily be drilled prior to tapping.

In general, the final finishing of internal work is done in advance of external finishing. The principal reason for this sequence is that internal surfaces are less likely to be damaged in material handling and subsequent processes so that their surfaces can be completed earlier in the processing. For internal work, the logical sequence of operations is drilling, boring, recessing, reaming, and tapping. The logical sequence of external work is turning, grooving, forming, and threading.

Rough work involving heavy cuts and liberal tolerances should be performed early in a sequence. Heavy cuts will reveal defects in castings or forgings more readily than light cuts, and it is advantageous to identify defective raw material as soon as possible. Because both tolerance and finish of external surfaces can be adversely affected by subsequent material handling and clamping, close-tolerance operations should be performed late in the processing sequence.

In considering the geometrical and dimensional control that can be maintained, the process engineer should keep in mind those operations that establish locating or mastering points in order to schedule them early in the process. For example, it is usually easier to maintain control from a large, plane surface than from a curved, irregular, or small surface; so the process engineer establishes a surface which is best qualified for the location of critical features, and schedules it for finishing at a point early in the sequence.

Specifications concerning tools. It is the responsibility of the process engineer not only to specify how a piece should be made but also to order all special tools to be used in conjunction with the specified manufacturing processes. Special tools refer to those jigs, fixtures, dies, and gages that are required to hold work, guide the cutting tool, and inspect results. The requisition for the special tools should specify (1) those areas or points best suited for locating the workpiece while it is being processed; (2) that portion of the workpiece suited for supporting or holding it while it is being processed; and (3) that portion or area best suited for clamping so that the workpiece is securely held during the processing.

In specifying the location and number of locators to contact the work, the analyst must keep in mind that three locators are needed to locate a plane, two are needed to determine a line, and one will determine a point. A workpiece can move in either of

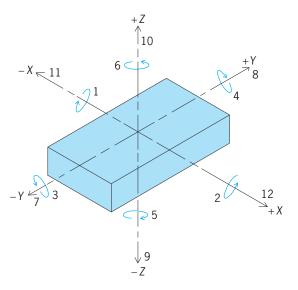


Fig. 1. Twelve degrees of freedom of a workpiece.

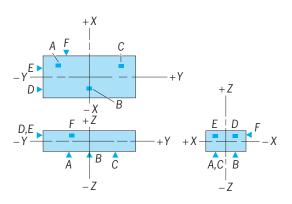


Fig. 2. Three views of a workpiece.

two opposed directions along three perpendicular axes (X, Y, and Z). Also, the work may rotate either clockwise or counterclockwise around each of these three axes. Each of these possible movements is a degree of freedom; hence 12 degrees of freedom exist. These 12 degrees of freedom are illustrated in **Fig. 1**. *See* DESCRIPTIVE GEOMETRY.

Work can be located positively by six points of contact in the tooling. These six points include three points on one plane. For example, in Fig. 2 the three locations A, B, and C on the bottom of the block prevent the work from moving downward and from rotating about the X and Y axes. By adding two locating points D and E on a plane parallel to the plane containing the X and Z axes, the work is prevented from rotating about the Z axis and also from moving negatively along the Yaxis. When the sixth and final locating point F is added on a plane parallel to the Y and Z axes, movement upward is prevented. Thus the first three locators prevented movements 1, 2, 3, 4, and 9, as shown in Fig. 1. The next locators prevent movements 5, 6, and 7, and the final locator prevents movement 11. This 3-2-1 locating procedure has prevented movement in 9 of the 12 possibilities. The three remaining degrees of freedom (8, 10, and 12) must not be restricted because they are needed in

order to provide clearance to load and unload the tooling.

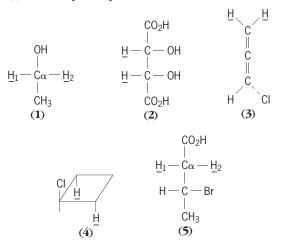
When requisitioning special tooling, the process engineer should indicate how the work is to be supported and specify the position of the locators. To assure correct position throughout the process, there must be rigid support so that the work does not deflect because of its static weight or the holding and tool forces applied to it, or both. Adding a support to avoid deflection during processing does not mean adding another locator. The support should not provide a point of location, and the workpiece should never contact the support until the tool or tooling forces are applied; the only purpose of the support is to avoid or limit deflection and distortion.

The holding force must be of sufficient magnitude to allow all locators to contact the workpiece during the processing cycle, but it is also important not to have holding forces so large that the work becomes marred. It is usually advantageous to place holding forces directly opposite locators. A nonrigid workpiece may require holding forces at several locations to hold the work against all locators. *See* PROCESS CONTROL. Benjamin W. Niebel

Bibliography. W. D. Fellers and W. W. Hunt, Manufacturing Processes for Technology, 2d ed., 2000; M. P. Groover, Fundamentals of Modern Manufacturing: Materials, Processes, and Systems, 2d ed., 2001; B. W. Niebel and A. B. Draper, Modern Manufacturing Process Engineering, 1989; R. Robin and D. Wield (eds.), Product Design and Technological Innovation, 1986; J. A. Schey, Introduction to Manufacturing Processes, 3d ed., 1999; Society for the Advancement of Material and Process Engineering, Processes: The Intercept Point, 1988.

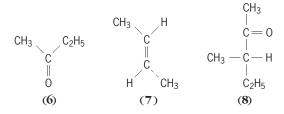
Prochirality

The property displayed by a prochiral molecule or a prochiral atom (prostereoisomerism). A molecule or atom is prochiral if it contains, or is bonded to, two constitutionally identical ligands (atoms or groups), replacement of one of which by a different ligand makes the molecule or atom chiral. Structures (1)-(5) are examples of prochiral molecules.



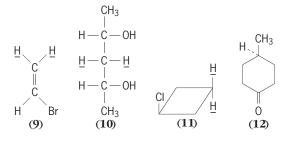
None of molecules (1)-(4) is chiral, but if one of the underlined pair of hydrogens is replaced, say, by deuterium, chirality results in all four cases. In compound (1), ethanol, a prochiral atom or center can be discerned (C α :CH₂); upon replacement of H by D, a chiral atom or center is generated, whose configuration depends on which of the two pertinent atoms (H₁ or H₂) is replaced. Molecule (**5**) is chiral to begin with, but separate replacement of H₁ and H₂ (say, by bromine) creates a new chiral atom at C α and thus gives rise to a pair of chiral diastereomers. No specific prochiral atom can be discerned in molecules (2)-(4), which are nevertheless prochiral [(**3**) has a prochiral axis].

Faces of double bonds may also be prochiral (and give rise to prochiral molecules), namely, when addition to one or other of the two faces of a double bond gives chiral products. Structures (6)-(8) have prochiral double-bond faces.

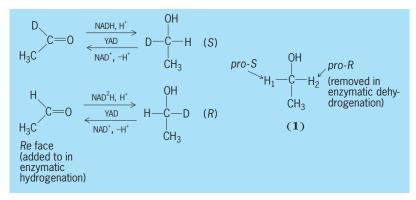


Hydrogenation of (6) gives chiral 2-butanol (addition to one face giving rise to one enantiomer, addition to the other to the opposite one), and HBr addition to one or the other face of (7) gives rise to enantiomeric chiral 2-bromobutanes, $CH_3CHBr CH_2CH_3$. Hydrogenation from one face or the other of (8) also leads to a new chiral center, but in this case diastereomers rather than enantiomers result by addition to the prochiral carbonyl faces.

Prostereoisomerism. Although the term prochirality is widely used, especially by biochemists, a preferred term is prostereoisomerism. This is because replacement of one or other of the two corresponding ligands (called heterotopic ligands; see below) or addition to the two heterotopic faces often gives rise to achiral diastereomers without generation of chirality. Examples are structures (9)-(12). Replacement



of one or another of the underlined pairs of hydrogens in (9) by bromine gives rise to *cis*- or *trans*-1,2dibromoethene; similar replacement in (10) by OH gives two different *meso*-2,3,4-pentanetriols; analogous replacement in (11) by Cl gives *cis*- or *trans*-1,3dichlorobutane; catalytic addition of hydrogen to the carbonyl double bond in (12) gives *cis*- or *trans*-4-



Reactions illustrating the ability of enzymes to distinguish between enantiotopic ligands or faces.

methylcyclohexanol. These are stereoisomeric pairs in all cases, but are devoid of chirality. Thus compounds (1)-(12) display prostereoisomerism, but only (1)-(7) display prochirality.

Heterotopic ligands and faces. The pertinent groups (H's) or faces (C=C, C=O) replaced or added to in (1)-(12) are called heterotopic (or, more precisely, stereoheterotopic) ligands or faces. If their-replacement [(1)-(4)] or an addition to them [(6) and (7)] gives rise to enantiomers, they are called enantiotopic. If replacement or addition results in diastereomers [(5), (8)-(12)], the ligand faces are called diastereotopic.

The term (stereo)heterotopic comprises both enantiotopic and diastereotopic ligands and faces. Ligands are called homotopic when their separate replacement by different ligands gives rise to identical products, as in CH₃Cl or CH₂Cl₂ (for example, replacing H by Br). Similarly, faces are called homotopic when addition to one or another of them gives rise to identical addition products, as in (CH₃)₂C=O or H₂C=CH₂.

Nomenclature. To name heterotopic ligands, a conceptual replacement is made of one of them by a heavier isotope. If the configuration of the resulting chiral molecule is R, the ligand replaced is called *pro-R*; if *S*-configuration results, the ligand is called *pro-S*. Thus replacement of H₁ by D in structure (1) gives rise to (*S*)-ethanol-1-*d*; hence H₁ is *pro-S* and, by default, H₂ is *pro-R*, a fact which can be established independently in that replacement of H₂ by D yields (*R*)-ethanol-1-*d*.

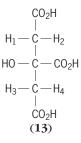
To name heterotopic faces, the Cahn-Ingold-Prelog system (sequence rule) is applied to the three ligands in the plane; if the resulting sequence (highest to lowest precedence) is clockwise, the face is called *Re*; if it is counterclockwise, *Si*. Thus the front face of the methyl ethyl ketone (6) is *Si* ($O \rightarrow C_2H_5 \rightarrow CH_3$) and the front face of the *trans*-2-butene (7) is $Re (=C \rightarrow CH_3 \rightarrow H)$; the rear faces of these two molecules would have the respective opposite designations [*Re* for compound (6), *Si* for compound (7)].

Significance. It is well known that enzymes can usually distinguish between enantiomers; thus liver alcohol dehydrogenase catalyzes the dehydrogenation of (*S*)-lactic acid, (*S*)-CH₃CHOHCO₂H, but not of

the R isomer, to pyruvic acid, CH₃COCO₂H. Similarly enzymes can distinguish between enantiotopic (as well as diastereotopic) ligands or faces, as illustrated by the classical work of F. Westheimer and B. Vennesland (see illus.). The fact that yeast alcohol dehydrogenase (YAD) mediates oxidation of (S)-ethanol-1-d by the coenzyme NAD⁺ with total retention of deuterium (that is, clean loss of hydrogen), but of (R)ethanol-1-d with integral loss of deuterium (despite the opposing isotope effect), is best explained by saying that the dehydrogenation of ethanol (labeled or not) by YAD-NAD⁺ always leads to removal of the pro-R hydrogen [H₂ in compound (1)]. Correspondingly, the reverse reaction, reduction of acetaldehyde or acetaldehyde-1-d by NAD²H and NADH, respectively, in the presence of YAD, invariably leads to addition of deuterium or hydrogen to the Re face. This type of stereochemical preference is due to the specificity of the fit of the substrate with the active site of the enzyme.

Diastereotopic, but not enantiotopic, groups are in principle distinct (anisochronous) in nuclear magnetic resonance (NMR) spectra. Thus ethanol (1) displays a single (though spin-split) methylene peak, but the underlined protons in (5) and (9)-(11) are distinct in their signals. This distinction must be kept in mind in the interpretation of NMR spectra.

In citric acid [structure (13)] all four marked H's



are heterotopic and therefore distinct toward enzymes; only one specific one is implicated in the aconitase-mediated dehydration of citric acid to cisaconitic acid, HO₂C-CH=C(CO₂H)CH₂ CO₂H. The CH₂CO₂H branch toward which the elimination occurs also can be specifically identified. Moreover, in the biosynthesis of citric acid from oxaloacetate and acetyl-CoA (aldol condensation), the oxaloacetatederived and acetate-derived branch can be distinguished: the former is *pro-R*, the latter *pro-S*. Since aconitase leads to dehydration toward the pro-R branch, it is the pro-S branch, that is, the acetyl-CoA derived one, which remains intact. This was demonstrated (by labeling) early in the investigation of the citric acid cycle, and appeared mysterious at the time; the concept of prochirality provides a logical explanation. See ENZYME; MOLECULAR ISOMERISM; NUCLEAR MAGNETIC RESONANCE (NMR); STEREOCHEMISTRY. Ernest L. Eliel

Bibliography. D. Arigoni and E. L. Eliel, *Top. Stere-ochem.*, 4:127, 1969; E. L. Eliel, Stereochemical non-equivalence of ligands and faces (heterotopicity), *J. Chem. Educ.*, 57:52-55, 1980; K. R. Hanson, Applications of the sequence rule, I: Naming the paired

ligands g,g at a tetrahedral atom X_{ggij} , II: Naming the two faces of a trigonal atom Y_{gbi} , *J. Amer. Chem. Soc.*, 88:2731, 1966; W. B. Jennings, Chemical shift nonequivalence in prochiral groups, *Chem. Rev.*, 75:307, 1975; K. Mislow and M. Raban, *Top. Stereochem.*, 1:1, 1967.

Prochlorophyceae

A class of prokaryotic organisms coextensive with the division Prochlorophycota in the kingdom Monera. Because prochlorophytes carry out oxygenevolving photosynthesis, they may be considered algae. They are distinguished from Cyanophyceae, the only other prokaryotic algae, by the presence in their photosynthetic lamellae of chlorophyll *b* in addition to chlorophyll *a* and the absence of phycobilin pigments. Otherwise, they resemble Cyanophyceae biochemically and ultrastructurally. The class comprises a single genus, *Prochloron*, with one species.

Prochloron didemni is a green, coccoid unicell 10-30 micrometers in diameter that lives only as an extracellular symbiont of colonial tunicates (didemnid ascidians) in the tropical Indo-Pacific. It has been intensely studied because advocates of the theory of origin of organelles by serial endosymbiosis believe that the precursor of green chloroplasts may have been an organism like *Prochloron*. Because *Prochloron* has not been grown in bacteria-free culture successfully, certain questions pertinent to its phylogeny have not been answered. *See* ALGAE.

Paul C. Silva; Richard L. Moe Bibliography. H. C. Bold and M. J. Wynne, *Introduction to the Algae: Structure and Reproduction*, 2d ed., 1997; R. A. Lewin, *Prochloron:* A status report, *Phycologia*, 23:203-208, 1984.

Procyon

The brightest star in the constellation Canis Minor, apparent magnitude +0.36. Procyon (α Canis Minoris) is among the stars nearest to the Earth, at a distance of only 3.5 parsecs (1.08×10^{14} km or 6.7×10^{13} mi). Its spectral type is F5, but Procyon is slightly overluminous compared to a main-sequence star of the same spectral type, which indicates that Procyon has already begun to evolve off the main sequence. Its intrinsic luminosity is about seven times that of the Sun, and the effective temperature is estimated to be 6500 K ($11,200^{\circ}$ F). The radius of the star is about 2.2 solar radii. *See* SPECTRAL TYPE; STELLAR EVOLUTION.

Irregularities in the proper motion of Procyon noticed in 1840 led to the prediction that a relatively massive unseen companion was altering the path of the star, causing it to move from side to side with a period of about 40 years. The faint 11th-magnitude companion was discovered in 1896 about 5" away, but is very difficult to see in the glare of the bright

primary. The mean physical separation is about 16 astronomical units. Procyon B was subsequently shown to be a white dwarf in the final stages of its evolution, with a luminosity only 1/2000 that of the Sun. From the astrometric orbit the masses of the primary and its companion have been computed as 1.75 and 0.62 solar masses, respectively. The progenitor of the white dwarf was originally the more massive of the two stars and underwent the final stages of its stellar evolution sooner than the original secondary, which is now seen as Procyon A. The age of the system is approximately 1 billion years. Procyon B is only about 1.5 times the size of the Earth, which implies that its mean density is enormous, probably exceeding 300,000 times the density of water, or 6 tons/in.³ (300 kg/cm³). It is the second nearest white dwarf, the nearest being the companion of Sirius. See BINARY STAR; STAR; WHITE DWARF STAR. David W. Latham

Product design

The determination and specification of the parts of a product and their interrelationship so that they become a unified whole. The design must satisfy a broad array of requirements in a condition of balanced effectiveness. A product is designed to perform a particular function or set of functions effectively and reliably, to be economically manufacturable, to be profitably salable, to suit the purposes and the attitudes of the consumer, and to be durable, safe, and economical to operate. For instance, the design must take into consideration the particular manufacturing facilities, available materials, knowhow, and economic resources of the manufacturer. The product may need to be packaged; usually it will also need to be shipped so that it should be light in weight and sturdy of construction. The product should appear significant, effective, compatible with the culture, and appear to be worth more than the price. The emphasis may differ with the instance. Durability in a paper napkin is different from durability in a power shovel.

To determine whether a design is well adjusted to the gross array, criteria are needed. Some are objective and measurable, such as clearances and efficiency, whereas others are quite subtle and even subjective. In a way, product design is an industrial art.

Ultimately the purpose of product design is to ascertain that the product will satisfy human wants and wishes either directly as consumer goods, or indirectly as capital equipment or components.

Except in the case of basic inventions, product design is a redesign to suit changed conditions, criteria, or enlightenment. Change may appear capricious as in some fashions and toys, may be the result of technological progress, or may result from a change in attitude toward the product or its function. In some areas trends can be discovered, but future preferences of buyers must be predicted, gambled on, or forced to occur. There are various steps in product design which are not necessarily in particular order. They are analytical studies, creative synthesis, drawings and models for appearance, function and specifications, plus calculations, experiments, and tests. *See* PRO-CESS ENGINEERING; PRODUCTION ENGINEERING; PRO-DUCTION PLANNING. Richard I. Felver

Product quality

Quality is the collection of features and characteristics of a product that contribute to its ability to meet given requirements. Though quality concepts date back to ancient times, the early work in modern times was on methods and processes to create standards for producing acceptable products. By the mid-1950s, mature methods had evolved for controlling quality, including statistical quality control and statistical process control, utilizing sequential sampling techniques for tracking the mean and variance in process performance. During the 1960s, these methods and techniques were extended to the service industry. During 1960-1980, there was a major shift in world markets, with the position of the United States declining while Japan and Europe experienced substantial growth in international markets. Consumers became more conscious of the cost and quality of products and services. Firms began to focus on total production systems for achieving quality at minimum cost. This trend has continued, and today the goals of quality control are largely driven by consumer concerns and preferences.

Characteristics. There are three views for describing the overall quality of a product. First is the view of the manufacturer, who is primarily concerned with the design, engineering, and manufacturing processes involved in fabricating the product. Quality is measured by the degree of conformance to predetermined specifications and standards, and deviations from these standards can lead to poor quality and low reliability. Efforts for quality improvement are aimed at eliminating defects (components and subsystems that are out of conformance), the need for scrap and rework, and hence overall reductions in production costs. Second is the view of the consumer or user, who can have different needs. To consumers, a high-quality product is one that well satisfies their preferences and expectations. This consideration can include a number of characteristics, some of which contribute little or nothing to the functionality of the product but are significant and influential in providing customer satisfaction. A third view relating to quality is to consider the product itself as a system and to incorporate those characteristics that pertain directly to the operation and functionality of the product. This approach should include overlap of the manufacturer and customer views.

An example of how the product quality of an automobile is examined is given in **Table 1**. From the manufacturer's viewpoint, this product could be rated as high quality, since units leave the final

Characteristic	Value
Warranty	3 years/36,000 mi (58,000 km)
Outgoing inspection	98%
Engine and power-train mean time to failure (MTTF)*	42,000 mi (67,000 km)
Acceleration 0–60 mi/h (0–100 km/h)	10 s
Braking distance 55-0 mi/h (88-0 km)	480 ft (140 m)
Fuel consumption	16 mi/gal (6.8 km/liter) in town, 20 mi/gal (8.5 km/liter) on highway

assembly processes with only a 2% observed rate of defects. Assuming that few failures occur during the early stages of use and that the major failures are likely to occur after the warranty has expired, the vehicle seems to be manufactured within the engineering and production specifications for which it was designed. Moreover, the warranty expenses should be relatively low. The customer, however, might view this product differently, particularly in the absence of any highly attractive features through options that might offset his or her satisfaction level. While the early failures might be few in number, the failures that occur after warranty can be quite expensive for the customer. In this example, from the viewpoint of a product system the product has more of a moderate level of quality based on the acceleration, braking distance, and fuel economy. See MANUFACTURING EN-GINEERING.

Elements. A number of dimensions or factors have evolved that contribute to overall product quality as it is presently defined. Some dimensions are more dominant than others in describing a particular product. The following elements are most common in describing quality.

Operating performance comprises measures of actual operating performance of the product. Examples include automobile gasoline consumption in miles per gallon (kilometers per liter), brake horse-power for an engine, range of a weapon, and hours of life for a battery.

Durability refers to the amount of use before product deterioration, normally measured in the time or hours of usage before a product fails.

Reliability refers to the probability of a product failing within a specified period of time, having survived to that point. Primarily applied to durable goods, reliability is assessed by the average rate of failure, mean time between failures (MTBF), and related failure time measures.

Conformance is the degree to which design and operating characteristics meet preestablished standards, as measured through defect count, scrap rate, and amount of rework.

Serviceability refers to the degree or ease of restoring service, measured in time to restore service, frequency of calls, and mean time to repair (MTTR).

Usability refers to the ease of use that a consumer experiences in developing the necessary preparation, skills, and proficiency for using a product as it was intended. The most common measure of usability is learning time.

Perceived quality concerns the overall image of a product among users and potential users. Though some objective measures are involved through surveys and opinions, perceived quality is generally quite subjective.

Esthetics concerns the way in which the product is sensed through looks, feel, sound, taste, and smell. *See* RELIABILITY, AVAILABILITY, AND MAINTAIN-ABILITY.

Many of the quality elements involve measurable product attributes that are easily quantified, but some are subjective and are influenced by selective group and individual preferences. This is particularly true of perceived quality and esthetics, which characterize the customer view of products.

Quality control. Quality control (QC) is the collection of methods and techniques for ensuring that a product or service is produced and delivered according to given requirements. This includes the development of specifications and standards, performance measures, and tracking procedures, and corrective actions to maintain control. The data collection and analysis functions for quality control involve statistical sampling, estimation of parameters, and construction of various control charts for monitoring the processes in making products. This area of quality control is formally known as statistical process control (SPC) and, along with acceptance sampling, represents the traditional perception of quality management. Statistical process control focuses primarily on the conformance element of quality, and to somewhat less extent on operating performance and durability. See PROCESS CONTROL; QUALITY CON-TROL.

Performance measures. It is convenient to think of product quality as a vector of attributes that span all important characteristics of the item relative to the overall performance according to the three bases—the manufacturer, the customer, and the overall product viewed as a system. Some of these elements are qualitative, but some are entirely quantitative. The challenge is to incorporate these elements into an overall measure of product performance suitable for monitoring progress in quality improvement and for making management decisions about the product quality and costs.

The overall performance of a product depends on each attribute separately. For example, if a product is considered with regard to two attributes (for example, quality and reliability), and if each of these attributes has three possible values (high, medium,

FIULUUL QUAILY 40	Product	quality	431
-------------------	---------	---------	-----

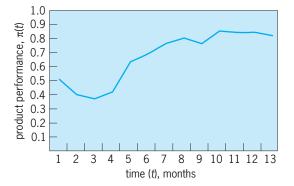
TABLE 2. Thresholds of overall quality and reliability							
Dimension	Threshold	Category					
Quality (defects/1000)	<5 5–20 >20	High Medium Low					
Reliability (mean time between failures, hours)	>200 100–200 <100	High Medium Low					

and low), then there are $3 \times 3 = 9$ possible pairs of values of the two attributes, and each pair can give rise to a different value of overall performance. Moreover, the overall performance can vary with time as the values of the attributes vary, and it is often useful to graph this variation.

For example, a new treadmill is brought to market, and the manufacturer has determined overall quality and reliability thresholds as shown in **Table 2**. The average time for customers to learn how to use all of the features on the unit ranges 2–5 h.

Here the most qualitative (subjective) element of interest is usability, because of the relative uncertainty in the training time. Therefore, the first step is to construct the set of combinations of low (L), medium (M), and high (H) attribute values for quality and reliability, giving (L,L), (L,M), (L,H), (M,L), (M,M), (M,H), (H,L), (H,M), (H,H). To incorporate the usability dimension, the learning time is categorized into intervals, such as "easy" for times up to 2.5 h and "normal" for times between 2.5 and 5 h. This forms an expanded range of attribute values. By assigning variables to each of these combinations and employing a process of pairwise comparisons among these attributes, relative weights can be assigned to establish an overall measure of product performance, $\Pi(t)$ [see illus.].

Quality, reliability, and warranties. Two types of product failures can generally be distinguished: those that occur relatively soon after they are produced and sold to consumers, and those that occur later after the product has been in use, perhaps even without fault for many cycles or hours of usage. From a manufacturer's point of view, the failures that occur early



Sample of monthly product performance, $\Pi(t)$. The variable $\Pi(t)$ is an index of the level of performance, with 1 corresponding to perfect and 0 corresponding to worst level.

in the product life are generally due to low quality and are attributed to a lack of conformance to manufacturing and production standards. Here quality of the product can be thought of as a state of acceptance that relates directly to the manufacturing processes, materials, and workmanship. Accordingly, quality is maintained and controlled through developed standards and specifications for characteristics such as dimensions and tolerances, composition and structure of materials, hardness and surface finish, and strength of linkages and mechanism.

Those failures that occur later in the life or usage cycle of a product are commonly due to low reliability, caused by problems in the design and engineering of the item. These failures typically occur as stress fractures, fatigue, and shearing, which relate to problems in the materials, basic design, and production processes. Sound planning early in the design process is necessary to overcome these and other problems in order to achieve high product reliability.

Warranties require the producer to pay all or some of the expenses involved in the repair or replacement of products that fail within a prescribed period of time. Whether planned by the manufacturer as part of a market strategy or stipulated by liability legislation, the costs for warranties can represent a significant portion of the cost to manufacture a product. Low product quality caused by poor conformance, low reliability, or both will lead to high warranty costs.

Operational measures of product performance. Warranty costs are expenses that occur because of low product quality. Therefore, measures that relate to the warranty expenditures also relate to product quality. Common operational measures that are used to track quality-related performance in production include the percentage of failures or defects, the number of warranty claims per unit sales, and the average warranty costs per unit sold. These measures, which involve warranty claims, also reflect customer attitude toward the product.

Concurrent engineering, quality function deployment, and total quality management (TQM) are modern management approaches for improving quality through effective planning and integration of design, manufacturing, and materials management functions throughout an organization. Quality improvement programs typically include goals for reducing warranty claims and associated costs because warranty data directly or indirectly impact most of the product quality dimensions. *See* ENGINEERING DESIGN.

Marlin U. Thomas

Bibliography. D. A. Garvin, Competing on the eight dimensions of quality, *Harvard Bus. Rev.*, 65(6):107-109, 1987; D. L. Goetsch and S. B. Davis, *Introduction to Total Quality*, 2d ed., Prentice Hall, 1997; C. L. Karnes, S. V. Sridharan, and J. J. Kanet, Measuring quality from the consumer's perspective: A methodology and its application, *Int. J. Prod. Econ.*, 39:215-225, 1995; N. A. Morgan and N. F. Piercy, Market-led quality, *Ind. Market Manag.*, 21:111-118, 1992.

Product usability

A concept in product design, sometimes referred to as ease of use or user-friendliness, that is related directly to the quality of the product and indirectly to the productivity of the work force. Customer surveys show that product quality is broken down into six components (in descending order of importance): reliability, durability, ease of maintenance, usability, trusted or brand name, and price. Ease of maintenance and usability both relate to product usability. Reliability also has a component of usability to it. If a product is too difficult to use and thus appears not to work properly, the customer may think that it has malfunctioned. Consequently, the customer may return the product to the store not because it is unreliable but because it does not work the way the customer thinks it should.

Since quality is often the main determinant in a customer's decision whether or not to buy the product, products that are highly usable are more likely to be purchased. Usability is especially important for computer products, both hardware and software, because poor products can be almost impossible to operate. As e-commerce grows on the Internet, the difference between a successful Web site and an unsuccessful one is often the usability of the site. If customers cannot easily make a purchase, an alternative site or method will be chosen. *See* HUMAN-COMPUTER INTERACTION; INDUSTRIAL ENGINEERING; PRODUCTIVITY.

Usability criteria. There are criteria by which a product's usability can be measured, including:

1. *Time to perform a task, or the execution time.* The ability to execute tasks quickly is a sign of good usability. Execution time is dependent on the number of steps that must be performed by the user. Intelligent layout of the keys, buttons, and other controls can also have a positive effect on execution time and usability.

2. *Learnability*. Usable products operate simply, and can be learned quickly. The steps are straightforward, intuitive, and easy to follow.

3. Mental workload, or the mental effort required to perform a task. A usable product will contain clues to using the product in the physical design instead of relying on people's memory of how to perform a task. Humans can retain about three to nine items in short-term memory at one time. If, due to a poor product design, the user is required to remember too much, then the usability of the product will suffer.

4. *Consistency in the design*. If several tasks must be performed while using the product, the operations should be consistent. Tasks are consistent if they involve similar steps or methods. In addition, product lines from the same company should operate consistently so that the user's learning experience is transferred from one product to the next. Learning time can be decreased because familiarity with one product can be applied to using a new product.

5. *Errors*. Errors can be categorized according to criteria 1–4 above. For instance, errors will occur if

a task requires so many steps that the user cannot remember the steps or their proper order; if a task is too difficult to learn; if the effort required to perform the task is too high; or if the steps in the task are so inconsistent with other steps they are difficult to remember.

The usability of a product usually cannot be optimized for all five criteria at the same time. Trade-offs will occur. As an example, a product that is highly usable in terms of fast execution times will often have poor usability in terms of the time needed to learn how to use the product. A product designer must be aware that it may not be possible for a product to be highly usable by all usability criteria, and so design according to the criteria that are most important to potential customers. Casual users of a product will have different demands on a product compared to expert users. *See* CONTROL SYSTEMS; HUMAN-FACTORS ENGINEERING; HUMAN-MACHINE SYSTEMS.

Laboratory methods for improving usability. Many companies, especially computer or consumer electronics companies, have laboratories in which to test the usability of their products. The methods of usability testing are formal experimentation, informal experimentation, and task analyses.

Formal experiments. Formal experiments take place in a laboratory in which different conditions can be controlled and manipulated. Usability is measured most often by counting the number of errors or recording the time needed to execute a task. Often, performance on different products or product features will be compared to determine which can be executed the fastest with the fewest errors. In a typical product design cycle, formal experimentation is often used to test users on early mock-ups, later on the product prototypes, and then on the final product.

The advantage of formal experimentation is that clear cause-and-effect relationships can be established. As an example, the experiment could show that a certain product design causes better usability performance either in faster use speed or reduced errors.

Formal experiments have some disadvantages when compared to other methods for testing usability. First, formal experimentation can be a costly procedure. Typically, each participant in the experiment is tested individually. In addition, setting up the experiment to control for possible confounding factors can be time-consuming. Another disadvantage is that a controlled setting may be good for establishing a cause-and-effect relationship, but it can also produce a sterile experimental environment that may not correspond to how the product will be used in practice. Finally, the data collected may not be very rich. The data will indicate that one product may be better than another, but it cannot necessarily be used to generate new ideas. In most cases, therefore, formal experimentation is used in the latter stages of the product development cycle. Informal experimentation in the early stages can help identify potential product features that may improve usability.

Informal experiments. Surveys, questionnaires, observations in natural settings, and focus groups are ways

to provide product designers with ideas on how a product can be improved. The surveys and questionnaires can be used throughout the design process to assess users' perceptions of the product, including what users like and dislike about the product. At the conception of the product, questionnaires can assess the need for a new product or the need to redesign an existing product. Once the product is placed on the market, product surveys can assess how the users have received the product. Questionnaires can also accompany formal experiments to measure users' satisfaction with the product.

Users can be observed in various settings performing various tasks with the product. Typically, products are tested inside the company designing the product. They can also be brought to consumer product shows or placed in stores. The people using the products in these settings are usually videotaped. The videotapes are often analyzed using a technique called content analysis. For content analysis, definable events, such as obvious displays of frustration, are counted when viewing the videotapes. If too many customers show high levels of frustration, then the problem tasks can be identified and corrected. The videotapes can be especially useful for identifying errors from which the users cannot recover. These kinds of errors are likely to cause product returns.

Focus groups, groups of usually four to eight people, are brought together to use a product and discuss its advantages and disadvantages. A focus group is designed to elicit comments from the participants that normally would not be elicited individually. A focus group moderator can direct the questions and discussions.

Informal experiments are particularly important at the beginning and end of the design process. Although clear cause-and-effect relationships cannot be established, the information received from these techniques can be very valuable in directing the product designers to potential usability problems.

Task analysis. Two kinds of task analysis are typically used: protocol analysis and cognitive analysis. For a protocol analysis, a single user will be brought to the lab to perform various tasks with the product. Generally, the user will be asked to verbalize his or her thought processes (provide verbal protocols) during execution of a task. The verbal protocols then can be analyzed to determine if the participant had the correct goal for accomplishing a task, if the goal was executed correctly, and what kind of confusion occurred. Verbal protocols provide an experimenter with a rich set of data that can be used to improve the product's usability.

In a cognitive model, the cognitive processes of a potential user can be modeled and the model can be executed, similar to a computer program. Some models are derived from the verbal protocols of subjects. Cognitive models can be used to make quantitative predictions of the usability of products. These predictions, which can be used in lieu of the more costly method of formal experimentation, have been shown to be very accurate. *See* COGNITION; MODEL THEORY; SIMULATION; SYSTEMS ANALYSIS; WORK MEA-SUREMENT.

Summary. Many of the best and most successful companies place great emphasis on usability of their products. Products that are not usable will lose market share, and customers will return these products to the store if they think that they are not working properly. In many cases, they are not working properly because the customer cannot understand how to use them. Usable products also have an effect on worker productivity. Excessive learning times, execution time, or errors can slow the accomplishment of tasks and reduce productivity. Products that are poor in usability force the worker to concentrate on how to use the product instead of concentrating on how to perform the task. Although laboratory methods for improving usability can increase the cost of the product design, the benefits (market share, productivity) will outweigh the costs. See METHODS EN-GINEERING; OPTIMIZATION. Ray Eberts

Bibliography. R. W. Bailey, *Human Performance Engineering*, Prentice Hall, Upper Saddle River, NJ, 1996; R. Eberts, *User Interface Design*, Prentice Hall, Englewood Cliffs, NJ, 1994; D. Norman, *The Psychology of Everyday Things*, Basic Books, New York, 1988; M. S. Sanders and E. J. McCormick, *Human Factors in Engineering and Design*, 7th ed., McGraw-Hill, New York 3d ed., 1993.

Production engineering

A branch of engineering that involves the design, control, and continuous improvement of integrated systems in order to provide customers with highquality goods and services in a timely, cost-effective manner. It is an interdisciplinary area requiring the collaboration of individuals trained in industrial engineering, manufacturing engineering, product design, marketing, finance, and corporate planning. In many organizations, production engineering activities are carried out by teams of individuals with different skills rather than by a formal production engineering department. Although production engineering originated in the manufacturing sector, recently it has found increasing application in the service industries, such as banking, airlines, and entertainment. Thus "product" denotes both a traditional manufactured product and a service product, such as an insurance policy or a bank loan. While traditional production engineering activities have taken place within a single factory, it is becoming more common for the production engineering team to include not only in-house personnel from the different skill sets mentioned above, but also personnel from several geographic locations and even other companies. It is quite common for a firm's design activity to take place in one country, production of components in another, and assembly of the final product in a third. Important components and subsystems of the finished product may be subcontracted to other firms, requiring production engineering teams to include personnel from different companies. Recent developments in information technology, such as the Internet and computersupported cooperative work, are important enablers of this process.

Product design. In today's global markets, where companies from all over the world can compete to sell in a given area, rapid response to changing market conditions is essential to a company's survival. The company must rapidly introduce new products to exploit market opportunities, and continuously improve its operations to deliver higher-quality products faster and at lower cost than its competitors. The introduction of a new product is critical for production engineering. Initially, the production engineering team works with the designers, helping them to develop a product that can be manufactured economically while preserving its functionality. Features of the product that will significantly increase its cost are identified, and alternative, cheaper means of obtaining the desired functionality are investigated and suggested to the designers. Decisions as to whether to manufacture a given component of the product internally or to purchase it from outside suppliers must be made, based on cost and quality considerations and the reliability of the production process or supplier. Similar considerations in the service industries involve decisions as to whether or not to subcontract certain services. The process of concurrently developing the product design and the production process is referred to by several names such as design for manufacturability, design for assembly, and concurrent engineering. See ACTIVITY-BASED COSTING; DESIGN STANDARDS; PRO-CESS ENGINEERING; PRODUCT DESIGN; PRODUCTION PLANNING.

Production process. The specification of the production process should proceed concurrently with the development of the product design. This involves selecting the manufacturing processes and technology required to achieve the most economical and effective production. The technologies chosen will depend on many factors, such as the required production volume, the skills of the available work force, market trends, and economic considerations. In manufacturing industries, this requires activities such as the design of tools, dies, and fixtures; the specification of speeds and feeds for machine tools; and the specification of process recipes for chemical processes. Deciding whether individual operations should be performed manually or using automated machinery is another recurring decision. In a service industry, the decisions might be for the specification of the steps required for a loan approval procedure or the processing of an insurance claim, along with the type of computer support to be provided at each step. The production engineering team must interact with other groups, such as research and development, to determine the viability of developing new technologies; machine and tooling design groups, to provide input on the needs of the process; information systems, to develop the necessary computer support; and finance and marketing groups, to evaluate the impact on costs and customer satisfaction of the decisions made. There is also interaction with human resource and industrial engineering groups to design workplaces in which workers can perform efficiently and safely, and to provide workers the necessary training. There should be a constant dialogue between the production engineering team and the product designers, enabling the latter to take advantage of the latest process knowledge to enhance the product design, and the former to adapt to changes in required functionality. *See* AUTOMATION; HUMAN-FACTORS ENGINEERING; MANUFACTURING ENGINEER-ING.

A critical factor in selecting the appropriate production process is the nature of the market for which the product is designed. If the product will serve a broad consumer market, such as consumer electronics or automobiles, the firm must produce a relatively standard product in very high volumes for a reasonably long time. In this situation, referred to as mass production, the firm can afford to develop highly specialized production equipment that is capable of performing specific operations for a particular product. Manual tasks can be divided into simple, repetitive elements which do not require high worker skill levels. Production engineering is active in the design of the production equipment and its integration with workers into an effective manufacturing process.

Pilot production. Actual production of physical products usually begins with a few prototype units being manufactured in research and development or design laboratories for evaluation by designers, the production engineering team, and sales and marketing personnel. Once a basic product design and production process have been agreed upon by all parties, relatively small volumes of the product will be produced on a pilot production line, which is composed of the same basic equipment and processes that will be used in the actual production facility. The goal of this pilot production phase is to give the production engineering team hands-on experience producing the product, allowing problems to be identified and remedied before investing in additional production equipment or shipping defective products to the customer. The pilot production process involves changes to the product design and fine-tuning of unit manufacturing processes, work methods, production equipment, and materials to achieve an optimal trade-off between cost, functionality, and product quality and reliability. Once a satisfactory degree of efficiency and predictability has been achieved, the production process is installed in a full-scale production facility capable of producing the product in sufficient volume to meet the market demand. See PILOT PRODUCTION; PROTO-TYPE.

Production facility. The production facility itself can be designed around the sequence of operations required by the product, referred to as a product layout (**Fig. 1**). In another scenario, the firm develops custom products to meet the specialized needs of individual customers. This requires the firm to produce a wide range of products in small volumes, which is referred to as job shop production. General-purpose production machinery is used, and often must be set up for each individual job, incurring significant changeover times while this takes place. This type of production facility is usually organized in a process layout (**Fig. 2**), where equipment with similar functions is grouped together. The highly job-dependent nature of the production operations also requires a skilled work force capable of evaluating different alternatives and making good choices. *See* HUMAN-MACHINE SYSTEMS; PRODUCTION METHODS.

Many firms fall between these two extremes, where a limited range of parts are produced in medium volumes. This type of environment is known as batch manufacturing, since products are made in batches to amortize changeovers of machines over an economic number of units of the product. Cellular or group technology layouts, where groups of machines are dedicated to the production of families of similar parts, are often used in these environments. An important production technology for the latter environment is flexible manufacturing systems, which are generally composed of a number of computer-controlled machines linked by a centralized computer controller and automated material-handling equipment. The goal of these systems is to be able to produce a variety of products with minimal changeover time, which is achieved by reprogramming the computer control as opposed to adjusting or reconfiguring the actual hardware. See FLEXIBLE MANUFACTURING SYSTEM.

The design of the production facility, an important part of production engineering, also overlaps significantly with industrial engineering. The number of each type of machine needed and how the machines are organized in the plant space available determines not only the production capacity of the facility but also to a great extent the degree of difficulty of the activities involved in coordinating the flow of material through the plant to manufacture the product. The day-to-day coordination of production activities (production planning and control), such as how to prioritize the different tasks requiring machine time or attention by personnel, are the domain of industrial engineering and operations management. However, the nature of these tasks and their difficulty is affected heavily by the plant layout and the choice of process technology. The effective performance of these tasks is a major part of any successful production engineering effort. See INDUSTRIAL ENGI-NEERING; INDUSTRIAL FACILITIES.

Process improvement. The production engineering process does not stop once the product has been put into production. A major function of production engineering is continuous improvement—continually striving to eliminate inefficiencies in the system and to incorporate and advance the frontier of the best existing practice. Production processes tend to evolve over time as more cost-effective methods become available. Market and product characteristics also change, requiring modification of the production processes. The task of production engineering is to identify potential areas for improving the performance of the production system as a whole,

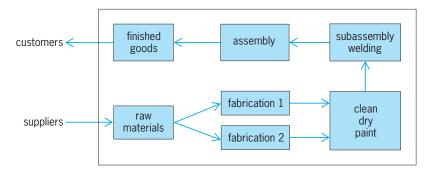
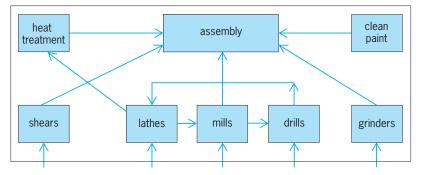


Fig. 1. Example of a process for a mass-production environment.





and to develop the necessary solutions in these areas. This often requires considerable involvement of the work force, with the production engineering team serving as facilitators and technical consultants. The solutions developed may range from modifications of production equipment and work methods to changes in product design or the revision of a quality specification. A successful continuous improvement program serves to eliminate waste in the company's operations, providing it cost and quality advantages over its competitors. *See* PROD-UCT QUALITY. Reha M. Uzsoy

Bibliography. R. G. Askin and C. R. Standridge, *Modeling and Analysis of Manufacturing Systems*, John Wiley, New York, 1993; J. Browne, J. Harhen, and J. Shivnan, *Production Management Systems*, Addison-Wesley, 2d ed., 1996; J. Heizer and B. Render, *Principles of Operations Management*, Allyn and Bacon, 6th ed., 2005; J. A. Tompkins and J. A. White, *Facilities Planning*, John Wiley, 3d ed., 2002; J. A. White (ed.), *Production Handbook*, John Wiley, 1987.

Production lead time

The total time required to convert raw materials into a finished product. More than the simple sum of machine operation and assembly times, it also includes time spent waiting for machines to be available, moving between machines, and performing quality inspections and other nonoperation functions. Short production lead times are ideal, as long times reduce the responsiveness of a company and result in high inventories. While customer lead time is the interval of time between placing the order and receiving the order, production lead time typically focuses on the time required to produce the order once all the necessary materials are present. It does not take into consideration the time required to obtain the necessary materials or the time the materials wait in the plant before production begins.

Production lead time is very important in maketo-order industries (such as the aerospace industry). There, lead time can be a deciding factor in obtaining orders. In addition to low cost and superior quality, companies win orders by promising to deliver them quickly. Lead time is not as much of an issue in make-to-stock industries (such as small appliances and other retail products) in that these industries are building inventories against future demand, thus providing a buffer against longer than expected production lead times.

Components. The primary components of production lead time are setup, processing, handling, and nonoperation. Setup time (t_s) is required to prepare the machine for the particular order of parts. Processing time (t_p) includes the time spent loading and unloading the part from the machine as well as actual machining time. Handling time (t_h) is the time an order spends being transported between machines. Nonoperation time (t_n) includes wait time (for both machines and transportation) and time spent on certain other activities such as inspection. Assuming that an order consists of Q identical parts and that there are N operations in the processing sequence (with *i* indicating the operation in the processing sequence), production lead time (LT) is expressed by Eq. (1).

$$LT = \sum_{i=1}^{N} [t_{si} + Qt_{pi} + t_{bi} + t_{ni}]$$
(1)

For example, consider an order that consists of 20 identical parts. Five operations must be performed to produce this part. Setup, processing, handling, and nonoperation times are provided in the **table**. The production lead time for this order as computed from Eq. (1) is 44.7 h. Assuming a 40-h work week, the production lead time for the order of 20 parts is 1.1 weeks.

Setting due dates. Knowledge of the lead time for a product is important for setting due dates for orders. When a customer places an order, either the customer states when the order is needed or asks

Data for lead time calculation				
Operation (<i>i</i>)	Setup time (t _s), h	Processing time (t _p), h/part	Handling time (t _h), h	Nonoperation time (<i>t_n</i>), h
1 2 3 4 5	2 3 1 2 4	0.10 0.05 0.07 0.15 0.20	0.25 0.25 0.25 0.25 0.25 0.25	3 4 2 6 5

for an estimated time of completion. In either case, a method of estimating lead time is required. If the customer requires a specific delivery date, the company needs to be able to determine if the customer due date is achievable given current production orders. If the customer requested date is not feasible, an estimate of when the order will be done is necessary for negotiating a new due date.

One method for setting due dates requires knowledge of the average production lead time and its standard deviation (σ_{LT}). Given this information, an estimated lead time is calculated such that there is a good probability of completing the order on time. The percentage of orders completed on time is known as the service level (s), where z_s denotes the value on the measurement axis (for which s of the area under the z curve lies to the right of z_s). The z curve is the standard normal curve. The standard normal table lists the values of z_s for values of s ranging from 0 to 100%. To be more precise, the table actually works in reverse in that for values of z_s ranging from -3.49 to +3.39, the value of s is given. Assuming that production lead time is normally distributed, the quoted lead time for an order can be computed by Eq. (2).

Estimated lead time =
$$LT + z_s \sigma_{LT}$$
 (2)

For example, suppose the average production lead time is 2 weeks with a standard deviation of 0.75 week. If a 95% service level is desired, an estimated lead time for the order is computed to be 3.23 weeks in Eq. (3). The value of 1.645 for z_s was obtained by

Estimated lead time = 2 weeks + (1.645)(0.75)= 3.23 weeks (3)

referencing a standard normal table. First, the value 0.95 is found in the body of the table (corresponds to s = 95%). The corresponding row and column label values are added together to obtain the value of z_s . The additional 1.23 weeks (1.645×0.75) is often called the safety lead time. Caution should be exercised when using this method to estimate lead times when the standard deviation is very high. A large standard deviation will lead to a large safety lead time. Since one-half of all orders are expected to complete in less than the average lead time, a large safety lead time can lead to a large finished goods inventory if orders cannot be delivered early.

Managing lead time. Production lead time can make the difference between receiving customer orders and going bankrupt. In all industries, a short production lead time is preferred to a long time. One way to reduce production lead time is to reduce the individual components, such as setup, handling, and nonoperation times. In some industries (such as metalworking factories), these three components constitute 95% of the total production lead time. Setup time can be reduced through the use of common fixtures and by scheduling similar orders to run back to back. Handling time can be reduced by improving the layout of the facility. Machines that follow one another

in the processing sequence of many products should be located near one another. Finally, nonoperation time can be reduced by efficient scheduling practices and the elimination of unnecessary inspections and other nonoperation activities.

Production lead time is also related to the amount of work in process (orders currently being worked on in the plant) on the production floor and the production rates of the machines. This relationship, shown in Eq. (4), is known as Little's law. For indus-

$$Production lead time = \frac{Work in process}{Production rate}$$
(4)

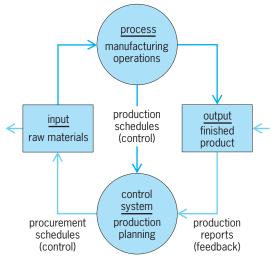
trial plants operating at less than 100% capacity, production lead time is largely unaffected by the amount of work in process in the plant. However, as the production rate reaches its maximum value (capacity), increasing the work in process merely serves to increase the production lead time. Therefore, controlling the release rate of materials to the plant is an important part of controlling production lead time. Little's law also shows that production lead time can be reduced by increasing the production rate of the plant. This approach typically requires an increase in the plant's capacity. Capacity can be increased by purchasing additional equipment or by working more hours (for example, authorizing overtime or adding another shift).

Production lead time is essential to the success of a company. In the global market place, customers demand low-cost, high-quality products in a short period of time. As a result, effective management of production lead time is as important as striving for low cost and high quality. Short production lead times allow a company to be responsive to its customers' needs. *See* INDUSTRIAL ENGINEERING; INVEN-TORY CONTROL; MANUFACTURING ENGINEERING; MA-TERIAL RESOURCE PLANNING; MATERIALS HANDLING; PRODUCTION PLANNING; QUALITY CONTROL in the McGraw-Hill Encyclopedia of Science & Technology. Elin M. Wicks

Bibliography. M. P. Groover, Automation, Production Systems, and Computer Integrated Manufacturing, Prentice Hall, 2d ed., 2000; W. J. Hopp and M. L. Spearman, Factory Physics, Richard D. Irwin, 2d ed., 2000.

Production planning

The function of a manufacturing enterprise responsible for the planning, scheduling, and coordination of all production activities. The planning phase involves forecasting demand and translating the demand forecast into a production plan that optimizes the company's objective, which is usually to maximize profit while optimizing customer satisfaction. The twin objectives of maximum profit and maximum customer satisfaction are not always compatible. Sometimes, a highly-valued customer may request a product that must be made under high-cost conditions. During the scheduling phase the production plan is translated into a detailed, sometimes hour-by-hour, schedule of



The production process as an input-output process.

products to be made. During the coordination phase, actual product output is compared with scheduled product output, and this information is used to adjust production plans and production schedules. *See* ACTIVITY-BASED COSTING; OPTIMIZATION.

If the production or manufacturing process is viewed as an input-output process, then the production planning function can be viewed as a control process with feedback (see **illus**.). The control is in the form of schedules and plans, while the feedback results from the comparison of the production reports with the production schedules. *See* CONTROL SYSTEMS.

Techniques. In the planning process, material requirements planning (MRP-I) or manufacturing resources planning (MRP-II) may be used. MRP-I is an inventory or materials planning tool, and MRP-II is a resources planning tool that includes planning for labor hours, machine hours, cash, floor space, and so forth. The primary difference is that MRP-II includes capacity resources planning (CRP). Essentially, MRP techniques are based on a master schedule that is created for delivery of product. A bill of materials (BOM) is used to establish the detailed schedule for all components, subassemblies, assemblies, and deliveries. Just-in-time (JIT) is both a technique and a philosophy of manufacturing planning and control. The objective is to have no inventory in the plant that is not being worked on actively. Ideally the work-inprocess (WIP) should be zero under a just-in-time system. Usually true just-in-time is not achieved but serves as a goal. See MATERIAL RESOURCE PLANNING.

The quantity of material ordered at any given time may be determined either by the MRP system or by use of economic order quantity (EOQ) or economic lot sizing (ELS) techniques, which consider cost of carrying inventory and the fixed costs associated with placing an order. The economic order point (EOP) and the reorder point (ROP) are used to determine when an order or a reorder is placed as a function either of time or of inventory level.

The scheduling function determines the routing of parts through the shop and the loading of machines. The objective should be to balance the work load on machines, departments, and the entire plant. Computer-aided process planning (CAPP) techniques are helpful in establishing the necessary operations, machines, and the sequences of operations. This information is useful for machine loading and part routing. *See* COMPUTER-AIDED DESIGN AND MANUFACTURING.

The operations on the shop floor may be controlled by visible records such as cards (in Japanese, kanbans) when the factory is using a just-in-time system. In a just-in-time system, every part is expected to arrive at its point of use just as the operation is ready to use it. Moving a kanban upstream is equivalent to placing an order, and moving a kanban downstream corresponds to the delivery of goods. Modern computer technology has put terminals on the shop floor. These terminals are used to download schedules and work assignments and to upload completion of product, hours used, and so forth; they tend to remove the paperwork from the shop floor. Many different types of charting techniques are available to accomplish these tasks. The Gantt chart, the basis of many charts, is the oldest such technique. See COM-PUTER; DIGITAL COMPUTER; GANTT CHART.

Constraints. Production planning generally has several constraints: demands must be met within reasonable limits inventories of raw materials, inprocess materials, and finished goods must not be excessively large or small; variations in personnel are usually limited; and the available production facilities are limited. A customer will not wait indefinitely for a product to be delivered. The length of time that is acceptable is dependent on the industry and the product. Inventories represent an investment in material that is not producing immediate revenue and therefore must be limited. Some inventories are necessary for efficient manufacturing in lot sizes and for balancing personnel and equipment requirements with customer demand. See INVENTORY CONTROL; PRODUCTION ENGINEERING; PRODUCTION SYSTEMS.

Applications. The functions of production planning are dependent upon the type of industry and type of product. If the manufacturing process is of an assembly-line nature, the planning is primarily of outputs and inputs and the avoidance of bottlenecks. If the manufacturing process is of the job-shop type, that is, if the product is made to the customer's order, the planning must involve a detailed schedule of individual parts through the manufacturing process. In both cases, the planning starts with a forecast of outputs or demands and proceeds to determine personnel and material inputs consistent with demands and company policy. Some of the readily identifiable tasks within the production planning function are planning, scheduling, dispatching, machine loading, expediting, inventory record keeping, and material requisitioning. All of these tasks involve the collection, analysis, interpretation, and dissemination of data. The sophistication of the data system is highly variable from company to company and also within departments or branches of a company.

The production control function does not exist independently of other functions within the company. Its activities must be coordinated with purchasing, sales, finance, production, tool engineering, design engineering, maintenance, and so on. Its primary responsibility is to coordinate the manufacture of the product to best satisfy the customer and the constraints imposed by the company management. *See* OPERATIONS RESEARCH. John E. Biegel

Bibliography. A. Ahmadian, R. Afifi, and W. D. Chandler, *Readings in Production and Operations Management: A Productivity Approach*, 1990; E. S. Buffa and R. K. Sarin, *Modern Production Operations Management*, 8th ed., 1987; J. L. Riggs, *Production Systems: Planning, Analysis, and Control*, 1987, reprint 1992; T. E. Vollmann, W. L. Berry, and D. C. Whybark, *Manufacturing, Planning and Control Systems*, 4th ed., 1997.

Production systems

A collection of workstations that perform operations such as manufacturing, assembly, inspection, finishing, and testing to create products. A workstation is a collection of resources that performs the same or similar operations for the same set of products. The resources at a workstation may be automated machines, machines operated by humans, or human operators performing manual operations. The storage and delivery of both raw materials and finished products are also part of the production system, but the focus here will be on the system that transforms materials and/or components into products.

The design or configuration of workstations in a production system is defined by the flow of products in the system. The flow of products is determined by how products move among the workstations during production, and is dictated by the volume and variety of products that the system will produce. In general, a simpler flow corresponds to lower product variety at higher volumes, and a more complex flow corresponds to higher product variety at lower volumes.

The specific production operations performed at workstations will vary greatly depending on the nature of the product and its requirements. For example, the production of automobile body components is performed through a series of metal-forming steps. The assembly of automobile bodies is generally accomplished in a set of automated welding process steps. The final assembly of the finished automobile is typically completed in a separate series of manual operations. Not all production systems are large and require the use of expensive machinery—for example, fast-food restaurants, office work, construction operations, and many others are production systems.

The design and operation of a production system can be quite complex and requires that many interrelated decisions be made. Examples of design decisions are the assignment of work to workstations (line balancing), the sizing of storage space for workin-process between workstations, the methods used to control production, and the methods used for moving product from workstation to workstation. Examples of operational decisions are setting staff

		Product variety					
		Very low	Low	Moderate	High		
	Low				Job shop	High	Pro
Product volume	Moderate			Group technology/ batch production		Moderate	Product flow
Product	High		Production line			Low	' complexity
	Very High	Continuous flow				Low	xity

The relationship between product variety and volume, production system type, and product flow complexity.

levels and shift lengths, production scheduling, and production system bottleneck isolation and improvement.

Production system configurations. The flow of products among the workstations in a production system will define its configuration or design. The type and complexity of product flow designed into a production system are determined by the volume and variety of products produced. At the two ends of the flow complexity spectrum are production lines with simple flow and job shops with complex flow. In between, there are hybrid systems called batch production systems and group technology systems (see **illustration**). In reality, production facilities may be a combination of several types of production system configurations.

A special case is that of a fixed-position production system, where the product stays fixed in position and various workstations move to the product to complete their operations. Examples of such products are ships, airliners, and buildings. These products are often unique or are produced this way because of the product size. Conceptually, these systems can viewed as very large job-shop and assembly systems.

Production lines and continuous-flow systems. Production lines and continuous-flow systems have very simple product flow. Product flow in these systems is conceptually a straight line, with all products moving through the same sequence of workstations. Examples are production lines such as automobile final assembly systems and fast-food preparation systems. In these systems, parts are distinct and countable (such as automobiles and sandwiches). These are referred to as discrete product production lines. Production systems with straight-line flow that produce fluids and powders are referred to as continuous-flow production systems, and are common in the chemical, food, and cosmetics industries.

Production lines and continuous-flow systems are normally high-volume systems. In early production lines, it was necessary that the operations performed at each workstation on each product were almost always the same to achieve high volume. Modern production lines can incorporate more product variety than in the past, but are still primarily limited to variations of a basic product architecture or "platform."

Early production lines consisted of workstations of manually operated machines or of human operators performing manual operations. The advent of computer-controlled machinery has permitted many workstations previously requiring manual operation to become fully automated. A modern automobile body assembly system is an example of where products are automatically moved from workstation to workstation, consisting of machinery and robots performing welding operations. As impressive as fully automated production lines are, many products are still produced most efficiently using manual operations. The prior examples of automobile final assembly and fast-food preparation come to mind.

Job-shop production systems. Job-shop production systems are normally intended for applications where the products are made either in single units or in limited production runs. The intent of job-shop production systems is to produce a high variety of products (in some cases an unlimited variety of custom products) at low volumes. Job shops are typically organized by process. For example, a manufacturing job shop may consist of separate areas for milling, drilling, and turning. The product flow in a job shop is very complex since each product will usually require a separate sequence of operations (called a product routing), and there is essentially no set limit on the number of times a product may have to visit the various processes. From a product flow viewpoint, a hospital is organized as a job shop. Patients (analogous to products in manufacturing) may need to visit a number of different processes, such as admissions, radiology, and orthopedics, but not necessarily in the same order or the same number of times.

Similar to what has been seen in modern production lines, computer-controlled machinery has also had an impact on job shops. Operations that previously required manually operated machinery and highly skilled labor have been replaced in many situations by computer-controlled machinery. This is more common in job shops that produce a number of products in larger volume, since there is relatively high overhead invested in programming of machines for different products. This requires higher volumes to be economical. Job shops strictly producing custom and prototype products are still commonly staffed with highly skilled workers, such as machinists, toolmakers, and diemakers, who perform operations using of a variety of generalpurpose machine tools.

Hybrid systems. Because many products or sets of related products are neither high-volume nor lowvolume, as would be found in production lines and job shops, hybrid systems have evolved. These systems strike a compromise between volume and variety, falling between that of production lines and job shops. Both batch and group technology production systems can be considered hybrid systems. The fundamental concept implemented in these types of systems is to group similar products and maintain a separate production system for each group. Differentiation of group technology from batch production is found in how product groups are formed, although there is no clear distinguishing boundary. In batch production, grouping of products is relatively simple. A criterion, such as size, is used to group products. An example of a batch production system is automotive sheet metal stamping (forming). Separate stamping lines, which accommodate tooling of approximately the same size and weight, are used to produce a group of similar-sized products. In group technology, the criteria for grouping products is generally more complex and is based on similarities in production operations and routings, tooling, and components. In this sense, group technology may be considered a special case of batch production. Grouping is done to minimize changeover time and cost when switching production to different products. In electronic circuit board assembly, a group of products may use components with similar circuit board insertion methods and similar size, thus permitting the use of the same automated machinery without the need of long setup times between product batches.

The production systems for a specific group of products within a hybrid system may fall anywhere between a job shop and production line, depending on the requirements of products in the group. In the sheet metal stamping example, the production system for a group is a production line and is relatively high-volume. In the electronic circuit board assembly example, the production system may contain automated machinery, which also can accommodate several different product routings. In this sense, the production system has some of the flexibility characteristic of a job shop.

Production control and scheduling. An important feature of production systems, regardless of type, is the method or methods used to control product flow into the production system, and how waiting products are processed at workstations. This is determined by the methods for initiating new production or production control, and how work at individual workstations is scheduled.

The two major categories of production control are "push" and "pull" production control systems. Production plants may use both types of production control at different areas or subsystems within the plant. In a push production control system, initiation of production is based on future orders and forecasted future demand. In a pull production control system, new production is initiated to replenish inventory that has been consumed. The decision on what method to use depends on many criteria, but demand variability, inventory and shortage costs, and production lead time are critical factors. As a simplified example, consider two different types of fast food: hamburgers and fried chicken. In anticipation of mealtime demand, a burger store may cook several burgers in advance (similar to a push system) and then cook more only to replace burgers purchased. This pull system works because the time to cook a burger is short. In contrast, the fried chicken store must cook a much larger quantity of chicken in anticipation of mealtime demand, as well as any large catering and phone orders received. More chicken is cooked in anticipation of demand and not necessarily because of chicken purchased. This push system is used since the time to cook chicken is much longer than the time to cook a burger. The shortage cost in terms of lost business and customer good will is much greater for the chicken store. See INVENTORY CONTROL

The scheduling of work at workstations is a complicated problem that affects the performance of the production system. In many production systems, scheduling equates to sequencing or ordering of the products waiting for work at a workstation. In a production line, scheduling may only be necessary before the first workstation since all products follow the same routing. Depending on how products are physically moved from workstation to workstation, changing the sequence in a production line may be very difficult. Since volume is a critical performance measure for production lines, scheduling of jobs is often executed to maximize production volume. Software packages are often used to help sequence jobs in production lines.

In job shops, scheduling is much more complicated. Because of high product variety and the number of different product routings, the products available for scheduling are different at different workstations and are constantly changing. The criteria used for scheduling products are also more complicated. Some common objectives of scheduling are to minimize the average time that products are in the production system, and to minimize the average lateness of products. Some software exists that can help schedule work in jobs shops. In many cases, scheduling is performed by employees based on years of experience and intuition. *See* SCHEDULING.

Performance measures. There are many ways to measure the performance of a production system. In today's competitive business environment, it is a given that a production system must produce the correct product at the required quality level. The efficiency at which this is achieved will vary. From a production systems perspective where the concept of product flow is applicable to many different products and situations, there are three main interrelated performance measures often used that answer the basic questions of how fast, how long, and how much.

1. Throughput (for example, products/hour), measures how much product a system can produce in a given time.

2. Time-in-system, flow time, or cycle time is a measure of how long it takes to produce a product

from start to finish, and reflects the responsiveness of the system to changes in customer demand.

3. Work-in-progress is a measure of how much partially completed product is typically found in the production system, and is related to the cost of production and the responsiveness of the system.

Total systems approach. The resources, processes, and methods used to produce, store, and deliver products often constitute a large and complex production system. This system exists to meet the goals and objectives of some organization. As such, the design and operation of the system and its components and subsystems should be evaluated in the context of the total production system.

Other critical components of the production system include the layout or arrangement of workstations in a facility, and the equipment and methods used to move and store material and products in the facility.

At the workstation level, many important decisions must be made about the operations and methods for completing the operations. In more typical or traditional production systems, this includes detailed decisions about manufacturing processes such as casting and molding, shearing and forming, machining/material removal, heat treating, finishing, assembly, and inspection. *See* INDUSTRIAL ENGINEER-ING; MANUFACTURING ENGINEERING. David S. Kim

Bibliography. R. G. Askin and C. R. Standridge, Modeling and Analysis of Manufacturing Systems, 1993; W. J. Hopp and M. L. Spearman, Factory Physics: Foundations of Manufacturing Management, 2d ed., 2001; S. Nahmias, Production and Operations Analysis, 5th ed., 2005.

Productivity

In a business or industrial context, the ratio of output production to input effort. The productivity ratio is an indicator of the efficiency with which an enterprise converts its resources (inputs) into finished goods or services (outputs). If the goal is to increase productivity, this can be done by producing more output with the same level of input. Productivity can also be increased by producing the same output with fewer inputs. One problem with trying to measure productivity is that a decision must be made in terms of identifying the inputs and outputs and how they will be measured. This is relatively easy when productivity of an individual is considered, but it becomes difficult when productivity involves a whole company or a nation.

Calculating productivity. A simple example of productivity is a paper carrier on a bicycle. If the paper carrier delivers 100 papers in 50 min, his productivity is 2 papers per minute. The output is the number of papers delivered. The input is the time required to deliver the papers (labor minutes). To calculate the productivity, the output is divided by the input (100 papers by 50 min). Later, the paper carrier analyzes the route taken to deliver the papers. Changes in the route are made to eliminate some backtracking and to take some additional shortcuts. Now when

the carrier delivers the same 100 papers, only 40 min are required. Thus, the new value for productivity is 100 papers divided by 40 min, or 2.5 papers per minute. The carrier has increased productivity by using the available resources (time) more effectively. Note that if productivity is calculated based on time, different units could be selected. Instead of minutes, the time units could be hours. The value 2.5 papers per minute could be converted to 150 papers per hour. Many industry and government productivity figures use hours as the time base.

For large organizations, it is difficult to determine every input that goes into making a product or providing a service. If the subject of a study is an automobile manufacturer, the many parts and services required to produce each automobile must be taken into consideration. Some of them are easier to quantify, such as the amount of steel, vinyl, and other materials that go into each automobile; the number of labor hours required to assemble each automobile (direct labor); and the cost of storing the parts in the factory while awaiting assembly. However, there are many factors that are not directly involved with the production of the automobile without which the company could not produce automobiles. These factors are more difficult to determine, but they are still inputs to the production process. Some of these factors include the cost of the required secretarial services, the cost of engineers that designed the automobile and set up the production line, and the cost to maintain and pay taxes on the factory.

Productivity measures. Industry and government officials have adopted three common types of productivity measures. However these measures are not exclusive. Many companies develop their own measures based on their unique situations. Therefore, it is often difficult to compare the productivity figures of different companies or nations. A comparison of productivity figures can be made only for those values that are calculated by using the same type of measure.

Partial productivity. In this simplest type of productivity measure, a single type of input is selected for the productivity ratio.

The previous example of paper carrier productivity uses a partial productivity measure where direct labor minutes is the input. Typically the company or organization selects an input factor that it monitors in daily activity. Direct labor hours is a factor that most companies monitor because they pay their employees based on hours worked. Therefore, the U.S. Department of Labor uses output per employee hour as a productivity measure in many of its published indices. Other factors that are commonly used in partial productivity ratios are capital (money invested in facilities and equipment), materials, or energy.

The major problem with a partial productivity measure is that it does not include all of the inputs required to make a product or provide a service. In the paper delivery example, if the carrier buys a car to deliver the papers, the delivery time can be reduced to 20 min. Based on the same partial productivity measure, the new productivity would be 5 papers per minute, or double what could be achieved by using a bicycle. However, if a partial, productivity measure based on capital is used, the carriers productivity would actually decline, because a car is much more expensive than a bicycle. It would now require a great deal more capital to deliver the papers. This example illustrates the problem with choosing a single factor upon which to base a productivity measure.

Total factor productivity. This measure combines labor and capital, two of the most common input factors used in the partial productivity measure. This measure is often used at the national level, because many governments collect statistics on both labor and capital. In calculating at the national level, the gross national product (GNP) is used as the output. A problem with this measure is that it does not include some types of inputs required to yield products.

Total productivity. This measure incorporates all the inputs required to make a product or provide a service. The inputs could be grouped in various categories as long as they determine the total inputs required to produce an output.

Because labor, capital, material, and energy all have different units, they cannot be directly added. Normally, they are all converted to constant standardcurrency amounts before they are added. An advantage of using total productivity is that it includes all the inputs required to achieve the product. A disadvantage is that companies or governments often do not have the data required to determine all the input factors.

Factors affecting productivity. Many factors affect productivity. Some general categories for these factors are product, process, labor force, capacity, external influences, and quality.

Product. The type of product and the design affects productivity. Ease of manufacture, assembly, and maintenance have come to be emphasized. Materials that are easier and cheaper to process have been developed. These factors have had a positive effect on productivity.

Process. The process that is used to manufacture a product or perform a service can affect the productivity. A simple example is a worker using a manual drill compared to worker using an electric drill. The capital difference between the two drills is small, but the worker with the electric drill will drill holes faster.

Labor force. Many factors in this category affect productivity. If a good working environment, with proper training and incentives, is provided, the employees will be more productive. The type of supervision and relationship with management can also affect productivity. Employees must also have the proper tools and supplies to perform their task.

Capacity. The capacity of the company will affect the productivity. If a company buys expensive equipment and builds a large factory but uses only a small part of it, productivity will be low. To raise productivity, the company must make more product with the same capital investment. The facility and equipment will be idle a smaller portion of the time, and productivity will increase. *External influences.* There are factors that a company cannot control that affect productivity. Consumer demand is one factor. Generally, a company will not produce more units than consumers will buy. If demand declines, the company will reduce the output. In this case the company will either have idle capacity or will switch capacity to different products. If idle capacity increases, productivity will decline.

Government regulation can affect productivity. In many countries, regulations require companies to provide safety equipment and safe conditions for employees. They also require companies to dispose of hazardous chemicals in environmentally safe ways. Some of these regulations increase costs significantly. Without these regulations, some companies would be irresponsible to both their employees and the environment. In the short term, total cost would be reduced and productivity would go up. However, in the long term, negative effects could be experienced by many.

Quality. Increasing quality can both help and hurt productivity. Quality awareness has caused many companies to adopt more productive processes and to design products that can be produced more efficiently. In turn, the cost has been lowered and customer demand raised. All of these factors help to raise productivity. However, sometimes more expensive processes or materials must be used to increase quality. This can increase total cost and reduce total productivity.

Productivity improvement plans. There are many different plans that companies develop in an attempt to improve productivity. Wage incentive plans and changes in management structure are two ways that companies focus on the labor force. Investment in research and development allows companies to develop new products and processes that are more productive. Quality improvement programs can reduce waste and provide more competitive products at a lower cost. *See* METHODS ENGINEERING; OPERATIONS RESEARCH; PRODUCTION PLANNING. Gregory L. Tonkay

Bibliography. D. I. Cleland and B. Bidanda (eds.), The Automated Factory Handbook: Technology and Management, 1990; W. K. Hodson (ed.), Maynard's Industrial Engineering Handbook, 1992; M. E. Mundel, Measures of productivity, Ind. Eng., 8(5):32-36, 1976; R. G. Schroeder, Operations Management: Decision Making in the Operations Function, 4th ed., 1993; D. S. Sink, Productivity Measurement Improvement, Evaluation, and Control, 1985.

Progesterone

A steroid hormone produced in the corpus luteum and placenta. The hormone has an important physiological role in the luteal phase of the menstrual cycle and in the maintenance of pregnancy. In addition, progesterone produced in the testis and adrenals has a key role as an intermediate in the biosynthesis of

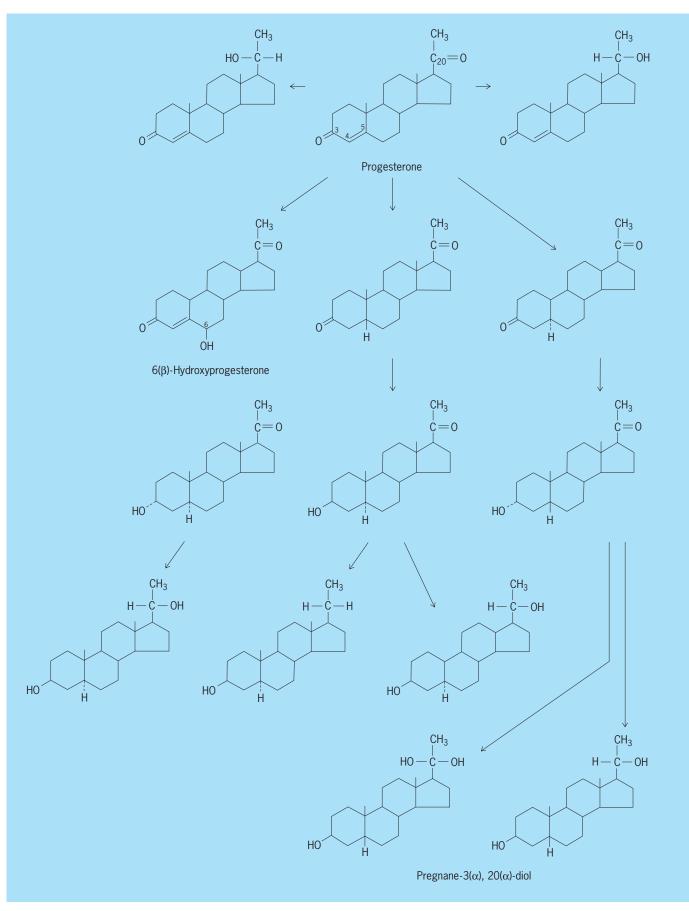


Fig. 1. Catabolic pathways of progesterone.

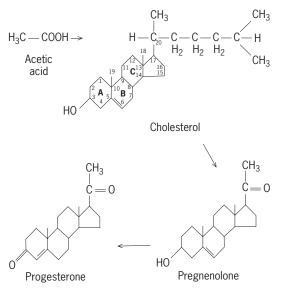


Fig. 2. Biosynthetic pathway for progesterone.

androgens, estrogens, and the corticoids (adrenal cortex steriod).

Catabolism of progesterone involves reductive reactions at carbon atoms 3, 4, 5, and 20, as well as hydroxylation carbon atom 6 (**Fig. 1**). The biosynthetic pathway to progesterone (**Fig. 2**) is common to all steroid-producing tissues and involves acetate \rightarrow cholesterol \rightarrow pregnenolone \rightarrow progesterone.

The blood contains, in addition to progesterone, the $20(\alpha)$ - and $20(\beta)$ -reduced metabolities, which have reduced biological activity.

Pregnane- $3(\alpha)$,20(α)-diol is, quantitatively, the most important metabolite in humans, and its determination in the urine is of clinical importance in conditions involving menstrual irregularities and also in abnormal pregnancies. *See* ANDROGENS; CHOLES-TEROL; ESTROGEN; PREGNANCY; STEROID; STEROL. Ralph I. Dorfman

Bibliography. C. W. Bardin et al. (eds.), *Progesterone* and *Progestins*, 1982; E. Gurpide (ed.), *Biochemical Actions of Progesterone and Progestins*, 1977.

Programmable controllers

Electronic computers, designed for durability in industrial environments, that are used for the control of machines and manufacturing processes through the implementation of specific functions such as logic, sequencing, timing, counting, and arithmetic. They are also known as programmable logic controllers (PLCs). Historically, process control of a single or a few related devices has been implemented through the use of banks of relays and relay logic for both the control of actuators and their sequencing. The advent of small, inexpensive microprocessors and single-chip computers, or microcontroller units, brought process control from the age of simple relay control to one of electronic digital control while neither losing traditional design methods such as relay ladder diagrams nor restricting their programming to that single paradigm. The computational power of programmable controllers and their integration into networks has led to their having capabilities approaching those of distributed control systems, and plantwide control is now a mixture of distributed control systems and programmable controllers. Applications for programmable controllers range from small-scale, local process applications in which as few as 10 simple feedback control loops are implemented, up to large-scale, remote supervisory process applications in which 50 or more process control loops spread across the facility are implemented. Typical applications include batch process control and materials handling in the chemical industry, machining and test-stand control and data acquisition in the manufacturing industry, wood cutting and chip handling in the lumber industry, filling and packaging in food industries, and furnace and rolling-mill controls in the metal industry. See DIGITAL COMPUTER; DISTRIBUTED SYSTEMS (CONTROL SYSTEMS); MICRO-PROCESSOR.

Although programmable controllers have been available since the mid-1970s, developments—such as the ready availability of local area networks (LANs) in the industrial environment, standardized hardware interfaces for manufacturer interchangability, and computer software to allow specification of the control process in both traditional (ladder logic) and more modern notations such as that of finite-state machines—have made them even more desirable for industrial process control. *See* LOCAL-AREA NET-WORKS.

Implementation. Programmable logic controllers are typically implemented by using commonly available microprocessors combined with standard and custom interface boards which provide level conversion, isolation, and signal conditioning and amplification. Microprocessors used in programmable controllers are similar or the same as those used in personal computers. A significant difference between personal computers and programmable controllers is the type of software operating system. The software of a programmable controller must respond to interrupts and be a real-time operating system, and these are characteristics which the typical operating system of a personal computer does not possess. See MICROCOMPUTER; OPERATING SYSTEM; REAL-TIME SYSTEMS; SOFTWARE.

Connectivity. Perhaps the biggest benefit of programmable controllers is their small size, which allows computational power to be placed immediately adjacent to the machinery to be controlled, as well as their durability, which allows them to operate in harsh environments. This proximity of programmable controllers to the equipment that they control allows them to effect the sensing of the process and control of the machinery through a reduced number of wires, which reduces installation and maintenance costs. The proximity of programmable controllers to processes also improves the quality of the sensor data since it reduces line lengths, which can introduce noise and affect sensor

calibration. Communication devices in programmable controllers, such as synchronous and asynchronous serial communications ports, allow these distributed programmable controllers to be connected through a local-area network to an operator interface for monitoring and system control, or connected to a workstation for program development and downloading of operational software into the programmable controllers. The versatility of a local-area network also allows for remote location of a central supervisory-monitor computer, as well as for field personnel to connect locally to the local-area network for local control while still monitoring signals that are passing through the local-area network-based system. The extensive computational capability of programmable controllers also allows for local monitoring of such conditions as low battery voltages, lack of sufficient cooling, or other failures of its own control capability. Depending on the extent of the failure, the programmable controller may trigger a local alarm or pass the information back to a central supervisory computer. This monitoring can also be extended to the monitoring of the process under control itself, with local analysis significantly reducing the amount of data that must be transferred over a network for centralized control and monitoring.

The availability of a local-area network also allows for the remote programming of local microcontroller units on the fly. That is, one control algorithm running in microcontroller unit can be replaced with another control algorithm from the remote site without local operator intervention. Hardware options can also be reconfigured and their status maintained in nonvolatile read-write memory, which is readily available in some programmable controllers, and as electrically erasable programmable read-only memory (EEPROM). Troubleshooting can be controlled from a remote site by downloading diagnostic software into the local programmable controller. In the most general sense, every programmable controller can be identical to every other one, with the remote supervisory computer downloading to each local programmable controller the program that is appropriate to the device or process that is being controlled. See COMPUTER STORAGE TECHNOLOGY; SEMICONDUCTOR MEMORIES.

Open architecture. Open architecture allows standardization of local-area-network interfaces, host software, operating systems, field buses, and programmable controller backplanes. This open architecture allows process-control engineers to integrate the programmable controllers from various manufacturers that are best suited to individual tasks into a comprehensive customized control system. This approach also allows for an easy upgrade path in which one manufacturer may have made a complete installation and, at a later date, a more advanced system from another manufacturer which has become available can be integrated with components of the first installation.

Beyond logic control. Because of the increased computational power associated with program-

mable controllers, control algorithms more complex than simple ladder logic can be implemented, and libraries of standard control algorithms are made readily available by manufacturers. This availability allows the process-control engineer to spend more time on the development of the control strategy and the associated parameters than on the debugging and verification of in-house-written software. This increased level of sophistication and partitioning of the design problem has led to software packages for cross-platform development (that is, development that can be carried out with various operating systems) of operating software for programmable controllers. In addition to libraries of control algorithms, signal-processing routines are available for the conditioning and compensation of sensor data from simple smoothing to spectral component analysis.

Software development. Cross-platform development of programmable controller software allows the use of more powerful computers such as personal computers or workstations in the designing, debugging, development, and downloading of this software. Additionally, the quality and user-oriented interface of the development software enables the user to develop the control program either by using traditional ladder-logic diagrams, which are shown graphically on the development computer's display, or through other methods such as state diagrams, sequential function charts, or function blocks. This improved userinterface capability makes programmable controllers more accessible to a broader range of process-control personnel who have different backgrounds and different working languages. Software for individual components of programmable controllers, such as a proportional-integral-derivative (PID) control algorithm for a robotic arm, are not programmed by the process-control engineer, but are indirectly programmed by the engineer inserting the relevant parameters into blanks in on-screen templates (Fig. 1). These parameters are then automatically transformed into fully functional software modules, which are interfaced to other components of the

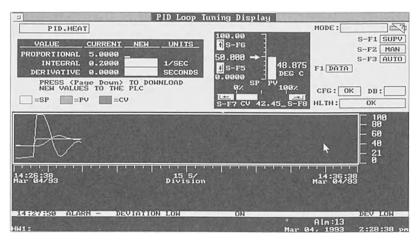


Fig. 1. Windows-based operator interface for the specification of parameters to be used for the tuning of a proportional-integral-derivative (PID) controller. (*After Allen-Bradley, Take Control of Your Application with Process Configuration and Operation (PCO)* Software, Publ. 6190-2.1, 1993)

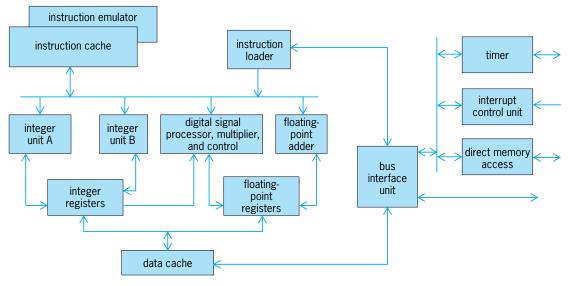


Fig. 2. Functional architecture of a typical high-performance single-chip microcontroller used for control and signal processing. (*National Semiconductor Corp.*)

control process. Another capability of cross-platform development software is the ability to simulate the control process before it is programmed into the programmable controller and connected to operating machinery.

Microcontrollers. Because of their self-contained input/output capability and the incorporation of customized on-chip coprocessors, single-chip microcontrollers are used extensively for the embedded control of processes in stand-alone equipment that is not associated with industrial control. These microcontroller units must be considered programmable controllers from a functional point of view, even though they are not used directly in an industrial setting. Microcontroller units are single-chip versions of general-purpose computers which have all five components that define a computer resident in a single integrated circuit. That is, a single integrated circuit contains the arithmetic logic unit (ALU), the control section, a local memory which may or may not be expandable off-chip, an input section, and an output section. An unusual feature of microcontroller units is their on-chip input/output section and signal coprocessors, which make them particularly suitable for small embedded control systems. For example, 32-bit-word-length microcontrollers are readily available which contain on-chip digital signal processing capability of sufficient power to implement a standalone facsimile send/receive system while simultaneously controlling the associated printer motors and user-interface requirements. These embedded processors are used to control devices from laser printers to home security systems, as well as automobile engine ignition timing and fuel injection for the reduction of pollution (Fig. 2). See EMBEDDED SYSTEMS; INTEGRATED CIRCUITS. Kenneth J. Hintz

Bibliography. Allen-Bradley, *Take Control of your Application with Process Configuration and Operation (PCO) Software*, Publ. 6190-2.1, 1993; G. L. Batten, *Programmable Controllers: Hardware, Software, and Applications*, 2d ed., 1994; K. J. Hintz and D. Tabak, *Microcontrollers: Architecture, Implemen-* *tation, and Programming*, 1992; C. T. Jones and L. A. Bryan, *Programmable Controllers*, 1983.

Programming languages

The different notations used to communicate algorithms to a computer. A computer executes a sequence of instructions (a program) in order to perform some task. In spite of much written about computers being electronic brains or having artificial intelligence, it is still necessary for humans to convey this sequence of instructions to the computer before the computer can perform the task. The set of instructions and the order in which they have to be performed is known as an algorithm. The result of expressing the algorithm in a programming language is called a program. The process of writing the algorithm using a programming language is called programming, and the person doing this is the programmer. *See* ALGORITHM.

In order for a computer to execute the instructions indicated by a program, the program needs to be stored in the primary memory of the computer. Each instruction of the program may occupy one or more memory locations. Instructions are stored as a sequence of binary numbers (sequences of zeros and ones), where each number may indicate the instruction to be executed (the operator) or the pieces of data (operands) on which the instruction is carried out. Instructions that the computer can understand directly are said to be written in machine language. Programmers who design computer algorithms have difficulty in expressing the individual instructions of the algorithm as a sequence of binary numbers. To alleviate this problem, people who develop algorithms may choose a programming language. Since the language used by the programmer and the language understood by the computer are different, another computer program called a compiler translates the program written in a programming language into an equivalent sequence of instructions that the

computer is able to understand and carry out. *See* COMPUTER STORAGE TECHNOLOGY.

Computers were first developed in the 1940s and have been used in industry since the early 1950s. As more is learned about the programming process and about what makes a good algorithm, the notation used by the programmer, the programming language, has evolved. The requirements of a good program today are vastly different from those in the 1950s.

Evolution of Programming Languages

Programming languages have progressed from machine languages, to assembly languages, to high-level languages.

Machine language. For the first machines in the 1940s, programmers had no choice but to write in the sequences of digits that the computer executed. For example, assume that a computer has 100,000 memory locations (numbered from 00,000 to 99,999) and has 10 registers (numbered from 0 to 9) for performing arithmetic calculations. The instructions for this machine are in an 8-digit format *WWXYYYYY*, where *WW* stands for an arithmetic operation, *X* signifies a register (called an operand register) to contain one argument for an operation, and *YYYYY* signifies the memory location. List (1)

10	Add contents of location YYYYY to X
11	Subtract contents of location YYYYY
	from X

- 12 Negate contents of register X. Ignore YYYYY
- 21 Move contents of location YYYYY to X 22 Move contents of location X to YYYYY
- 22 Move contents of location X to YYYYY
- 31 Next instruction to execute is at YYYYY if X is positive

(1)

(2)

defines a few simple instructions. Assume we want to compute the absolute value of A + B - C, where *A* is the value at machine address 3012, *B* is the value at address 3013, and *C* is the value at address 3014, and then store this value at address 3015. Sequence (2) of instructions, beginning at location 02345, would be:

Address	Instruction	Interpretation
02345	21103012	Get contents of
		03012 (A) and put
		into register 1
02346	10103013	Add contents of 03013
		(B) to register 1
02347	11103014	Subtract contents of
		03014 (C) from
		register 1
02348	31102350	Execute at address
		02350 if register 1
		is positive;
		Otherwise go
		to next instruction
02349	12100000	Make negative contents
		of register 1 positive
02350	22103015	Move $ A + B - C $
		to location 03015

It should be clear that programming in this manner is difficult and fraught with errors. Explicit memory locations must be written, and it is not always obvious if simple errors are present. For example, at location 02347, writing 101... instead of 111... would compute |A + B + C| rather than what was desired. This is not easy to detect.

Assembly language. Since each component of a program stands for an object that the programmer understands, using its name rather than numbers should make it easier to program. By naming all locations with easy-to-remember names, and by using symbolic names for machine instructions, some of the difficulties of machine programming can be eliminated. A relatively simple program called an assembler converts this symbolic notation into an equivalent machine language program. Example (2) could be programmed in assembly language as sequence (3):

Label	Operation	Operand	Comments
	Load	R1, A	Place operand A into register 1
	Add	R1, B	Compute A + B by adding B to
	Subtract	R1, C	register 1 Compute $A + B - C$ by subtracting C
	Ifplus	R1, SKIP	from register 1 Goto location SKIP if register 1 is
	Negate	R1, 0	positive Compute $ A + B - C $ by negating register 1
SKIP	Store	R1, D	Save answer in D by moving register 1 to location D
Α	Define	03012	Define A as location 03012
В	Define	03013	Define B as location 03013
С	Define	03014	Define C as location 03014
D	Define	03015	Define D as location 03015
			(3)

The symbolic nature of assembly language greatly eased the programmer's burden, but programs were still very hard to write. Mistakes were still common. Programmers were forced to think in terms of the computer's architecture rather than in the domain of the problem being solved.

High-level language. The first programming languages were developed in the late 1950s. The concept was that if we want to compute |A + B - C|, and store the result in a memory location called *D*, all we had to do was write D = |A + B - C| and let a computer program, the compiler, convert that into the sequences of numbers that the computer could execute. FORTRAN (an acronym for Formula Translation) was the first major language in this period; in

it, the above example would be written as (4)

$$D = abs (A + B - C)$$
(4)

where *abs* represents a function that computes the absolute value of its argument. If a language does not have the function *abs* to invoke, we could have written the program as (5)

$$D = A + B - C;$$

IF D < 0 THEN
$$D = -D;$$

WRITE (D);
(5)

which consists of three statements: an assignment statement, an IF-THEN statement to take care of the case where the initial calculation was negative and the absolute value requires a positive result, and a WRITE statement to write out the answer.

Although FORTRAN statements were patterned after mathematical notation, the last statement in (5), D = -D, shows the basic difference between mathematics and a program. In mathematics the = symbol implies that both sides of the equation have the same value. However, in FORTRAN and some other languages, the equal sign is known as the assignment operator. The action carried out by the computer when it encounters this operator is, "Make the variable named on the left of the equal sign have the same value as the expression on the right." Because of this, in some early languages the statement would have been written as $-D \rightarrow D$ to imply movement or change, but the use of \rightarrow as an assignment operator has all but disappeared.

The compiler for FORTRAN converts that arithmetic statement into an equivalent machine language sequence, such as in (2). In this case, we did not care what addresses the compiler used for the instructions or data, as long as we could associate the names A, B, C, and D with the data values we were interested in.

At first there was some resistance to these higherlevel languages (called higher-level because they were deemed more complex or higher than the assembly languages of that day). Computers were extremely expensive, and it was important to keep them busy doing useful work. Early compilers often generated inefficient sequences of instructions in compiling a program, and it was believed that even though programmers made mistakes in assembly code, it was still cheaper and more efficient to fix those errors in order to produce programs that were more efficient on these expensive machines.

We take a different view today of high-level languages. Computers have become orders-ofmagnitude cheaper and faster; even if a program is not as efficient as assembly code, the extra running time is not important. Moreover, compiler technology has become so good that compiled programs often run as fast or faster than hand-written assembly code. Finally, we have learned, to our dismay, that programmers still make errors. Significant research has gone into developing programming languages that minimize the occurrence of those errors. We have learned that it is more important to get the algorithm right rather than to try to execute every machine cycle on a modern computer. Most new language design features today are concerned with their effects on the programming process rather than on their ultimate efficiency on the underlying machine hardware.

Early high-level languages. The period from 1957 through 1962 was the most significant in the development of programming languages. Almost all modern concepts first appeared in the four important languages developed in that period. Those languages and the major features they introduced are as follows.

FORTRAN. FORTRAN was developed by IBM under the leadership of John Backus in 1957. It became the model for most current languages, including BASIC, C, C++, and Java. A program executes a sequence of statements in which each statement looks like a mathematical expression. Example (4), where D =abs(A + B - C), enables mathematicians to easily translate mathematical relationships into programming language algorithms.

In FORTRAN, the basic program unit is the subroutine or function. Each subroutine is a self-contained unit that computes some value. The goal is to design and develop each subroutine independently such that each can be used repeatedly at multiple places in a program. For example, if a function to compute the standard deviation of a list of numbers was needed, a single *StandDev* function could be written and used wherever the standard deviation was required.

ALGOL-60. The name of this language, developed in 1960 by an international committee led by Peter Naur and John Backus, is an acronym for ALGOrithmic Language, 1960. This international effort extended the ideas present in FORTRAN. A notation called BNF (Backus-Naur Form), which is a grammatical form called a context-free grammar, defined the language. Because of advances in compiler technology, defining languages using BNF grammars allows for efficient compilers to be developed almost automatically. ALGOL-60 further developed the concept of the subroutine (called a procedure in ALGOL-60) to permit the easy communication of data from the calling procedure to the called procedure.

ALGOL-60 was the forerunner of programming language control structures. It also included a primitive form of inheritance, although it was not until the early 1970s that the work of David Parnas on information hiding and Alan Kay on Smalltalk led to the development model we now call object-oriented design.

COBOL. The U.S. Defense Department under the leadership of Grace Hopper organized the development of COBOL in 1960. The name is an acronym for Common Business Oriented Language. This language has been a mainstay of the business community since its conception. COBOL introduced the concept of record to provide access to data for business applications. During the 1970s, Niklaus Wirth further refined this concept with the type declaration and the record type in the Pascal programming language.

LISP. About the same time that FORTRAN was being designed, John McCarthy at the Massachusetts Institute of Technology was investigating software for problem solving. He developed LISP (for LISt Processing). This form of programming, called applicative programming, is the basis for almost all artificial intelligence and expert system developments since then. *See* ARTIFICIAL INTELLIGENCE; EXPERT SYSTEMS.

Later high-level languages. Figure 1 presents an overview of the development of the major programming languages. Well over 500 additional languages have been identified, but the languages in Fig. 1 generally have made the most lasting contribution to language design. *See* COMPUTER PROGRAMMING; DIG-ITAL COMPUTER.

Structure of Programming Languages

Programs written in a programming language contain three basic components: (1) a mechanism for declaring data objects to contain the information used by the program; (2) data operations that provide for transforming one data object into another; (3) an execution sequence that determines how execution proceeds from start to finish.

Data declarations. Data objects can be constants or variables. A constant always has a specific value. Thus the constant 42 always has the integer value of forty-two and can never have another value. On the other hand, we can define variables with symbolic names. The declaration of variable *A* as an integer informs the compiler that *A* should be given a memory location much like the way the variable *A* in example (2) was given the machine address 03012. The program is given the option of changing the value stored at this memory location as the program executes.

Each data object is defined to be of a specific type. The type of a data object is the set of values the object may have. Types can generally be scalar or aggregate. An object declared to be a scalar object is not divisible into smaller components, and generally it represents the basic data types executable on the physical computer. In a data declaration, each data object is given a name and a type. The compiler will choose what machine location to assign for the declared name. For example, in C, declaration (6) does

the following: (1) *A* and *Myvalue* are both declared as scalar variables. (2) *A* and *Myvalue* are both declared with an integer data type. (3) The compiler allocates separate machine memory locations for each variable.

Scalar types. Programming languages usually include the following scalar types:

Integer. As in mathematics, integers represent the whole numbers, 0, 1, 2, 3, and so forth, and the negative numbers, -1, -2, -3, and so forth. However, since words in a computer memory are of a fixed size [for example, the 8-digit machine language from (1)], integers in a programming language will have a minimum and a maximum value.

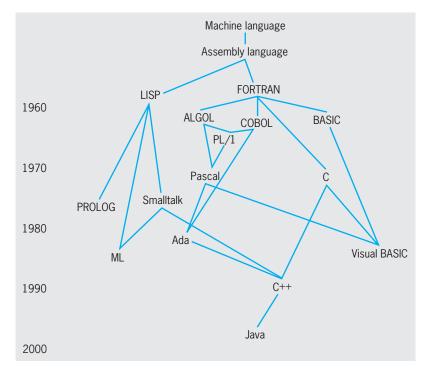


Fig. 1. Evolution of languages.

Float. Float or real is the computer representation of rational numbers written in scientific notation. Thus $1.2E^2$ represents 1.2×10^2 or 120. As with integers, because of the fixed size of a machine's memory, not all rational numbers can be represented. Typically a real data object can have from 8 to 15 significant digits, and can represent exponents with values extending from approximately 10^{-300} to 10^{+300} . It must be emphasized that even though these data objects are called real variables, they are mathematically rational numbers. The expression (1.0/3.0) * 3.0 is not equal to 1 because 1.0/3.0 is equal to .333333... for as many places as the machine is capable of storing. Multiplying this value by 3.0 gives a result of only .99999... for approximately 8 to 15 digits and is not equal to 1.0.

Boolean. This scalar type represents the boolean values *true* and *false*. Operations on boolean data types follow the rules of boolean algebra. *See* BOOLEAN ALGEBRA.

Character. Each individual memory location in most modern computers represents a single character. Most computers today use the 8-bit byte as a memory location; thus there are 2⁸ or 256 possible byte values. Character codes such as ASCII (American Standard Code for Information Interchange) or IBM's EBCDIC (Extended Binary Coded Decimal Interchange Code) define the externally printable character for each internal bit representation. For example, the ASCII bit pattern 01000001 (the number 65 in decimal) represents the character A; 01000010 (or the number 66) represents the character B. Defining data objects as characters allows for manipulating data that represent written text.

Enumerated. Languages today also include a programmer-defined scalar type called an

enumerated type. For example, to define the seasons of the year, declaration (7)

Enum {*spring*, *summer*, *fall*, *winter*}
$$A$$
; (7)

may be used in C. Data object *A* can take on one of the values *spring*, *summer*, *fall*, or *winter* and no other value. This allows the programmer to define types relevant to the problem being solved, and allows the compiler to find some common errors, such as if the programmer then tried to write (8).

$$A = 7; \tag{8}$$

The value 7 would not be of the appropriate type.

References. A reference is a scalar data object that is the address of another object. These are also called pointer variables. If *A* is an integer data object and *Ptr1* and *Ptr2* are both reference variables whose values are the address of *A*, then the value of *A* may be changed by referencing either *Ptr1* or *Ptr2* in conjunction with a dereferencing operator as done in the C-like family of languages.

Aggregate types. Aggregate or composite data types are data types whose objects can be divided into smaller components, usually scalar objects or other aggregate types. The following are the usual aggregate types:

Arrays. An array is an ordered sequence of data objects. The array has a lower and an upper bound; thus we can refer to the first member of the array, the second member, and so forth. Each array member is itself a data object of some type. Example (9)

$$Int A[10]; \tag{9}$$

would define an array of 10 elements (from A[0] through A[9]) in C, with each array element of the array being an integer. We can also defined multiple dimensions. A two-dimensional array called *B* with 10 rows and 10 columns can be defined as B[10,20]. This array consists of 10×20 different elements. Each element is uniquely identified by the value of its row and column. The first element of row 1 is B[1,1], the second element of row 1 is B[1,2].... The last element of the last two elements of the last row of the array are B[10,19] and B[10,20].

An array element can usually be of any valid type. Therefore we can define an array of integers, an array of characters, or an array of an array of reals.

Strings. A string is an ordered sequence of characters. In many programming languages a string is simply an array of characters. For example, in Pascal a string of length 20 could be defined as example (10).

$$Type \ string20 = array[1..20] \ of \ char; \qquad (10)$$

Records. A record is a named collection of other data objects. Whereas in an array each of the objects is of the same type, in a record each object may be of a different type and is individually named. In Pascal a

record that defines an address could be example (11).

Тур	e address = record	
	Name: string20;	
	Street: string20;	
	City: string20;	(11)
	State: string20;	
	PostalCode: integer	
	end;	

Var MyAddress: address; (12)

Pascal declaration (12) declares *MyAddress* to be of type address, and allows the program to access the data objects *MyAddress.Name*, *MyAddress. Street*, *MyAddress.City*, *MyAddress.State*, and *MyAddress.PostalCode*.

Data operations. Data operations provide for setting the values into the locations allocated for each declared data variable. In general this is accomplished by a three-step process: a set of operators is defined for transforming the value of each data object, an expression is written for performing several such operations, and an assignment is made to change the value of some data object.

Operations. For each data type, languages define a set of operations on objects of that type. For the arithmetic types, there are the usual operations of addition, subtraction, multiplication, and division. Other operations may include exponentiation (raising to a power), as well as various simple functions such as modula or remainder (when dividing one integer by another). There may be other binary operations involving the internal format of the data, such as binary and, or, exclusive or, and not functions. Usually there are relational operations (for example, equal, not equal, greater than, less than) whose result is a boolean value of true or false. There is no limit to the number of operations allowed, except that the programming language designer has to decide between the simplicity and smallness of the language definition versus the ease of using the language.

Other data types include other relevant operations. Boolean operations would include the *and*, *or*, *exclusive or*, and *not* operations. For strings we usually have only the relational operations for comparing character data.

Expressions. The expression is the primary mechanism for transforming one data value into another. Most languages use an *infix* notation, similar to common mathematics, where the operation symbol is between the operands; however, there are three common notations for presenting expressions. (A * B) - (C * D) is expressed in examples (13), (14), and (15).

<i>Expression</i> <i>notation</i> Infix	A Example A * B – C * D	<i>Meaning</i> Operator between operands	(13)
Prefix	- * A B * C D	Operator precedes operands	(14)

Postfix
$$A B * C D * -$$
 Operator
follows (15)
operands

Infix notation depends upon an implied ordering of operations. Thus (13) is assumed to mean $(A * B) - (C^* D)$ rather than an interpretation of A * (B - C) * D. This is based the concept of precedence. Precedence defines the order of evaluation. We normally give multiplication and division highest precedence, meaning do these operations first, and then do addition and subtraction. With additional operators, such as character operations, logical operations such as *and* and *or*, binary operations on bits, and so forth, the C language defines 17 levels of precedence.

Since A * B - C * D means (A * B) - (C * D), the expression A * (B - C) * D requires the use of parentheses. The Polish mathematician Jan Lukasiewicz discovered in the 1920s that one could unambiguously write expressions without using parentheses, as is required with the usual infix notation. This notation, originally called Polish notation, now goes by the name of prefix notation. By placing the operator immediately before its operands, the order of evaluation of the operations becomes unambiguous. Rather than the expression in (13), the prefix for the expression A * (B - C) * D is given by * * A - BCD. Both expressions have different and unambiguous prefix formats.

Polish notation remained a mathematical curiosity until the 1960s, when it was employed in the ALGOL compiler. By converting all expressions into postfix notation, a variation on Lukasiewicz's prefix notation with the operator following its operands, extremely fast compilers could be built. The evaluation of an expression in postfix uses algorithm (16).

If the next object is an operand, then save it.

If the next object is an operator, (16) then apply operator to saved objects and save result

For the expression in (15), (A B * C D * -), its evaluation proceeds as in sequence (17).

Input	Action	Saved values
Α	Save A	Α
В	Save B	A, B
*	Apply * to A and B and save (A * B)	(A*B)
С	Save C	(A * B), C
D	Save D	(A * B), C, D
*	Apply $*$ to last 2 saved values, C and D, and save (C - D)	(A * B), (C – D)
—	Apply – to last 2 saved values, and save result	$(\mathbf{A} * \mathbf{B}) - (\mathbf{C} * \mathbf{D})$
		(17)

The use of postfix notation radically changed the design of programming language compilers. It allowed compilers to generate machine language that was comparable or better than the best humandesigned machine language, and it opened the door for future languages to be more concerned with improving the program design process than with just generating efficient machine language instructions.

Assignment statement. The assignment statement is usually the primary mechanism for an expression to have an impact on the data variables in a program. It is usually given by the syntax *variable* = *expression*, and means to evaluate *expression* and assign its value to the storage location defined by the symbolic name *variable*.

Execution sequence. The purpose of a program is to manipulate some data in order to produce an answer. While the data operations provide for this manipulation, there must be a mechanism for deciding which expressions to execute in order to generate the desired answer. That is, an algorithm must trace a path through a series of expressions in order to arrive at an answer. Programming languages have developed three forms of execution sequencing: (1) control structures for determining execution sequencing within a procedure; (2) interprocedural communication between procedures; and (3) inheritance, or the automatic passing of information between two procedures.

Control structures. Corrado Böhm and Giuseppi Jacopini showed in 1966 that a programming language needs only three basic statements: an assignment statement (discussed above), an IF statement, and a looping construct. Anything else can simplify programming a solution, but is not necessary. If we add an input and an output statement, we have all that we need for a programming language. Languages execute statements sequentially with the following variations to this rule.

IF statement. Most languages include both common variations, (18) and (19),

IF expression
$$THEN Statement_1;$$
 (18)

IF expression THEN Statement₁ (19) ELSE Statement₂;

of the IF statement. In the IF-THEN statement (18), the expression is evaluated, and if the value is *true*, then Statement₁ is executed next. If the value is *false*, then the statement after the IF statement is the next one to execute. The IF-THEN-ELSE statement (19) is similar, except that specific true and false options are given to execute next. After executing either the THEN or ELSE part, the statement following the IF statement is the next one to execute.

Statement₁ and Statement₂ may actually be several statements. Most languages include a bracketing construct to group sequences of statements. Thus, many languages use BEGIN Statement₁; Statement₂;... END to indicate a single grouping of statements.

Looping constructs. The usual looping constructs are the WHILE statement and the REPEAT statement. While only one is necessary, languages usually have both. The WHILE statement is given by syntax (20).

This has the interpretation of evaluating *expression*. If *true*, execute *Statement*, and then reexecute the WHILE. In order for this process to terminate, something in *Statement* has to eventually cause the expression to become false when it is next evaluated. The REPEAT statement is given by syntax (21).

It is similar to the WHILE statement except that *Statement* is executed first and then *expression* is next evaluated. The process repeats if the expression is false, and exits to the following statement if true. Although the continuation process between the WHILE and REPEAT statements is different (the WHILE continues if the expression is true and the RE-PEAT continues if the expression is false), the major difference between the two statements is that the WHILE statement may not execute *Statement* at all (that is, if the initial value for expression was *false*), whereas the REPEAT statement will execute *Statement* at least once.

Other control structures. Most languages include other control structures to ease the programming process. Among the most common are the following.

Controlled repetition. The WHILE and REPEAT statements both iterate for an undetermined number of times. If the number of times is already known, a FOR statement can be used. The FOR statement specifies the explicit number of times a statement is executed, such as FOR I := 1 to 20. The single statement following this construct will be repeated 20 times. The control variable I is incremented after each execution of the statement until it reaches the limit value of 20.

<u>Case statement.</u> The CASE or SWITCH statement is like a multiway IF statement. Control passes to one of many possible statements, depending upon the value of the initial expression.

<u>Go to statement.</u> The GOTO had the simple syntax of GOTO *label*, where *label* is the symbolic name attached to a statement. It provides for a specific branch point to be specified after testing for an expression value. Example (5) using GOTO statements could be written example (22).

$$D = A + B - C;$$

IF $D < 0$ THEN GOTO NEXTSTMT;
 $D = -D;$
NEXTSTMT: WRITE (D);

(22)

This construct has been found to be very error-prone and totally unnecessary to the writing of good programs. Until the work of Böhm and Jacopini, it was assumed that the GOTO statement was a requirement for programming languages. They showed that a GOTO was unnecessary, since any program function could be written using only the IF and the WHILE constructs. Various software products are sold with the stated purpose of "structuring bad programs into good," with the Böhm and Jacopini theorem as the justification. But this is using their result backwards. A poorly designed, unstructured program will just become a poorly designed "structured" one. What the Böhm and Jacopini theorem does say is that you will be able to solve your problem using only IF and WHILE. It is an existence proof that a solution does exist. It is up to the programmer to find a good solution.

Interprocedural communication. The various control structures provide a mechanism for executing within a given procedure. However, the essence of good program design is to build a large system consisting of multiple procedures and functions, design each one independently, and then have them communicate necessary information among themselves. The procedure call is the primary communication mechanism, but there are others as well.

<u>Procedure call</u>. The procedure call is typically given by the syntax *FunctionName(arguments)*, such as the example abs(A + B - C) in (23).

In this case, the expression A + B - C, called an argument, is evaluated, and that value is passed to the function *abs*. The function *abs* uses that value to compute an answer, which is passed back to the procedure that invoked the procedure call.

In designing the called procedure, the passed argument is referred to as an actual parameter and given some symbolic name, as in example (23). Within the

Calling			
procedure	Argument		
abs(A + B - C)	A + B - C	\leftrightarrow	Called
	Parameter		procedure
	Х		abs(X)
			(23)

procedure, X behaves as if it is a new data declaration.

The relationship between the formal parameter [X in (23)] and its actual argument [A + B - C in (23)] can differ. Several mechanisms for communication between the argument and the parameter have been developed. The following four are the most common:

1. In call by reference, the parameter is actually the address of the argument. If one writes abs(A) and the formal parameter in abs is named X, then changes to the value of X are actually reflected as references to A. X is an alias to the variable A, and both names refer to the same actual data object. This method is the most common and efficient form of argument transmission in programming languages.

2. In call by value, the parameter is actually the value of the argument. If one writes abs(A) and the formal parameter in abs is named X, then upon entry to the procedure, X is set to the value of A. Changes to the value of X do not affect the value of A. The programming language creates a new data object for the parameter. This method is the most efficient since all references to the parameter are data variables local to the called procedure. It is limited, however, in

that the purpose of a called procedure is to compute several values and return them to the calling procedure. Call by value does not permit those values to be returned.

3. Call by result is similar to call by value except that upon returning to the calling procedure, the final value of the parameter is passed back as the value of the argument. This provides the flexibility of returning values to arguments, as is the case with call by reference, with some of the efficiency of the call-by-value implementation. It is a useful model, but one that is not implemented often.

4. Call by name was first described in the report defining ALGOL-60. It has the simplest theoretical definition. In example (23), the function *abs* is simply a replacement for the definition of the actual absolute value function. In practical terms, it means that if *X* is a call-by-name parameter for the argument A + B - C, then each time *X* is referenced in *abs*, the actual argument A + B - C is reevaluated. This can be extremely inefficient, since in most programs the reevaluation of A + B - C will not change each time.

More seriously, call by name can lead to anomalous situations. Consider example (24)

Calling procedure	Called procedure	
A = 1;	PROCEDURE F(NAME X,	
	REFERENCE Y);	
B = 2;	INTEGER Z;	(24)
CALL F $(A + B, B)$;	Z = X;	
	Y = 1;	
	IF $Z = X$ THEN WRITE	

where parameter *X* is a call-by-name parameter and parameter *Y* is a call-by-reference parameter. Procedure *F* is passed the arguments A + B and *B*. In *F* the assignment to *Z* evaluates to 3 (1 + 2). The following assignment to *Y* changes the value of *B* to 1 since it is a call-by-reference parameter. The IF statement evaluates to false, since *X* now is 2 (1 + 1) and *Z* is the old value of *X* or 3. Because of these issues, call by name is rarely used as a parameter mechanism today.

The above four classical parameter mechanisms are defined by their implementation. In each case, the definition describes how the variables will be implemented. The Ada language uses a difference approach, involving *in-out* parameters. It describes the behavior of the parameter. A value, which is passed into a procedure and used within the procedure, will be an *in* parameter. Upon exit from the procedure, the value is not passed back. This has the same behavior as the call-by-value parameter and is often implemented that way. However, the language definition does not specify the implementation, only the behavior.

A parameter that has a value passed back to the calling argument is an *out* parameter, and one that is passed in and then the final value passed back to the calling argument is an *in-out* parameter. *In-out* parameters are often implemented as call by result, or even call by reference.

Functional operations. For statement-oriented languages such as C, ALGOL-60, or Java, the discussion parameters above describes the transmission of information into procedures. But languages such as LISP, and more recently ML, have a totally different execution model. These languages are based upon a functional model of execution that does not explicitly declare data objects. One creates the function that is the result of the computation that one wants.

Example (22) is the standard statement-oriented version of the simple *abs* example given in (5). We can write these solutions functionally in LISP or ML as in example (25).

LanguageExpressionLISP(write (abs
$$(-(+A B) C)))$$
(25)MLwrite (abs $(A + B - C))$

LISP uses a prefix notation as given in (14) and ML uses the more common infix notation (13). What is important to notice, however, is that neither language uses intermediate data objects. The argument A + B - C is evaluated and passed to the *abs* function. The value returned from the *abs* function is then passed to the *write* function. In no case does this value need to be stored as an intermediate value with its own name, as is the case with other variable transmission mechanisms.

The actual code for computing example (25) [using the algorithm in (5)] could be written in ML as example (26).

write (Let
$$D = A + B - C$$

in
if $D < 0$ then D
else -D):
(26)

This states that *write* will write the value of its argument, which is the value of the IF expression. The expression A + B - C is assigned the name D and, in the IF statement, either D or its negation is returned. In no case is there an explicit variable storage for any data object.

Messages. In 1974 the Smalltalk language introduced a new model for describing information between procedures. This is the concept of the message. The mental image being portrayed is of a stationary object being passed messages to procedures called methods. The object responds to the message and waits for the next message. Rather than a program being a collection of procedures communicating with one another, a program is now a collection of objects, each defining its own environment, and each responding to queries about its environment from other objects. By adding the concept of inheritance, we have the principal idea behind objectoriented languages.

In Smalltalk we would have integer objects. Example (25) would be implemented with the *print* method and *abs* method for integers. Operations in Smalltalk follow their operands, and execution proceeds from left to right with no operation hierarchy as is the case with FORTRAN or C. Since the value A + B - C is an instance of an integer object, the program equivalent for (25) would be (27).

$$(A + B - C)$$
 abs print (27)

Execution proceeds as in sequence (28).

Smalltalk	
expression	Execution
Α	A is an integer object
+B	The $+$ (addition) method for integer
	A is passed the value B.
	The integer object $(A + B)$ is returned
	by the addition method of A.
-C	The – (subtraction) method for
	integer object $(A + B)$ is passed the
	value C. The integer object (A + B
	− C) is returned.
abs	The abs method for integer object
	(A + B - C) is called and it returns
	the integer value object that is the
	value of abs.
print	The print function for the integer
	value result of the abs function
	is called. The value of the
	expression is printed.
	(28)

Although we still have functions, procedures, and data variables, by describing programming objects with this model we get a design process that looks very much like the problem we are trying to solve. This model, built in Smalltalk in 1974, is the one major feature that was not present in the early days of programming language design.

Inheritance. Inheritance is the third major form of execution sequencing. In this case, information is passed automatically between program segments. The concept of inheritance began with ALGOL-60 and its static scope rules, and was extended with the information hiding concepts of Parnas and implemented in Smalltalk. This became the basis for the current models used in today's object-oriented languages C++ and Java.

Scope. The scope of a variable declaration is the set of procedures where it can be used. In general, a variable is declared; the variable name can be used throughout that procedure. However, in some languages (ALGOL-60, C, Pascal), a segment can be declared within another segment. For example, in C we can declare a block using the symbols { and }. Consider example (29).

C program	Comment
{ int I, J;	I and J declared integer variables
I = J+1;	
{ int J, K;	A new J and K declared integer
	variables. I is also available in this
	block.
}	
}	
	(29)

In this case, the scope for I is the entire example, both the outer block and the inner block. The scope for the J declared on the first line is only the outer block since there is a declaration for a new variable J in the inner block. The use of scope was one of the first examples of passing information between program segments.

Encapsulation. In 1972, Parnas introduced the concept of encapsulation. The idea was that good program design would consist of a data object definition, and only those functions that needed to know the data object definition would have access to it. For example, the Pascal *MyAddress* record was defined in example (11). Only functions using the type *MyAddress* could have access to the actual structure of the record. A module for an address list could have the functions *CreateAddress*, *PrintAddress*, *RetrieveSpecificAddress*, or any other function required to manage an address list. All other parts of the program would access an address list by calling one of these specific address list functions.

Encapsulation was a clear description of how to divide a program into manageable parts for effective construction. The problem was that there was no current language that effectively implemented that structure. Simula, a variation of ALGOL-60 used for simulation problems, did contain the concept of a data class for declaring data objects. The development of Smalltalk solved the encapsulation problem.

Smalltalk developed the concept of a class object. There were integer class objects, string class objects, file class objects, and so forth. Data objects would then be instances of these class objects. Therefore, we could have an instance of an integer object, an instance of a string, an instance of a file. However, the real power in Smalltalk was that objects would inherit the properties of the objects from which they were created. Thus, if an integer object were designed with the methods (that is, functions) of addition and subtraction, each instance of an integer object would inherit those same functions. One would only need to develop these operations once and then the functionality would pass on to the derived object.

The real power of Smalltalk derives from the fact that all objects are derived from one master object called an Object. An Object is the parent class of objects such as magnitude, collection, and stream. Magnitude now is the parent of objects that have values, such as numbers, characters, and dates. Collections can be ordered collections such as an array or an unordered collection such as a set. Streams are the parent objects of files. From this structure an entire class hierarchy can be developed.

If we develop a method for one object (for example, *print* method for *object*), then this method gets inherited to all objects derived from that object. Therefore, there is not the necessity to always define new functionality. If we create a new class of integer that, for example, represents the number of days in a year (from 1 to 366), then this new integerlike object will inherit all of the properties of integers, including the methods to add, subtract, and print values. It is this concept that has been built into C++, Java, and current object-oriented languages.

Given a function call (or message in Smalltalk), because of inheritance, we often cannot tell where the method is actually defined. This allows us to develop polymorphic functions. A polymorphic function is one which each different set of arguments can call an entirely different function to execute. Thus the Smalltalk message *print* in each of examples (30)

Smalltalk		Value
message	Interpretation	printed
2 print	Integer variable print method	2
'ABC' print	String variable print method	'ABC'
Object print	Object class print method	An Object
Integer print	Integer class print method. This is the same method inherited from Object print	An Integer
	<i>,</i> , ,	(30)

calls an entirely different method to execute.

Once we build concepts around a class definition, we have a separate package of functions that are self-contained. We are able to sell that package as a new functionality that users may be willing to pay for rather than develop themselves. This leads to an economic model where companies can build add-ons for existing software, each add-on consisting of a set of class definitions that becomes inherited by the parent class. *See* OBJECT-ORIENTED PROGRAM-MING.

Computing Environments

Since the 1950s, the use of computers has undergone several radical shifts. In response to changing computer hardware and computing environments, programming language development has tried to keep pace with the needs of those environments.

Mainframe era. The period from the 1940s to 1970 can be regarded as the mainframe era. Computers were large, expensive machines used to solve corporate problems. Data entry for these machines utilized punched cards (the 80-column Hollerith card). Languages were needed so that programmers could submit their program one day and get their results the next. This was the era of FORTRAN, ALGOL-60, COBOL, and similar languages.

Personal computer era. The 1970s saw the introduction of terminals to mainframe computers. There was still a single large mainframe, but users could access computers from their desks. Programming started to become more interactive, and quick turnaround from 1 day to several minutes became feasible. Languages that encouraged quick turnaround, such as BASIC, became popular.

The situation was to change in 1976 with the introduction of the personal computer (PC). Users no longer needed to be connected to a mainframe, and standalone computing on an inexpensive machine was now possible. Interactive languages, which provided immediate feedback to the user, such as BASIC, became the standard on such machines.

The personal computer entered the mainstream of business computing in 1981 with the introduction of the IBM PC. The role of the mainframe computer became that of a corporate data server, and computing moved to the desktop, especially as desktop computers became faster and cheaper.

During the 1970s, many of the modern principles of the graphical user interface (GUI) [for example, windows and the use of a mouse for point and click data entry] were developed at the Xerox Palo Alto Research Center. These features became widely available with the introduction of the Apple Macintosh (1984) and subsequently with Microsoft Windows for IBM-compatible PCs. *See* HUMAN-COMPUTER IN-TERACTION.

With the graphical user interface, building window applications became important, and objectoriented program design became an economic necessity. Class libraries could be sold, allowing users to easily build applications that include windows and features such as menus, scroll bars, and buttons. Visual Basic evolved with predefined methods for including these window-class objects.

Internet era. Beginning in 1970, the U.S. Advanced Research Projects Agency (ARPA) began research in linking machines together. Initially a project to allow machines to continue to communicate and survive in the event of an attack on the United States, the initial Defense Department ARPANET evolved into the worldwide Internet by the mid-1980s. Today the Internet has millions of computers connected to it, mostly PCs. *See* INTERNET.

The Internet was a researcher's domain until the late 1980s. CERN, the European nuclear research facility in Geneva, Switzerland, was developing a method whereby researchers could share technical documents. In 1989, Tim Berners-Lee at CERN came up with the concept of the HTML (hypertext markup language) hyperlink as a way to navigate around the Internet and retrieve documents in a system known as the World Wide Web. With the development of the MOSAIC web browser in 1993, the general populace discovered the Internet, and by the end of the 1990s millions of users were connecting to it each day to search for information. *See* WORLD WIDE WEB.

Use of the World Wide Web led yet again to a change in the role for programming languages. Early web pages were static: A report could be displayed, a picture shown, or control could navigate to another web site. However, as an industry started to develop around the Web (e-commerce), a need developed for active web pages. Web developers wanted the ability to obtain data from users (personal information, ordering information) and then process that information in some meaningful way. This led to the need for web languages, such as Java.

Programming Paradigms

Today, a great deal is understood about the programming process. There are many languages that satisfy different needs. The following describes the basic programming models in use today and the various languages that use those programming models.

Imperative programs. An imperative program is a sequence of commands or statements telling the computer how to execute a particular algorithm.

This was the model for FORTRAN, ALGOL, and COBOL in the 1960s, and is still the major model in use for languages such as C, C++, BASIC, and Java. Execution of a program proceeds from statement to statement, each statement accessing and modifying specific data objects in a machine's memory. We can think of an imperative program as developing the steps that transform a set of data items from initial values into their final values; each statement performs one step in this complex series of transformations. It reflects the basic architecture of a computer and seems most natural for describing the programming process. However, it is not always the best model for solving a particular problem.

Applicative programs. Whereas imperative programming is concerned with the series of steps needed to transform data from some initial set of values into their final set of values, in an applicative program one constructs the function that transforms all the data from initial value to final value. Because of its close association with functions, it is also called functional programming. While this sounds similar to the imperative model, it leads to a very different model of programming. The goal is to transform all the data, function by function, until the final answer is obtained.

Examples (31) and (32) show the difference between the two models. Here, given a sequence of numbers, create a sequence of the same numbers in the reverse order.

An imperative solution, such as written in FOR-TRAN or C, would proceed as in (31).

- 1. Set Last to be end of the first sequence
- 2. Set New to be 1 (Location of the new sequence)
- 3. Move the object at *Last* to the object at *New* (Move one member of sequence)
- 4. Subtract 1 from Last and add 1 to New
- 5. Repeat steps 3 and 4 for all members of the sequence

As shown in **Fig.** 2a, the imperative solution proceeds by moving each item in the sequence, one at a time, from last to first.

(31)

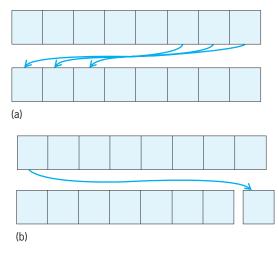


Fig. 2. Diagrams of (*a*) imperative solution and (*b*) applicative solution.

The applicative solution is given in Fig. 2*b* and is specified as (32).

- 1. Remove the first member from the list.
- 2. Reverse the remaining members of the list.
- 3. Add the first member to the reverse remainder of the list. (32)

In the applicative case, we are unconcerned with specific data items; we consider the switching operation on the entire sequence at one point. In this case, the applicative solution applies to the entire sequence. The first member of the sequence is moved to the end and then the process is repeated for the remaining members, with the second members now being the beginning of the list.

This is an example of a recursive solution, since the solution, in step 2, required using the solution we were still developing. In general, recursive solutions are shorter but harder to write than imperative solutions. However, once written, they can be shown to be correct more easily than the other method.

Applicative programming is the basis for LISP. Problems requiring searching large databases often can be expressed this way, and applicative programs have been useful in developing artificial intelligence applications. Each function in an applicative solution to a problem represents some fact about the solution. The structure of a functional program then becomes (33).

 $Fact1 = statements \ describing \ fact 1$ $Fact2 = statements \ describing \ fact 2$ $Fact3 = statements \ describing \ fact 3$

The solution to the problem then becomes (34).

Solution = fact1 and fact2 and fact3 and fact4
(34)

Executing the program is simply the statement: *answer* = *solution(input data)*.

Declarative programs. In both applicative and imperative programs, the programmer presents the algorithm for computing a solution. In the imperative case, the program is a step-by-step sequence of statements of computing the answer from the initial set of data items. In the applicative case, the program is a sequence of functions that are executed to transform the input data into the answer.

Prolog represents a third method. In this case, the structure of the solution is given, and the programming system determines how to find a solution. The above example could be modified to state: Given a sequence of numbers, create a sequence of the same numbers in ascending order.

Then, we could represent a solution according to Prolog-like algorithm (35).

- 1. Let I represent the Ith member of the sequence.
- 2. A (I,I + 1):- greaterthan (I, I + 1), switch
- (I, I + 1). 3. A (I, I + 1) :- write (A).

Although not true Prolog, this represents how Prolog

(35)

searches for a solution. In this case, line 2 of (35), we have a relation A. If I and I + I have the property that the *greaterthan* function is true (that is, member at I is greater than the member at I + I), then the *switch* function is called (I and I + I reverse order). However, if *greaterthan* always fails, then no variable is out of order, and statement 3 is next executed. The only way that 2 fails is for A to now be in order, so the *write* function prints out the value of the sorted array.

In this case we simply specified constraints on the output: For any member of the output sequence, the following member must have a greater value. We did not specify how the computer would provide for this condition, only that when the program executed the *write* function in line 3 of (34), the array *A* would be sorted.

Applications requiring searching large databases are amenable to this technique. It has also been found to be useful is solving logic problems; thus the method is also called logic programming. Much of the work in a compiler that automatically translates a BNF description of a programming language into some internally useful form is another example use of this declarative model. A compiler writer simply gives the BNF description of a language, and the compiling system automatically generates a program that will recognize programs written in that language.

Current Programming Language Models

Of the languages given in Fig. 1, and of the hundreds not presented here, a few stand out as currently significant notations for which to solve computer programs.

C. C was developed by AT&T Bell Laboratories during the early 1970s. At the time, Ken Thompson was developing the UNIX operating system. Rather than using machine or assembly language as in (2) or (3) to write the system, he wanted a high-level language. Starting with a language named BCPL, he created a small subset named B. By adding data types and additional operations, he renamed B as C. *See* OP-ERATING SYSTEM.

C has a structure like FORTRAN. A C program consists of several procedures, each consisting of several statements, that include the IF, WHILE, and FOR statements. However, since the goal was to develop operating systems, a primary focus of C was to include operations that allow the programmer access to the underlying hardware of the computer. C includes a large number of operators to manipulate machine language data in the computer, and includes a strong dependence on reference variables so that C programs are able to manipulate the addressing hardware of the machine.

The relationship between C arrays and string variables illustrates the close relationship of C programs to the underlying hardware. Declaration (36)

$$cbar A[10];$$
 (36)

creates an array of 10 characters, A[0] through A[9].

However, the C compiler actually creates a reference variable for *A* pointing to the first element of the array. Therefore, *A* is just a pointer to A[0]. The value at this location (called dereferencing *A*) can be accessed by the array reference A[0] or by the pointer dereference operation (that is, *) as in **A*. Accessing the *I*^{tb} element of the array can be denoted as either A[I] or as *(A + I). This duality between C types, such as strings and the underlying integer representation for those types, makes C both an extremely efficient language in which to write systems programs and a very difficult language in which to write correct programs.

C++. C++ was developed in the early 1980s as an extension to C by Bjarne Stroustrup at AT&T Bell Labs. The idea was to extend it with Smalltalk-like classes. Each C++ class would include a record declaration as well as a set of associated functions. In addition, an inheritance mechanism similar to Smalltalk was included in order to provide for a class hierarchy for any program. The Pascal record declaration from (11) could be extended with class functions via declaration (37).

Class	address {	
	String20 Name;	
	String20 Street;	
	String20 City;	
	String20 State;	(37)
	int PostalCode;	
	<pre>void CreateAddress();</pre>	
	<pre>void PrintAddress();</pre>	
	<pre>void RetrieveSpecificAddress(); }</pre>	

In order to access *address* records, one of the functions *CreateAddress*, *PrintAddress*, or *RetrieveSpecficAddress* would have to be called. This provides for the encapsulation so important for good program design.

C++ also included inheritance. A new class, *MyAddress*, could be derived from *address* via declaration (38).

Class MyAddress : *address*{... *additional functions*} (38)

Here *MyAddress* inherits all of the properties (that is, record description and class functions) of *address* as well as creation of any new functions specific for that class. C++, then, includes the efficiency of a C program with the encapsulation, inheritance, and flexibility of object-oriented programming in Smalltalk.

Java. By the early 1990s, the World Wide Web was becoming a significant force in the computing community, and web browsers were becoming ubiquitous. However, for security reasons, the browser was designed with the limitation that it could not affect the disk storage of the machine it was running on. All computations that a web page performed were carried out on the web server accessed by web address (its Uniform Resource Locator, or URL). That was to prevent web pages from installing viruses on user machines or inadvertently

(or intentionally) destroying the disk storage of the user.

The effect of this was that all information had to be transferred from the user to the web server. James Gosling at Sun Microsystem led development of a product called Oak, which evolved into Java in 1995, that could execute on the user's machine. Thus, instead of transferring all of the user's data to the server machine, the server would transfer the application program to the user.

In order to make this efficient, the transferred program had to be small, and it had to be capable of running on every browser on every computer connected to the Internet. Sun decided to transfer a version of the source program (the bytecode), rather than the usually longer machine language program. A program built into the browser would execute the bytecode as if it were a machine language program for that particular machine. Such a program is typically called an interpreter.

The constraint on this bytecode language was that it could not affect the disk storage of the user's machine. Initially, C++ was the programming language of choice due to its class hierarchy and efficient execution model. But because of pointer references and binary operations on machine addresses in the underlying C base language, it would be impossible to assure adherence to the web security model. C++ had to be dropped as the base language, and it has been replaced in this context by Java. Java bears a strong similarity to C++, but has eliminated many of the problems of C++. The three major features addressed by Java are:

1. There are no reference variables, thus no way to explicitly reference specific memory locations. Storage is still allocated by creating new class objects, but this is implicit in the language, not explicit.

2. There is no procedure call statement; however, one can invoke a procedure using the member of class operation. A call to *CreateAddress* for class *address* would be encoded as *address.CreateAddress*().

3. A large class library exists for creating web-based objects.

The Java bytecodes (called applets) are transmitted from the web server to the client web site and then execute. This saves transmission time as the executing applet is on the user's machine once it is downloaded, and it frees machine time on the server so it can process more web "hits" effectively. *See* CLIENT-SERVER SYSTEM.

Visual Basic. Visual Basic, first released in 1991, grew out of Microsoft's GW Basic product of the 1980s. The language was organized around a series of events. Each time an event happened (for example, mouse click, pulling down a menu), the program would respond with a procedure associated with that event. Execution happens in an asynchronous manner.

Each displayed screen is called a form. Forms are the visual components, and they are event controllers. Controls (for example, push buttons, radio buttons, check boxes, labels, text boxes, and scroll bars) are added to the form. Each control creates events, the set of actions that occur as a result of an action in the user interface, for example, pressing a key on the keyboard or clicking a mouse button.

Visual Basic provides a different program design model. Rather than looking at a program as a series of statements that form an algorithm, in Visual Basic the approach is to determine what actions occur in response to a series of events.

Prolog. Although Prolog development began in 1970, its use did not spread until the 1980s. Prolog represents a very different model of program execution, and depends on the resolution principle and satisfaction of Horn clauses of Robert A. Kowalski at the University of Edinburgh. That is, a Prolog statement is of the form p:-q.r, which means p is true if both q is true and r is true [as in example (35)].

A Prolog program consists of a series of Horn clauses, each being a sequence of relations concerning data in a database. Execution proceeds sequentially through these clauses. Each relation can invoke another Horn clause to be satisfied. Evaluation of a relation is similar to returning a procedure value in imperative languages such as C or C++.

As an example, assume there is a relation *parent_of*. *Parent_of* (X,Y) means that X is the parent of Y. Assume that there is a database of relationships (39).

parent_of (John,Mary)	parent_of	
	(Susan,George)	
parent_of (Susan,Mary)	parent_of	(39)
	(Henry,John)	(57)
parent_of (John,George)	parent_of	
	(George,Alice)	

In Prolog, one queries the database. If all the facts are specified, then Prolog responds either *true* (40) or *false* (41), depending upon whether the data are in the database. However, some facts are omitted, Prolog will search the database to see if a match can be found (42).

Prolog query	Prolog response	
parent_of (John,George)	true	(40)
parent_of (Alice,Mary)	false	(41)
parent_of (X,John)	X = Henry	(42)

Prolog programs are built by creating new relationships. A grandparent_of (X, Y) relationship can be defined by saying there is some person Z who is the parent of Y, and X is the parent of Z. This is specified in Prolog as (43).

$$\begin{array}{l} \text{grandparent_of (X,Y) :- parent_of (X,Z),} \\ \text{parent_of (Z,Y)} \end{array}$$
(43)

Prolog executes this by first trying to satisfy the first *parent_of* relationship. That defines a value for *X* and *Y*. It then sees if that association meets the second relation. If so, a true match is found. As with the single relation, if any operand is missing, Prolog tries

Prolog query	Prolog response	
grandparent_of (John,George)	false	(44)
grandparent_of (John,Alice)	true	(45)
grandparent_of (X.Alice)	X = John	(46)

Unlike the other languages mentioned, Prolog is not a complete language. That means there are algorithms that cannot be programmed in Prolog. However, for problems that are amenable for searching large databases, Prolog is an efficient mechanism for describing those algorithms. *See* SOFTWARE; SOFT-WARE ENGINEERING. Marvin V. Zelkowitz

Bibliography. T. Pratt and M. Zelkowitz, *Programming Languages: Design and Implementation*, Prentice Hall, 4th ed., 2001; R. Sebesta, *Concepts of Programming Languages*, Benjamin Cummings, 5th ed., 2001; R. Sethi, *Programming Languages: Concepts and Constructs*, Addison Wesley, 2d ed., 1996.

Progression (mathematics)

Ordered, countable sets of numbers, x_1, x_2, x_3, \ldots , not necessarily all different. In general such sets are called sequences, whereas the term progression is usually confined to the special types: the arithmetic, in which the difference $x_k - x_{k-1}$ between successive terms is constant; the geometric, in which the ratio x_k/x_{k-1} is constant; and the harmonic, in which the reciprocals of the terms are in arithmetic progression.

Arithmetic progressions. If the first term is a and the common difference b,

$$x_1 = a, x_2 = a + b, x_3 = a + 2b, \dots,$$

 $x_n = a + (n + 1)b, \dots$ (1)

In the sum of *n* terms S_n , two terms equidistant from the ends always have the same sum $x_1 + x_n$; hence $2S_n = n(x_1 + x_n)$ and

$$S_n = n\frac{x_1 + x_n}{2} = n\left(a + \frac{n-1}{2}b\right)$$

If x_1 , x_2 , x_3 are in arithmetic progression, $x_2 = (x_1 + x_3)/2$ is called the arithmetic mean of x_1 and x_3 .

Geometric progressions. If the first term is a and the common ratio r,

$$x_1 = a, x_2 = ar, x_3 = ar^2, \dots,$$

 $x_n = ar^{n-1}, \dots$ (2)

Excluding the case r = 1 (when all terms are the same), the sum S_n of n terms satisfies

 $S_n - rS_n = a - ar^n$; hence,

$$S_n = a \frac{1 - r^n}{1 - r}$$

If |r| < 1, $r^n \to 0$ as $n \to \infty$; hence the sum of the infinite geometric series

$$\sum_{n=1}^\infty ar^{n-1}=rac{a}{1-r}\qquad |r|<1$$

If x_1 , x_2 , x_3 are in geometric progression, $x_2 = \sqrt{x_1x_3}$ is called the geometric mean of x_1 and x_3 . Since

$$(\sqrt{x_1} - \sqrt{x_3})^2 = x_1 + x_3 - 2\sqrt{x_1 x_3} \ge 0$$
$$\frac{x_1 + x_3}{2} \ge \sqrt{x_1 x_3}$$

the arithmetic mean of two unequal positive numbers exceeds their geometric mean.

The arithmetic mean *A* and the geometric mean *G* of *n* positive numbers are defined as

$$A = \frac{x_1 + x_2 + \dots + x_n}{n}$$
 and $G = \sqrt[n]{x_1 x_2 \cdots x_n}$

Also, $A \ge G$.

and

Harmonic progression. The reciprocals of sequence (1) form a harmonic progression. There is no compact expression for the sum of *n* terms. If x_1 , x_2 , x_3 are in harmonic progression,

$$x_2 = \frac{2x_1x_3}{x_1 + x}$$

is called their harmonic mean.

Sum sequence. A general method of summing a sequence of n terms depends upon a theorem in the difference calculus which is the analog of the fundamental theorem of the differential calculus.

If $x_n = f(n)$ is defined for n = 0, 1, 2, ..., the difference f(n) is defined as

$$\Delta f(n) = f(n+1) - f(n)$$

 Δ is a linear operator; that is,

$$\Delta[af(n) + bg(n)] = a\Delta f(n) + b\Delta g(n)$$

In the difference calculus the factorial powers $n^{(k)}$ for integral k are defined by

$$n^{(k)} = egin{cases} n!/(n-k)! & k < n \ n! & k = n \ 0 & k > n \end{cases}$$

Thus

$$n^{(k)} = n(n-1)(n-2)\cdots(n-k+1)$$
$$0 < k \le n$$
$$(-k) = \frac{1}{(n+1)(n+2)\cdots(n+k)} \qquad k > 0$$

In particular

n

$$n^{(0)} = 1, n^{(1)} = n, n^{(-1)} = \frac{1}{n}$$

٦	TABLE 1. Differences	
	<i>f</i> (<i>n</i>)	$\Delta f(n)$
	constant	0
	r ⁿ	(r -1)r ⁿ
	n ^(k)	<i>kn</i> ^(k-1)
	cos nα	$-2 \sin \frac{1}{2} \alpha \sin (n + \frac{1}{2}) \alpha$
	$\sin n \alpha$	$2 \sin \frac{1}{2} \alpha \cos (n + \frac{1}{2}) \alpha$

TABLE 2. Antidiffer	rences	
<i>f</i> (<i>n</i>)	Δ^{-}	⁻¹ f(n)
0	constant	
r ⁿ	$\frac{r^n}{r-1}$	(<i>r</i> ≠ 1)
{ <i>n</i> ^(k)	$\frac{n^{(k+1)}}{k+1}$	(<i>k</i> ≠ −1)
1	п	(k = 0)
cosnα	$\frac{\sin{(n-1/2)}}{2\sin{\alpha/2}}$	
sin <i>n</i> α	$\frac{\cos{(n-1)}}{2\sin{\alpha}}$	

Moreover, $n^{(k)}$ satisfies the functional equation

$$n^{(k)} \cdot (n-k)^{(b)} = n^{(b)} \cdot (n-b)^{(k)} = n^{(b+k)}$$

This leads to the general definition of the factorial power

$$x^{(k)} = \frac{\Gamma(x+1)}{\Gamma(x+1-k)}$$

for all real values of *x* and *k* for which the gamma functions exist. When both of the gammas are infinite (for arguments 0, -1, -2, ...), define $x^{(k)} = \lim (x + \epsilon)^{(k)}$ as $\epsilon \to 0$; thus

$$(-1)^{(2)} = \lim(\varepsilon - 1)^{(2)} = \lim \Gamma(\varepsilon) / \Gamma(\varepsilon - 2)$$
$$= \lim(\varepsilon - 1)(\varepsilon - 2) = 2!$$

The last two entries of Table 1 follow from

$$\Delta e^{in\alpha} = (e^{i\alpha} - 1)e^{in\alpha} = 2i \sin \frac{\alpha}{2e^{i[n+(1/2)]\alpha}}$$

on taking real and imaginary parts.

If $\Delta F(n) = f(n)$, F(n) is called the antidifference of f(n) and written $\Delta^{-1}f(n)$. Two antidifferences of

TABLE 3. Sequence of cubes						
n	0	1	2	3		
P(n)	0	1	8	27	<i>P</i> (O) = 0	
$\Delta P(n)$	1	7	19		$\Delta P(0) = 1$	
$\Delta^2 P(n)$	6	12			$\Delta^2 P(0)/2! = 3$	
$\Delta^3 P(n)$	6				$\Delta^{3}P(0)/3! = 1$	

f(n) differ at most by a constant (or by a periodic function of period 1 for functions f(x) of a continuous variable). Table 1 implies the table of antidifferences (**Table 2**).

Just as antiderivatives are used to compute definite integrals, antidifferences are used to compute definite sums:

$$\sum_{n=p}^{q} f(n) = \Delta^{-1} f(n) \Big|_{p}^{q+1} = F(q+1) - F(p)$$

The proof is immediate on replacing f(n) by $\Delta F(n) = F(n + 1) - F(n)$ and performing the indicated summation.

Example 1. The arithmetic progression, as in notation (1):

$$\sum_{n=1}^{N} a + b(n-1) = (a-b)n + \frac{1}{2}bn^{(2)} \Big|_{1}^{N+1}$$
$$= (a-b)N + \frac{1}{2}b(N+1)N$$

Example 2. The geometric progression, as in notation (2):

$$\sum_{n=1}^{N} ar^{n-1} = \frac{ar^{n-1}}{r-1} \Big|_{1}^{N+1} = a\frac{r^{N}-1}{r-1}$$

Example 3. The cosine sequence:

$$\sum_{n=1}^{N} \cos n\alpha = \frac{\sin(n-1/2)\alpha}{2\sin^{1}/2\alpha} \Big|_{1}^{N+1}$$
$$= \frac{\sin(N+1/2)\alpha - \sin^{1}/2\alpha}{2\sin^{1}/2\alpha}$$

In order to sum a polynomial P(n) of degree k, P(n) may be expressed in terms of factorial powers $n^{(1)}$, $n^{(2)}$, ..., $n^{(k)}$ by Newton's theorem:

$$P(n) = P(0) + \frac{\Delta P(0)}{1!} n^{(1)} + \frac{\Delta^2 P(0)}{2!} n^{(2)} + \dots + \frac{\Delta^k P(0)}{k!} n^{(k)}$$

The differences of P(n) when n = 0 are computed as in **Table 3**.

In Table 3, $P(n) = n^3$ and hence

$$n^{3} = n^{(1)} + 3n^{(2)} + n^{(3)}$$
$$\sum_{n=1}^{N} n^{3} = \frac{1}{2}n^{(2)} + \frac{1}{4}n^{(3)} + \frac{1}{4}n^{(4)} \Big|_{1}^{N+1}$$
$$= \frac{1}{4}n^{2}(n-1)^{2} \Big|_{1}^{N+1} = \frac{1}{4}(N+1)^{2}N^{2}$$

Summation by parts,

$$\Delta^{-1} \left[f(n) \,\Delta g(n) \right]$$

= $f(n) \,g(n) - \Delta^{-1} \left[g(n+1) \,\Delta f(n) \right]$

is frequently useful in finding antidifferences. With $g(n) = (-1)^n$, $\Delta g(n) = 2(-1)^{n-1}$, this gives

$$2\Delta^{-1} \left[(-1)^{(n-1)} f(n) \right] = (-1)^n f(n) - \Delta^{-1} \left[(-1)^{n-1} \Delta f(n) \right]$$

When P(n) is a polynomial of degree k, $\Delta^{k+1}P(n) = 0$, and this formula repeatedly applied gives

$$\Delta^{-1}[(-1)^{n-1}P(n)]$$

and hence the sum

$$\sum_{n=1}^{N} (-1)^{n-1} P(n)$$

The following tabulation permits the summation of alternating powers $(-1)^{n-1}n^k$:

Summation by parts also gives antidifferences such as $\Delta^{-1}[r^n P(n)]$. Thus $\Delta^{-1}(nr^n) = r^{n-1}p(n-p)$, where p = r/(r-1). See SERIES. Louis Brand

Bibliography. W. G. Kelley and A. C. Peterson, *Difference Equations*, 2d ed., 2001; L. M. Milne-Thomson, *The Calculus of Finite Differences*, updated edition, 2004; P. K. Rees, F. W. Sparks, and C. S. Rees, *College Algebra*, 10th ed., 1990.

Projective geometry

A geometry that investigates those properties of figures that are unchanged (invariant) when the figures are projected from a point to a line or plane. Although isolated theorems appeared earlier, the first systematic treatment of the subject was given by the French army officer. J. V. Poncelet (1788–1867), who began his *Traité des propriétés projectives des figures* in 1812 while a prisoner of war in Russia.

Two features of plane projective geometry are (1) introduction of an ideal line that each ordinary line g intersects (the intersection being common to all lines parallel to g), and (2) the principle of duality, according to which any statement that is obtained from a valid one (theorem) by substituting for each concept involved, its dual, is also valid. ("Line" and "point" are dual, "connecting two points by a line" is dual to "intersecting two lines," and so on.) The subject has been developed both synthetically (as a logical consequence of a set of postulates) and analytically (by the introduction of coordinates and the application of algebraic processes). The characteristic properties of geometrical projection are that it defines a one-to-one correspondence that preserves cross ratio between the two sets of entities involved (for example, the points of two lines, the lines of two pencils, where a pencil of lines consists of all the lines on a point). Thus if P_i , P'_i (i = 1,2, 3, 4) are corresponding points of two lines, and x_i, x'_i (i = 1, 2, 3, 4) are their respective coordinate, then relation (1) holds. All such correspondences be-

$$\frac{(x_3 - x_1)(x_4 - x_2)}{(x_3 - x_2)(x_4 - x_1)} = \frac{(x_3' - x_1')(x_4' - x_2')}{(x_3' - x_2')(x_4' - x_1')}$$
(1)

tween one-dimensional forms are given by the linear transformation x' = (ax + b)/(cx + d), $ad - bc \neq 0$; between two-dimensional forms (for example, two

planes) by relations (2), with nonvanishing determi-

$$x' = (a_{11}x + a_{12}y + a_{13})/(a_{31}x + a_{32}y + a_{33})$$

$$y' = (a_{21}x + a_{22}y + a_{23})/(a_{31}x + a_{32}y + a_{33})$$
(2)

nant $|a_{ij}|$ (*i*, *j* = 1, 2, 3). The latter transformations form an eight-parameter group, the projective group of the plane. It contains, as a proper subgroup, the three-parameter group (rotations and translations) that defines euclidean geometry of the plane; and so euclidean geometry is a subgeometry of projective geometry.

Conics are defined as the class of points of intersection of corresponding lines of two projective pencils of lines (on distinct points). If these intersections are collinear, the pencils are perspective and the conic is degenerate. Otherwise, the conic is nondegenerate. Any two nondegenerate conics are equivalent; that is, a projective transformation exists that carries one into the other.

One of the well-known theorems of projective geometry was proved by the French architect G. Desargues (1593-1662): If the lines joining corresponding vertices of two triangles are concurrent, the corresponding sides intersect in collinear points. The converse of the theorem is its dual, and hence is also valid. *See* CONFORMAL MAPPING; EUCLIDEAN GEOMETRY. Leonard M. Blumenthal

Prokaryotae

Prokaryotic cells are usually microscopic, but some recently discovered prokaryotes are visible to the naked eye. A taxonomic kingdom comprising microorganisms that lack a true nucleus or any other membrane-bound organelle. Instead of a nucleus, prokaryotic cells contain a nucleoid, which is a DNA-dense region of the cell where DNA replication and transcription occur. The chromosome of these cells is circular and, in contrast to eukaryotic cells, continuously codes for cellular proteins and other molecules. Some cells also contain a plasmid, which is a smaller separate piece of circular DNA. The ribosomes of prokaryotes are found floating loose in the cytoplasm, whereas in eukaryotic cells the ribosomes are found in or on the rough endoplasmic reticulum. See CLASSIFICATION (BIOLOGICAL); EU-KARYOTAE.

Bacteria and Archaea. Prokaryotae include the Bacteria and the Archaea. Although the Archaea and the Bacteria look identical under the microscope, they differ in their chemical composition and are unrelated. Bacteria typically have a peptidoglycan cell wall and a cytoplasmic membrane, while Archaea do not have peptidoglycan in their cell wall. Archaea can often be found in extreme environments, such as hot springs, in which most Bacteria cannot survive. *See* BACTERIA.

Characteristics. Endospores are survival structures made by some prokaryotes; some have been found to be viable after thousands of years. Prokaryotic cells undergo binary fission, where one cell divides into two equal daughter cells. DNA may also be passed from one cell to another by conjugation, transduction, or transformation. Conjugation involves the formation of a long hollow tube, known as the sex pilus, down which DNA passes from the donor cell to the recipient cell. Transduction occurs after one cell becomes infected with a bacterial virus or bacteriophage, which can sometimes carry pieces of bacterial DNA to a new, uninfected cell. Transformation is the uptake of naked DNA released by a cell that has lysed. This DNA may contain genes useful to the prokaryote, such as antibiotic-resistance genes. *See* BACTERIOPHAGE; DEOXYRIBONUCLEIC ACID (DNA); VIRUS.

Motility. Motile prokaryotic cells move by gliding over surfaces or by flagella. Most flagella are external, found at one end of the cell (polar), at both ends of the cell (bipolar), or over the entire surface of the cell (peritrichous). The spirochetes, a group of prokaryotes that are spirally shaped, contain internal flagella known as axial filaments. Most prokaryotes are single-cell microorganisms; however, a few species are filamentous, mycelial, or colonial. *See* CELL MOTILITY.

Respiration. Different species of prokaryotes have different oxygen requirements. Some may be aerobic, requiring atmospheric oxygen; others are anaerobic, functioning in the absence of oxygen. A few prokaryotes are microaerophilic, requiring lowered oxygen and increased carbon dioxide concentrations to grow and reproduce. Prokaryotes respire to produce energy. Respiration is associated with the cytoplasmic membrane, where the electron transport chain is located. Energy is released from the reactions that occur as electrons are passed along carriers of the electron transport chain. *See* RESPIRATION.

Fermentation. Some prokaryote species use fermentative pathways for metabolism. Humans take advantage of fermentation by using these prokaryotes to produce cheese, yogurt, and pickled vegetables. Eukaryotic yeast cells also ferment and are used to produce wine and beer. *See* FERMENTATION. Marcia Pierce

Bibliography. M. Hoppert and F. Mayer, Prokaryotes, *Amer. Scientist*, 87:518–525, 1999; J. Lederberg, *Encyclopedia of Microbiology*, 2d ed., Academic Press, San Diego, 2000; E. Nester et al., *Microbiology: A Human Perspective*, 4th ed., McGraw-Hill, New York, 2004; L. M. Prescott, J. P. Harley, and D. A. Klein, *Microbiology*, 6th ed., McGraw-Hill, New York, 2005; H. N. Schulz and B. B. Jørgensen, Big bacteria, *Annu. Rev. Microbiol.*, 55:105–137, 2001.

Prolacertiformes

A group of diapsid reptiles classified within the Archosauromorpha and recorded from Upper Permian to Lower Jurassic deposits around the world (Australia, Antarctica, India, China, Algeria, South Africa, Israel, Europe including Russia, and North America). Prolacertiforms (also known as pro-

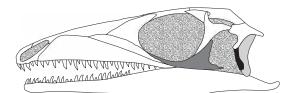


Fig. 1. Skull of *Prolacerta* in lateral view. The lower temporal bar is incomplete, with the jugal shown in dark gray, the reduced quadratojugal in black, and the emarginated quadrate in light gray. (*Redrawn and simplified from C. E. Gow, The morphology and relationships of Youngina capensis Broom and Prolacerta broomi Parrington, Palaeontol. Afr.,* 18:89–131, 1975)

torosaurs) share a suite of distinctive characters, including elongate neck vertebrae with long low neural spines and long slender ribs, reduction of the lower temporal bar (a bone forming a portion of the lateral aspect of the skull and part of the base of the cranium), and emargination of the quadrate (part of the upper jaw joint) to support a tympanum (**Fig. 1**). Despite the similarities, however, there is now some doubt that prolacertiforms form a natural (monophyletic) group since some (such as the archetypal *Prolacerta*) seem to lie closer to archosaurs than others (such as *Protorosaurus* and *Tanystropheus*). *See* ARCHOSAURIA; DIAPSIDA; REP-TILIA.

Characteristics. Before the adoption of modern methods of phylogenetic analysis (cladistics), prolacertiforms were one of the problematic groups for reptile systematists. Many prolacertiforms were classified with lizards or protolizards on the basis of a small set of skull characters (loss of the lower temporal bar, emargination of the quadrate to support an eardrum), despite major differences in the rest of the skeleton. Postcranial characters provide strong support for a relationship with archosauromorphs. The presence of strong rib processes on the vertebrae is an archosauromorph character, as are features of the hindlimb. In the ankle, the astragalus and calcaneum were movably articulated with one another so that the foot could face forward throughout the stride. In addition, the calcaneum had a posterior process, like the heel in mammals, which served as a lever in plantar flexion of the foot.

Diversity. Although many Late Permian and Triassic reptiles have been attributed to the prolacertiforms, many are fragmentary and/or incompletely described. Among the best-known are the primitive *Protorosaurus* (Upper Permian, Europe), *Prolacerta* (Lower Triassic, South Africa and Antarctica), *Tanystropheus* (Middle to Upper Triassic, Europe and the Middle East), and *Tanytrachelos* (Upper Triassic, North America).

Protorosaurus. Protorosaurus from the Upper Permian of Britain and Germany was up to 6 ft (2 m) in length, but most specimens are smaller than this. It was one of the first fossil reptiles to be formally described (in 1710), being interpreted first as a crocodile and then as a lizard. In life, *Protorosaurus* may have resembled a large monitor lizard, but there is no relationship with true lizards.

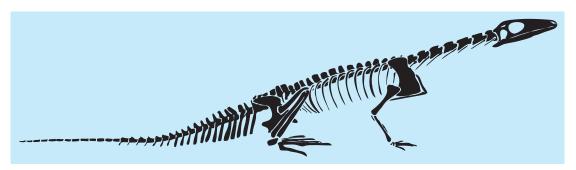


Fig. 2. Skeleton of Prolacerta. (Simplified from C. E. Gow, The morphology and relationships of Youngina capensis Broom and Prolacerta broomi Parrington, Palaeontol. Afr., 18:89–131, 1975)

Prolacerta. Prolacerta is smaller and more derived than *Protorosaurus*. The first specimen to be recovered (Lower Triassic, South Africa) was an isolated skull. It was originally misidentified as an ancestor of lizards. With the discovery of the postcranial skeleton, it became clear that *Prolacerta* was more closely related to archosaurs. *Prolacerta* is now known from numerous specimens (**Fig. 2**).

Tanystropheus. The most distinctive prolacertiform was *Tanystropheus* from the Middle to Upper Triassic of Europe and the Middle East. Individuals of the largest species of *Tanystropheus* could reach 9 ft (3 m) in length, but much of this consisted of an extraordinary neck that was at least three times the length of the trunk region. Not only were there more neck vertebrae than in other prolacertiforms (12 versus 8–9), but each vertebra was exceptionally elongated. Adults of *Tanystropheus* were aquatic, but juveniles may have been at least semiterrestrial, as they have sharply pointed tricuspid teeth more suited to insectivory.

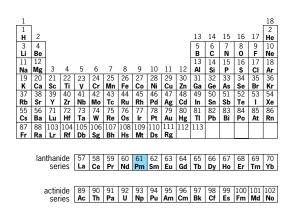
Tanytrachelos. Tanytrachelos was closely similar to *Tanystropheus*, but smaller, from the Upper Triassic/Lower Jurassic of North America. On current evidence, *Tanytrachelos* was the last surviving prolacertiform. Susan E. Evans

Bibliography. M. J. Benton, Vertebrate Palaeontology, 3d ed., Blackwell, Oxford, 2004; M. J. Benton (ed.), The Phylogeny and Classification of the Tetrapods, vol.1: Amphibians, Reptiles and Birds, Clarendon Press, Oxford, 1988; R. L. Carroll, Vertebrate Paleontology and Evolution, W. H. Freeman, New York, 1997; A. S. Romer, Osteology of the Reptiles, University of Chicago Press, 1956.

Promethium

A chemical element, Pm, atomic number 61. Promethium is the "missing" element of the lanthanide rareearth series. The atomic weight of the most abundant separated radioisotope is 147. *See* PERIODIC TABLE.

Although a number of scientists have claimed to have discovered this element in nature as a result of observing certain spectral lines, no one has succeeded in isolating element 61 from naturally occurring materials. It is produced artificially in nuclear reactors, since it is one of the products that results



from the fission of uranium, thorium, and plutonium.

All the known isotopes are radioactive. Its principal uses are for research involving tracers. Its main application is in the phosphor industry. It has also been used to manufacture thickness gages and as a nuclear-powered battery in space applications. *See* RARE-EARTH ELEMENTS. Frank H. Spedding

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; K. A. Gschneidner Jr., J.-C. Bünzli, and V. K. Pecharsky (eds.), *Handbook on the Physics and Chemistry of Rare Earths*, 2005.

Pronghorn

An antelopelike animal, *Antilocapra americana*, the sole representative of the family Antilocapridae, and of uncertain taxonomic affinities (see **illustration**). This animal is reputed to be the fastest ungulate in North America, sprinting as fast as 60 mi/h (96 km/h) and sustaining speeds of 30 mi/h (48 km/h) for about 15 mi (24 km). It is the only hollow-horned ungulate with branched horns present in both sexes. Like the deer, the pronghorn sheds its horns each fall; the new growth is complete by midsummer. The adult is about 3 ft (0.9 m) high, weighs about 100 lb (45 kg), is covered with coarse brittle hair, and has 32 narrow teeth with the dental formula I 0/3 C 0/1 Pm 3/3 M 3/3. *See* ANTELOPE; DENTITION.

The pronghorn lives in small herds in rather wild, rocky desert country. It blends well into the environment because of its cream-colored fur. Cartilaginous pads on the feet allow the animal to travel quickly



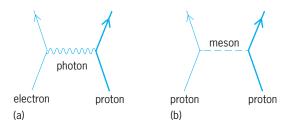
Pronghorn (Antilocapra americana), found in western North America. (Photo by James C. Leupold/U.S. Fish and Wildlife Service)

and quietly on the rocky terrain. Pronghorns feed on cactus, sagebrush, and other vegetation. In late summer, a buck begins to accumulate a harem of about 10–15 does. Two young are usually born in the spring after a gestation period of 35 weeks. Average lifespan for a pronghorn is about 8 years. *See* ARTIODACTYLA. Charles B. Curtin

Bibliography. R. M. Nowak, *Walker's Mammals of the World*, Johns Hopkins University Press, 1999; D. E. Wilson and S. Ruff, *The Smithsonian Book of North American Mammals*, 1999.

Propagator (field theory)

The probability amplitude for a particle to move or propagate to some new point of space and time when its amplitude at some point of origination is known. The propagator occurs as an important part of the probability in reactions and interactions in all branches of modern physics. Its properties are best described in the framework of quantum field theory for relativistic particles, where it is written in terms of energy and momentum. Concrete examples for electron-proton and proton-proton scattering are provided by the Feynman diagrams in the illustration. The amplitude for these processes contains the propagators for the exchanged photon and meson, which actually specify the dominant part of the probability of each process when the scattering occurs at small angles. In similar fashion, for any electromagnetic process, a propagator for each internal line of the Feynman diagram (each line not connected directly to the outside world) enters



Feynman diagrams for scattering processes. (a) Electron-proton scattering via photon exchange. (b) Proton-proton scattering via meson exchange. the probability amplitude. *See* FEYNMAN DIA-GRAM; PROBABILITY (PHYSICS); QUANTUM ELECTRO-DYNAMICS; QUANTUM FIELD THEORY; QUANTUM MECHANICS.

Given the wave equation for a particle or field, the corresponding Green's function is obtained by solving the wave equation when a point source for the particle or field is turned on for an instant. This Green's function solution is the probability amplitude or propagator for the particle or wave to propagate from the source to the point of observation. This procedure can readily be applied to the Schrödinger equation in nonrelativistic quantum mechanics, to Maxwell's equations for the electromagnetic field, to the meson field described by the Klein-Gordon equation in Yukawa theory, to the relativistic electron as described by the Dirac equation, and so on. The point source acting for an instant is described by the Dirac delta function $\delta(x' - x) \,\delta(y' - y)$ $\delta(z' - z) \delta(t' - t)$, abbreviated as $\delta^3(\mathbf{x}' - \mathbf{x})$ $\delta(t' - t)$, which says that a source of unit strength at (x, y, z) in three-dimensional space is turned on at the instant t. See GREEN'S FUNCTION; MAXWELL'S EQUA-TIONS; NONRELATIVISTIC QUANTUM THEORY; RELA-TIVISTIC QUANTUM THEORY; WAVE EQUATION.

An illustration of this calculational procedure can be given by use of the relativistic version of the Schrödinger equation for a free particle, Eq. (1), and its Fourier transformation, Eq. (2).

$$-\hbar^2 \frac{\partial^2}{\partial t^2} \phi(\mathbf{x}, t) = (-\hbar^2 c^2 \nabla^2 + m^2 c^4) \phi(\mathbf{x}, t) \quad (1)$$

$$-\hbar^2 \frac{\partial^2}{\partial t^2} \tilde{\phi}(\mathbf{p}, t) = (\mathbf{p}^2 c^2 + m^2 c^4) \tilde{\phi}(\mathbf{p}, t) \quad (2)$$

The point source delta function can be inserted in Eq. (1) and $\phi(\mathbf{x},t)$ relabeled as $G(\mathbf{x}' - \mathbf{x}, t' - t)$, yielding Eq. (3). The result for each space Fourier component, from Eq. (2), is given by Eq. (4). When the corresponding time-energy Fourier components are inserted, this equation yields Eq. (5). The propagator in momentum space is formally obtained by division of Eq. (5), to yield Eq. (6), where, for illus. (*b*), *E*

$$\left[-\hbar^2 \frac{\partial}{\partial t^2} + \hbar^2 c^2 \nabla^2 - m^2 c^4\right] G(\mathbf{x}' - \mathbf{x}, t' - t)$$
$$= \frac{\delta^3 (\mathbf{x}' - \mathbf{x}) \delta(t' - t)}{2} (2)$$

$$\left(-\hbar^2 \frac{\partial^2}{\partial t^2} - \mathbf{p}^2 \mathbf{c}^2 - m^2 \mathbf{c}^4\right) \tilde{G}(\mathbf{p}, t' - t) = \delta(t' - t)$$
(4)

$$(E^{2} - \mathbf{p}^{2}c^{2} - m^{2}c^{4})\tilde{G}(\mathbf{p}, E) = 1$$
 (5)

$$\tilde{G}(\mathbf{p}, E) = \frac{1}{(E^2 - \mathbf{p}^2 c^2 - m^2 c^4)}$$
(6)

and **p** are the energy and momentum carried by the exchanged meson. The quantity $G\dot{w}$ is defined when this denominator goes to zero by specifying boundary conditions which *G* must satisfy; for example, for the retarded propagator, G = 0 if *t* is less than *t*. Kenneth E. Lassila

Bibliography. J. D. Bjorken and S. Drell, *Relativistic Quantum Fields*, 1965; J. D. Bjorken and S. Drell, *Relativistic Quantum Mechanics*, 1964; C. Itzykson and J. B. Zuber, *Quantum Field Theory*, 1980; M. Kaku, *Quantum Field Theory: A Modern Introduction*, 1993.

Propellant

Usually, a combustible substance that produces heat and supplies ejection particles, as in a rocket engine. A propellant is both a source of energy and a working substance; a fuel is chiefly a source of energy, and a working substance is chiefly a means for expending energy. Because the distinction is more decisive in rocket engines, the term propellant is used primarily to describe chemicals carried by rockets for propulsive purposes. *See* AIRCRAFT FUEL; THERMODYNAMIC CYCLE.

Propellants are classified as liquid or as solid. Even if a propellant is burned as a gas, it may be carried under pressure as a cryogenic liquid to save space. For example, liquid oxygen and liquid hydrogen are important high-energy liquid bipropellants. Liquid propellants are carried in compartments separate from the combustion chamber; solid propellants are carried in the combustion chamber. The two types of propellants lead to significant differences in engine structure and thrust control. For comparison, the effectiveness of either type of propellant is stated in terms of specific impulse. *See* ROCKET PROPULSION; SPECIFIC IMPULSE; THRUST.

Liquid Propellants

A liquid propellant releases energy by chemical action to supply motive power for jet propulsion. The three principal types of propellants are monopropellant, bipropellant, and hybrid propellant. Monopropellants are single liquids, either compounds or solutions. Bipropellants consist of fuel and oxidizer carried separately in the vehicle and brought together in the engine. Air-breathing engines carry only fuel and use atmospheric oxygen for combustion. Hybrid propellants use a combination of liquid and solid materials to provide propulsion energy and working substance. Typical liquid propellants are listed in the **table**. Physical properties at temperatures from storage to combustion are important. These properties include melting point, boiling point, density, and viscosity. *See* METAL-BASE FUEL.

The availability of large quantities and their high performance led to selection of liquefied gases such as oxygen for early liquid-propellant rocket vehicles. Liquids of higher density with low vapor pressure (see table) are advantageous for the practical requirements of rocket operation under ordinary handling conditions. Such liquids can be retained in rockets for long periods ready for use and are convenient for vehicles that are to be used several times. The high impulse of the cryogenic systems is desirable for rocket flights demanding maximum capabilities, however, such as the exploration of space or the transportation of great weights for long distances.

Performance. Jet propulsion by a reaction engine, using the momentum of the propellant combustion products ejected from the engine, is not limited to atmospheric operation if the fuel reacts with an oxidizer carried with the engine. Performance of the propellant in such an engine depends upon both the heat liberated and the propellant reaction products. Combustion with air of effective fuels for air-breathing engines gives approximately 18,000 Btu/lb (42 megajoules/kg), whereas fuels which are more effective in rocket engines may give only 15,000 Btu/lb (35 MJ/kg). A high heat of reaction is most effective with gaseous products which are of low molecular weight.

Performance is rated in terms of specific impulse (occasionally specific thrust), the thrust obtained per pound of propellant used in 1 s. An alternate measure of performance is the characteristic exhaust velocity. The theoretical characteristic exhaust velocity is determined by the thermodynamic properties of the propellant reaction and its products. Unlike the specific impulse, the characteristic exhaust velocity is independent of pressure, except for second-order

Propellant	Boiling point, °F (°C)	Freezing point, ° F (°C)	Density, g/ml	Specific impulse,* s
Monopropellants				
Acetylene	-119 (-84)	-115 (-82)	0.62	265
Hydrazine	236 (113)	35 (2)	1.01	194
Ethylene oxide	52 (11)	-168 (-111)	0.88	192
Hydrogen peroxide	288 (142)	13 (-11)	1.39	170
Bipropellants				
Hydrogen	-423 (-253)	-433 (-259)	0.07	
Hydrogen-fluorine	-306 (-188)	-360 (-218)	1.54	410
Hydrogen-oxygen	-297 (-183)	-362 (-219)	1.14	390
Nitrogen tetroxide	70 (21)	12 (-11)	1.49	
Nitrogen-tetroxide-hydrazine	236 (113)	35 (2)	1.01	290
Red nitric acid	104 (40)	-80 (-62)	1.58	
Red fuming nitric acid-uns-dimethyl hydrazine	146 (63)	-71 (-57)	0.78	275

effects, such as reactions modifying the heat capacity ratio of the combustion gases.

The relationship between these parameters is given by Eq. (1), in which I_s is specific impulse in

$$I_s = \frac{F}{\dot{w}} = \frac{c^* C_F}{g} \tag{1}$$

seconds, *F* is thrust, and w⁻ is flow rate of propellant. The characteristic exhaust velocity c^* is given in feet per second. C_F is the thrust coefficient, and *g* is the gravitational constant.

The actual exhaust velocity of the combustion gases is given by the product of the characteristic exhaust velocity and the thrust coefficient, c^*C_F . The thrust coefficient is a function of the heat capacity ratio of the combustion gases (the ratio of the heat capacity at constant pressure to the heat capacity at constant volume) and of the ratio of the chamber pressure to the exhaust pressure. The heat-capacity ratio of common propulsion gases varies from 1.1 to 1.4.

Increase in the combustion-chamber pressure increases the specific impulse (**Fig. 1**). Variation in the stoichiometry of the propellant reaction (the oxidizer-fuel ratio) also affects performance. A slightly fuel-rich reaction gives higher performance with common liquid propellants despite the lower heat of reaction because of more favorable workinggas composition. Increase in chamber pressure usually moves the optimum performance point toward the stoichiometric reaction ratio. A nonstoichiometric ratio may be used to give low combustion temperatures if required by the structural materials.

A properly designed engine can give 95-100% of the theoretical performance shown in the table.

Combustion. The energy of liquid propellants is released in combustion reactions which also produce the working fluid for reaction propulsion. The liquids in a bipropellant system may ignite spontaneously on contact, or they may require an ignition device to raise them to ignition temperature. In the first case they are called hypergolic liquids; in the second case, anergolic liquids. Combustion can be initiated with a spark, a hot wire, or an auxiliary hypergolic

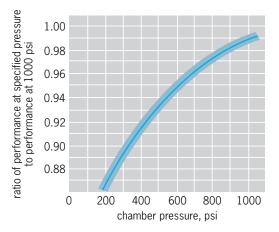


Fig. 1. Approximate effect of chamber pressure on specific impulse. 1 $\ensuremath{\mathsf{psi}}=6.9\ensuremath{\,\mathsf{kPa}}$.

liquid. Monopropellant combustion, or more properly decomposition, can also be ignited by catalysis with an active surface or by a chemical compound in solution. Ignition of common hypergolic bipropellants occurs in a period of 1–100 ms following initial contact of the liquids. Catalytic quantities of detergents or of certain compounds of metals with several oxidation states, such as vanadium pentoxide, decrease the ignition delay period of specific bipropellants.

The combustion chamber in operation contains a turbulent, heterogeneous, high-temperaturereaction mixture. The liquids burn with droplets of various sizes in close proximity and traveling at high velocity. Larger masses of liquid may be present, particularly at the chamber walls. Very high rates of heat release, of the order of 10^5-10^6 Btu/(min)(ft³)(atm) [10^3-10^4 joules/(s)(m³)(pascal)] are encountered.

Oscillations with frequencies of 25–10,000 Hz or more may accompany combustion of liquids in jetpropulsion engines. Low-frequency instability (chugging) can result from oscillations coupling the liquid flowing into the combustion chamber with pressure pulses in the chamber. Higher frequencies (screaming) can result from gas oscillations of the acoustic type in the chamber itself.

Engine performance, in contrast to theoretical propellant performance, depends upon effective combustor design. Mixing and atomization are essential factors in injection of the propellants into the combustion chamber. Injector and chamber design influence the flow pattern of both liquid and gases in the chamber. The characteristic chamber length L^{*} is given by Eq. (2), in which V_c is the chamber vol-

$$L^* = \frac{V_c}{A_T} \tag{2}$$

ume and A_T is the area of the nozzle throat. In general, monopropellants require larger L^* than bipropellants to provide an equal fraction of theoretical performance in a rocket engine, as expected from the slower combustion exhibited by monopropellants. Stanley Singer

Solid Propellants

A solid propellant is a mixture of oxidizing and reducing materials that can coexist in the solid state at ordinary temperatures. When ignited, a propellant burns and generates hot gas. Although gun powders are sometimes called propellants, the term solid propellant ordinarily refers to materials used to furnish energy for rocket propulsion.

Composition. A solid propellant normally contains three essential components: oxidizer, fuel, and additives. Oxidizers commonly used in solid propellants are ammonium and potassium perchlorates, ammonium and potassium nitrates, and various organic nitrates, such as glyceryl trinitrate (nitroglycerin). Common fuels are hydrocarbons or hydrocarbon derivatives, such as synthetic rubbers, synthetic resins, and cellulose or cellulose derivatives. The additives, usually present in small amounts, are chosen

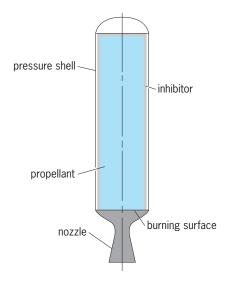


Fig. 2. End-burning grain loaded in rocket.

from a wide variety of materials and serve a variety of purposes. Catalysts or suppressors are used to increase or decrease the rate of burning; ballistic modifiers may be used for a variety of reasons, as to provide less change in burning rate with pressure (platinizing agent); stabilizers may be used to slow down undesirable changes that may occur in long-term storage.

Solid propellants are classified as composite or double base. The composite types consist of an oxidizer of inorganic salt in a matrix of organic fuels, such as ammonium perchlorate suspended in a synthetic rubber. The double-base types are usually highstrength, high-modulus gels of cellulose nitrate (guncotton) in glyceryl trinitrate or a similar solvent.

Propellants are processed by extrusion or casting techniques into what are often intricate shapes that are commonly called grains, even though they may weigh many tons. The double-base types and certain high-modulus composites are processed into grains by casting or extrusion, and are then loaded by insertion of the cartridgelike grain or grains into the combustion chamber of the rocket. This technique requires some type of mechanical support to hold the propellant in place in the chamber. Certain types of composite propellants, bonded by elastomeric fuels, can be cast directly into the chamber, where the binder cures to a rubber and the grain is then supported by adhesion to the walls. Most highperformance rockets are made by this case-bonding technique, permitting more efficient use of weight and combustion-chamber volume.

Burning rate. The thrust-time characteristic of a solid-propellant rocket is controlled by the geometric shape of the grain. Often it is desired that burning not take place on certain portions of the grain. Such surfaces are then covered with an inert material called an inhibitor or restrictor. Neutral-burning grains maintain a constant surface during burning and produce a constant thrust. Progressive burning grains increase in surface and give an increas-

ing thrust with time. Degressive or regressive grains burn with decreasing surface and give a decreasing thrust.

An end-burning grain is shown in **Fig. 2**. This type of configuration is neutral, because the surface stays constant while the grain burns forward. For most applications, radial-burning charges which burn outward from the inside perforation are superior because most of the wall area of the chamber can be protected from hot gas generated by combustion. Such protection is a built-in feature of the case-bonded grain; with the cartridge-loaded, inhibited charge, protection is provided by the addition of obturators to prevent gas flow around the outside of the grain.

Figure 3 shows a progressive design called an internal-burning cylinder. Various star-shaped perforations can be used to give neutral or degressive characteristics. By ingenious use of geometry, the thrust-time characteristic can be designed to meet almost any need. Another important neutral-grain design, the uninhibited, internal-external-burning cylinder, is used widely in short-duration applications such as the bazooka rocket weapon, which contains many such grains.

Propellant burns at a rate r proportional to a fractional power n of the pressure P as expressed by Eq. (3) in which K_1 is the coefficient

$$Y = K_1 P^n \tag{3}$$

of proportionality. This rate may be determined at various pressures by measurements of the burning rate of propellant strands in a strand burner (**Fig. 4**). If the propellant is to operate properly, the exponent *n* in the burning-rate equation must be less than 1. As illustrated by **Fig. 5***a*, if n < 1, there is a stable operating pressure at which the lines of fluid generation and fluid discharge intersect. Pressure cannot rise above this value because gas would then be discharged at a faster rate than

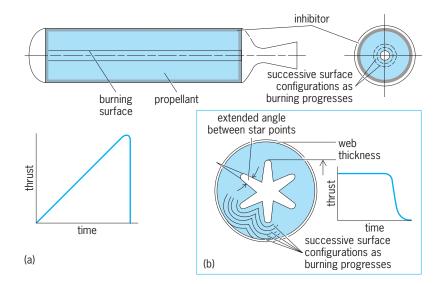


Fig. 3. Internal-burning solid-propellant charge configurations with typical thrust-time characteristics. (a) Cylindrical cavity. (b) Star-shaped cavity for level, or neutral, thrust-time characteristic.

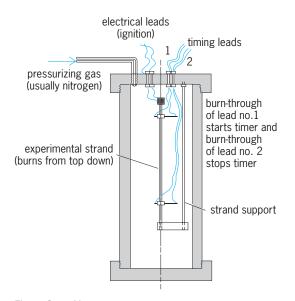


Fig. 4. Strand burner apparatus.

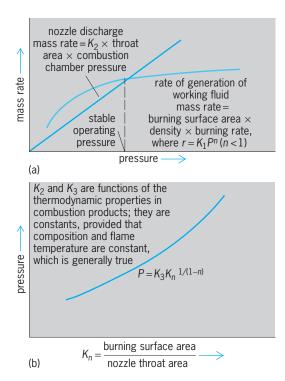


Fig. 5. Solid propellant. (a) Stable burning condition. (b) Typical pressure characteristic.

it is generated by burning of propellant. When the propellant meets the requirement n < 1, a relationship exists between operating chamber pressure and a design parameter known as K_n , which is the ratio of the propellant burning area to the nozzle throat area. This ratio is illustrated in Fig. 5*b*.

Specific impulse of solid propellants is normally rated with the rocket operating at a chamber pressure of 1000 lb/in.² (6.9 megapascals) and exhausting through an optimum nozzle into sea-level atmosphere. Under these conditions, solid propellants in use today can give an impulse of about 250, which is near the ceiling imposed by compositions based

on ammonium perchlorate and hydrocarbons, and is 5-10% lower than impulses obtainable from liquid oxygen and gasoline.

The lower specific impulse of solid propellants is partly overcome by their densities, which are higher than those of most liquid propellants. In addition, solid-propellant rockets are easy to launch, are instantly ready, and have demonstrated a high degree of reliability. Because they can be produced by a process much like the casting of concrete, there seems to be no practical limit to the size of a solid-propellant rocket. H. W. Ritchey Divergenture to be a be a be a better

Bibliography. A. Davenas (ed.), Solid Rocket Propulsion Technology, 1992; L. DeLuca and M. Summerfield (eds.), Nonsteady Burning and Combustion Stability of Solid Propellants, 1992; P. G. Hill and C. R. Peterson, Mechanics and Thermodynamics of Propulsion, 2d ed., 1992; G. A. Olah and D. R. Squire (eds.), Chemistry of Energetic Materials, 1997; G. P. Sutton and D. M. Ross, Rocket Propulsion Elements: An Introduction to the Engineering of Rockets, 6th ed., 1992; M. Timnat (ed.), Advanced Chemical Rocket Propulsion, 1987.

Propeller (aircraft)

A hub-and-multiblade device for changing rotational power of an aircraft engine into thrust power for the purpose of propelling an aircraft through the air (**Fig. 1**). An air propeller operates in a relatively thin medium compared to a marine propeller, and is therefore characterized by a relatively large diameter and a fairly high rotational speed. It is usually mounted directly on the engine drive shaft in front of or behind the engine housing. *See* PROPELLER (MA-RINE CRAFT).

Propellers range in size from diameters of 15 cm (6 in.) for model airplanes up to 5.8 m (19 ft) for

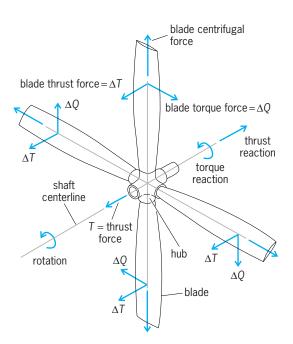


Fig. 1. Typical four-bladed propeller system.

some post-World War II bomber-type airplanes. Usually propellers have two, three, or four blades. For high-speed or high-powered airplanes, six or more blades are used. In some cases these propellers have an equal number of opposite rotating blades on the same shaft, and are known as dual-rotation propellers. The engine power converted to thrust power by propellers ranges from a fraction of a horsepower to values in excess of 10,000 hp (1 hp = 746 W). The usual operating forward speed of propellers ranges from 0 to 450 knots (230 m/s). The rotational tip speed is in the range of 198 to 290 m/s (650 to 950 ft/s), corresponding to a Mach number range of 0.58 to 0.85. For low propeller noise output, it is important to operate at low values of rotational tip speed.

Until the development of turbine-powered straight-jet or turbofan engines, the propeller was the only efficient device available for propelling an airplane through the air. Except for a few special-purpose military and cargo airplanes, the propeller is used mainly on general aviation airplanes having reciprocating and turbo shaft engines with power ratings of 75 to 1500 hp (55 to 1100 W). *See* GENERAL AVIATION.

The performance of a propeller is measured in terms of its efficiency η , which is the ratio of the power output over the power input. Since the power output is equal to the thrust force of the propeller times the speed *V* of the airplane, the efficiency is given by Eq. (1). The efficiency of well-designed pro-

$$\eta = \frac{\text{power out}}{\text{power in}} = \frac{\text{thrust} \cdot \text{velocity}}{\text{power}} = \frac{TV}{P} \quad (1)$$

pellers is 85-90% at cruise flight conditions.

Principles of operation. A propeller blade advances through the air along an approximate helical path which is the result of its forward and rotational velocity components (Fig. 2). This reaction is similar to a screw being turned in a solid surface, except that in the case of the propeller a slippage occurs because air is a fluid. Because of the similarity to the action of a screw, a propeller is also known as an airscrew. To rotate the propeller blade, the engine exerts a torque force. This force is reacted by the blade in terms of lift and drag force components produced by the blade sections in the opposite direction (Fig. 2). As a result of the rotational forces reacting on the air, a rotational velocity remains in the propeller wake with the same rotational direction as the propeller. This rotational velocity times the mass of the air is proportional to the power input. The sum of all the lift and drag components of the blade sections in the direction of flight is equal to the thrust produced. These forces react on the air, giving an axial velocity component opposite to the direction of flight. By the momentum theory, this velocity times the mass of the air going through the propeller is equal to the thrust

The axial momentum efficiency of a propeller η_m is inversely proportional to the axial velocity component v in the final wake. At an aircraft velocity *V*

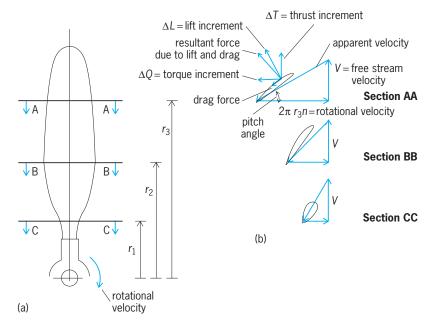


Fig. 2. Propeller blades with velocity and force components. (a) Front view. (b) Cross sections. Here n = angular rotational velocity, in rpm.

the axial momentum efficiency is given by Eq. (2).

$$\eta_m = \frac{2V}{2V + v} \tag{2}$$

Thus, for high efficiency, the slip velocity in the final propeller wake should be small. Since the thrust is equal to the change in axial momentum, mv, it is apparent that for high efficiency the mass of air, m, should be as high as possible. This implies that, for a given power input, high efficiency is obtained with the highest-diameter propeller possible and is a typical propeller design characteristic. In a jet engine the thrust is produced with a low mass flow and a high wake or slip velocity, thus giving a lower efficiency.

To develop lift, and therefore thrust, on a blade, airfoil sections similar to those used on a wing are placed along the blade span from the hub to the tip. These sections have a blade angle distribution that decreases from hub to tip approximately as the pitch of the helix of the screw (Fig. 2). The size of the airfoil and the number of blades on the propeller are chosen so that the angle of attack of the section is low, and so that the sections are operating with the lift required to absorb the given power input. When the airfoils are operating at low angles of attack relative to the apparent velocity vector (Fig. 2), the high lift drag ratios needed for high efficiency are obtained. As with a wing, the drag of the airfoil depends on the operating lift coefficient, section type, and speed relative to the speed of sound (the Mach number). When the Mach number is high, large losses can be encountered that result in low values of efficiency. Propellers with swept-back blades, like wings, to reduce the compressibility losses may be used to obtain high efficiency at high flight speeds. See AIRFOIL; WING.

Propeller blades. A propeller blade is essentially a twisted wing with a radius six to seven times its

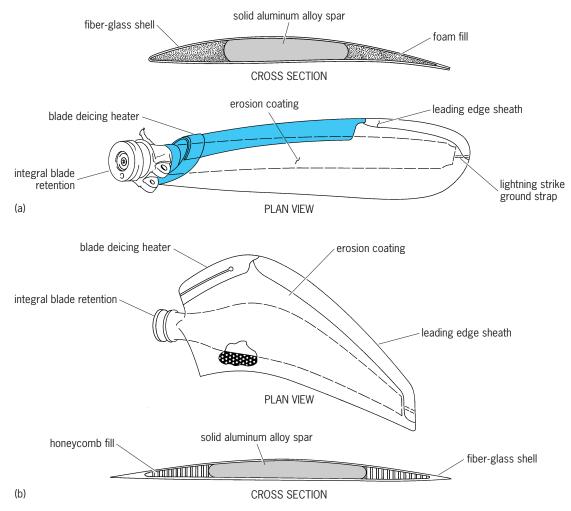


Fig. 3. Composite propeller blade construction. (a) Fiber-glass shell with solid aluminum alloy spar and foam fill. (b) Fiber-glass shell with solid aluminum alloy spar and honeycomb fill.

chord. The blade (Fig. 2 and **Fig. 3**) uses airfoil sections that have a thickness-to-chord ratio of about 4% at the tip and 25% at its root. For best efficiency, the airfoil sections operate at a small angle of attack, producing lift and drag forces.

A propeller blade must be designed to withstand very high centrifugal forces, which increase from tip to root and can be of the order of 25 tons (222 kilonewtons) where the blade attaches to the hub. The blade also must withstand the thrust force produced plus any vibratory forces generated, such as those due to uneven flow fields. Such flow fields exist when the propeller shaft axis is at an angle of attack relative to the flight direction. Vibratory forces are also transmitted to a propeller by the gas pressure forces generated by reciprocating engines and by the flow field produced by the various components of the airplane. The vibratory forces can be especially high for pusher propeller installations.

To withstand the high stresses due to rotation, propeller blades such as those in Fig. 3 have been made from a number of materials, including wood, aluminum, hollow steel, plastic composites, and the construction illustrated. The most common material used has been solid aluminum. However, the composite blade construction illustrated, plus blades having a fiber-glass shell with a foam filler, are being used for new turboprop installations because of their very light weight and high strength characteristics.

Propeller types. For a small, low-power airplane, very simple, fixed-pitch, single-piece, two-blade propellers are used. The rotational speed of these propellers depends directly on the power input and forward speed of the airplane. Because of the fixed-blade angle of this type of propeller, it operates near peak efficiency only at one condition and will limit the engine rotational speed at the low forward speeds of the aircraft. Usually such a propeller is designed with a blade set for good climb or cruise performance.

To overcome the limitations of the simple fixedpitch propeller, configurations that provide for variable blade angles are used. The blades of these propellers are retained in their hub so that they can be rotated about their centerline while the propeller rotates. For the normal range of operation, the blade angle varies from the low blade angle needed for takeoff to the high blade angle needed for the maximum speed of the airplane. The lowest blade angle is that angle which allows the engine to develop full power at the rated engine rpm (revolutions per minute). As the speed increases, the blade angle increases to maintain the desired rpm for good engine performance and to operate at conditions for high efficiency.

The operating speed of the propeller is controlled by the propeller governor, which is set to the desired rpm by the pilot. With an increase in power, the propeller rpm increases; if it exceeds the desired value, the governor causes the blade angle to increase, which causes the propeller to absorb more power and thus decrease the rpm. A decrease in rpm results in a decrease in blade angle, with the opposite effect.

Propellers operating at blade angles between those needed for takeoff and high-speed flight are used on single-engine aircraft. On multiengine airplanes the blade angle range is increased so that the propeller can develop negative thrust for braking during landing. As a safety feature, a full-feathering blade angle is also provided. Here the blade angle is increased to almost 90° so that the propeller will not windmill and its drag will be a minimum when an engine fails.

There have been many different means used for changing the blade angle of a propeller. These include the use of mechanical, electrical, and hydraulic actuators. The hydraulic actuator controlled by an rpm governor is the system most commonly used today. The relatively simple nonfeathering propeller uses hydraulic pressure to actuate a piston that is attached to the blade and increases the blade angle with an increase of pressure. When the pressure on the piston is reduced, the centrifugal twisting moment of the blade is in the direction to reduce blade angle. With a full-feathering and reversing propeller blade, counterweights and a spring are used instead of hydraulic pressure to cause the angle to increase. With this type of propeller (Fig. 4), the hydraulic pressure increase causes the blade angle to be reduced by overcoming the twisting moment of the counterweights and the spring. The governor controls the hydraulic pressure to the actuator, and thus

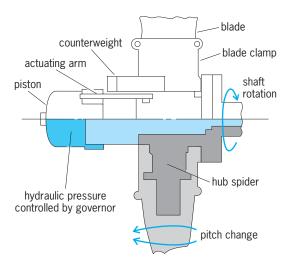


Fig. 4. Actuating mechanism to change propeller pitch.

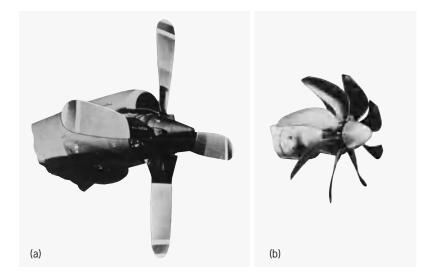


Fig. 5. Size and configuration comparison of (a) conventional turboprop installation and (b) advanced turboprop fan. (*Hamilton Standard Division of United Technologies*)

the blade angle. This system has important inherent safety features since it increases the blade angle in the event of a loss of hydraulic pressure, thus preventing a dangerous overspeeding of the engine and propeller.

With large propellers, the addition of counterweights and a spring results in a large weight increase. To reduce weight and eliminate the counterweights, hydraulic pressure is used to increase and decrease the blade angle of the propeller. With such a system, safety devices must be used to control the system and prevent unwanted changes of blade angle in the event of failures.

A typical four-blade conventional propeller for a high-powered turbo propeller is illustrated in **Fig. 5***a*. This propeller has a power loading of 80 to 120 kW/m² or 10 to 15 SHP/ D^2 , where SHP is the shaft horsepower input to the propeller with a diameter of *D*feet. Compared to the conventional propeller is a propeller fan scaled to reflect the higher power loading of 300 kW/m² (37.5 SHP/ D^2) used for this configuration (Fig. 5*b*). The propeller fan, with light, swept-back blades, is designed for use with 0.8-Mach cruise high-speed airplanes. *See* AIRCRAFT PROPULSION; AIRPLANE; HELICOPTER. Henry V. Borst

Bibliography. F. Delp, *Aircraft Propellers and Controls*, 1979; H. Glauert, *The Elements of Aerofoil and Airscrew Theory*, 1944, reprint 1983; M. Kroes and T. W. Wild, *Aircraft Powerplants*, 7th ed., 1994.

Propeller (marine craft)

The component of a ship propulsion power plant that converts engine torque into propulsive force or thrust, thus overcoming the ship's resistance to forward motion by creating a sternward accelerated column of water. Early steamship propulsion forms, such as the jet propeller (1782) and the paddle wheel (1801), were gradually replaced by the screw propeller (1804), which since 1860 has been the only propeller type used in ocean transport, mainly because of the evolution of the marine engine toward higher rotative speed.

The advantages of the screw propeller include light weight, flexibility of application, good efficiency at high rotative speed, and relative insensitivity to ship motion. The fundamental theory of screw propellers is applicable to all forms of marine propellers. In its current form, a screw propeller consists of a streamlined hub, attached outboard to a rotating engine shaft, on which are mounted two to seven blades. The blades can be integral with the hub, detachable, or movable. The screw propeller indeed has the characteristic motion of a screw: It rotates about the shaft axis as the ship advances through the water. The screw blades are approximately elliptical in outline. Propellers are being developed in which the major axis of the ellipse is an arc, which can give the propeller blades the appearance of a scimitar. One or more propellers are usually fitted as low as possible at the ship's stern to act as thrust-producing devices. The low position of the propellers affords good protection and sufficient immersion during the pitching motions of the ship. The diameter of the propeller usually is not larger than 0.80 of the loaded draft in seagoing single-screw vessels. See SHIP DE-SIGN

Geometry. The side of a propeller blade away from the hull is called the face or pressure side. The opposite side is called the back or suction side. In its simplest form the blade face is part of a helicoid surface, and its axial advance per revolution defines the propeller's pitch which, in the case of a constant-pitch propeller, is constant over the radius, *R*. Many modern propeller designs which are adapted to the complex flow pattern behind the vessel's stern, called the wake, employ a radially varying pitch (variablepitch propeller). The pitch occurring at 0.7R is then taken to be representative for the nominal pitch. The dimensionless ratio of the pitch, *P*, to the propeller diameter, *D*, is the pitch ratio. It is an important characteristic of the screw propeller.

The difference between the pitch velocity, nP (where n is revolutions per second), and the mean axial speed of advance, V_A , of the screw through the surrounding medium is called the slip velocity. Expressed as a fraction of nP, this yields a dimensionless quantity called the real slip ratio, S_R , as given in Eq. (1),which is one of the fundamental quantities

$$S_R = \frac{nP - V_A}{nP} = \frac{1 - V_A}{nP} \tag{1}$$

determining propeller action and efficiency. In propeller technology this relationship is expressed as propeller advance ratio, Eq. (2). The relationship be-

$$J_A = \frac{V_A}{nD} \tag{2}$$

tween real slip ratio and advance ratio is expressed in Eq. (3).

$$J_A = \frac{P}{D}(1 - S_R) \tag{3}$$

Blade shape. The propeller blade outlines can take a number of different forms depending upon the design objective. The majority are essentially elliptical with a small amount of skew. Skew is the lateral shift of the centerline of the blade from a straight reference line radiating from the centerline of the blade at the hub. This skew line can take any continuous shape from the root to the tip of the blade. On a righthand-turning propeller, most of the skew is to the left of the reference line. The magnitude of the skew at the tip is usually defined as the angle in degrees that the tip makes with the reference line at the centerline axis of the propeller. Conventional propellers usually have less than 10° of skew. Propellers with more than 25° of skew are usually considered "highly" skewed and require special care in design to ensure adequate strength in the blades. Figure 1 shows an example of a highly skewed propeller. Propellers with this blade shape produce significantly less propellerinduced blade-frequency vibration. Some propeller blades, such as those used in controllable-pitch propellers with high skew, skew the blade to the right of the reference line in the root region and then heavily skew the blade in the tip region to the left of the reference line. The purpose is to produce a blade that has low spindle torque about the blade axis in order to minimize the blade turning torque and obtain the advantages of the high-skew propeller in reducing the blade-frequency vibratory forces.

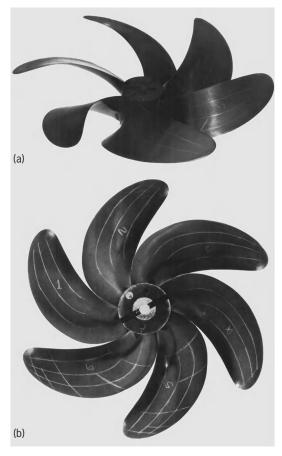


Fig. 1. Highly skewed model propeller. (a) Side view. (b) Front view.

Theory. Propeller theory developed along two main lines: the momentum theory and the circulation theory. In the former, the origin of the thrust is explained by the change in momentum that takes place in the fluid. This theory views the propeller as an actuator disc (infinite number of blades with no thickness) in an ideal fluid, thus offering no indication as to the proper shape of the blades. It does provide useful information on the ideal efficiency, contraction of the slipstream (the bounding cylindrical streamline which contains the flow accelerated by the propeller when creating thrust), and the velocities inside and outside the propeller slipstream.

In order to design the detailed shape of the blades for the desired thrust, it became necessary to utilize a theory based on the lift-drag characteristics of propeller blade sections. Initially, various blade element theories were developed, but these required empirically derived corrections to provide the desired results. These older theories have been superseded by circulation theory, which treats the additional axial and rotational velocities in the slipstream as being induced by vortex systems distributed radially in the lifting line theory and over the whole blade in the lifting surface theory.

The lifting line representation uses a radial lifting line to represent the radial distribution of the propeller blade loading. It provides the overall propeller performance at the design condition. The theory has been extended to represent the loading on the whole surface of the blade, thus providing the detailed shape of the propeller blade sections to create the desired thrust. The theory has been augmented by the panel methods, which use appropriate combinations of sources, sinks, vortices, and dipoles distributed over the blade surface. The panel methods are particularly useful in predicting off-design and cavitation performance. The use of these theories, particularly the panel methods, requires the aid of digital computers but does permit the designer to achieve the desired performance at both the design and off-design conditions.

Model tests. To aid in propeller design, open-water experiments are done in towing tanks using small-scale model propellers. Self-propulsion tests with a corresponding model of the ship establish the interaction of the propeller with the hull. These experiments are used to verify design calculations, particularly the thrust, torque, and revolutions per minute (rpm). In some instances, to assist in vibration analysis, a specialized dynamometer is used to measure the blade-frequency forces generated by the propeller. *See* TOWING TANK.

Number of propellers. The choice of the number of propellers to incorporate into a vessel design is based on several factors. In general, a single-screw arrangement yields a higher propulsive efficiency than multiple screws, particularly when most of the propeller is operating in the boundary layer of the ship and can recover some of the energy loss. In addition, single-screw propulsion systems generally result in savings in machinery cost and weight in comparison to multiple-screw arrangements.

Shaft powers in excess of 90,000 horsepower (67,100 kW) can be successfully transmitted by a single screw propeller in large, fast oceangoing ships. Greater power output necessitates the use of multiple (mostly twin or quadruple) screw arrangements. However, operational considerations may also result in a decision to opt for a multiple-screw system. In some cases, as in powerful shallow-draft towboats, the maximum available propeller diameter limits the efficient conversion of engine power per shaft. In others, preference for a certain number of diesel engine units to produce the propulsive power needed determines the number of propellers. Other considerations, such as safety or operability of the vessel in the event of engine breakdown (especially in the case of large passenger vessels), may favor multiplescrew arrangements.

Number of blades. Both the maximum open-water efficiency and the optimum blade diameter for the highest propulsive efficiency may decrease slightly with a large number of blades. In practice the number of blades varies from two to seven, with four, five, and six the most common. Two-bladed propellers are used in fast racing craft and in sailing vessels with auxiliary engine power where the propeller can be locked in an upright position behind the keel to reduce drag when sailing. In oceangoing ships, the natural frequencies of vibration of the propeller shaft, the hull, and the propulsive plant and their major harmonics in the operating speed range are the governing factors in determining the number of blades. Excitation of hull or propeller shafting resonance at blade frequency should be avoided.

Propeller vibrations. The propeller exciting forces may be of the once-per-revolution frequency or of blade frequency or its multiples. Once-per-revolution forces are produced by mechanical imbalance or inequalities of blade spacing, pitch, or contour, or by blade damage. Blade frequency forces are produced by the interaction of the propeller with the irregular flow pattern in which it operates and with the adjacent hull surface. Vertical and horizontal transverse forces acting at the stern excite vibration in the hull. Torque and thrust variations excite vibration in the propeller shaft and the machinery. A hull stern shape conducive to more uniform flow into the propeller and generous clearance between the propeller and the hull are desirable to reduce vibratory effects.

Under special circumstances, propeller blade vibration can occur that produces an audible noise ("singing propeller"). In most instances, this can be prevented by the application of special chisel-shaped edges to provide fixed flow separation points along the trailing edge of the propeller blades. *See* SHIP POWERING, MANEUVERING, AND SEAKEEPING.

Propeller cavitation. The formation and collapse of vapor-filled bubbles, or cavities, causes noise, vibration, and often rapid erosion of the propeller blade material, especially in fast, high-powered vessels. This phenomenon is known as cavitation. As long as the rotational and translational speeds of the

propeller are not too high, the onset of cavitation can be delayed or limited to an acceptable amount by utilization of special blade sections. The application of circulation theory in design is of great help, but it is highly advisable to have a model of the propeller tested in a variable-pressure water tunnel, water channel, or vacuum tank under simulated operating conditions and at the proper cavitation index to assure good performance. *See* WATER TUNNEL (RE-SEARCH AND TESTING).

To minimize the effects of cavitation, propellers are usually made of a corrosion-resistant zinc-free nickel aluminum bronze. Certain stainless steel alloys are also finding increasing application for heavy-duty propellers. *See* CAVITATION.

Supercavitating and superventilated propellers. Supercavitating and superventilated propellers are designed to have fully developed blade cavities which spring from the leading edge of the blade, cover the entire back of the blade, and collapse well downstream of the blade trailing edge. The blade of such propellers has unique sections which usually are wedge-shaped with a sharp leading edge, blunt trailing edge, and concave face. Supercavitating and superventilating propellers are distinguished by the nature of the gases in the cavity. Supercavitating propellers have cavities filled with water vapor and small amounts of gases dissolved in the fluid medium. Superventilated propellers have cavities filled primarily with air or gases other than water vapor. They may be fully submerged propellers with a gas supply system through the propeller shaft and propeller blades into the cavity, or partially submerged propellers draw-

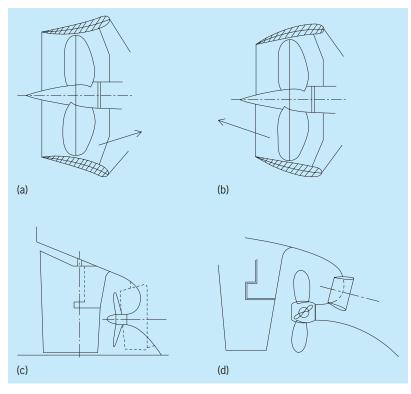


Fig. 2. Ducted propeller arrangements. (a) Accelerating or Kort nozzle. (b) Decelerating nozzle or pumpjet. (c) Mitsui duct ahead of propeller. (d) Wake-equalizing duct.

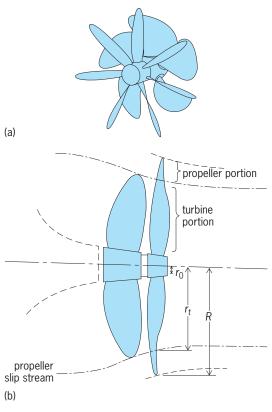


Fig. 3. Typical propeller with vane wheel. (a) Oblique view. (b) Side view.

ing air from the water surface as the blade enters the water.

Both of these propeller types are intended for use in high-speed craft (speeds usually greater than 45 knots or 83 km/h) such as hydrofoil boats, surfaceeffect ships, and the higher-speed planing craft with high propeller-shaft rotational speeds. The application of supercavitating propellers has been very limited. They were applied to a number of high-speed hydrofoil patrol vessels. *See* AIR-CUSHION VEHICLE; HYDROFOIL CRAFT.

The partially submerged propeller is widely used in racing hydroplanes (prop-riders). It has evolved by trial and error, and the blade sections are shaped similar to those developed for the supercavitating propeller. They are very efficient two-bladed propellers with thicker sections than would be needed on a superventilating propeller to sustain the high fatigue loads. Speeds up to 175 knots (320 km/h) have been attained. Interest in the partially submerged propeller is being extended to other vessels, such as surface-effect ships, using more blades and blade sections of the type that were developed for the superventilating propeller.

Controllable-pitch propellers. For ships which normally operate at widely varying speeds and propeller loadings (for example, towboats, rescue vessels, trawlers, and ferryboats), the application of controllable-pitch (rotatable-blade) propellers permits the use of full engine power at rated rpm under all operational conditions, ensuring maximum thrust production and utmost maneuverability. Since these

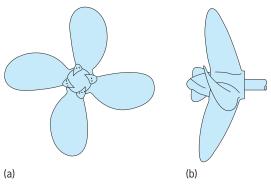


Fig. 4. Propeller with boss cap fins. (a) Front view. (b) Side view.

propellers are also reversible, they permit the use of nonreversible machinery (such as gas turbines). The hydraulic or electric servomotor used to adjust the pitch of the blades requires a hollow tailshaft for its operation. The propeller pitch can be controlled directly from the ship's bridge. In each case the operational advantages of the controllable-pitch propeller must be weighed against the disadvantages of more complex construction, higher manufacturing cost, and the possibility of a small decrease in efficiency due to the larger diameter of the propeller hub and the exposed shafting.

Ducted propellers. It was found in the early 1930s on river towboats that the tow force could be increased up to 40% by enclosing the propeller in a nonrotating, coaxially converging streamlined nozzle which accelerated the flow into the propeller. With time this has spawned a number of applications which use a duct in conjunction with the propeller to improve the propulsive performance.

Accelerating duct (Kort nozzle). This coaxial nozzle around the propeller, named after its inventor, is

shaped such that it accelerates the flow into the propeller (**Fig.** 2*a*). At low speeds and heavy propeller loadings, the accelerated flow increases the propeller thrust, with the duct itself providing an additional component of thrust. Research has shown that the Kort nozzle has wide application whenever the propeller is heavily loaded. It has been installed on all types and sizes of ships, from small coastal fishing trawlers, ice breakers, and tugboats, to very large crude-oil tankers. The Kort nozzle in a tunneled stern is widely used on powerful river towboats of modern design.

Decelerating duct. This coaxial nozzle around the propeller is shaped such that it decelerates the flow into the propeller (Fig. 2b). Decelerating the flow delays the onset of cavitation, thus reducing the radiated noise generated by the propeller. It is intended for use on vessels that need to minimize radiated noise at as high a speed as possible in order to accomplish their mission. These are primarily military vessels engaged in stealth operations and antisubmarine warfare. The propulsive efficiency of this type of duct is lower than that of a conventional propeller.

Mitsui duct. This coaxial nozzle is placed a short distance forward of the propeller so that the propeller no longer operates within the nozzle, thus substantially reducing the complexity of installation (Fig. 2*c*). The duct is shaped to accelerate the flow into the propeller. However, it does not increase the propeller efficiency to the same extent as the Kort nozzle, nor does the nozzle produce as much thrust. In modifying the flow around the stern of full-formed vessels such as tankers and bulk carriers, it tends to reduce the hull resistance. Although the gains in propulsive efficiency are smaller than from an accelerating nozzle, the benefits are such that many large tankers have been converted to this form of propeller propulsion.

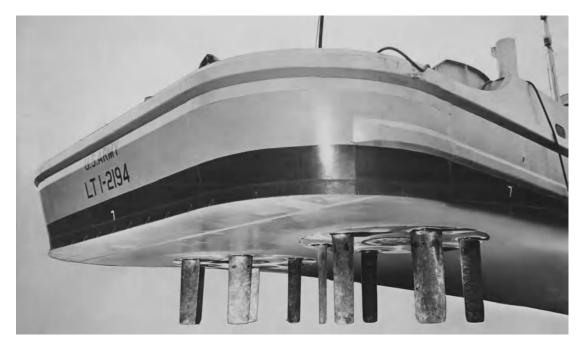


Fig. 5. Twin cycloidal propellers which are rated at 1100 hp (800 kW) each on a U.S. Army towboat. The length of the spadelike vertical blades is 4.5 ft (1.4 m). (*Pacific Car and Foundry Co.*)

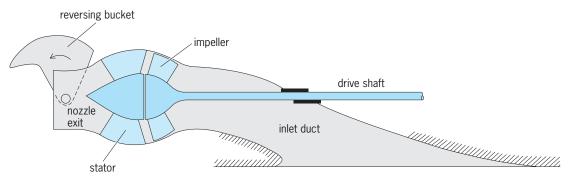


Fig. 6. Typical waterjet arrangement.

Wake equalizing duct (WED). This nozzle consists of two semicircular ducts located on the port and starboard sides of the hull forward of the propeller in the upper part of the propeller disc (Fig. 2d). The nozzle is shaped to accelerate the flow into the upper part of the propeller disc where the velocity behind the hull is lowest. This tends to make the total flow into the propeller more uniform, thereby reducing the magnitude of the propeller-induced vibratory forces and increasing propeller efficiency. Like the Mitsui duct, the WED can modify the flow around the stern of full-formed ships in the region of the duct and reduce the hull resistance a small amount. The WED is a relatively inexpensive device which can be applied during or after ship construction with prospects of a short amortization period. It is being applied to a wide variety of single-screw commercial ships.

Contrarotating propellers. Since the rotational energy in the propeller race represents an energy loss,

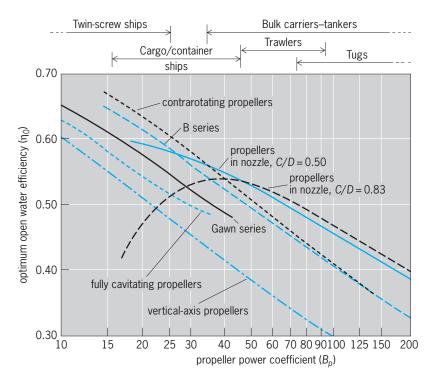


Fig. 7. Typical optimum open-water efficiencies for different propeller types. C = length of duct; D = diameter of propeller. (*After J. D. van Manen, The choice of the propeller, Mar. Technol., April 1966*)

the idea of recovering the lost energy is attractive and has led to the development of contrarotating propellers working on coaxial, contraturning shafts. If the propellers are properly designed, there are minimum rotational losses in the race behind the second propeller. Such propellers have been used for many years on torpedoes, achieving high efficiency, balancing the reaction torque, and preventing the body from rotating. The need to conserve energy has revived interest in this form of propulsion for commercial vessels. Installations on Japanese vessels have shown reductions in shaft horsepower as great as 15%. Although the gains in propulsive efficiency are great, the increased complexity and capital cost of the propulsion system have inhibited its general application.

Propeller and vane wheel. Another approach to recovering some of the rotational energy loss in the propeller race utilizes a freely rotating vane wheel mounted on the same shaft but not mechanically connected to the propeller (**Fig. 3**). The vane wheel is somewhat larger in diameter than the propeller; thus the part of the vane wheel inside the slipstream acts as a turbine and the part outside as a propeller producing additional thrust. The vane wheel usually rotates at about 40% of the speed of rotation of the propeller and generally consists of seven or more relatively narrow blades. Application has been primarily to single-screw commercial vessels with gains in thrust of 5-10% reported.

Propeller boss cap fins (PBCFs). This device is designed to recover some of the rotational energy loss generated at the root of the propeller blades (**Fig. 4**). It consists of fins located on the propeller hub cap equal to the number of propeller blades. In addition to a small increase in propulsive efficiency, the PBCF reduces the strength of the hub vortex, thereby reducing the likelihood of cavitation-induced erosion on the leading edge of the rudder when installed behind the propeller shaft. Gains in propulsive efficiency of 2–8% have been reported. PBCFs have been installed primarily on commercial ships.

Pod propulsion. Pod propulsion is a specialized application of propeller propulsion in which the propeller is mounted on a pod which, in turn, is attached to the hull of the vessel by means of a strut. Marine propellers are usually located at the stern of the vessel and are driven by line shafting directly connected to the main engine or its reduction gears. In the pod

configuration, the propeller is mounted on the trailing or leading end of the pod; in the former case it is usually called a pusher and in the latter a tractor. If the vessel has electric propulsion, the electric motor can be mounted in the pod and connected directly to the propeller. Units with electric motors up to 15 kW are now being manufactured. The pod may be either fixed or trainable in azimuth, thus eliminating the need for a rudder. In the latter configuration the pod propeller provides a very high degree of maneuverability to the vessel. In applications in which the propeller is heavily loaded, a Kort nozzle is added to the configuration.

Cycloidal propeller. This type of propulsor rotates about a vertical axis with the blades rotating in a cycloidal path. The cycloidal propeller was introduced in the 1920s by Kirsten-Boeing and has evolved into the Voith-Schneider design. It consists of a circular disk set flush in the vessel's bottom, carrying near its periphery a number (usually six or eight) of spadelike rotating vertical blades which undergo a cycloidal movement about the vertical axes (Fig. 5). Depending on the phasing of the cycloidal motion, the propeller's thrust can be directed in any direction, thus providing unusual flexibility and maneuverability at low ship speed without need for a rudder. However, the vulnerability of the blades to damage and the complicated mechanism limit application to special inland-waterways craft.

Waterjet propulsion. This type of propulsor is rapidly finding use in special types of craft (**Fig. 6**). It is a reaction propulsor in contrast to the screw propeller and its derivatives, which derive their thrust as lifting devices. In operation, water is drawn from under the craft through an inlet duct by an internal pump, usually an axial-flow pump, where energy is added to the water, which is then expelled at higher velocity through an outlet nozzle in the stern. The thrust is generated by the increase in momentum supplied to the water.

Comparative performance. The performance of some of the propulsors discussed above in comparison to conventional commercial propellers is dependent upon the particular application, how the propulsor interacts with the hull, and how it responds to the flow field created by the hull. Nominal comparisons can be made if consideration is limited to the propulsor efficiency as a singular entity, ignoring the interaction with the hull. Typical comparative propeller efficiencies of the identified propulsors as a function of the propeller power coefficient, B_P , are shown in **Fig. 7**. This coefficient is defined in Eq. (4),

$$B_P = P^{1/2} N / V_A^{2.5} \tag{4}$$

where B_P is the power coefficient, P is the power (ft-lb/s), N is the speed of the propeller (rmp), and V_A is the speed of advance of the propeller (ft/s). The line identified as "B series" in the figure is the conventional merchant ship propeller, in this case having four blades. The line identified as "Gawn series" is a special three-bladed series intended for high-speed vessels.

The supercavitating, superventilating, partially submerged propellers, and the waterjet are omitted from this figure, as their comparitive performance is more a function of speed than of propeller loading. They can outperform the conventional propeller at speeds exceeding 30–35 knots (55– 65 km/h). *See* MARINE ENGINE; MARINE MACHINERY. Jacques B. Hadler

Bibliography. R. A. Cumming, W. B. Morgan, and R. J. Boswell, Highly skewed propellers, *Soc. Nav. Arch. Mar. Eng. Trans.*, 80:98-135, 1972; E. V. Lewis (ed.), *Principles of Naval Architecture*, Society of Naval Architects and Marine Engineers, vol. II, pp. 164-240, 1988; E. R. May, *The Theory and Practice of Controllable Pitch Propellers*, 1979; F. R. Young, *Cavitation*, 1989.

Proprioception

The sense of position and movement of the limbs and the sense of muscular tension. The awareness of the orientation of the body in space and the direction, extent, and rate of movement of the limbs depend in part upon information derived from sensory receptors in the joints, tendons, and muscles. Information from these receptors, called proprioceptors, is normally integrated with that arising from vestibular receptors (which signal gravitational acceleration and changes in velocity of movements of the head), as well as from visual, auditory, and tactile receptors. Sensory information from certain proprioceptors, particularly those in muscles and tendons, need not reach consciousness, but can be used by the motor system as feedback to guide postural adjustments and control of well-practiced or semiautomatic movements such as those involved in walking. See MOTOR SYSTEMS.

Sensory discriminations. Position and movement of limbs can be sensed either when the muscles are relaxed and movement is passive or during active muscular contraction (the latter sense called kinesthesis). In the former case, the threshold for sensing small changes in the angle of the joints, as measured experimentally, depends upon the extent and velocity of the change in angle and is considerably better at proximal joints such as hip or shoulder joints than at more distal joints such as those in the fingers. For example, thresholds vary roughly between 0.2° for the hip to 0.7° for the main joint of the big toe when the speed of displacement is 10°/min. If passive movements of a joint are superimposed, by the experimenter, on active movements of the same joint, discriminatory capacity is improved. Human observers can also identify precisely the absolute position of a limb (placed either actively by the subject or passively by an experimenter) by actively matching the position with the corresponding limb on the opposite side of the body. Humans are also accurate in judging the amount of force exerted by their muscles in making a movement or in maintaining the position of a joint against an opposing force. The latter capacity can be assessed from discriminations between the

active forces required to compress springs of different stiffness held between the thumb and finger. In fact, discriminations analagous to these can be made to a lesser degree without any information from proprioceptors, indicating that there is a "sense of effort" by which the motor system sends a copy of information sent to the muscles to higher sensory processing centers. In normal conditions such central feedback enables the nervous system to determine which proprioceptive information resulted from intended movements and which came from stimuli externally applied.

Proprioceptors. Receptors for proprioception are the endings of peripheral nerve fibers within the capsule or ligaments of the joints or within muscle. These endings are associated with specialized end organs such as Pacinian corpuscles, Ruffini's cylinders, and Golgi organs (the latter resembling histologic Golgi structures in the skin), and muscle spindles. *See* CUTANEOUS SENSATION.

Pacinian corpuscles are found mainly in the connective tissue surrounding the joint and are sensitive to vibrations or abrupt changes in position of the joint. Joint receptors, such as the Ruffini and Golgi types, within the capsule are compressed or stretched when the joint moves, and signal the direction and velocity of movement as well as the steady position of the joint. There are also Golgi organs located at the junction between a muscle and its tendon. When the muscle is lengthened and the ends of a muscle are fixed, then active muscle contraction will shorten the contractile part of the muscle, thereby stretching the inelastic tendon fascicles and exciting the Golgi receptors. Thus, these receptors are excited by muscle tension or force.

Two other stretch receptors are located within the muscle spindle, the latter consisting of small (intrafusal) muscle fibers attached at each end to the main (extrafusal) fibers of skeletal muscle. Each spindle is supplied with two types of sensory nerve endings. One type is excited mainly by changes in the length of the muscle, while the other is sensitive not only to changes in length but also to the velocity of muscle elongation. In addition, the muscle spindle is supplied with axons from neurons called gamma motoneurons, excitation of which causes the intrafusal fibers to contract. Whenever the intrafusal fibers are contracted more than the extrafusal fibers, the sensory endings of the spindle are excited. Stretch receptors are included in many reflexes involving the spinal cord and higher centers in the central nervous system. These reflexes have roles in the control of movement and posture. For example, when a static muscular contraction (such as that occurring in the legs during standing) is subjected to an added load, muscle spindles are stretched, resulting in a reflexive increase in contraction of extrafusal fibers (and relaxation of intrafusal fibers) to maintain the load. Activity in the gamma motoneurons then serves to readjust the length of the muscle spindle (via contraction of the intrafusal fibers) to its previous value in order to maintain the muscle spindle's sensitivity to subsequent varying conditions.

One or more types of muscle receptors may also contribute to the sense of movement and position of the joints since this sense is still possible, although impaired, following local anesthetization of joint receptors or after replacement of the joint with a prosthesis. Conversely, the sense of movement and position also remains, albeit impaired, after elimination of sensory signals from muscle receptors.

Cutaneous receptors may provide some relevant information (particularly in signaling that limb movement has occurred), although the results of experiments in which the skin is anesthetized indicate that this information is not essential. Thus, the sense of movement and position of the limbs probably depends upon the central integration of information from several different types of proprioceptors. *See* NERVOUS SYSTEM (VERTEBRATE); SENSATION; SOMESTHESIS. Robert Lamotte

Bibliography. E. Autrum et al. (eds.), *Progress in* Sensory Physiology, vol. 9, 1989; S. I. Fox, *Human* Physiology, 7th ed., 2001; R. F. Schmidt (ed.), *Fun*damentals of Sensory Physiology, 3d ed., 1986.

Propulsion

The process of causing a body to move by exerting a force against it. Propulsion is based on the reaction principle, stated qualitatively in Newton's third law, that for every action there is an equal and opposite reaction. A quantitative description of the propulsive force exerted on a body is given by Newton's second law, which states that the force applied to any body is equal to the rate of change of momentum of that body, and is exerted in the same direction as the momentum change. *See* NEWTON'S LAWS OF MOTION.

In the case of a vehicle moving in a fluid medium, such as an airplane or a ship, the required change in momentum is generally produced by changing the velocity of the fluid (air or water) passing through the propulsion device or engine. In other cases, such as that of a rocket-propelled vehicle, the propulsion system must be capable of operating without the presence of a fluid medium; that is, it must be able to operate in the vacuum of space. The required momentum change is then produced by using up some of the propulsive device's own mass, called the propellant. *See* AERODYNAMIC FORCE; AIRFOIL; FLIGHT CHARACTERISTICS; FLUID MECHANICS; PRO-PELLANT.

Propulsion principle. Any change in momentum, according to Newton's second law, must be exactly equal to the propulsive force exerted on the body. For example, one may examine the simple propulsive means consisting of a hollow sphere with a small hole of area A (**Fig. 1**). If the sphere is filled with gas whose pressure is higher than that of the surrounding medium, the gas will exert pressure on the inside surface of the hollow sphere, with the exception of the hole, through which gas will rush out into the surroundings. Directly opposite the hole is an area A of the sphere on which the compressed gas does

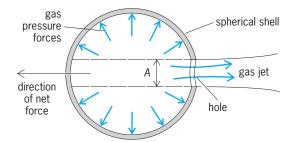


Fig. 1. A shell containing an aperture and saturated with gas at higher pressure than any outside fluid illustrates the principle of propulsion.

exert a force. In all other radial directions inside the sphere, pressure forces balance each other. Consequently, the net force on the sphere is in the direction shown. To compute the size of the unbalanced force, one can either measure and add or subtract all the pressures at the shell surface (including the gas pressure in the hole), or, more simply, in accordance with Newton's second law, calculate the momentum of the escaping gas jet.

This simple propulsive device shows that a rocket gives more thrust in space than in the atmosphere, because atmospheric pressure tends to increase the pressure in the hole, reducing the net force in the forward direction and, of course, reducing by an equal amount the effluent gas momentum. Also, it shows that although the amount of net thrust can be calculated by using the momentum change of the fluid medium, the thrust itself is produced, as it must be, by something pushing against the body— this case, the pressure exerted by a gas. *See* ROCKET PROPUL-SION.

Most propulsion engines are more complicated than the simple spherical rocket discussed here. Consider, for example, an airplane in level flight at constant speed (**Fig. 2**). Within a rectangular region (the control volume) around the airplane, net forces must be in equilibrium. The sum of all momentum changes occurring across the surfaces of the control volume is balanced by the total of all the forces on the airplane, in accordance with Newton's law.

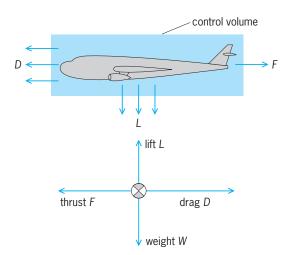


Fig. 2. All forces, including momentum forces, within a control volume about a vehicle are in equilibrium.

Some air passing over the wings is given downward momentum, resulting in upward reaction forces, called lift, on the airplane. This air is also given some forward momentum as it is accelerated downward, resulting in a rearward force on the airplane called induced drag. Viscous drag is also produced (in the rearward direction) because the airplane carries some air forward with it through viscous adherence, giving the air forward momentum. Finally, the airplane propulsive means (such as an engine-driven propeller or a turbojet engine) takes in quiescent air, which is thereby accelerated to the same speed as the airplane, providing a rearward force called ram drag, and then the propulsive means discharges the air at a much higher rearward velocity (with high rearward momentum), resulting in a forward force called gross thrust. If all these forces are balanced, the airplane flies at constant speed in a straight line; if not, the airplane is accelerated in whichever direction the forces are unbalanced, thereby providing another momentum change to again bring net forces within the control volume into equilibrium.

Thrust. Propulsion capability is measured in terms of the thrust delivered to the vehicle. In general, the net thrust delivered by any propulsive means, neglecting the effect of ambient pressure, is given by Eq. (1), where F = net thrust in lb; $V_c =$ velocity of

$$F = (W_a + W_f)V_e - W_aV \tag{1}$$

gas leaving the propulsive means relative to the vehicle in ft/s; V = flight velocity of the vehicle through still air in ft/s; $W_a =$ mass flow rate of the ambient medium (air or water) through the propulsive means in (lb-s²/s)/ft; and $W_f =$ mass flow rate of fuel (or onboard propellant) through the propulsive means in (lb-s²/s)/ft.

The first term is the gross thrust; the second is the ram drag. In a rocket engine, no ambient fluid (air or water) is taken into the propulsive means; thus there is no ram drag. The net thrust of a rocket in vacuum is thus given by Eq. (2), where W_f represents the total

$$F = W_f V_e \tag{2}$$

mass flow rate of on-board propellant. *See* TURBINE ENGINE SUBSYSTEMS.

Efficiency. Propulsion efficiency was formerly defined as the useful thrust power divided by the propulsion power developed. Unfortunately, because useful thrust power can only be determined as the thrust multiplied by the flight velocity (*FV*), and because flight velocity is basically a relative term rather than an absolute term, propulsion efficiency has little real meaning as a measure of propulsion effectiveness.

The two terms most generally used to describe propulsion efficiency are thrust specific fuel consumption (SFC) for engines using the ambient fluid (air or water), and specific impulse (I_{sp}) for engines which carry all propulsive media on board. Thrust specific fuel consumption is given by Eq. (3), where

$$SFC = \frac{W_f g_0}{F} = g_0 \frac{W_f}{W_a} \frac{1}{\left(1 + \frac{W_f}{W_a}\right) V_e - V}$$

$$\approx g_0 \frac{W_f}{W_a} \frac{1}{V_e - V} \tag{3}$$

SFC is expressed in pounds of fuel per second per pound of thrust, g_0 is acceleration due to gravity in ft/s^2 , and, usually $W_f/W_a < 1$.

Specific impulse is determined similarly (although inversely) from Eq. (4). Obviously, effective propul-

$$I_{\rm sp} = \frac{F}{W_f g_0} = \frac{V_e}{g_0} \tag{4}$$

sion performance is indicated by low values of SFC or high values of $I_{sp.}$

The energy source for most propulsion devices is the heat generated by the combustion of exothermic chemical mixtures composed of a fuel and an oxidizer. An air-breathing chemical propulsion system generally uses a hydrocarbon such as coal, oil, gasoline, or kerosine as the fuel, and atmospheric air as the oxidizer. A non-air-breathing engine, such as a rocket, almost always utilizes propellants that also provide the energy source by their own combustion. Here the choice is wider: Fuels may be hydrocarbons again, but may also be more efficient low-molecularweight chemicals such as hydrazine (N₂H₄), ammonia (NH₃), or even liquid hydrogen itself. Oxidizers are usually liquid oxygen, nitrogen tetroxide (N₂O₅), or liquid fluorine.

Where nuclear energy is the source of propulsive power, the heat developed by nuclear fission in a reactor is transferred to a working fluid, which either passes through a turbine to drive the propulsive element such as a propeller, or serves as the propellant itself. Nuclear-powered ships and submarines are accepted forms of transportation. In the case of nuclear-powered vehicles, the concept of specific fuel consumption is no longer valid, because the loss in weight of the nuclear fuel is virtually zero. However, the specific impulse is still a meaningful measure of performance, inasmuch as it is based on thrust and propellant flow rate, not on mass consumption of fuel. See AIRPLANE; SHIP NU-CLEAR PROPULSION; SHIP POWERING, MANEUVERING, AND SEAKEEPING; TURBINE PROPULSION. Jerry Grey

Prosobranchia

The largest and most diverse subclass of the molluscan class Gastropoda. The group includes mostly marine snails but with a few fresh-water and land genera, all retaining an anterior mantle cavity and internal evidence of torsion. Adult prosobranchs always retain the streptoneurous (twisted-commissure) condition of the central nervous system, with the commissures to the visceral ganglia in the characteristic figure-of-eight pattern. This pattern reflects the torsion through 180° during larval (or embryonic) development which has brought the mantle cavity to a position above the head and facing forward. This contrasts with conditions in the other two gastropod subclasses, Opisthobranchia and Pulmonata, in which the effects of torsion are reduced or obscured in adults by secondary processes of development and growth. *See* OPISTHOBRANCHIA; PULMONATA.

The diversity of functional morphology exhibited by the prosobranchs is not equaled by any comparable subclass in the entire animal kingdom. From two-gilled forms with symmetrical cardiac and renal structures (the "Diotocardia") which can be numbered among the most primitive of all living mollusks, evolution within prosobranch stocks has involved increasing asymmetry of pallial, cardiac, and renal systems and greater hydrodynamic efficiencies. Torsion and the anterior mantle cavity create locomotory, circulatory, sanitary, and hydraulic problems which have been solved in a variety of ways in different prosobranchs. Four orders are commonly recognized in subclass Prosobranchia: Archaeogastropoda, Neritacea, Mesogastropoda, and Neogastropoda.

Archaeogastropoda. All primitive two-gilled (or zygobranch) snails belong to the order Archaeogastropoda, as do all forms retaining a right functional kidney or other symmetric renal and cardiac structures. To deal with the problems of sanitation created by the mantle cavity opening anteriorly over the head, various shell slits (*Pleurotomaria*) or serial shell openings (*Haliotis*, abalones, and ormers) are developed. Other archaeogastropods, which have a single aspidobranch gill but still some risk of pallial clogging, include the acmaeid limpets and the elegant top-shells (Trochidae). All of these forms are ecologically limited to clean waters over hard substrata.

Neritacea. The Neritacea also possess a single aspidobranch gill, but have the functional kidney on the left side, thus allowing the genital system (on the right) to develop more elaborate secondary structures. The most primitive neritaceans are littoral snails, largely tropical, living on coral and similar rock surfaces. However, this group has evolved completely independently of (though occasionally parallel to) the rest of the gastropods. The morphological separation of renal function allows the development of a penis and of a uterus capable of retaining large shelled eggs, so that neritaceans have successfully invaded land and fresh waters (the family Helicinidae shows the independent evolution of a lung for air breathing). Neritacea show a patchy zoogeographic distribution; over one-third (about 120 species) of terrestrial neritaceans live in Jamaica, and about 70% of these are endemic to that island.

Mesogastropoda and Neogastropoda. In the two other prosobranch orders, Mesogastropoda and Neogastropoda, there is usually a one-sided combshaped or pectinibranch ctenidium or "half-gill" with its axis fused to the mantle wall, and always a simplified mantle cavity which is hydraulically efficient and from which both feces and nitrogenous excretory wastes are discharged to the right-hand side of the snail. Such snails can live on, and even in, the softer substrata of the sea bottom, and their gill systems can operate in silty waters. There is also a functional separation of genital from renal ducts, allowing the development of internal fertilization, of large eggs, and even of viviparity.

Mesogastropoda. Among the mesogastropods, a wide variety of marine life-styles are found, from that of the slipper limpet, *Crepidula*, which filter-feeds by using the pectinibranch gill (with considerable elongation of individual ctenidial leaflets into filaments), to those of the naticid moon snails like *Polinices* which are predatory carnivores on clams. Another group, *Littorina* and its allies, are among the most successful of seashore animals, with some high-level species becoming physiologically almost terrestrial. Fresh-water mesogastropod genera include *Valvata*, which is unusual in retaining an aspidobranch gill, and *Viviparus*, in which females are ovoviviparous, producing "near-adult" young after a 10-month "gestation."

Neogastropoda. The most highly specialized prosobranch order is the Neogastropoda, comprising marine carnivores and carrion feeders. In addition to the efficient simplified gill and mantle cavity (now with a characteristic inhalant siphon with a chemoreceptive osphradium for detecting prey), there is a simplified gut and a flesh-tearing stenoglossan radula borne on a long extensible proboscis. The order is anatomically rather uniform despite a wide diversity of habitats and a large number of species, including many carnivores which specialize on a single prey species (such as some species of *Conus* which eat flatfish after paralyzing them with a neurotoxin). *See* NEOGASTROPODA.

Evolution. As a group, the prosobranchs constitute the most diverse and largest subclass in the whole phylum Mollusca. The most primitive stem stock of prosobranchs is represented by the extant genus *Pleurotomaria* as a living "archetype" with gills, heart auricles, kidneys, gonads, and pallial sense organs all as paired structures. This neontological statement based on functional anatomy is reinforced by the fossil record: the Pleurotomariacea date from the mid-Cambrian, while fossil siphonate genera (corresponding to the Neogastropoda) appear later and increase markedly in the Mesozoic and Cenozoic. *See* GASTROPODA; LIMPET; MOLLUSCA.

W. D. Russell-Hunter

Bibliography. V. Fretter and A. Graham, *British Prosobranch Molluscs*, 1962; W. D. Russell-Hunter, *A Life of Invertebrates*, 1979; K. M. Wilbur (ed.), *The Mollusca*, vols. 1–8, 1983, 1984.

Prospecting

Exploration for mineral deposits; also referred to as ore search. The result of prospecting is the discovery of potentially economic mineralization, that is, the prospect. Mineral exploration also continues beyond prospecting to include the delineation and evaluation of the prospect to determine its minability as an orebody or economic mineral deposit. A successful prospect is developed into a mine. *See* MINING.

Prospecting generally pertains to the search for deposits of metallic ore minerals, but it also includes the search for nonmetallic or industrial minerals and rocks such as sulfur, potash, and limestone, and mineral fuels such as petroleum, coal, and oil shale.

Prospecting may be carried out by individuals, but is generally done by well-supported multidisciplinary teams or groups of geologists and specialists working for mining and exploration companies or for government mining organizations.

In a successful program of regional prospecting, a dozen or more target areas may be chosen for the final stages of investigation, including drilling. A complete regional prospecting program is expected to take several years, and it requires an expenditure of several million dollars.

With much of the Earth's readily accessible surface having been investigated for minerals, prospecting is increasingly directed toward the discovery of deeper mineralization in recognized mining districts; mineralization hidden beneath overlying rocks, sediments, and soils; and mineralization in the less-known jungle, arctic, and offshore parts of the world. *See* MARINE MINING; OIL AND GAS, OFFSHORE.

Characteristics

The traditional approach to prospecting involves an examination of surface exposures for direct geologic evidence of ore mineralization and a subsequent examination of the subsurface by pitting, drilling, or underground prospect workings. The traditional approach commonly has become integrated with less direct methods that involve projections and interpretations from geologic mapping as well as geochemical, geobotanical, remote-sensing, and geophysical surveys for finding anomalous or extraordinary conditions that can be evaluated as evidence of underlying mineralization.

Much of modern prospecting is done in a regional exploration sequence of programmed stages. The early or strategic stages in regional exploration constitute broad geologic studies and field reconnaissance to identify specifically favorable subregions or districts. Additional reconnaissance takes place in the more favorable terrain, and follow-up or tactical stage prospecting is centered in the target areas recognized as having a high mineral potential. Specific exploration targets are identified; more detailed surface and subsurface prospecting work then provides for the actual discovery of ore mineralization and the designation of a prospect.

Some prospecting programs may take place entirely within a recognized target area, a mining district where current or past mining operations have already demonstrated a geologic potential for additional deep or hidden ore mineralization. This kind of program calls for intensive subsurface investigation through geologic mapping, geochemical sampling, and geophysical surveys in existing mines, accompanied by deep drilling from surface or underground sites, and the driving or sinking of new underground workings.

A central element in all prospecting programs and activities is the acquisition of mineral rights, some form of concession or ownership that will provide for the eventual mining of a discovered orebody. Most governments consider the stagewise sequence in regional exploration and thus provide reconnaissance permits or hunting permits, exploration licenses that convey preference rights within a defined area, and ultimate mining concessions. In the United States, many minerals on federal lands are subject to "location" by the staking of mining claims; other minerals such as coal, petroleum, and certain leasable industrial minerals on federal lands, offshore, and on most state lands are subject to prospecting permits and to mining lease arrangement by competitive bidding. Mining claims that have been patented-that is, purchased from the government-and privately owned mineral rights on older private land in the United States are obtained by direct lease or purchase.

Geologic Guides

Prospecting is carried out on the basis of the guides to ore associated with a conceptual image of the anticipated orebody. The image is referred to as an exploration model, and it is drawn from the characteristics of known orebodies in similar terrain. The exploration model and its guides to ore are expressed in terms of the regional and local geologic pattern; it has a certain diagnostic mineralogical character, it will commonly have a halo or envelope of associated guide minerals, and it will be expected to have a recognizable geochemical and geophysical expression. *See* ORE AND MINERAL DEPOSITS.

An exploration model actually has two appearances, one being the fundamental subsurface appearance of primary mineralization at depths below the influence of weathering, and the other being the near-surface weathered derivative. Except in recently glaciated or actively eroding terrain with a shallow zone of oxidation, it is the weathered derivative that will appear in geologic evidence. Where the targeted mineralization relates to a concealed or blind orebody, interpreted geophysical and geochemical anomalies provide signatures representing the geological condition at depth.

Metallogenic associations. The broadest geologic guidelines are applied to an exploration model during the reconnaissance stage in prospecting; they deal with the probable location of an undiscovered mineral deposit relative to regional trends of known and related deposits. Where a belt or an area of leadzinc or copper-molybdenum deposits, for example, is of a certain geologic age or is associated with a major geologic feature such as the edge of an ancient depositional basin or a zone of tectonic disturbance, a metallogenic trend is indicated. Spatial projections of the trend or other comparable geologic trends in

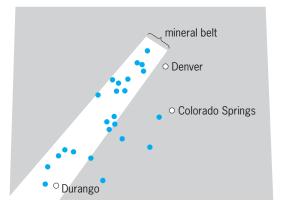


Fig. 1. Major metal mines (color dots) in the Colorado Mineral Belt, constituting a metallogenic guide in regional prospecting.

linear zones may be selected as favorable terrain for prospecting.

The largest of metallogenic trends are along the edges of lithospheric plates where interactions between continental and oceanic crustal segments have provided conditions for the genesis of ore deposits; the copper, iron, and tin belts of the Central Andes are of this kind. More localized metallogenic trends involved in prospecting are represented by the Colorado Mineral Belt (**Fig. 1**), a prominent zone of associated mines and prophyry intrusions; by the Mother Lode of California, a fault zone extending along the west side of the Sierra Nevada; and by the Carlin Trend, an alignment of gold deposits in northeastern Nevada. *See* PLATE TECTONICS.

In further defining favorable areas during prospecting, characteristic associations between certain rock types, geologic ages, and mineral deposits are used as guides. The association may relate to syngenetic ore deposition and processes of volcanism, sedimentation, or igneous intrusion, as with Precambrian sequences of banded iron formation in many parts of the world, the copper-bearing Precambrian ore shale of Zambia, and the layers of ultrabasic rock and chromium ore in the Precambrian Bushveld Complex of South Africa. An association may also pertain to the occurrence of epigenetic orebodies in certain preexisting host rocks, as with the Mississippi Valley-type lead-zinc ore in recognized lithologic sequences of limestones and dolomites; or it may pertain to spatially associated rocks such as the granitic intrusions that accompany many tin, copper, tungsten, and molybdenum deposits. Associations between rocks and mineral deposits have been specified in artificial intelligence models such as the Prospector expert system, and used in assessing the probability of new ore deposits in chosen areas. See EXPERT SYSTEMS.

Structural associations. The relation of ore deposits to structural features such as faults, brecciated areas, unconformities, and the borders of intrusions is used as a guide in prospecting. Mineralizing solutions will have commonly migrated along fractures and will have deposited ore in veins, breccia pipes (**Fig. 2**), stockworks, and contact zones in older

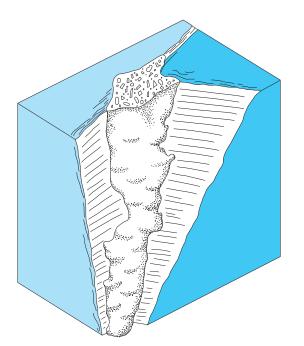


Fig. 2. Breccia pipe orebody at a fault intersection, representing a structural association and a guide in prospecting.

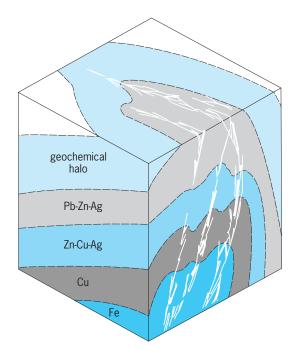


Fig. 3. District metal zonation pattern in a vein complex, constituting a geologic guide to prospecting for deeper ore.

rocks or at exhalative centers on an ancient sea floor. Thus, the overall structural pattern in a region is studied during reconnaissance, and significant centers of volcanism and structural deformation are considered as target areas for follow-up prospecting. During the more detailed stage of prospecting, weakly mineralized or geochemically anomalous structural features and their geophysical patterns are traced toward potentially favorable intersections and lithologic contacts. *See* ORE AND MINERAL DEPOSITS; STRUCTURAL GEOLOGY. **Zoning.** One closely followed guide to ore is the expected pattern of mineral zoning. Ore and gangue minerals change in character with distance from the center of mineralization, and there is commonly an additional outer halo or envelope of alteration minerals and trace amounts of index elements in the enclosing rock. The entire zonal sequence may occur within a distance of a few meters, or it may involve groups of orebodies in a system that extends for several kilometers. Regional zoning may extend even farther, and it may incorporate a series of orebodies that are in themselves zoned.

A single mineral deposit or a group of closely associated orebodies with a central iron (Fe)-rich zone of pyrite, magnetite, and quartz may be surrounded by a more or less complete copper-iron (Cu-Fe) zone of chalcopyrite, pyrite, and quartz and by an outer zinc-copper-silver (Zn-Cu-Ag) zone characterized by sphalerite, tetrahedrite, and ankerite (Fig. 3). In addition, there may be a still more outlying zone of scattered lead-zinc-silver (Pb-Zn-Ag) mineralization with galena and sphalerite as the principal ore minerals and with calcite as the principal gangue mineral. The host rock beyond the outermost lead-silver (Pb-Ag) zone may have a geochemical signature; and it may have a bleached or discolored appearance due to the presence of alteration minerals such as mica and chlorite that can be identified by microscopic study or laboratory analysis. The outlying ore, gangue, and alteration minerals may also have a recognizable geophysical signature. In prospecting, the expected zonal pattern of target rings in an exploration model is used in relating information from widely dispersed outcrops to a more complete system of mineral deposits that may still be hidden.

Outcrops. Gossan, the residue of red, brown, and yellow iron oxides and silica that remains from the weathering and near-surface leaching of sulfide ore minerals, is used in prospecting as a guide to hidden primary mineralization (**Fig. 4**). The color and textural characteristics of gossan are related to the minerals in the orebody from which they were derived; outcrops of gossan are therefore examined for indications of the nature of the underlying deposit.

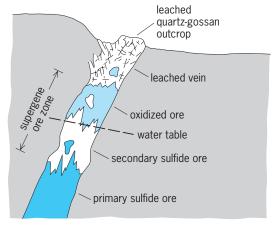


Fig. 4. Leached, supergene, and primary ore in an outcropping copper sulfide–quartz vein.

Due to the weathering of other iron-bearing minerals and due to the general complexity of weathering processes, it is often hard to discriminate between true gossan and other forms of limonitic iron, and it may be difficult to make a precise evaluation of gossan. Even so, the appearance of a cellular and heavily iron-stained outcrop, sometimes with additional staining in green or blue from oxidized copper minerals, black from manganese, or pink from cobalt, is an important factor in selecting targets for details prospecting. *See* WEATHERING PROCESSES.

In some gossans, secondary accumulations of gold and silver minerals may actually be higher-grade than the underlying primary deposit. As a result, some exposures of gossan that were mined for precious metals in earlier times only down to the primary sulfide mineralization can be evaluated again as guides to currently economic ore deposits. Oxidized and secondarily enriched (supergene) ores of copper, lead, nickel, and zinc also accumulate at shallow depth in the zone of weathering; a characteristic gossan remains at the surface.

Prospecting Methods

Both direct and indirect methods of prospecting make use of an appropriate exploration model and its guides to ore. Conventional or direct methods involve the recognition of the guides to ore in their surface expression. Indirect methods make use of geochemical, geobotanical, remote-sensing, and geo-

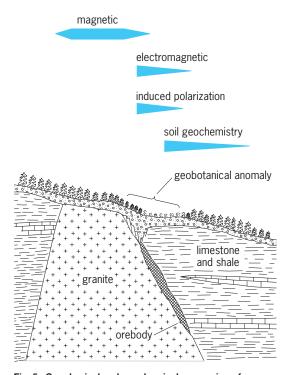


Fig. 5. Geophysical and geochemical expression of a granitic intrusion and a sulfide orebody beneath a cover of glacial sediments and soil, with topographic and geobotanical indications of mineralization. The geobotanical anomaly is revealed in stunted trees, smaller bushes, and grass. Each wedge-shaped symbol indicates one of the other specified anomalies, with the height corresponding to its intensity at each point, and the length, therefore, marking the extent of the anomaly.

physical surveys for the recognition of anomalies, geologic incongruities that can be related to a hidden deposit or to a favorable structural or lithologic condition in the terrain (**Fig. 5**). In a prospecting program, the sets of data acquired by direct and indirect methods are integrated in the development of successively defined target areas. The guiding exploration model is now a complex association in many levels of geomorphology, geology, geochemistry, geophysics, and economics. The spatial data are collected, studied, and interpreted using geographic information systems (GIS), computer-based technologies for managing and visualizing exploration programs. *See* GEOGRAPHIC INFORMATION SYSTEMS.

Direct methods. The topography itself may give evidence of abrupt depressions related to the leaching and collapse of sulfide ore minerals, or it may show boldly exposed silicified zones associated with ore. Some of the latter expressions of ore mineralization represent outcrops of siliceous iron formation host rocks, quartz-filled breccia pipes, and the prospector's classic quartz reefs that indicate vein deposits. Aerial photography and satellite imagery are valuable in searching for the topographic expression of potential ore mineralization; they are also used in the reconnaissance study and mapping of potentially oreassociated structural patterns and in the search for color and texture manifestations of altered areas and gossan.

Outcrops of gossan are examined in the field for evidence of underlying ore mineralization, and trains of float (fragments of ore and gossan) are traced toward their apparent topographic origin. In glaciated terrain, trains of ore boulders are mapped and traced systematically toward their apparent sources; the copper deposit at Outokumpo, Finland, was located in this way from a boulder train 30 mi (50 km) long.

Placer gold and placer accumulations of other minerals such as platinum, cassiterite, rutile, and diamonds are sought as economic deposits in themselves and are used as guides to upstream deposits of associated minerals. In addition, resistant and relatively dense minerals in stream gravels and residual heavy minerals in soil are collected by the long-established prospector's method of panning the loose material, and these are traced to a source area.

Geochemical and geobotanical methods. Geochemical prospecting is based on two characteristics of orebodies: an association with anomalous concentrations of chemical elements within primary halos in the surrounding rock, and an association with secondary dispersal patterns of chemical elements in the surficial products of their weathering and erosion. Geochemical methods involve the field and laboratory analysis of sampled rock, soil, vegetation, and other natural materials for trace amounts (in parts per million or billion) of the principal indicator elements that provide more recognizable or farther-reaching anomalies. *See* GEOCHEMICAL PROSPECTING.

Geochemical sampling. In rock sampling, primary alteration halos and leakage mineralization may be identified for several meters away from the principal mineralization and may be detected for a considerable distance from large ore deposits. In the lead-zinc deposits of Ireland, manganese pathfinder halos extend for as much as 4 mi (7 km), and in the western United States leakage anomalies have been used to locate blind silver-lead-zinc orebodies beneath some 200 ft (60 m) of overlying rock.

The sampling of soil, glacial debris, and stream sediments for the identification of secondary geochemical patterns is a modern correlative to the tracking of visible minerals in gossan, float, ore boulders, and alluvium. Halos or downslope fans of indicator and pathfinder elements occur in residual soil above weathered ore deposits, and longer dispersion trains of trace elements occur in stream and glacial sediments. The discovery of La Caridad, a major copper deposit in Mexico, was aided by the recognition of a diagnostic geochemical train that extended for 12 mi (20 km) in stream sediments.

Lake sediments and water, especially in such terrain as that of the Canadian Shield, reflect the geochemical nature of a local catchment area, and they are sampled in geochemical reconnaissance. Dissolved metal salts and dissolved gases occur in the waters of streams, wells, and springs; these sites are sampled as guides to mineralization in surface- and ground-water drainage systems. Snow and meltwater have also been sampled for entrained mercury vapor and for gaseous complexes given off during the weathering of orebodies, and gases have been sampled directly from the soil and in the atmosphere.

Vegetation sampling. Also known as biogeochemical prospecting, this method is an indirect means of collecting information from deeper soil, rock, and ground water. Roots of trees and bushes accumulate certain elements from the subsurface and transmit them to the foliage; the root systems of phreatophytes in desert areas can reach to ground-water levels at depths of 160 ft (50 m) or more. Samples, commonly twigs or leaves, are taken for analysis from a relatively well-distributed type of vegetation such as pine trees or juniper bushes.

Geobotanical prospecting. In this method the presence of morphological or mutational changes in vegetation is taken as a guide to local geochemical conditions. The Central African copper flower, the European zinc violet, and the selenium-associated poison vetch of the western United States are examples of indicator plants used in prospecting. Areas with significant numbers of indicator plants and areas with dwarfed, discolored, or characteristically metalstressed plant growth can be detected in surface reconnaissance, but their detection is aided considerably by the study of color aerial photographs and aircraft or satellite imagery. *See* GEOCHEMICAL PROS-PECTING.

Remote sensing. Imagery provided by remote sensing from aircraft and orbiting satellites is of fundamental importance in prospecting and in the patterns of exploration data associated with GIS. Photogeology, the interpretation and mapping of terrain from aerial photographs, is in wide use,

especially in reconnaissance as an aid to the direct identification of guides to ore and as a means of broader geologic investigation. Aerial photography in spectral bands of the near-infrared and nearultraviolet frequencies is also used in photogeology for discriminating between types of exposed rock and soil and for emphasizing the appearance of bleached and stained areas as well as geobotanical anomalies. Airborne remote-sensing systems have provided radar imagery of terrain in the prospecting of cloud-covered jungle regions, and they have furnished thermal-band infrared surveys for recognizing anomalously warm areas that may be associated with mineralization. Airborne multispectral sensors with the capability of identifying some of the specific kinds of minerals in altered zones have been tested for use in prospecting. See AERIAL PHOTOGRAPH; RE-MOTE SENSING.

The most widely used satellite imagery in prospecting is from the American *Landsat* series and the French *SPOT* series of spacecraft. The resolution of terrain features is less than that of airborne remote sensing, but the global coverage of major geologic features in multiple passes and seasonal conditions affords an especially suitable base for reconnaissance work in prospecting. Higher-resolution color photography, multispectral imagery, and radar data from newer spacecraft, including Canada's *Radarsat*, Japan's *JERS*, and the European Space Agency's *ERS*, have been applied to reconnaissance prospecting in those regions where coverage is available.

Geophysical methods. Geophysical exploration is based on the measurement of physical properties associated with geologic features. As a means of both airborne and ground prospecting for mineral deposits, it involves the recognition of contrasts in properties between the deposit and the adjacent rock, generally to depths on the order of 330-660 ft (100-200 m), and the definition of deeper structural and lithologic features to be used as guides to ore mineralization. The signals gathered in geophysical exploration are separated as far as possible from irrelevant noise and resolved into patterns associated with known and assumed geologic conditions; departures from the regional and local patterns are geophysical anomalies, and these are interpreted in respect to appropriate exploration models.

Magnetic, electrical, electromagnetic, and radioactive methods are the most widely used in prospecting for ore and industrial minerals deposits. Seismic and gravity methods are employed to some extent but not nearly so much as in petroleum prospecting. Ground-penetrating radar can provide images of shallow features for environmental investigations. Geophysical surveys are often made by several methods, so that more than one physical property can be taken into account. *See* GEOPHYSICAL EXPLO-RATION.

Magnetic surveys. Surveys are made with magnetometers—sensitive instruments placed at successive ground stations along traverse lines, carried in a helicopter, or carried in a fixed-wing aircraft in a pattern of flight lines (**Fig. 6**).

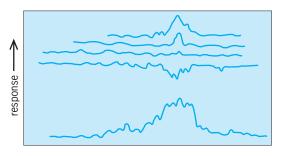


Fig. 6. Response curves for airborne electromagnetic (upper) and magnetometer (lower) prospecting system over a magnetic (pyrrhotite) and conductive stockwork deposit. The peak represents the location of the orebody.

The measurements in magnetic surveys are related to the degree of induced magnetization imposed on a rock or a mineral deposit by the Earth's magnetic field. The most common magnetic minerals are magnetite, ilmenite, pyrrhotite, and specular hematite. Magnetite, the mineral with the highest magnetic susceptibility, is a major component of iron orebodies, an abundant mineral in lithologic zones associated with contact metamorphic deposits, a fundamental component of mafic intrusive and volcanic rocks, and a minor mineral in many other rocks. A magnetic anomaly could therefore have several geologic explanations. A distinct and sharp anomaly, for example, may be associated with a body of iron ore, a contact metamorphic copper deposit, or a mafic dike; alternatively, it may represent a nickel orebody with pyrrhotite as the principal mineral, or a titanium orebody composed mainly of ilmenite. A subtle and broad magnetic anomaly, taken into account with an expected geologic pattern, may be the expression of a granitic intrusion or of a contact between basaltic volcanic rocks and limestone. See GEOMAGNETISM; MAGNETOMETER.

Electrical surveys. In electrical methods of prospecting, pairs or series of electrodes are placed in the Earth, and measurements are made of natural or induced currents associated with geologic features. The self-potential or spontaneous potential method involves traversing an area to find the patterns of natural currents set up by electrochemical action in conductive orebodies, especially sulfide mineral bodies. With resistivity methods, an electrical field supplied to the current electrodes during traversing is distorted in the Earth by variations in resistivity between geologic features, and the resulting pattern is measured at receiving or potential electrodes. A conductor, a zone of anomalously low resistivity, is interpreted in relation to an appropriate exploration model and alternative geologic conditions. A conductor may represent a mineral deposit, but it may also be related to such things as a conductive body of graphitic schist or clay. See ROCK, ELECTRICAL PROP-ERTIES OF

Induced polarization, the most popular of electrical methods, makes use of resistivity measurement and also an electrochemical polarization effect that occurs between two modes of electrical conduction, ionic (in pore fluids) and electronic (in metallic minerals). Mineralized rocks, where both kinds of conductors are present, can be detected by time domain surveys in which a chargeability or characteristic flow of current is measured between pulses of applied current; mineralized rocks can also be detected by frequency domain surveys in which the resistivity differs with changes in the frequency of the applied current. In unmineralized rock, where the current is carried only by ionic conduction, there is no flow between pulses, and the resistivity is independent of the applied frequency. The apparent resistivity and the chargeability, the frequency effect, or an inherent phase shift are measured at successive stations along a traverse; these values and a derived parameter, a metal factor, are determined for a series of depth points below the surface.

Electromagnetic surveys. Electromagnetic methods provide a means of determining differences in the electrical conductivity of rocks without having to make direct contact with the Earth. An electric current in a transmitting coil or at a distant source such as an atmospheric disturbance or a very low frequency communication station generates a primary electromagnetic field in the Earth's crust; eddy currents are induced in conductive bodies, and the resulting secondary field is measured with a detector coil (**Fig.** 7). As with electrical methods, conductors must be evaluated in relation to the appropriate exploration model and to other conductive geological conditions.

Airborne electromagnetic methods have been quite successful in locating anomalies associated with massive sulfide orebodies, especially in areas such as the Canadian Shield where conductive mineralization is hidden at shallow depth beneath glacial debris. Airborne electromagnetic surveys (Fig. 6), generally flown in combination with magnetic and radioactive systems, are also used as an aid to geologic mapping in regions where outcrops are scarce.

Ground electromagnetic traverses are commonly made in selected target areas with various configurations of transmitting and detecting coils. In some

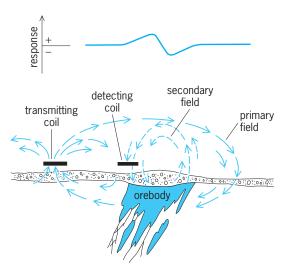


Fig. 7. Effect of a conductor on a ground electromagnetic system and the associated response amplitude curve.

electromagnetic methods, the transmitter is a ground cable or loop several miles long, and the small detector coil is carried across the area or is flown above the area in a helicopter.

Radiometric surveys. The principal methods in airborne and ground radioactivity surveying are gammaray spectrometry and total-radiation radiometrics. In both methods, the detector is a crystal that emits a flash of light, a scintillation, when struck by a gamma ray. The scintillation is converted to a voltage pulse that is read or recorded as a count rate. Gamma rays are emitted by uranium-, thorium-, and potassiumbearing minerals; the gamma-ray spectrometer is a specifying detector, equipped with separate channels for measuring the diagnostic radiation from each of three sources. *See* GAMMA-RAY DETECTORS.

Radioactivity surveys are the principal approach to prospecting for uranium deposits; the deposit or its uranium-bearing weathered equivalent must be near the surface, inasmuch as gamma rays are completely absorbed by a few inches of water or a few feet of barren soil and rock. Radioactivity surveys are also used in prospecting for deposits of phosphate rock, titanium-bearing sand, and rare-earth minerals, many of which contain minor amounts of associated uranium and thorium.

Airborne radioactivity surveys are made during regional exploration as an aid to geologic mapping; this application is based on the presence of potassium minerals and minor amounts of uranium and thorium in granitic intrusions and in several types of sedimentary rock.

Subsurface Prospecting

Direct subsurface investigations are a part of prospecting within target areas and in the detailed stages of target investigation. Subsurface prospecting is done by pits, trenches, and drill holes, and by underground workings such as shafts and tunnels.

Drilling is the principal method of subsurface prospecting where evidence of ore mineralization and geophysical or geochemical anomalies indicates a target for prospecting at a depth of more than a few feet. The common methods are diamond drilling, rotary drilling, and percussion drilling. In diamond drilling (**Fig. 8**), a solid core is taken as a sample,

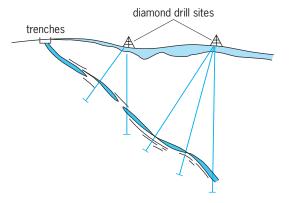


Fig. 8. Subsurface exploration pattern with vertical and angle diamond-drill holes.

and holes may be drilled at various orientations to depths of several thousand feet. In rotary and percussion drilling, rock chips or cuttings come to the surface in a stream of air, foam, or mud, and these are used as samples of the material at depth. *See* DRILLING AND BORING, GEOTECHNICAL; ENGINEER-ING GEOLOGY.

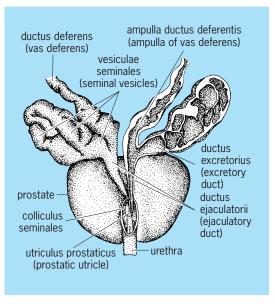
Geophysical information is obtained by the probing or logging of drill holes. Electrical and electromagnetic logging is done in holes drilled in search of metallic orebodies; with these methods, the radius of search is extended considerably beyond that of the small-diameter cylinder of sampled rock. Gamma-ray methods of geophysical drill-hole logging have become standard practice in prospecting for uranium ore, and there are several methods of nuclear activation in which other rocks and ore minerals in the walls of drill holes are subjected to neutron bombardment to provide a secondary radioactivity signature. *See* WELL LOGGING.

The investigation of exploration target areas by sinking shafts and driving tunnels is a subsurface prospecting approach with a long history. However, underground prospect workings are more often used in the delineation and evaluation of orebodies after a discovery has been made in drill holes. Even so, there are situations in which an irregularly mineralized orebody can still be identified best in larger-scale underground sampling and in drilling from mine workings, in comparison with drill holes from the surface. William C. Peters

Bibliography. J. Chaussier and J. Morer, *Mineral Prospecting Manual*, 1987; A. M. Evans (ed.), *Introduction to Mineral Exploration*, 1995; R. J. Goldfarb and R. L. Nielson (eds.), *Integrated Methods for Discovery: Global Exploration in the 21st Century*, Society of Economic Geologists, 2002; M. Kuzvart and M. Böhmer, *Prospecting and Exploration of Mineral Deposits*, 2d rev. ed., 1986; W. C. Lacy (ed.), *Mineral Exploration*, 1983; R. Marjoribanks, *Geological Methods in Exploration and Mining*, 1997; W. C. Peters, *Exploration and Mining Geology*, 2d ed., 1987; J. Tilton, R. G. Eggert, and H. Landsberg (eds.), *World Mineral Exploration*, 1988.

Prostate gland

A triangular body in men, the size and shape of a chestnut, that lies immediately in front of the bladder with its apex directed down and forward. It is found only in the male, having no female counterpart. The prostatic portion of the urethra extends through it, passing from the bladder to the penis. This organ contains 15–20 branched, tubular glands which form lobules. The gland ducts open into the urethra. Between the gland clusters, or alveoli, there is a dense, fibrous connecting tissue, the stroma, which also forms a tough capsule around the gland, continuous with the bladder wall. Penetrating the prostate to empty into the urethra are the ejaculatory ducts from the seminal vesicles which are



Prostate gland and seminal vesicles. (After W. A. N. Darland, ed., American Illustrated Medical Dictionary, 19th ed., Saunders, 1942)

located above and behind the organ (see **illus.**). The prostatic gland secretes a viscid, alkaline fluid which aids in sperm motility and in neutralizing the acidity of the vagina, thus enhancing fertilization.

A midline groove, palpable on rectal examination, indicates the two lateral lobes of glandular tissue; a medial lobe lies posteriorly. Small at birth, the prostate enlarges to adult size during puberty. After middle age, the prostate is sometimes subject to new tissue growth, usually benign, that may result in interference with urine flow through the compressed urethra. *See* PROSTATE GLAND DISORDERS. Walter Bock

Waller Ducr

Prostate gland disorders

Disorders of the prostate gland that can cause voiding symptoms and sexual dysfunction. The three most common diseases are prostatitis, benign prostatic hyperplasia, and prostate cancer. Prostatitis is an inflammatory condition of the gland, causing a variety of voiding symptoms and pain in the area around the prostate. Benign prostatic hyperplasia is a benign growth of the inner portion of the gland that causes progressive obstruction of the urinary channel (urethra) at the base of the bladder. Prostate cancer is the most common nonskin malignancy in men and is a major health concern for the aging population. *See* CANCER (MEDICINE).

Benign prostatic hyperplasia. Benign prostatic hyperplasia is a common disease of the aging male. Approximately 30–50% of all men develop urinary symptoms because of an enlarged prostate. While the exact cause of this disease is not known, it is believed that the male hormones testosterone and dihydrotestosterone play a significant role in the de-

velopment and maintenance of benign prostatic hyperplasia.

The disease process results in the development of outlet obstruction indicated by urinary hesitancy, weak stream, dribbling, and the sensation of incomplete bladder emptying. In addition, for reasons that are not completely understood, a number of individuals also develop irritative voiding symptoms: frequency, nocturia (excessive urination at night), and urgency.

Not all individuals have outlet obstruction. As many as 25% may have a coexisting neurogenic bladder and, if treated solely for prostatism, usually will not improve.

Treatments. Transurethral resection is the most common treatment offered, but there are other forms of therapy. The transurethral resection is associated with a 70–80% success rate and a 3–10% complication rate. All other forms of therapy have a lower complication rate but also a lower success rate. The alternate approaches include watchful waiting, medication, and minimally invasive procedures. Watchful waiting is an appropriate form of management in individuals with only mild symptoms. The majority of these individuals will not progress in the short term and will suffer no damage to the lower urinary tract if observation is elected. Those who have moderate symptoms can be offered a number of options ranging from medical therapy to surgery.

Medication. Two types of medication are used to treat the symptoms associated with prostatism (the symptom complex caused by urethral obstruction). Because the enlarged prostate is dependent on androgens, pharmacologic manipulation that lowers levels of dihydrotestosterone results in prostate gland shrinkage and improvement in the urinary symptoms.

Another form of medical therapy reduces the tone of the prostatic urethra by relaxing the prostate. This effect is achieved by an agent that blocks the alpha receptors which, when stimulated, cause contractions of the prostate gland.

The decision of which drug to use is based upon the clinical condition of the individual.

For individuals with severe symptoms or obstruction, surgery remains the best alternative. Individuals with very large prostates are best treated with an open prostatectomy. Those with moderately enlarged glands can have a transurethral resection. The smaller prostates can be treated by a transurethral incision. These operations will effectively cure 70–80% of the patients. Approximately 10% of the individuals will require a reoperation within 10 years. Complications are greater with this form of treatment in comparison to medical therapy or the minimally invasive techniques.

Prostate cancer. In men, prostate cancer is the most common cancer and the second leading cause of cancer death. In the United States the number of new cases has dramatically increased owing in part to the use of prostate specific antigen for the early detection of prostate cancer. When the antigen is used in screening or for early diagnosis, the detection rate

increases significantly. The American Cancer Society recommends an annual rectal examination starting at age 40 and an annual screening study beginning at age 50.

Stages of prostate cancer. Once a diagnosis of prostate cancer is made, the treatment options depend on the stage of the disease. Several factors influence therapeutic choices prior to making recommendations for treatment. These include the individual's age and health, and the stage and grade of the lesion.

Individuals whose disease is confined to the prostate (stage A or B) may be managed by watchful waiting, radical prostatectomy, or radiation therapy. In those with locally advanced disease (stage C), radiation or hormonal therapy may be indicated. In individuals with metastatic disease (stage D), hormonal therapy is used.

With the use of prostate specific antigen testing, a greater number of individuals show localized disease. Thus the number of individuals treated by radical prostatectomy has significantly increased. Radical prostatectomy involves removing the pelvic lymph nodes. If they are free of metastatic disease, the entire prostate is removed, from the bladder to the urinary sphincter.

Major modifications of this operation have greatly decreased the complications. The nerve-sparing or anatomic radical prostatectomy has resulted in a decrease in the risk of urinary incontinence from 10% to less than 3% of individuals, and an increase in preservation of potency to as high as 70%.

Noninvasive techniques. Conformal therapy utilizes three-dimensional computerized images to reconstruct the prostate in planning the external beam therapy. This technique sculpts the beam to the prostate and eliminates irradiation to noncancerous contiguous tissues (small bowel, rectum, and bladder). Conformal therapy may also allow for increased dose and possibly improved results. Real-time ultrasound guided brachytherapy allows for the placement of radioactive sources (either iodine-125 or palladium-103) directly into the substance of the prostate gland without spreading the radiation to the contiguous tissues.

The optimal treatment for individuals with stage C prostate cancer is not known. Younger men might best benefit from hormonal therapy given for 3-4 months prior to radical prostatectomy.

Individuals with metastatic prostate cancer are best managed with hormonal therapy, which deprives the prostate cancer cells of androgens, causing the cancer to regress. Complete hormonal therapy is accomplished by combining castration with medication that blocks the adrenal androgens. This medication, called an antiandrogen, works at the level of the androgen receptor within the prostate cancer cell. The combination of the antiandrogen with castration effectively removes all available androgens from the prostate cancer cell. The result is a longer time in remission and a prolonged survival. *See* CHEMOTHER-APY AND OTHER ANTINEOPLASTIC DRUGS; PROSTATE GLAND; RADIOLOGY. Nelson N. Stone

Prosthesis

An artificial replacement of a body part. It may be an internal replacement such as an artificial joint or an external replacement such as an artificial limb. A prosthesis may be individually manufactured for each client, such as a limb prosthesis, or may be manufactured in large numbers but in different sizes, such as artificial hip joints. Prostheses of all types are lighter and more functional than their predecessors; the broad field of prosthetics has benefited from advances in materials, miniaturization, and computergenerated fabrication.

Limb prosthetics. A standard nomenclature is used to refer to level of amputation and related prostheses. The term trans is used when an amputation goes across the axis of a long bone, such as transtibial (across the tibia of the leg) or transhumeral (across the humerus of the arm). When there are two bones together such as the tibia and fibula, the primary bone is identified. Amputations between long bones or through a joint are referred to as disarticulations and identified by the major body part, such as knee disarticulation. The term partial is used to refer to a part of the foot or hand distal to the ankle or wrist that may be amputated. Artificial limbs are designed to enable the individual to participate in all aspects of life, including sports and leisure activities.

The majority of lower limb amputations are performed as a result of vascular disease; more than 50% are performed on individuals over the age of 60, who have peripheral vascular disease secondary to diabetes. The next most frequent cause of amputation is trauma, with cancer and congenital deformities following. The major causes of juvenile amputations are trauma, congenital abnormalities, osteomyelitis, and cancer. In the upper extremity, the major cause is trauma followed by disease. *See* VASCULAR DISOR-DERS.

Upper and lower extremity prostheses need to be comfortable, functional, and cosmetic. If the prosthesis is not comfortable, the individual will not wear it; pain or discomfort can be the greatest impediment to successful prosthetic rehabilitation. The prosthesis must also allow the individual to perform desired activities that could not be done without a prosthesis, and to do them with the lowest possible expenditure of energy.

For most individuals, a well-fitting prosthesis will allow a greater range of mobility activities with less expenditure of energy than crutches or a wheelchair. This may not always be true for elderly individuals with transfemoral amputations or for individuals with bilateral amputations. While many individuals of all ages with bilateral transtibial amputations regain functional mobility with prostheses, adults with bilateral transfemoral amputations usually find ambulation too fatiguing. When the energy demands for prosthetic mobility are greater than for mobility without a prosthesis, the prosthesis is rarely worn.

The upper extremity has always posed a major challenge since the great complexity of hand

function is difficult to duplicate mechanically. The loss of sensation limits the function of the hand or hook, and replacement for sensory function has not been developed.

Sockets. Socket design varies with the level of amputation and the configurations of the individual residual limb. The prosthetic socket must support body weight and hold the residual limb firmly and comfortably during all activities. Each area of the residual limb tolerates pressure differently; therefore the tissues are selectively loaded so that the most weight is borne by pressure-tolerant tissue (such as wide and flat bony areas or tendons) and the least by pressure-sensitive tissue (such as nerves that are close to the skin and sharp bony prominence). This is accomplished by relief (socket concavity) over pressure-sensitive areas and socket convexity over the pressure-tolerant areas. Additionally, the socket needs to grip the residual limb firmly to reduce movement between the socket and the skin. The more movement between the residual limb and socket, the less secure the client will be during activities and the more risk for skin abrasions. Total contact between the distal end of the residual limb and the socket is required to aid in sensory feedback and to prevent swelling and skin problems. Sockets are individually constructed for each client from a cast made of the residual limb. The socket may be constructed with computer-aided design and computer-aided manufacture techniques. An electronic scanner is used to determine the shape of the amputation limb, and then a test socket is fitted to ensure comfort and to provide for appropriate weight-bearing and stabilizing pressures. The socket may be constructed by using a machine that molds plastic directly over the positive model of the residual limb. Sockets may be hard and rigid, or flexible and supported by a rigid frame.

Suspension. There are several methods of suspending each type of prosthesis. Suction sockets allow suspension without belts, sleeves, or cuffs. Negative pressure is created between the skin and socket by putting the residual limb in the socket in a manner to eliminate all air and create an intimate fit. In the transtibial prosthesis a rubber sleeve or a cuff that fits on the thigh may be used. In the transfemoral prosthesis a belt around the pelvis provides suspension.

Feet. Foot and ankle function is complex, and a variety of prosthetic feet are designed to respond dynamically to the pressure of walking and running. They store energy at the moment of heel contact, then return it at toe-off. The degree of dynamic response varies with the materials used in the foot, with carbon graphite providing the greatest degree of response.

Knee mechanisms. Replacement of knee function requires, at a minimum, control of both stance and swing. There are mechanical knee joints that provide some degree of each but not both. Knee mechanisms using hydraulic control can provide some degree of control of both stance and swing phases. Complex knee mechanisms allow an athlete to run



Fig. 1. Lower-limb prosthesis. Flexible plastic socket, combined with hydraulically controlled knee, plus ankle-foot mechanism permits controlled movement and sufficient speed to allow this above-knee amputee to continue many sports activities. Cosmetic outer covering has been removed to display the internal components. (Mauch Laboratories, Inc.)

by using a regular step-over-step pace. (Fig. 1).

Terminal device. Of paramount importance is the prosthetic component that will substitute for the missing hand. There is no device that completely replaces the appearance or function of the anatomic hand. Nevertheless, the prosthetic counterpart, called a terminal device, can contribute to the wearer's function and self-esteem. The two types of terminal device are the hand and the hook. Either is secured to a plastic socket encasing the forearm.

The hand may be active or passive: passive hands have no moving parts; active hands have a mechanism that permits the client to control finger position by appropriate action in the proximal part of the amputated limb. Both active and passive hands are made in a range of sizes to suit most children and adults, and are covered with a flexible plastic glove that matches the individual's skin tone.

The most popular active hand is operated myoelectrically (Fig. 2). The individual wears a socket with one or more skin electrodes that contact appropriate muscle groups. When the muscle contracts, the electrode transmits the microvoltage generated by the muscle to a mechanism that causes the electrical signal to operate a small motor that, in turn, enables the hand mechanism to open or close the fingers. The motor is powered by a battery worn inside the prosthetic socket or, in the case of wrist disarticulation, elsewhere on the body. Myoelectric control can provide excellent grasp force. The alternative type of prosthetic hand has a steel cable that is attached to a trunk harness. By flexing the shoulder, the individual puts tension on the cable that pulls on the hand mechanism, causing the fingers either to open (voluntary opening) or to close (voluntary closing). Relaxing tension on the cable allows the fingers to revert to the opposite position.

A hook is the other basic type of terminal device. Made either of aluminum or steel, hooks have two fingers that the client can open and close. Myoelectrically controlled hooks are available; however, most individuals who wear hooks have cable-operated ones that are either voluntary-opening or voluntaryclosing. Cable-operated hooks are much lighter in weight than any type of hand and are less expensive. Hooks are also more durable.

Other components. Sockets are custom-made of plastic and are intended to fit snugly to distribute pressure over the largest area. The proximal portion of the transradial socket terminates in the vicinity of the humeral epicondyles. A self-suspending socket supports the weight of the terminal device and wrist unit by a secure hold at the epicondyles. A harness is necessary to suspend the socket that is not supported by the epicondyles, to transmit shoulder motions to a cable, or to accomplish both functions. The customary design, figure-of-8, has a loop around each shoulder that connects in back.

The prosthesis for the individual whose amputation is at the elbow disarticulation level or higher includes a terminal device, wrist unit, forearm shell to replace the length of the absent forearm, an elbow unit, socket, and suspension. All elbow units have a hinge for elbow flexion and a locking mechanism to enable the wearer to retain the desired flexion angle. The elbow unit may incorporate a turntable that allows the individual to rotate the forearm shell medially and laterally. The figure-of-8 harness suspends the prosthesis and transmits shoulder and shoulder girdle motions to the cable system. Ordinarily, one cable operates the terminal device and the elbow hinge, and a second cable operates the elbow lock.

Joint replacements. Artificial replacements of joints, such as hip, knee, or shoulder, are another type of prosthesis. The knee and hip joints are the most frequently replaced, occasionally for trauma but usually for degenerative disease. Lower limb joint replacements must have load-bearing capabilities and are fabricated in different sizes of metal. Joint replacements may sometimes be secured by polymer adhesive. Hip and knee joint replacements are common, and a high degree of restoration of function is usually obtained. Most individuals experience major reduction or elimination of joint pain and 90–100% return of function.

In the upper extremity, shoulder and metacarpal phalangeal joints are most frequently replaced. The joint replacement in the upper limb needs to be light and to allow a great range of movement.

Other prostheses. There are a great variety of other types of replacements of body parts. Women who lose a breast to cancer are fitted with a prosthesis that can be fabricated of a number of lightweight materials and molded to resemble in shape and texture the remaining breast. In some instances an artificial breast made of biocompatible gel may be permanently inserted subcutaneously.

Prosthodontics are devices used to replace teeth. Lost teeth may be replaced by the insertion of a bridge containing several teeth resembling in shape

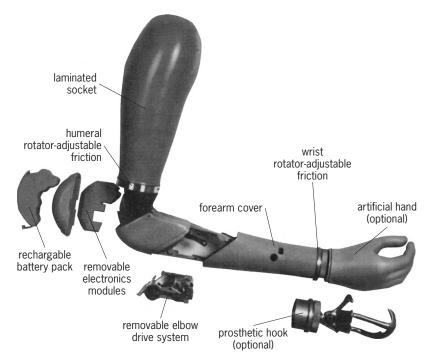


Fig. 2. Upper-limb prosthesis. Miniature electronic circuitry uses muscle activity from the upper-arm remnant to control the electric elbow. When the elbow is locked in position, control is automatically transferred to the electronic wrist and hand. (*Motion Control, Inc.*)

and size the missing ones. Dentists can also implant a single artificial tooth permanently into the jaw. *See* DENTISTRY.

The cardiac pacemaker is a form of a prosthesis replacing the natural electrical stimulation of the heart with a battery-operated device inserted within the body. Damaged heart valves are replaced with artificial valves attached directly to the heart muscle and to the major blood vessels. Artificial valves may be of disk or ball-and-seat design and allow normal flow of the blood from one chamber to the other. The artificial valve also prevents backflow. Prosthetic eyes, fabricated to resemble the remaining eye in color and configuration, serve a cosmetic function only. Bella J. May

Biologic-prosthetic systems. Although mechanical or electromechanical prosthetic devices represent the main option in many situations, regeneration of destroyed or resected tissue is the preferred goal. The use of inert materials to provide a basis for regenerated tissue growth offers intermediate options for reconstruction.

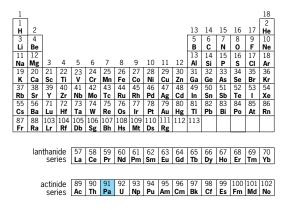
A porous material—metal, ceramic, or polymer of good biocompatibility will host the ingrowth of specialized tissue and its associated vascularity. But this ingrowth may require many months for full investment of the available void spaces, and so the prosthetic system must provide adequate support and protection during this period. The ingrowth phenomenon may be utilized, for example, to develop attachment of a prosthesis to a bone site by investment with normal bony cells. Clinical trials have been conducted for several designs of "cementless" total hip and total knee prostheses. On a more modest scale, porous metal systems are used for the reconstruction of cancer resections of the long bones and for the fusion of vertebrae.

Specialized tissue growth can utilize porosity, in the form of a fabriclike implant, as a substrate which stimulates and guides the growth. A development of this nature is an artificial skin system, which involves an underlying mesh of a biodegradable polymer into which fibroblasts and keratinocytes, two types of human skin cells, can be made to grow. *See* BIOMEDICAL ENGINEERING; SURGERY. W. Rostoker

Bibliography. J. H. Bowker and J. W. Michael (ed.), Atlas of Limb Prosthetics: Surgical and Prosthetic Principles, 1992; J. P. Paul, J. D. Baylor, and J. M. Courtney (eds.), Biomaterials in Artificial Organs, 1985; R. B. Salter, Textbook of Disorders and Injuries of the Musculoskeletal System, 3d ed., 1999.

Protactinium

A chemical element, Pa, atomic number 91. Isotopes of mass numbers 216, 217, and 222–238 are known, all of them radioactive. Only ²³¹Pa, the parent of actinium, ²³⁴Pa, and ²³³Pa occur in nature. The most important of these is ²³¹Pa, an α -emitter with a half-life of 32,500 years. The artificial isotope, ²³³Pa, is important as an intermediary in the production of fissile ²³³U. Both ²³¹Pa and ²³³Pa can be synthesized by neutron irradiation of thorium. *See* ACTINIUM; PE-RIODIC TABLE; RADIOACTIVITY; URANIUM.



Protactinium is, formally, the third member of the actinide series of elements and the first in which a 5*f* electron appears, but its chemical behavior in aqueous solution resembles that of tantalum and niobium more closely than that of the other actinides. *See* NIOBIUM; TANTALUM.

Metallic protactinium is silver in color, malleable, and ductile. The crystal structure is body-centered tetragonal. Samples exposed to air at room temperature show little or no tarnishing over a period of several months. The numerous compounds of protactinium that have been prepared and characterized include binary and polynary oxides, halides, oxyhalides, sulfates, oxysulfates, double sulfates, oxynitrates, selenates, carbides, organometallic compounds, and noble metal alloys. *See* ACTINIDE ELE-MENTS. H. W. Kirby Bibliography. P. W. Atkins et al., *Inorganic Chemistry*, 4th ed., 2006; F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., 1999; S. Cotton, *Lanthanide and Actinide Chemistry*, 2d ed., 2006.

Protandry

That condition in which an animal is first a male and then becomes a female. It occurs in many groups, in addition to oysters and cyclostomes. The reverse condition is protogyny. *See* PROTOGYNY.

Tracy I. Storer

Proteales

An order of flowering plants, division Magnoliophyta, in the eudicots. Consisting of three families, the order is one of the most controversial in current classifications, with a lack of obvious morphological characters linking the family Nelumbonaceae (two species) to the other two families, Platanaceae (seven species) and Proteaceae (about 1350 species). However, DNA sequences indicate that the three families form a natural group, and detailed studies have revealed that Platanaceae and Proteaceae share similar wood anatomy. *See* MAGNOLIOPHYTA; MAGNOLIOP-SIDA; WOOD ANATOMY.

Nelumbonaceae are aquatic, rhizomatous herbs with peltate leaves held above the water on long petioles, and have often been thought to be related to the true waterlilies (Nymphaeales). The fruit is highly characteristic with loose seeds embedded in the hard receptacle. Nelumbo (sacred lotus and American lotus) is used as a source of food (seeds, receptacle, and rhizome), as an ornamental, and as a sacred plant in several Asian countries. Platanaceae, from the Northern Hemisphere, are deciduous, monoecious trees with simple, palmately lobed leaves, flaking bark, and branched hairs. The fruits are hairy achenes in globose clusters which hang down. Plane trees (Platanus) are common street trees, especially in London, due to the regular loss of bark (which gives some resistance to pollution), and they also provide timber. See BARK; DECIDUOUS PLANTS; FRUIT; LEAF; NYMPHAEALES.

Proteaceae (predominantly tropical and subtropical, in the Southern Hemisphere) are evergreen shrubs and trees that often accumulate aluminum. The majority of species are from the Cape of South Africa and Australia. The roots often have short lateral branches (proteoid roots). Leaves are normally spirally arranged, simple (sometimes pinnate), and with xerophytic adaptations. Flowers often have specialized modes of pollen presentation. Fruits are of various types, often releasing seeds only after exposure to fire. Several genera (including Banksia, Protea, and Leucospermum) are widely cultivated for cut flowers; Grevillea and other genera are used for timber; and Macadamia yields edible nuts. See EV-ERGREEN PLANTS; ROOT (BOTANY). Michael F. Fay

Proteasome

A large proteolytic complex that degrades intracellular proteins. Intracellular proteins are continuously synthesized and degraded, and their levels in cells reflect the fine balance between these two processes. The rate of breakdown of individual proteins inside the cell varies widely and can be altered according to changes in the cellular environment. In eukaryotic cells, the site for degradation of most intracellular proteins is a large proteolytic particle termed the proteasome. Proteasomes are a major cell constituent, constituting up to 2% of cellular protein, and are essential for viability. They are found in the cytoplasm and nucleus of all eukaryotic cells. Simpler but homologous forms of the proteasome are also present in archaea and bacteria. Much of the knowledge about the structure and function of eukaryotic proteasomes has been gained from studies of these simpler systems.

The form responsible for most intracellular protein breakdown is termed the 26S proteasome, which functions as a component of the ubiquitinproteasome pathway (**Fig. 1**). This system catalyzes the rapid turnover of many critical regulatory proteins (for example, transcription factors, cell-cycle regulators, oncogenes); thus the proteasome plays a pivotal role in controlling a wide variety of cellular processes, ranging from cell division to circadian rhythms, gene transcription, and immune responses. Most normal long-lived proteins (which make up the bulk of proteins in cells) and damaged, misfolded, or mutated proteins (which could cause disease if they accumulated in cells) are also degraded by this pathway.

Proteasome versus protease. The proteasome differs from a typical proteolytic enzyme in many important respects. The typical protease is a single subunit enzyme of 20,000 to 40,000 daltons (Da). By contrast, the proteasome is up to 100-fold greater in size (2.5 million Da) and contains about 50 proteins with multiple enzymatic functions, some of which require adenosine triphosphate (ATP). Unlike typical protease pathways, the ubiquitin-proteasome pathway is ATP-dependent and uses the energy stored in ATP to mark, unfold, and transfer the protein substrate into the degradative chamber of the proteasome. Traditional proteases simply cleave a protein and release the partially digested fragments, whereas the proteasome binds and cuts the protein substrate into small peptides ranging from 3 to 25 residues in length, ensuring that partially digested proteins do not accumulate within cells. Peptides released by the proteasome are rapidly hydrolyzed to amino acids by peptidases in the cytosol, which are then reutilized for new protein synthesis.

Ubiquitin conjugation. Ubiquitin is a small (76 amino acids) protein found in all eukaryotic cells but not in bacteria and archaea. Most proteins degraded by the 26S proteasome are first marked by the covalent linkage to multiple ubiquitin molecules (Fig. 1). Ubiquitin conjugation represents a means of providing selectivity and specificity to the degradation process. Long chains of ubiquitin molecules are attached to proteins through the action of three enzymes called E1, E2, and E3. The E1 enzyme activates ubiquitin in an ATP-dependent manner and transfers it to one of the cell's approximately 15 E2's, which function as ubiquitin carrier proteins. The cell has hundreds of distinct E3 enzymes, each of which binds a specific group of proteins destined for degradation and catalyzes the transfer of the activated ubiquitin molecules from the E2 to a lysine side chain in the substrate protein. Proteins tagged

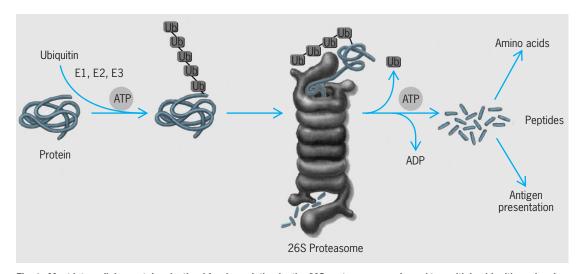


Fig. 1. Most intracellular proteins destined for degradation by the 26S proteasome are bound to multiple ubiquitin molecules through the action of three enzymes: a ubiquitin-activating enzyme (E1), a ubiquitin-conjugating enzyme (E2), and a ubiquitin-protein ligase (E3). Degradation of ubiquitinated proteins occurs within the central chamber of the proteasome. Ubiquitin conjugation, unfolding of a substrate, and its translocation into the inner cavities of the proteasome require energy obtained by ATP hydrolysis to ADP. The great majority of peptides released by the proteasome are further digested into amino acids by other peptidases in the cytosol, but a small fraction of peptides escape destruction and serve in MHC class I antigen presentation.

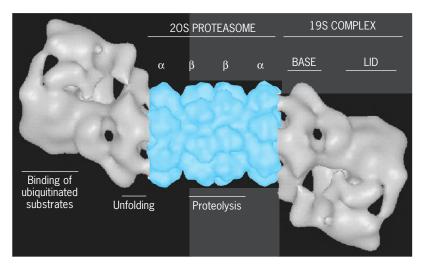


Fig. 2. Three-dimensional computerized reconstruction from electron microscopic images of the 26S proteasome. This large proteolytic complex is composed of about 50 different subunits that form subparticles, the 20S proteasome and the 19S regulatory particle. The lid of the 19S particle binds polyubiquitinated protein substrates and removes ubiquitin chains from the substrate. The substrate is then unfolded by the six ATPases located in the base of the 19S particle and translocated into the inner cavities of the 20S proteasome. Degradation occurs inside the central chamber of the 20S proteasome, formed by the two adjacent β -rings which contain six proteolytic sites. (After W. Baumeister et al., The proteasome: Paradigm of self-compartmentalizing protease, Cell, 92:367–380, 1998)

with ubiquitin chains are then rapidly bound and destroyed by the 26S proteasome.

Structure and function. The 26S proteasome is a complex composed of two very different components-the core 20S particle and one or two 19S regulatory particles (Fig. 2). Proteins are degraded within the 20S (720-kDa) particle, a 28subunit barrel-shaped structure composed of four stacked rings, each containing seven distinct but homologous subunits surrounding a central cavity. The two inner rings (formed by the β -subunits) contain the proteolytic active sites and together form a central chamber in which proteins are degraded. The two outer rings (formed by the α -subunits) together with the β -rings form two antechambers through which protein substrates must pass. The proteolytic active sites of the β -subunits face the central cavity so that cleavage of proteins can occur only within the cavity. Only three of the seven β -type subunits possess catalytic activity. Each is specific for a different type of amino acid sequence: two cleave preferentially after basic amino acids, two after large hydrophobic amino acids, and two after acidic residues. Thus, the 20S particle can cleave nearly all the peptide bonds in proteins.

Substrates can enter the 20S particle only at the center of the α -ring. The N-terminal sequences of the α -subunits form a gate that is normally maintained in a closed position. Proteins or peptides cannot, therefore, freely enter the inner cavity of the 20S proteasome from the cytosol. This gate must be opened by a 19S regulatory particle for protein substrates to gain access to the catalytic chamber of the proteasome. The small diameter of the entry channel—even in the opened state—requires polypeptides to be unfolded in order to enter the chamber and be degraded. The 19S particle not only binds ubiquitinated proteins

but also unfolds them and promotes their translocation into the 20S particle. The architecture of the particle, in which proteolysis is isolated to a separate chamber and access is tightly regulated, prevents inappropriate destruction of cell proteins.

The 19S (890-kDa) regulatory particle associates with one or both ends of the 20S proteasome and is composed of at least 17 different subunits. It consists of two different functional entities, the lid and the base (Fig. 2). The lid provides specificity to proteolysis and contains the binding site for ubiquitin chains and enzymes that break down the polyubiquitin chains so that the ubiquitin can be recycled for use in subsequent rounds of proteolysis. The base, which touches the α -rings of the 20S core particle, contains six different ATPases (a type of enzyme that hydrolyzes ATP and uses the energy released to drive biological processes). These ATPases form a ring that uses ATP to unfold protein substrates. They also have the capacity to open the gate in the α -ring of the 20S proteasome. In doing so, they promote translocation of substrates into the inner chambers of the 20S particle, where they are degraded. These structural features of the proteasome, the need for ubiquitination prior to degradation, and the energy requirement for proteolysis clearly evolved to provide a remarkable degree of selectivity and regulation to the degradative process and to safeguard against the nonspecific digestion of essential cellular proteins.

This complex system evolved from earlier proteolytic complexes found in archaea and bacteria, which lack ubiquitin and 26S proteasomes. Archaea, for example, contain 20S proteasomes that are arranged as in eukaryotes but appear to function in protein breakdown in association with an ATPase ring complex called PAN (proteasome-activating nucleotidase). PAN stimulates the ATP-dependent degradation of proteins by archaeal proteasomes, but it does not stimulate hydrolysis of small peptides that appear to freely enter into these particles. In prokaryotes, proteins bind directly to PAN. This hexameric-ring structure is homologous to the six ATPases found in the 19S base of the eukaryotic proteasome and has the capacity to unfold globular proteins and promote their entry into the 20S particle. Many features suggest that this PAN complex was the evolutionary precursor of the 19S regulator base of the eukaryotic 26S proteasome. The lid of the 19S regulator evolved from a distinct enzyme complex that allows degradation of ubiquitinated proteins. Thus, with evolution, the ATP-dependent function of the eukaryotic proteasome became linked to ubiquitin conjugation to provide greater selectivity and regulation, and lid subunits became necessary to bind the ubiquitinated protein substrates.

Catalytic mechanism. Proteolytic enzymes are normally classified according to which catalytic amino acids are in their active sites; proteins undergo a different nucleophilic attack depending on whether a serine, cysteine, aspartic, or metal group is in the active site of the enzyme. The sequences of the proteasome's catalytic β -subunits are not homologous to those of known proteases, and the pattern of sensitivity to various inhibitors differs from that of any known protease family. X-ray diffraction studies and mutagenesis of different amino acids in the proteasome have uncovered a new type of proteolytic mechanism. The active-site nucleophile of the proteasome is the hydroxyl group of a threonine at the amino terminus of the β -subunit. Because no other protease of this type is known, it has been possible to synthesize or isolate inhibitors that specifically block proteasomes, without affecting other cellular enzymes. These inhibitors have been valuable tools in clarifying the proteasome's mechanism and determining its role in cells.

Antigen presentation. In mammals, a small portion of peptides generated by proteasomes escapes destruction in the cytosol and is transported into the endoplasmic reticulum (ER). In the ER, peptides of defined lengths (between 8 and 11 amino acids) are bound by specialized membrane-bound proteins, known as major histocompatibility complex (MHC) class I molecules, and are subsequently transported to the cell surface. The immune system is continually surveying the surface of all cells for nonself peptides originating from viral or mutated proteins. If cells carrying such nonnative peptides are encountered, they are rapidly destroyed by cytotoxic T lymphocytes. Specialized proteasomes, called immunoproteasomes, play an important role in the generation of most MHC class I peptides. They are normally found in lymphoid tissues (for example, the spleen and thymus) but are induced in other cells of the body during infection or in response to the cytokine interferon- γ . Immunoproteasomes contain three alternative proteolytic β -type subunits that replace constitutive β subunits. These alternative subunits change the way that peptide bonds are cleaved within proteins in order to favor the production of peptides that can preferentially bind MHC class I molecules.

20S proteasomes can also associate with an activator called PA28, or the 11S regulator. PA28 is a heptameric ring composed of two types of subunits that can be attached to one or both ends of the 20S proteasome. PA28 greatly enhances the proteasome's degradation of small peptides (but not of intact or ubiquitinated proteins) by opening up the channel in the base of the particle. Like the alternative β -subunits of the immunoproteasome, PA28 is induced by the cytokine interferon- γ . This cytokine also enhances the formation of a hybrid proteasome particle that consists of a 20S core particle capped with the 19S regulator on one side and with PA28 on the other. It is thought that within these hybrid particles ubiquitinated substrates initially bind to the 19S regulator and peptide products are released through PA28. These hybrid particles appear to play a major role in the generation of MHC class I antigens.

Role in disease. The novel catalytic mechanism of the proteasome has permitted the synthesis of selective inhibitors that can be used to study the function of proteasomes in intact cells. These studies resulted in many of the present insights about the role of the ubiquitin-proteasome pathway in various biological processes. Besides being involved in the rapid clearance of mutated and abnormal proteins which accumulate in various inherited diseases, this pathway is implicated in regulated degradation of transcription factors, oncogenes, tumor-suppressor genes, and cell-cycle regulators. Therefore it is not surprising that several prevalent diseases are caused by the gain or loss of function of this proteolytic system. For example, carcinomas arising from cells lining the uterine cervix are caused by certain strains of human papilloma virus. Upon infecting cells, these viruses produce a protein (E6) that in concert with an E3 (called E6-AP) causes rapid ubiquitination and proteasome degradation of the major tumor suppressor p53. The lack of p53 dramatically increases the tendency of cells to become malignant. Mutation of another E3, the von Hippel-Lindau tumor suppressor, is responsible for some types of kidney tumors. Among its many functions, the suppressor regulates the levels of a specific transcription factor that controls the growth of blood vessels into tissues. This protein is normally maintained at a low level by rapid ubiquitination and degradation by proteasomes; however, when the suppressor is mutated, the transcription factor accumulates in cells, and tumors are able to form new blood vessels and grow rapidly.

The ubiquitin-proteasome system is also very important in the onset of inflammation. The transcription factor known as nuclear factor-kappa B (NF- κ B) is the key regulator of expression of many inflammatory mediators. Normally, it is localized to the cytoplasm through binding of an inhibitory protein, termed I κ B, which prevents its translocation to the nucleus. However, bacterial surfaces, oxidative stress, cytokines, and a cytokine tumor necrosis factor- α cause rapid phosphorylation, ubiquitination, and degradation of I κ B, allowing NF- κ B to enter the nucleus, bind to DNA, and activate inflammatory gene expression.

A general acceleration of the ubiquitin-proteasome pathway is responsible for the excessive breakdown of cell proteins in muscle during various disease states characterized by muscle wasting, such as cancer cachexia, diabetes mellitus, sepsis, inflammatory diseases (for example, AIDS), chronic renal failure, and nerve injury or disuse atrophy. There is appreciable interest in using proteasome inhibitors to treat various pathological states because of the key role of the proteasome in human disease. Promising results have been reported in animal models of various diseases, including cancer, rheumatoid arthritis, and stroke. Human trials with these agents are in progress. See ADENOSINE TRIPHOSPHATE (ATP); AMINO ACIDS; CYTOLYSIS; ENZYME; HISTOCOMPATI-BILITY; NUCLEOPROTEIN; PROTEIN; TUMOR VIRUSES. Tomo Saric; Stewart Lecker; Alfred L. Goldberg

Bibliography. A. Ciechanover, A. Orian, and A. L. Schwartz, Ubiquitin-mediated proteolysis: Biological regulation via destruction, *Bioessays*, 22(5):442-451, 2000; D. H. Lee and A. L. Goldberg, Proteasome inhibitors: Valuable new tools for cell biologists, *Trends Cell Biol.*, 8:397-403, 1998; J. M. Peters,

J. R. Harris, and D. Finley (eds.), *Ubiquitin and the Biology of the Cell*, Plenum Press, New York, 1998; K. I. Rock and A. L. Goldberg, Degradation of cell proteins and the generation of MHC class I-presented peptides, *Annu. Rev. Immunol.*, 17:739–779, 1999; A. L. Schwartz and A. Ciechanover, The ubiquitinproteasome pathway and pathogenesis of human disease, *Annu. Rev. Med.*, 50:57–74, 1999; D. Voges, P. Zwickl, and W. Baumeister, The 26S proteasome: A molecular machine designed for controlled proteolysis, *Annu. Rev. Biochem.*, 68:1015–1068, 1999.

Protective coloration

A strategy that organisms use to avoid or deflect the attacks of predators by misleading the latter's visual senses. Protective devices directed at the hearing or the sense of smell of the predators are harder for researchers to detect, and are known in only a small number of species.

Types. Protective coloration can be classified according to whether the functioning or malfunctioning of the vertebrate visual system is exploited. Exploiting the malfunction of the system means simply "not being seen": the prey fails to attract the attention of the predator, usually because it is the same color as the general background (the green of leaves, the mottling of lichen, or the brown of soil) or because it fails to cast a shadow. The organism avoids producing shadow by flattening itself against the substrate, or by countershading, in which the lower parts of a cylindrical prey such as a caterpillar are more lightly colored than the upper parts. As shadows normally form on the underside of cylinders, the shading cancels the shadow and makes the caterpillar optically flat. Animals that match their background often have an ability to select the appropriate background to rest on, or much less frequently (although widely known in the case of the chameleon) can change their own color to match

Exploiting the functioning of the vertebrate perceptual system takes many forms: the vertebrate visual cortex decodes the image on the retina in a hierarchical process starting with the detection of edges, and then fills the shapes so defined with colors. A predator may thus detect the shape of a moth which it will fill with the color of the moth; this process enables detection. A moth may counter this by possessing strikingly contrasted patches of color on its wings, arranged in a random way that resembles no interesting shape. What the predator then sees is a random arrangement of, for example, black and white, which it cannot decode as a significant shape; the outline of the moth is broken up.

The prey may alternatively exploit the learning capacity of the predator. Many objects in the world are probably accurately perceived but ignored as of no interest. For example, insectivorous birds assuredly see leaves but do not attack them because they have learned (or perhaps know innately) that these are not edible. Resemblances to leaves, twigs, thorns, flowers, parts of flowers, and more bizarre objects like fresh turds (usually bird droppings) or discarded head capsules of ants are very widespread. This type of camouflage is termed mimetic camouflage. Camouflage in general is often termed cryptic coloration.

Mimetic coloration and warning color. Coloration may be considered mimetic if protection is achieved by a resemblance to some other existing object, which is recognized by the predator but not associated in its mind with feeding. Usually this negative or neutral association is learned, but in a minority of instances it is almost certainly innate. Small birds have an innate flight response to large eyes in closeup (which normally indicate that a cat or a predatory bird is dangerously close). This reaction is exploited by many moths and other insects, which have eyelike markings, sometimes very convincing in their shading and highlighting, on concealed parts of the wings. Attack by a predator causes such a moth not to flee but to change its posture rapidly to reveal the fake eyes, thus frightening away the attacker. Motmots (birds which habitually prey on snakes) have a similar innate fear of the red, black, and yellow striping patterns of the deadly coral snakes. These patterns are mimicked by various nonvenomous snakes. The fear that most birds have of snakes in general results in effective mimicry of even cryptically colored snakes by caterpillars with snakelike coloring and defensive behavior that usually combines the display of large eyelike markings with the posture of a snake about to strike. The ancestor of the domestic silkworm had such coloring, and the now colorless silkworm continues to use the display.

Flash coloration. Flash coloration seems to be an intermediate between a warning color and large eyespots. The prey is cryptic when at rest, but reveals brilliantly colored parts while escaping. This behavior seems to function simply by startling the predator. Very small eye marks at the tips of the wings, or a false head at the wrong end of the body (shown by some coral reef fish, for example) may cause the predator to misdirect its attack.

Aposematic defense. Protection through the possession of a chemical or physical defense that is dangerous to one's potential predator, accompanied by a strikingly conspicuous pattern known as warning coloration (often black, red, yellow, and white), is widespread-the ensemble of defense and color is termed aposematic. Examples occur in Hymenoptera, Lepidoptera, and sea slugs. The actual defense ranges from toxic venoms through stings (in wasps, for example), to the oozing of noxious foams or hemolypmph (as in ladybirds), to the possession of toxic chemicals (cyanides, cardiac glycosides, alkaloids) that will poison the predator or simply produce a revolting taste. The defense may be synthesized de novo or may be obtained from a host plant (as with many butterflies and moths) or from the animal's own prey (as with the use of nematocysts by some sea slugs). The function of the warning color is to remind the predator of its previous unpleasant experience, and is, for a variety of reasons connected

with the neurology and cognitive aspects of learning, easier to remember than a pattern which resembles the background.

Bright coloration must have evolved in circumstances in which the disadvantages of being greatly conspicuous were outweighed by the advantages conferred by more rapid learning on the part of the predators. Related factors involve the prey's being so tough that it can often survive an attack (when the predator will readily learn not to attack it again), or the prey's being gregarious, so that individuals, although killed when attacked, confer benefits on their immediate relatives through their death (gregarious groupings usually being siblings from the same batch of eggs).

The warningly colored life-style involves the development of some distinctive behavior, such as remaining motionless and oozing liquid (or otherwise threatening the predator) when attacked; and in butterflies, habitually flying more slowly and in more shady habitats (where the body temperature remains lower on account of the absence of direct solar heating). Camouflaged butterflies in comparison fly rapidly in open sunny areas and have very high body temperatures. These flight differences lead to the aposematic butterflies having wings of greater aspect ratio, and thus affect the evolution even of the butterflies' morphology.

Batesian and Müllerian mimicry. Warningly colored species can be exploited as models for mimicry, just as can inedible objects such as thorns and leaves. Sometimes the term mimicry is restricted to such resemblances between edible species and actively defended and warningly colored models. Much is known about the evolution of this kind of mimicry in butterflies (and to a lesser extent, in bees and flies). If the mimic is entirely edible, the relationship is parasitic; the mimic benefits from the resemblance, but as every encounter with a mimic reduces the predator's aversion, the model suffers some increase in the rate of attack; such mimicry has traditionally been termed Batesian mimicry. Alternatively, the mimic may be almost or fully as defended as the model, leading to a mutualistic relationship known as Müllerian mimicry, in which both the model and the mimic species suffer a decreased rate of predation (compared with what would happen if each had a different color pattern).

The evolution of Müllerian mimicry has been a powerful force in shaping the appearance of tropical insects. In most areas of the tropics (at least within the forest habitats), the numerous species of warningly colored butterflies tend to share only five or six patterns (known as mimicry rings); this number is also known from tropical wasps, and from bumblebees in the temperate zone (where butterfly mimicry is much less extensive, probably because of a reduced number of species).

Müllerian rings in any one area are parasitized by the local Batesian mimics. They can frequently be distinguished from the Müllerian mimics by retaining their camouflage on the underside of their wings (the part seen when the butterfly is resting). Sometimes mimicry is confined to the female (so that the male retains an ancestral camouflaged or sexual display pattern). Occasionally, Batesian mimics can be distinguished from the Müllerian mimics by being polymorphic, that is, by exhibiting several different patterns within the one species, each pattern belonging to a different mimicry ring.

Evolution of mimicry. In general, the evolution of mimicry occurs in two distinct phases. In the first phase, a mutation of rather large effect produces a rough resemblance between the future mimetic species and its model. This occurs partly by happenstance (out of the great range of mutations which a species may produce, one will eventually resemble something else), and partly because similar processes of development occur in the embryogenesis of the pattern: all butterflies probably potentially share certain physicochemical processes during the laying down of the pattern on their wings. If the rough resemblance is close enough, the new mutation spreads.

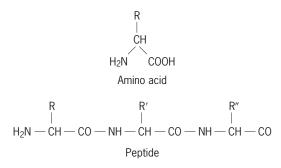
During the second phase, the pattern of the mimic gradually changes to resemble the model ever more closely, as the fine discriminatory powers of the predators eliminate poor resemblances at the expense of better ones. This modification phase is thought to involve the establishment of genes of individually rather small effect on the pattern, which cumulatively refine the resemblance. During this phase, the Batesian mimic places the model at a disadvantage; the model therefore evolves away from the pattern of the mimic. It is, however, inevitably built into the dynamics of the system that the mimic can evolve faster than the model; this is the only case in which scientists know in evolutionary terms why the host is unable to shake off its parasite.

Evolution in these two phases occurs in both Batesian and Müllerian mimicry; Müllerian mimics join, or even switch between, mimicry rings in this fashion. It is possible that evolution in general is not always entirely gradual, but may involve small discontinuous changes. *See* ANIMAL EVO-LUTION; MACROEVOLUTION; POPULATION ECOLOGY. John R. G. Turner

Bibliography. W. B. Cott, *Adaptive Coloration in Animals*, 1940; M. Edmunds, *Defence in Animals*, 1974; E. B. Ford, *Ecological Genetics*, 4th ed., 1975; P. M. Sheppard, *Natural Selection and Heredity*, 4th ed., 1975; J. R. G. Turner, Butterfly mimicry: The genetical evolution of an adaptation, *Evol. Biol.*, 10:163–206, 1977; R. L. Vane-Wright and P. R. Ackery (eds.), *The Biology of Butterflies*, 1984.

Protein

Polymeric compounds made up of various monomeric units called amino acids. Amino acids are joined together in a chain by peptide (amide) bonds between the α -carboxyl groups and the α -amino groups of adjacent amino acids. The first amino acid in a protein usually contains a free α -amino group, as shown below. Proteins generally



contain from 50 to 1000 amino acids per chain. Small chains of up to 50 amino acids are usually referred to as peptides or polypeptides. *See* AMINO ACIDS; PEPTIDE.

Occurrence. Proteins are central to the processes of life. They are fundamental components of all biological systems, performing a wide variety of structural and functional roles. For example, proteins are primary constituents of structures such as hair, tendons, muscle, skin, and cartilage. Several hormones, such as insulin and growth hormone, are proteins. The substances responsible for oxygen and electron transport (hemoglobin and cytochromes, respectively) are conjugated proteins that contain a metalloporphyrin as the prosthetic group. Chromosomes are highly complex nucleoproteins, that is, proteins conjugated with nucleic acids. Viruses are also nucleoprotein in nature.

Protein enzymes are the catalysts of nearly all biochemical transformations. Pepsin and rennin are examples of digestive enzymes involved in breaking down food. DNA polymerases are enzymes that duplicate DNA for cell division, other enzymes are needed to repair damaged DNA, and gene expression is carried out by RNA polymerases. The chemical reactions used in metabolic pathways for carbohydrates (citric acid cycle), lipids, amino acids, and energy production (oxidative phosphorylation) are catalyzed by protein enzymes. All living things contain proteins because they serve as the molecular tools and machines of life. *See* ENZYME.

Of the more than 200 amino acids that have been discovered either in the free state or in small peptides, only 20 amino acids, or derivatives of these 20, are present in mammalian proteins. They are alanine, arginine, asparagine, aspartic acid, cysteine, glutamic acid, glutamine, glycine, histidine, isoleucine, leucine, lysine, methionine, phenylalanine, proline, serine, threonine, tryptophan, tyrosine, and valine. Each has a unique side chain (represented by R, R', R" in the peptide structure above) that defines the chemical properties of the amino acid. Except for glycine, amino acids have an inherent asymmetry referred to as chirality. In all natural proteins, they are of the L-configuration. Amino acids with the mirrorimage configuration are p-amino acids. Both p- and L-amino acids occur in the free state and in small peptides. The amino acids can be grouped according to similarities in size, shape, charge, aromaticity, and polarity. Amino acids with nonpolar, hydrophobic

side chains are alanine, isoleucine, leucine, methionine, tryptophan, phenylalanine, proline, and valine. Another group with polar, uncharged side chains is asparagine, cysteine, glutamine, serine, threonine, and tyrosine. Finally, five amino acids have side chains that can be charged at physiological pH values. Arginine, lysine, and histidine can be protonated to bear a positive charge, whereas aspartic acid and glutamic acid can be deprotonated to bear a negative charge. Commonly occurring derivatives of these 20 amino acids are cystine, hydroxyproline, and hydroxylysine. Cystine results from the oxidation of two cysteines to form a disulfide bridge. Hydroxyproline and hydroxylysine are formed by the enzymatic hydroxylation of specific proline and lysine, respectively. Modification of these two amino acids takes place while the polypeptide chain of collagen is being synthesized on the ribosome. Other examples of biological modification of parent amino acids are O-phosphoserine in phosvitin, ³N-methylhistidine in actin, and ε -N-methyllysine in histones.

Biosynthesis. The linear arrangement of amino acids in a protein is termed its sequence (primary structure). The sequence is highly specific and characteristic for each particular protein. It is determined by the DNA sequence of each protein's gene that is expressed in the form of messenger RNA. Elucidation of the mechanism by which proteins are built up from free amino acids has been one of the key problems of molecular biology. *See* DE-OXYRIBONUCLEIC ACID (DNA); MOLECULAR BIOLOGY; RIBONUCLEIC ACID (RNA).

Although a few proteins such as collagen are stable indefinitely in adulthood, most body proteins are in a continual process of degradation and synthesis (turnover). For example, the half-life of serum proteins in humans is about 10 days. Each amino acid is activated by adenosine triphosphate (ATP) by a specific enzyme called an aminoacyl-tRNA synthetase. These enzymes establish the relationship between specific nucleotides and amino acids, which is the basis for the genetic code. This activated amino acid is covalently attached to a transfer RNA that contains a triplet of nucleotides (anticodon) at one end. Each triplet of nucleotides is unique for one amino acid. Protein synthesis, the joining of amino acids to form polypeptides, takes place on the ribosome, a large ribonucleoprotein complex that translates mRNA into proteins. On the ribosome, aminoacyl-tRNA binds specifically to a segment on mRNA through bonding between the anticodon on the tRNA and a complementary codon of three nucleotides on the mRNA. Two activated aminoacyl-tRNA complexes bind adjacent mRNA codons and then react together to form a peptide bond. The process continues until the synthesis of the entire protein is complete. See RIBO-SOMES.

Structure. The task of determining the sequences of proteins was a major preoccupation of many scientists in the twentieth century. Frederick Sanger was awarded a Nobel Prize in 1958 for determining the first protein sequence—insulin, a 51-amino acid protein. Sanger's work proved for the first time

that a protein was a pure chemical entity comprising a single amino acid sequence and not a collection of closely related amino acid sequences (sometimes referred to as statistical proteins). Later developments in DNA sequencing led to a Nobel Prize for Walter Gilbert and Frederick Sanger (his second). The relative ease of DNA sequencing combined with many additional technical innovations has permitted the whole-scale sequencing of genomes. Thus, most protein sequences are now deduced from the corresponding gene sequence. With this wealth of sequence information, researchers in the fields of protein chemistry and molecular and cellular biology are focused on understanding the functional roles and properties of proteins.

Proteins of similar function from different species have a common genetic origin and thus have related sequences. When differences exist in the sequence, changes are usually conservative, such as the replacement of one small hydrophobic amino acid by a different small hydrophobic amino acid. The degree of similarity between one protein in two different animals has been used to evaluate the relative genetic distance of the two animals. This type of comparison was made for the β -globin chain of hemoglobin, the blood protein that carries oxygen and carbon dioxide, for chimpanzees, monkeys, orangutans, gorillas, and humans. The relatively close relationship for all these animals is reflected by their nearly identical amino acid sequences for β -globin. However, some amino acid differences exist between β -globin chains. For example, chimpanzee β -globin is identical to gorilla β -globin at 146 of the 147 amino acids. The single difference is an arginine in one sequence and a lysine in the other. In this example, arginine and lysine are both positively charged amino acids and β -globin functions with either. Thus, this change is conservative. See HEMOGLOBIN; PROTEINS, EVOLU-TION OF

Nonconservative changes can alter the ability of a protein to function, as is evident by the single amino acid difference between normal and sickle human β -globin. Sickle β -globin contains valine at a position normally occupied by glutamic acid. This single amino acid change alters the overall structure of hemoglobin which, in turn, alters the shape of the red blood cell. When a person inherits the gene for sickle β -globin from both parents, it results in the disease sickle cell anemia. *See* SICKLE CELL DISEASE.

Proteins are not stretched-out polymers; rather, each adopts a specific extended or compact and organized structure called its native structure. It is still not completely understood how proteins "fold" into their structures, nor can we accurately predict the complete structure from its amino acid sequence. Pioneering studies that won Christian Anfinsen the Nobel Prize in 1972 showed that proteins fold themselves—that all the information needed to reach the correct three-dimensional structure is contained in the linear sequence of amino acids. Thus, after synthesis on the ribosome, a polypeptide chain folds to its native structure. In order to prevent the misfolding of certain proteins, chaperone proteins

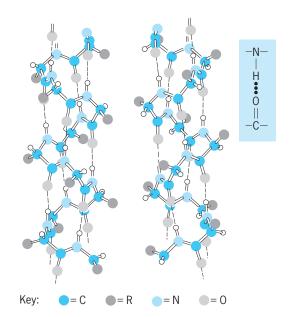


Fig. 1. The α -helix proposed by L. Pauling and R. B. Corey. The repeating —NH—CO—C units form the backbone which spirals up in a left-handed or a right-handed fashion. (The right-handed configuration is adopted in natural proteins with L-amino acids.) Hydrogen bonds are indicated by the broken lines. Note that the side chains (R) are all directed out from the helix. The pitch of the helix (axial displacement per complete turn of the helix) is about 0.54 nm, and there are 3.7 amino acids contained in one complete turn. The structure at the right shows the hydrogen bond. (After J. T. Edsall and J. Wyman, Biophysical Chemistry, vol. 1, Academic Press, 1958)

intervene to facilitate proper folding. Proteins that become misfolded or that have lost their native structure pose a major problem in cells. Diseases such as cystic fibrosis and Alzheimer's disease are characterized by misfolded proteins. In cystic fibrosis, a mutation in the chloride ion channel protein prevents the final stages of protein folding and leads to deficiencies in ion transport across cell membranes. In Alzheimer's, a protein known as amyloid precursor protein takes on a misfolded shape that causes the protein to clump together, ultimately forming plaques in the brain. *See* MOLECULAR CHAPERONE.

The polypeptide backbone of protein can fold in several ways by means of hydrogen bonds between the carbonyl oxygen and the amide nitrogen (Fig. 1, inset). Structural elements created by backbone hydrogen bonding interactions in the polypeptide are called secondary structures and include features like α -helices, β -sheets, and turns. In helices, the backbone is coiled in a regular fashion that brings peptide bonds separated by several amino acids into close spatial approximation. The stability of a helix is attributed to hydrogen bonds between these peptide bonds. The configuration of an α -helix, which is the most abundant secondary structural element in globular proteins, is described by a backbone amide hydrogen of one peptide unit bonded to a carbonyl oxygen (four amino acids away) that is part of another peptide unit (Fig. 1). Helical structures are often made up of peptide segments of around 10 amino acids, which corresponds to about three turns of helix. In keratin, the fibrous protein found in hair and nails, much longer helices are formed and pack with helices from other keratin polypeptides to give tensile strength. *See* HYDROGEN BOND.

In addition to α -helices, polypeptides form β -sheet structures that are made of two or more segments that run parallel or antiparallel to each other and connect through backbone hydrogen bonds. The amino acid side chains stick out on alternating sides of the β -sheet surface. The human porin protein is made almost entirely of β -strands arranged to form a central pore that allows sugars, amino acids, ions, and other small metabolites in and out of cells.

The third level of folding in a protein (tertiary structure) comes through interactions between different parts of the molecule. At this level of structure, various secondary structure elements are brought together and interact through many types of associations. Hydrogen bonds between different amino acids and peptide bonds, hydrophobic interactions between nonpolar side chains of amino acids such as phenylalanine and leucine, and salt bridges such as those between positively charged lysyl side chains and negatively charged aspartyl side chains, all contribute to the tertiary structure specific to a given protein. Disulfide bridges formed between two cysteines at different linear locations in the molecule can stabilize parts of a three-dimensional structure by introducing a covalent bond as a cross-link. The result is a unique architecture that is predetermined by the particular sequence of amino acids in the protein.

Finally, some proteins contain more than one polypeptide chain per molecule. This feature is referred to as the quarternary structure. There is usually a high degree of interaction between each subunit, for example, between the α - and β -globin chains of hemoglobin. The tetrameric structure of hemoglobin exemplifies how four individual polypeptide chains (two α - and two β -chains) function cooperatively. The binding of oxygen to one subunit causes a conformational change in the protein that facilitates the binding of oxygen to the other subunits. Thus, the cooperative binding of oxygen to hemoglobin is coordinated to facilitate oxygen binding in the lungs and its release in the tissues where oxygen pressure is slightly lower.

Among proteins with a known three-dimensional structure, there are a few recurring structures that predominate. These structural superfolds include closely related proteins, which have a high level of identical amino acids and perform similar functions, as well as unrelated proteins, which have independent amino acid sequences and function despite overall similarity in architecture. Nine superfolds dominate the known protein structures. These are the globin, trefoil, up-down, immunoglobulin, $\alpha\beta$ sandwich, jelly roll, doubly wound, UB $\alpha\beta$ roll, and TIM barrel folds (Fig. 2). Hemoglobin and myoglobin are related and are representative of members of the globin fold family, which is characterized by eight α -helices arranged in a Greek key motif (the topological signature of many β -barreis and a majority of β -sandwich structures). Antibodies contain

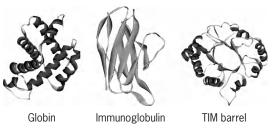


Fig. 2. Three commonly occurring protein folds.

several repeating segments that have eight-stranded β -sheet structures belonging to the immunoglobulin fold family. Several enzymes that carry out the chemical reactions of metabolism are members of the TIM barrel fold family. This fold is made of eight β -strand/ α -helix structures wrapped into a barrel. The namesake enzyme TIM (triose-phosphate isomerase) is used in the breakdown of sugars. Another member of the TIM barrel fold family is ribulose 1,5-bisphosphate carboxylase oxygenase (Rubisco), a plant protein that fixes carbon dioxide from the air into organic carbon needed to sustain life and is the most abundant protein on Earth. While many proteins are found to adopt one of the superfolds, many others have a unique architecture.

Preparation. Many of the structural and functional studies of proteins are performed with purified preparations of proteins. Historically, proteins were obtained from their tissue of origin, but modern DNA technology allows proteins to be produced in high yield in bacteria, yeast, or isolated cells from insects and mammals. The isolation of a protein from any of these sources is accomplished by a variety of techniques. The primary concern is that the protein be isolated in its native form—in other words, not denatured or degraded during preparation.

All purification methods seek to exploit differences in solubility, charge, size, and resin-binding specificity in order to enrich the solution for the desired protein. The first step in preparation is usually to mechanically disrupt the tissue or cells containing the protein. Next, any of several methods are applied, often sequentially, to purify the protein. A few of the commonly applied methods are described.

Salting out. High concentrations of neutral salt tend to precipitate proteins; the salt concentration necessary for such precipitation varies from protein to protein. The most commonly used salt for this purpose is ammonium sulfate, which has the advantages of high solubility at the low temperatures (around freezing temperature to room temperature) used to isolate proteins. By adding salt stepwise to a protein solution and separating the precipitate by centrifugation at each step, fractions are obtained in which the desired protein is concentrated in one of them.

lon-exchange chromatography. Ion-exchange resins and modified cellulose derivatives are powerful tools for protein purification. A solid resin is chemically modified to introduce charged groups for anion (negative charge) or cation (positive charge) binding. Proteins, which typically contain many positive and negative charges on their surfaces, will interact with a specific resin with different binding affinities. The strength of the protein-resin interaction can be modulated by altering the pH of the solution and thus the protonation state of the protein, as well as the ionic strength of the buffer. Buffers, mixed to produce a gradient of pH and ionic strength, are used to differentially bind and elute proteins from the resin. Ion-exchange chromatography has the capacity to handle gram quantities of proteins. Low-molecular-weight proteins, especially those with markedly acidic (positively charged) or basic (negatively charged) properties, such as ribonuclease, histones, and lysozyme, can be separated from most other proteins in a mixture in a single experiment on an ion-exchange column. See CHROMATOGRAPHY.

Affinity chromatography. The resolving power of ionexchange chromatography can be remarkably enhanced by the covalent attachment of a small molecule or of a protein that has an affinity for the target. For example, streptavidin is a protein from Streptomyces bacteria that tightly binds biotin. It is readily purified by binding to biotin-conjugated resins, which has thus led to its use as a handle for purification of other proteins. Therefore, through protein engineering, streptavidin is covalently linked to a second protein of interest. The two linked proteins can be purified by biotin affinity chromatography. Immunoglobulins, or antibodies, are tightly bound by protein A, made by Staphylococcaus bacteria. Purification of immunoglobulins by affinity chromatography on protein A-coupled resin is an important method of isolating commercially produced immunoglobulins for medical therapies in diseases ranging from arthritis to cancer. In some cases, total purification in one step can be achieved with affinity chromatography. See PROTEIN ENGINEERING.

Countercurrent distribution. Countercurrent distribution is an effective system for purification that can separate peptides based on their differential solubilities in two immiscible liquids. However, the range of application has been narrow. The power of the method was shown by the successful separation of insulin into two components differing by only a single amino acid. This method has been much more widely utilized in fractionation of peptides than in fractionation of proteins. *See* EXTRACTION.

Preparative ultracentrifugation. Proteins of high molecular weight can be concentrated by applying a high gravitational field to the protein, contained in a tube, for a prolonged time. An ultracentrifuge produces a high gravitational field in which the protein sediments through the field, concentrating protein at the bottom of the tube. However, it is possible to achieve significant purification only for the heaviest or the lightest components in a mixture, and the degree of purification will depend on the range of sedimentation velocities represented. Solutions with densities less than that of the solvent allow proteins to float to the surface under the influence of a strong gravitational field. This principle has been successfully applied to the study of serum lipoproteins which, by virtue of their lipid content, have solution densities less than those of other, non-lipid-containing serum proteins. Sufficient salt is added to the serum to raise the solution densities of non-lipid-containing proteins above those of the lipoproteins. Centrifugation to equilibrium brings the lipoproteins to the surface, where they can be collected. This technique has been refined by the use of repeated centrifugation steps at progressively raised salt densities, allowing subfractions of lipoproteins of different solution density classes to be obtained on a preparative scale. *See* ULTRACENTRIFUGE.

Gel filtration. Gel filtration is a technique for the purification of proteins that separates mainly on the basis of molecular weight. Dextran beads, just micrometers in diameter, are used to exclude proteins of high molecular weight but to permit smaller proteins to diffuse in and out. With a column of such beads, a continuous redistribution and separation occurs as the protein mixture moves through the column. Beads of different porosities are available to separate proteins of different molecular weights. For example, with one type of dextran bead, proteins with molecular weights of about 10,000 can be easily separated from those whose molecular weights are greater than 50,000. *See* FILTRATION.

Analysis. Methods for protein analysis include amino acid analysis, protein sequencing, mass spectrometry, isoelectric focusing, gel electrophoresis, and specific activity.

Amino acid analysis. Because of the relatively high molecular weights of proteins, the classical methods of organic chemistry are not adequate to establish composition or structure. A more meaningful empirical formula for a protein is given in terms of the total number of amino acids and the relative proportion of each amino acid. Thus, a protein can be hydrolyzed in hydrochloric acid to liberate the individual amino acids, which can be quantitatively estimated by amino acid analysis. These data, together with an independent measurement of molecular weight (that can be converted into the total number of amino acids), yield the amino acid composition, that is, the number and kind of each amino acid. In those few cases in which a protein is known to lack a particular amino acid (for example, tryptophan is absent from ribonuclease, and isoleucine is absent from human adult hemoglobin), the amino acid composition can be used to estimate the purity of a sample.

Protein sequencing. Developed by Pehr Edman in the 1950s, sequential degradation (one amino acid at a time from one end of the chain) of proteins and peptides was used for decades to determine the sequence. The procedure was done by coupling the free N-terminal α -amino acid with a chemical (particularly phenylisothiocyanate) that removes the amino acid (under appropriate conditions) and leaves the remainder of the polypeptide chain intact. It is then possible to identify the cleaved derivative of the first amino acid and repeat the process on each truncated peptide that is successively generated.

Protein sequencing is rarely used anymore to determine the complete sequence of amino acids

in a protein, because deduced protein sequences are available from the DNA sequences of many organisms. However, it remains a frequently used method to determine protein identity. Identity can be uniquely established from the sequence of the first six to ten amino acids and comparison of that sequence to those predicted from the sequences of the genes in the organism from which the protein was isolated.

Not all proteins generate useful sequence information from the amino terminus, based on reaction with phenylisothiocyanate. Sometimes the amino terminus of some proteins is blocked, or covalently modified, so the Edman method cannot work. To circumvent this problem, proteins can be randomly broken down into smaller polypeptides or peptides. The sequences of the individual peptides are then determined. However, the sequences of these pieces have to be combined in the correct order, and this can be challenging. Alternatively, stepwise removal of amino acids from the carboxyl end by enzymes called carboxypeptidases is used to determine the amino acid sequence at the opposite end of the polypeptide. *See* PROTEIN DEGRADATION.

Mass spectrometry. Powerful new applications of mass spectrometry have catalyzed its use in biology. Mass spectrometry has become an important tool for protein sequencing and identification, and for localization of modifications that occur after the protein has been synthesized. Mass-spectrometric analysis yields the molecular weight of the protein. The masses of the peptides in a proteolytic digest of a protein can also be determined. Then by comparison to protein sequence databases, the protein can be identified.

lsoelectric focusing. The net charge of a protein is largely a function of the relative number of basic (lysine, histidine, and arginine) and dicarboxylic amino acids (aspartic and glutamic acids). This net charge strongly influences the solubility of a protein at different pH values because the solubility depends in part on the proportion of polar groups. When the hydrogen-ion concentration is high (low pH), the net charge is positive; when the hydrogen-ion concentration is low (high pH), the net charge is negative. The pH at which the net charge of the protein is zero is defined as the isoelectric point (pI). Isoelectric focusing-the focusing by electrophoresis of individual proteins according to their respective pI's-is usually carried out in a vertical column that contains a sucrose concentration gradient or in a semisolid phase of dextran. Synthetic polymers of amino acids (polyamino acids) of varying isoelectric points are added to the protein mixture that is to be resolved. In an electric field, each synthetic polymer migrates to the point where it has zero net charge. Thus, a stable gradient of these polyamino acids is established that, in effect, is a pH gradient. The added proteins then are separated according to the intrinsic pI (isoelectric point) of each one. See ISOELECTRIC POINT; PH

Gel electrophoresis. Protein samples are analyzed for purity and size by various methods of polyacrylamide

gel electrophoresis (PAGE). Under native conditions, that is, in the absence of denaturing agents, proteins migrate according to their charge density and size and shape. At slightly basic pH, most proteins carry a net negative charge. The lower the charge density (less charge per molecular mass), the slower a protein migrates. The frictional force of the gel matrix creates a sieving effect, retarding the movement of proteins according to size.

Under denaturing conditions, mixtures of proteins are separated by PAGE based on size. Sodium dodecyl sulfate (SDS) is added to denature the protein. (Regardless of the shape of the native, undenatured protein, all proteins assume a rod shape when denatured. As a consequence, the shape differences between proteins are eliminated.) SDS is negatively charged and has an inherent tendency to stick to the protein backbone in an amount proportional to a protein's molecular weight. Thus, each SDSprotein complex migrates through the gel according to molecular weight. The size of a protein is calculated from the distance it migrates in the gel, when compared to proteins of known molecular mass.

Two-dimensional electrophoresis combines the resolving power of isoelectric focusing and denaturing PAGE to separate complex mixtures of proteins. To separate proteins that migrate at a single pI, SDS-PAGE is run at a right angle to the direction of the first (isoelectric focusing) electrophoresis. By these methods, thousands of proteins can be separated into a two-dimensional array of "spots" composed of single pure proteins. This technique is widely used in the field of research known as proteomics.

Specific activity. Function is the ultimate outcome of structure. Therefore, the function or activity of a protein should be measured whenever possible. The activity of a protein refers to the specific function of the native protein. In the case of an enzyme, the activity is the catalysis of a chemical transformation of substrate into product. Enzyme activities are usually reported in activity units such as micromoles of product formed per minute. The specific activity of an enzyme is the total activity divided by the amount of protein. During the purification of an enzyme, specific activity increases because contaminating proteins are removed.

For example, luciferase is an enzyme found in the light-producing organ of fireflies, which create light to signal other fireflies. The enzyme catalyzes a chemical reaction that converts luciferin, adenosine triphosphate, and oxygen (O_2) into oxyluciferin, adenosine monophosphate, pyrophosphate, and light. Thus, one way to measure the activity of firefly luciferase is to measure the amount of light produced.

A pure, completely native enzyme will have the highest specific activity. As the protein becomes degraded or denatured, the specific activity will be reduced. Mutations that affect (negatively or positively) the functioning of an enzyme can be assessed by measuring the specific activity and comparing it to that of the wild-type enzyme. Those mutations that disrupt activity are important because their locations in the structure can identify the residues needed for catalysis. Karla L. Ewalt; Paul Schimmel; Gertrude E. Perlmann; James M. Manning

Bibliography. M. C. Anson et al. (eds.), Advances in Protein Chemistry, vol. 40, 1990; C. F. Barbas III et al., Semisynthetic combinatorial antibody libraries: A chemical solution to the diversity problem, Proc. Nat. Acad. Sci. USA, 89:4457-4461, 1992; S. J. Benkovic, Catalytic antibodies, Annu. Rev. Biochem., 61:29-54, 1992; R. A. Bradshaw et al. (eds.), Molecular Biology of Intracellular Protein Sorting and Organelle Assembly, 1988; J. R. Jacobsen et al., An efficient antibody-catalyzed aminoacylation reaction, Science, 256:365-367, 1992; P. Schimmel, Alanine transfer RNA synthetase: Structure-function relationships and molecular recognition of transfer RNA, Adv. Enzymol. Related Areas Mol. Biol., 63:233-270, 1990; G. E. Schultz and R. H. Schirmer, Principles of Protein Structure, 1988; W. Voelter et al. (eds.), Chemistry of Peptides and Proteins, vol. 3, 1985.

Protein degradation

The enzymatic destruction of proteins. Like other macromolecular (large) components of an organism, proteins are in a dynamic state of synthesis (generation) and degradation (destruction, proteolysis). During proteolysis, the peptide bond that links the amino acids to one another is hydrolyzed, and free amino acids are released. The process is carried out by a group of enzymes called proteases. *See* ENZYME; PROTEIN.

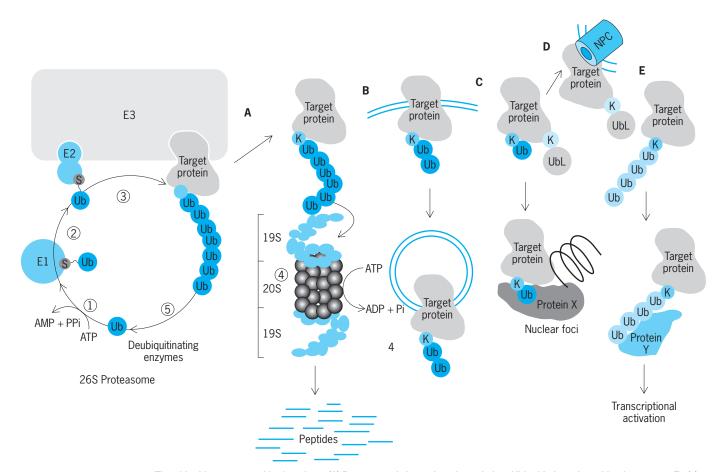
Mechanisms. Distinct proteolytic mechanisms serve different physiological requirements. It is necessary to distinguish between destruction of "foreign" and "self" proteins. Foreign proteins contained in the diet are degraded "outside" the body, in the digestive tract. This is necessary in order to prevent a response of the immune system to the intact proteins. The released amino acids are then absorbed and serve to build the organism's self proteins. Self proteins can be divided into extracellular and intracellular. The two groups are degraded by different mechanisms. Extracellular proteins that in solution in the plasma (the blood fluid phase), such as blood coagulation factors, immunoglobulins, cargo carrier proteins (such as the low-density lipoprotein, LDL, that carries cholesterol; or transferrin that carries iron), and peptide hormones (such as insulin) are taken up by pinocytosis (engulfment of small droplets) or following binding to specific receptors. They are carried via a series of vesicles (endosomes) that fuse with lysosomes, where they are degraded by the proteases contained in these organelles. During this process, the extracellular proteins are never exposed to the intracellular environment (cytosol). Degradation of proteins in lysosomes is not specific, and all engulfed proteins are degraded at similar rates.

Several observations lead to the prediction that

degradation of many intracellular proteins must be carried out by a completely different and nonlysosomal mechanism. In contrast to lysosomal degradation, the destruction of many cellular proteins is specific, and different proteins have half-life times that range from a few minutes (such as the tumor suppressor protein p53) to several days (such as the muscle proteins actin and myosin). Furthermore, the stability of certain proteins changes under different conditions. For example, cell-cycle regulators such as mitotic cyclins are stable throughout the cell cycle and are degraded only during the short mitotic time window, an event essential for the cell in order to exit mitosis. Other proteins are degraded only following a specific signal that is transmitted from the extracellular environment, and their degradation serves to turn off the signaling event. Also, inhibitors of lysosomal degradation do not affect these processes, a finding that suggests it occurs in the cytosol. Taken together, it has become clear that the proteolytic system involved in intracellular protein degradation must be distinct from lysosomal proteolysis. The discovery of the ubiquitin-proteasome proteolytic pathway has resolved this enigma.

Degradation of a protein via the ubiquitin pathway involves two steps: tagging of the substrate by covalent attachment of multiple ubiquitin molecules, and degradation of the tagged protein by the 26S proteasome with release of free and reusable ubiquitin. In addition to serving as a proteolytic signal, ubiquitination can regulate other processes. Depending on the character of the internal linkage between the ubiquitin moieties, polyubiquitination can also lead to activation of transcriptional regulators. Modification by a single moiety of ubiquitin can target proteins for degradation in the lysosome. In addition to ubiquitin, other modifiers that belong to the ubiquitin family of proteins have been discovered, and designated ubiquitin-like proteins, UbLs. Modification by these proteins can serve a variety of nonproteolytic functions, such as activation of enzymes or routing of the tagged proteins to their subcellular destination. Thus, ubiquitination has emerged as a critically important posttranslational modification that plays major roles in regulating a broad array of basic cellular processes. Aberrations in the system have been implicated in the pathogenesis of many diseases, converting the system into a platform for drug targeting (for a scheme of the ubiquitin system and its multiple functions, see the illustration).

Ubiquitination. Conjugation of ubiquitin, a 76residue polypeptide, to the protein substrate proceeds via a three-step cascade mechanism catalyzed by these enzymes: E1, E2, and E3. E1 activates ubiquitin, and an E2 transfers it from E1 to the substrate that is bound to an E3. E3 enzymes catalyze the last critical step—covalent attachment of ubiquitin to the substrate. Following conjugation of the first ubiquitin moiety, other ubiquitin moieties are added to one another to generate a polyubiquitin chain. The chain is recognized by the downstream 26S proteasome. E3's play a key role in the ubiquitin system, as they serve as the specific substrate recognition elements



The ubiquitin system and its functions. (A) Proteasomal-dependent degradation. Ubiquitin is activated by the enzyme E1 (1), followed by its transfer to a ubiquitin-carrier protein E2 (2). E2 transfers the activated ubiquitin to the protein substrate that is bound specifically to a unique ubiquitin ligase E3 (3). Successive conjugation of ubiquitin moieties to one another (via Lys⁴⁸ of the previous ubiquitin moiety) generates the polyubiquitin chain that serves as the degradation signal for the downstream 26S proteasome (4). The substrate is degraded to short peptides (4), and free and reusable ubiquitin is released by deubiquitinating enzymes (5). (B) Mono- or oligoubiquitination targets membrane proteins to degradation in the lysosome/vacuole. (C) Monoubiquitination or (D) a single modification by a UbL protein, SUMO for example, can target proteins to different subcellular destinations such as nuclear foci or the nuclear pore complex. (E) Generation of a polyubiquitin chain in which the moieties are bound to one another in a different manner than in the chain recognized by the proteasome (via Lys⁴⁸) can activate transcriptional regulators. Ub denotes ubiquitin, K denotes Lys, and S denotes Cys (to which ubiquitin binds on E1, E2, and certain E3's, prior to its transfer to the substrate).

of the system. Approximately 1000 different E3's have been identified in the human genome. UbLs are activated in a similar, albeit not an identical, mechanism.

Degradation. The proteasome is a multicatalytic protease that degrades polyubiquitinated, but not untagged, proteins. It is composed of two subcomplexes: a 20S core particle (CP) that carries the catalytic activity, and a 19S regulatory particle (RP). The 20S core particle is a barrel-shaped structure composed of four stacked rings: two identical outer seven-subunits-each α rings and two identical inner seven-subunits-each β rings. The catalytic sites are localized to some of the β subunits. Each extremity of the 20S can be capped by a 19S regulatory particle. One important function of the 19S regulatory particle is to recognize ubiquitinated proteins. Other functions of the 19S regulatory particle are to unfold the substrate and to open an orifice in the α ring to allow entry of the substrate into the proteolytic chamber. Following degradation, short peptides are

generated, as well as reusable ubiquitin that is released by deubiquitinating enzymes; one of them is a subunit of the 19S regulatory particle.

Substrate recognition. A major problem is how the ubiquitin system recognizes its innumerable substrates in a specific manner. Why are certain proteins stable while others are extremely short-lived? Why are certain proteins degraded only at a particular time point along the cell cycle or following specific stimuli while they are stable under other conditions? Specificity of the ubiquitin system is determined by two distinct groups of proteins: E3's, and modifying enzymes and ancillary proteins. Within the ubiquitin system, substrates must be recognized and must bind to a specific E3 as a prerequisite to their ubiquitination. In a few cases, recognition is constitutive (the targeting motif can be a single amino acid residue such as the N-terminal residue). In most cases, however, the substrates must undergo a posttranslational modification such as phosphorylation (for example, the two neighboring Ser residues in the case of $I\kappa B\alpha$, an inhibitor of nuclear factor κB) or oxidation (hydroxyproline in the hypoxia-inducible factor- 1α subunit, HIF- 1α), or must associate with a molecular chaperone in order to be recognized. In some instances, it is the E3 that must be "switched on" by undergoing posttranslational modification (phosphorylation is required, for example, to activate the proto-oncogene c-Cbl).

Functions of the ubiquitin system. The ubiquitin system is involved in regulating numerous basic cellular processes. Among them are cell division, differentiation and development, the cellular stress response, DNA repair, transcriptional regulation, and modulation of the immune and inflammatory responses. The list of cellular proteins that are targeted by ubiquitin is growing rapidly. The proteins include cell cycle regulators, tumor suppressors, transcriptional activators and their inhibitors, cell surface receptors, and endoplasmic reticulum proteins. Mutated/denatured/misfolded proteins are recognized specifically and removed efficiently. Thus, the system plays a key role in the cellular quality control and defense mechanisms.

Regulation of the ubiquitin system. The ubiquitin system can be regulated at the level of ubiquitination or of proteasome activity. Since conjugation and proteasomal degradation are required for a multitude of cellular functions, regulation must be specifically tuned. In a few cases, general rather then specific components of the pathway are modulated. For example, upregulation of the pathway is observed during massive degradation of skeletal muscle proteins that occurs following long fasting, and also under pathological conditions such as malignancy-induced cachexia, severe sepsis, metabolic acidosis, or following denervation. In most cases, however, regulation is specific, affecting defined substrate(s) and a distinct ligase.

Human diseases. Inactivation of a major enzyme such as E1 is obviously lethal. Yet, mutations or acquired changes in enzymes or recognition motifs in substrates that do not affect vital pathways or that affect the involved process only partially, may result in a broad array of diseases. The pathological states associated with aberrations in the system can be classified into two groups: those that result from loss of function—mutation in a ubiquitin system enzyme or target substrate that results in stabilization of the latter; and those that result from gain of function—abnormal or accelerated degradation of the substrate. A few examples can highlight the two states—accelerated and inhibited degradation.

Malignancies. Alterations in ubiquitination and deubiquitination reactions have been directly implicated in the etiology of several malignancies. In general, cancers can result from stabilization of oncoproteins or destabilization of tumor suppressors. *See* CANCER (MEDICINE).

In the case of uterine cervical carcinoma, the level of the tumor suppressor protein p53 is extremely low. Most of these malignancies are caused by high-risk strains of the human papillomavirus. It was shown that p53 is targeted for destruction by the virally encoded oncoprotein E6. Degradation is mediated by the naturally occurring E3, E6associated protein (E6-AP), where E6 serves as an ancillary protein that allows recognition of p53. E6-AP will not recognize p53 in the absence of E6. In the absence of p53, the virus can integrate its genome into the host genome, an event that would have otherwise caused p53 to induce apoptosis. While the native substrates of E6-AP are still obscure, a mutation in the enzyme has been implicated as the cause of Angelman syndrome, which is characterized by mental retardation, seizures, out-of-context laughter, and abnormal gait.

 β -Catenin plays an important role in differentiation of the colorectal epithelium, and aberrations in ubiquitin-mediated regulation of its levels play an important role in the development of colorectal tumors. Under resting conditions, β -catenin is phosphorylated, which leads to recruitment of the β -TrCP ubiquitin ligase ε 3, followed by its ubiquitination and degradation. Signaling promotes dephosphorylation, stabilization, and subsequent activation of β catenin via complex formation with other inactive subunits. β -Catenin interacts with the tumor suppressor APC (adenomatous polyposis coli) to generate a complex that appears to regulate its level. Aberrations in degradation of β -catenin that lead to its stabilization, accumulation, and subsequent oncogenic activation can result from two different mechanisms: mutations in the phosphorylation motif of the protein and mutations in the APC/axin machinery (APC and axin are interacting proteins that regulate the cytoplasmic levels of β -catenin by binding to it).

Aberrations in components of the ubiquitination machinery can also cause malignancies. Mutations in the von Hippel-Lindau (VHL) gene predisposes individuals to a wide range of highly vascularized malignancies such as renal cell carcinoma and cerebellar hemangioblastomas. The tumors arise from constitutive expression of the hypoxia inducible factor subunit HIF-1 α , which results in overexpression of hypoxia-inducible gene products such as the vascular endothelial growth factor, the erythropoietin receptor, and glycolytic enzymes. pVHL is a subunit in a ubiquitin ligase that targets HIF-1 α for ubiquitination and degradation. Under normoxic conditions, HIF- 1α is hydroxylated specifically on Pro residue 564 to generate hydroxyproline. The hydroxylated Pro is recognized by the pVHL E3 complex that targets the molecule for ubiquitination and degradation. Under hypoxic conditions, HIF-1 α is stable, as the efficiency of the hydroxylation reaction under these conditions is low. Loss of VHL function due to a mutation stabilizes HIF- α , which can explain the stimulation of vascular growth in these tumors.

Neurodegenerative disorders. Accumulation of ubiquitin conjugates of certain proteins in inclusion bodies have been reported in a broad array of chronic neurodegenerative diseases, such as Alzheimer's disease and Parkinson's disease. However, in all these cases a direct pathogenetic linkage to aberrations in the ubiquitin system has not been established. *See* ALZHEIMER'S DISEASE; PARKINSON'S DISEASE.

The case of Parkinson's disease highlights the complexity of the involvement of the ubiquitin system in the pathogenesis of neurodegeneration. Mutations in several proteins have been described that link the system to the pathogenesis of the disease. One important player is Parkin, which is an E3 ubiquitin ligase. Mutations in the gene appear to be responsible for the pathogenesis of autosomal recessive juvenile parkinsonism (AR-JP), a rare disease by itself, but one of the most common familial forms of the disease. Parkin ubiquitinates and promotes the degradation of several substrates, such as CDCrel-1, a synaptic vesicle-enriched septin GTPase, or the Pael receptor, a putative G protein-coupled endoplasmic reticulum transmembrane polypeptide. It is possible that aberration in the degradation of one or more of these substrates that leads to its accumulation is neurotoxic and underlies the pathogenesis of AR-JP.

Cystic fibrosis. The cystic fibrosis gene encodes the cystic fibrosis transmembrane regulator (CFTR), which is a chloride channel. One frequent mutation in the channel is Δ F508. The mutation leads to a multisystem disorder characterized by chronic obstruction of airways and severe maldigestion due to exocrine pancreatic dysfunction. CFTR^{Δ F508} does not reach the cell surface and is retained misfolded in the endoplasmic reticulum, where it is subsequently degraded.

Immune and inflammatory disorders. As described above, the human papillomavirus evolved a mechanism for proteolytic removal of p53. Another interesting example of a virus that evolved a mechanism that exploits the ubiquitin system is the human cytomegalovirus. The virus encodes two endoplasmic reticulum proteins, US2 and US11, that downregulate the expression of MHC (major histocompatibility complex) class I heavy-chain molecules. The MHC molecules are synthesized and transported to the endoplasmic reticulum, but shortly thereafter, in cells expressing US2 or US11, they are transported back to the cytosol, deglycosylated, and degraded by the proteasome following ubiquitination. Thus, the virus inactivates the cell's ability to present antigens to cytotoxic T cells, which enables it to replicate and propagate.

Drug development. Because of the central role that the ubiquitin system plays in a broad array of basic cellular processes, development of drugs that modulate the activity of the system may be difficult. Inhibition of enzymes common to the entire pathway, such as the proteasome, may affect many processes nonspecifically, although a narrow window between beneficial effects and toxicity can be identified for a short-term treatment. Indeed, such a drug, Velcade, has been shown to be beneficial in multiple myeloma, a malignancy that affects bone marrow.

A completely different approach to drug targeting can be the development of molecules that interfere with the activity of specific E3's. Thus, a drug that will bind specifically to the p53 pocket in Mdm2, the E3 of p53 that is upregulated in many tumors and acts as an oncogene, will prevent association of the enzyme with p53, which results in upregulation of the tumor suppressor. Aaron Ciechanover; Kazuhiro Iwai

Bibliography. A. Ciechanover, From the lysosome to ubiquitin and the proteasome, Nat. Rev. Mol. Cell Biol., 6:79-86, 2005; C. de Duve and R. Wattiaux, Functions of lysosomes, Annu. Rev. Physiol., 28:435-492, 1966; M. H. Glickman and A. Ciechanover, The ubiquitin-proteasome pathway: Destruction for the sake of construction, Pbysiol. Rev., 82:373-428, 2002; A. L. Goldberg and A. C. St. John, Intracellular protein degradation in mammalian and bacterial cells: Part 2, Annu. Rev. Biochem., 45:747-803, 1976; G. E. Mortimore and A. R. Poso, Intracellular protein catabolism and its control during nutrient deprivation and supply, Annu. Rev. Nutr., 7:539-564, 1987; R. T. Schimke and D. Doyle, Control of enzyme levels in animal tissues, Annu. Rev. Biochem., 39:929-976, 1970; R. Schoenheimer, The Dynamic State of Body Constituents, Harvard University Press, Cambridge, MA, 1942.

Protein engineering

The design and construction of new proteins or enzymes with novel or desired functions, through the modification of amino acid sequences using recombinant DNA (deoxyribonucleic acid) technology. The sizes and three-dimensional conformations of protein molecules are also manipulated by protein engineering. The basic techniques of genetic engineering are used to alter the genes that encode proteins, generating proteins with novel activities or properties. Such manipulations are frequently used to discover structure-function relationships, as well as to alter the activity, stability, localization, and structure of proteins. *See* GENETIC ENGINEERING; PROTEIN.

Point mutants. Many subtle variations in a particular protein can be generated by making amino acid replacements at specific positions in the polypeptide sequence. Each protein is unique by virtue of the sequence of its amino acids. At any position in the sequence, an amino acid can be replaced by another to generate a mutant protein that may have different characteristics by virtue of the single replaced amino acid. For example, pancreatic ribonuclease A is an enzyme comprising 124 amino acids that cleaves the covalent bonds that join ribonucleic acids (RNA). If at position 119 in the sequence the naturally occurring histidine is replaced with an alanine, the mutant protein is referred to as a histidine $119 \rightarrow alanine$ (H119A) mutant of ribonuclease A. This mutant protein is expected to have little or no biological activity, because histidine 119 is important for that activity. Other mutations have very little effect on their proteins. This is particularly true when the amino acid being changed is substituted with other closely related amino acids and when the amino acid is not conserved in the same protein found in other organisms. Typically, site-directed protein engineering targets amino acids that are involved in a particular biological activity. See AMINO ACID; MUTATION.

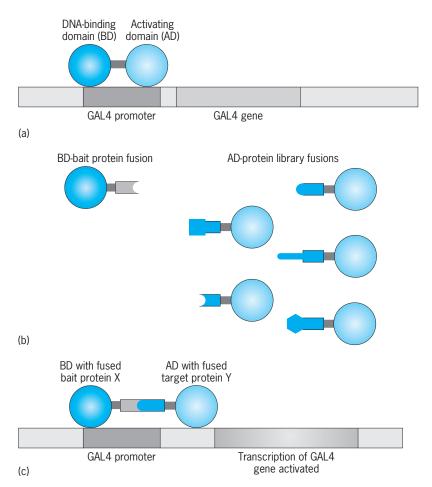
Because there are 20 different amino acids, 20

different variants of ribonuclease A can be created just by having a set that differs only by the amino acid at position 119. If changes are also made at another position, for example, position 41 (where normally there is a lysine), then in principle there are 400 different variants that can be created by making all possible combinations of substitutions at just these two positions in the sequence. As more positions are varied, the number of combinations becomes enormous so that, with just six different positions subjected to all possible variations, 64 million different proteins can be generated. These proteins will share the identical sequence, except at the six varied positions.

Deletion mutants. In addition to the substitution of amino acids at specific positions, amino acids can be deleted from the sequence, either individually or in groups. These proteins are referred to as deletion mutants. Deletion mutants may or may not be missing one or more functions or properties of the full, naturally occurring protein. Proteins from eukaryotes, such as mammals, plants, and fungi, can have very long amino acid sequences and as a consequence tend to be organized as several modular protein domains linked together. Deletion mutants have been a useful way to create smaller proteins that contain only one or a few of these domains so that the individual properties can be studied.

Hybrid/fusion proteins. Protein sequences can be joined or fused to that of another protein. The resulting protein is called a hybrid, fusion, or chimeric protein, which generally has characteristics that combine those of each of the joined partners. Protein fusions have been extensively used to study interactions between two or more proteins. For example, an application of protein fusion methods called the yeast two-hybrid screen was developed to identify proteins that interact with each other (see illustration). The system was developed by separating a yeast transcription-activating protein into two functional domains. The first domain of the transcription-activating protein is fused to a protein that is being studied for interactions. The second domain of the transcription-activating protein is fused to a collection of proteins encoded in a protein library. The first fusion protein and one from the protein library are expressed together in yeast. When the two fusion proteins interact, they can bring together the two halves of the transcription activator and turn on the expression of a reporter gene (see illustration). Another example is the fusion of a protein to an inherently fluorescent protein, such as the green fluorescent protein (GFP) from the jellyfish Aequorea victoria. Fluorescent fusion proteins have been used to study the location, movement, appearance, and degradation of proteins in living cells. See FLUORESCENCE.

Another common application for fusion proteins is to facilitate purification by affinity chromatography. A short sequence of amino acids or a protein domain is fused to one end of the polypeptide. Six histidines can be used to purify a protein by affinity to nickel. The Fc region of an antibody (region of an antibody molecule that binds to cell-surface antibody recep-



Yeast-two-hybrid screen. (a) The yeast transcription-activating protein GAL4 has two separate domains: a DNA-binding domain (BD), which binds the promoter region of the GAL4 gene, and a transcription-activating domain (AD), which stimulates production of the GAL4 RNA transcript. (b) These two domains are modular and can be fused onto other proteins while retaining their function. This forms the basis of the yeast two-hybrid assay (screen). The first hybrid, or fusion, protein consists of a protein of interest (bait protein) fused to the GAL4 AD and a library of many proteins. (c) If a protein from the library is able to interact with the first fusion protein, visualized in the figure as a protein with a complementary shape, the GAL4 transcription factor is reconstituted, and transcription of the gene is activated. Proteins that do not interact, those with noncomplementary shapes, fail to activate the system.

tors) can be fused to a protein to purify it by affinity to protein A, a protein from bacteria that specifically binds antibodies. *See* LIQUID CHROMATOGRAPHY.

Novel activities or properties. Point and deletion mutants and hybrid proteins are constructed to obtain polypeptides with new properties. These proteins are either created individually, by site-directed mutagenesis, or they are generated as a large pool or library of millions of variants. The library is then screened or subjected to a special selection procedure to obtain the protein or proteins with the desired characteristics.

Proteins generally have stable and unique threedimensional structures and are active at physiological temperatures of 37° C (98.6°F). They usually lose their three-dimensional shapes as the temperature is raised more than 5 or 10° above 37° C and, as a consequence, lose their biological activities. In some instances, a protein may be cold-sensitive; that is, it may lose its conformation and activity at lower temperatures. The effect of temperature on protein stability can be modified by protein engineering, in particular, by introducing amino acid replacements that enhance or destabilize the molecular packing interactions in the core of the protein structure, such as has been done with the enzymes T4 lysozyme and staphylococcal nuclease.

Protein engineering has been used to produce therapeutic proteins with improved properties such as increased solubility and stability. For example, insulin was engineered through mutagenesis to create monomeric forms that are fast acting (insulin lispro and insulin aspart). Conversely, another form (insulin glargine) was created by mutagenesis to precipitate upon injection and give a sustained release of insulin. Mutation of a free cysteine in aldesleukin (a synthetic version of interleukin-2 used to treat some forms of cancer) or interferons beta-1b (used to treat relapsing-remitting multiple sclerosis) was used to produce therapeutics with decreased aggregation. *See* CANCER (MEDICINE); INSULIN; INTERFERON; INTERLEUKIN; MULTIPLE SCLEROSIS.

Proteins can be engineered to acquire new biological activities. For example, a catalytic antibody is a variant of an antibody. Antibodies are proteins that normally bind to a specific molecule but do not alter the bound molecule in any way. A catalytic antibody is one which has been changed by mutations to have a novel sequence that folds into a structure that catalyzes a specific reaction (such as amide bond formation, ester hydrolysis, and decarboxylation). Catalytic antibodies function like enzymes, and are created to catalyze reactions for which there are no naturally occurring enzymes. Fifty or more different reactions have been catalyzed by the action of catalytic antibodies that were obtained individually by methods of protein engineering. See ANTIBODY; CATALYTIC ANTIBODY.

Structure-function relationships. With over 180 genomes sequenced, including human and mouse, the amino acid sequence of a particular protein is now generally available from many different organisms. This advance during the past decade has created situation in which protein sequences are available even when there is no information on the biological activity of the protein. Thus, the focus has shifted from sequencing to understanding the function of all of these proteins. Techniques such as the yeast two-hybrid system have been applied to identify protein interactions, and crystallographic structures are being determined to elucidate biological information about all the proteins in an entire organism. Still, the chemical basis for activity is often not completely understood, even after the biological activity and high-resolution structure have been determined. The construction of mutant proteins can elucidate the role of a particular amino acid at a specific position in the sequence. See GENE.

Protein engineering has been applied to understand how enzymes catalyze reactions, from identifying which amino acids are essential for catalysis to analyzing how amino acid changes alter particular aspects of a reaction such as substrate specificity. Antibiotics such as erythromycin are made by large multidomain proteins called polyketide synthases. Site-directed mutagenesis has been used to modify the substrate specificity of one polyketide synthase reaction so that the product contains a malonate unit, whereas the product of the original enzyme contained a methylmalonate unit. In addition to site-directed mutagenesis, the order of the polyketide synthase domains have been shuffled to create proteins that could catalyze the synthesis of new antibiotics. *See* ANTIBIOTIC; ENZYME.

An extension of site-directed mutagenesis allows nonnatural amino acids to be incorporated into proteins. Nonnatural amino acids are not naturally encoded by the genome, but instead include a wide variety of amino acids that are present in cells or produced by synthetic methods. This protein engineering method allows the chemical properties of a particular amino acid to be changed beyond that naturally encoded by a gene. For example, nonnatural amino acids incorporated into proteins may contain sugars, nucleophiles (electron donors), electrophiles (electron acceptors), crosslinkers (a chemical compound that forms covalent bonds between adjacent polymer chains that lock the chains in place), or altered shapes and sizes. Detailed studies on ion channels with over 60 nonnatural amino acids incorporated illustrate the chemical power of this approach.

Minimalist proteins. Proteins are usually large molecules composed of several hundred or even more than a thousand amino acids. Yet, the portion of the protein responsible for a specific biological activity is usually concentrated in a small part of the structure. In these circumstances, there are advantages to creating a minimalist protein having enough of the structure to retain the desired activity. A smaller protein is often superior for analyzing and manipulating structure-function relationships, and the reduced cost of materials for manufacturing a small protein can also be an important motivation for reducing the size.

To obtain minimalist proteins with novel properties, libraries with large numbers of different amino acid sequences have been created. These sequences are often so short (less than 50 amino acids) that they are referred to as peptides, oligopeptides, or polypeptides, and not as proteins. (A protein is regarded as a long polypeptide, generally at least over 80 amino acids in length and usually much longer.) Those with novel properties, such as an ability to bind to a specific ligand (a unique molecule that is usually small) are then selected. In these instances, the sequences of the peptides are created de novo and are not based on that of a naturally occurring protein. See GENE; GENE ACTION; GENETIC CODE; PEP-TIDE. Karla L. Ewalt; Paul Schimmel

Bibliography. J. A. Branningan and A. J. Wilkinson, Protein engineering 20 years on, *Nature Rev. Mol. Cell Biol.*, 3:964–970, 2002; J. Lippincott-Schwartz and G. H. Patterson, Development and use of fluorescent protein markers in living cells, *Science*, 300:87– 91, 2003; S. A. Marshall et al., Rational design and engineering of therapeutic proteins, *Drug Discovery Today*, 8:212–221, 2003; P. G. Schultz et al., The

Protein folding

The process by which a long, flexible, unbranched chain of amino acids (the unfolded state of a protein) spontaneously changes shape to form a stable three-dimensional conformation (the native or folded state). This specific folding process is required for the protein's biological function. The sequential arrangement of amino acids is encoded by the genetic material, deoxyribonucleic acid (DNA) [**Fig. 1**]. The 20 most frequently used amino acids have a region in common that forms the covalent peptide bonds in the chain, termed the backbone, and a region that is chemically distinct, identified as the side chain. Amino acid side chains vary in size

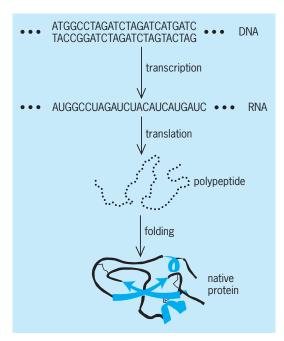


Fig. 1. An integrated view of protein production with folding as the terminal step. The instructions for building proteins are stored in the DNA sequence, which is transcribed into RNA and translated into an amino acid chain. The varying chemical properties of amino acid side chains drive protein folding. The final structure is characterized on different levels: primary structure is the sequence of amino acids; the local structure depicted below as arrows (β -strands) and coils (α -helix) is termed secondary structure; and the arrangement of secondary structural elements is tertiary structure. Many proteins consist of multiple polypeptide chains that work together as a single unit; the number and arrangement of these chains define the quaternary structure for these proteins. Some proteins must noncovalently bind heterologous chemical groups in order to complete the folding process. These groups (many of which are dietary vitamins and minerals) stabilize the protein's structure and may expand the chemical reactions the protein can perform. In contrast, several amino acids in a previously folded protein can be modified by covalent addition of chemical moieties (for example, sugars, lipids, phosphate or acetyl groups) that may alter the structure and/or function of the protein.

and shape, as well as the propensity to interact with water, oil, and ions. As a result, the side chains are differentially able to participate in noncovalent chemical bonds (hydrophobic interactions, electrostatic interactions, and hydrogen bonds) with each other and with the solvent. In addition, the amino acid cysteine can form covalent disulfide bonds with other cysteines. Together, this chemical interaction code determines the network of bonds that drives protein folding, stabilizes the native structure, and mediates protein function for any given sequence of amino acids.

Since a single protein chain can consist of thousands of amino acids, and 20 different amino acids are commonly used as assembly units, the number of possible protein sequences and configurations is immense. The function of a protein is determined by its three-dimensional structure, and hence protein functions are diverse: among many functions, these complex molecules act as efficient and specific biological machines to catalyze biochemical reactions for metabolism, provide structural support, regulate biological processes, and transport other molecules. The process of protein folding places the protein's atoms in the correct three-dimensional arrangement required for that protein's unique function. Loss of the protein's native structure through denaturation involves a simultaneous loss of the protein's function as well. Therefore, a comprehensive understanding of protein folding may explain protein malfunctions in vivo, as well as suggest a means to design artificial proteins with novel functions. See AMINO ACIDS; BIOORGANIC CHEMISTRY; PEPTIDE; PROTEIN.

Process. Understanding the folding process and the forces that maintain the native structure is one of the oldest and most challenging problems in biochemistry. The native state of a protein is generally 5-10 kcal/mol more stable than the unfolded state, which is actually an ensemble of loose structures. Levinthal's paradox theorizes that a random search of all possible configurations to find the one with lowest energy (most stable) would take a minimum of 10⁷⁷ years. In reality, local formations can occur in 10-100 nanoseconds, loop formations can fold in 500 ns, and the minimum time for the folding of a small protein is 1 millisecond. Therefore, protein folding must be a directed process. Folding intermediates have been proposed to limit the conformational search by defining one or more pathways to the native state. In addition, intermediates are essential for certain processes such as translocation across or insertion into membranes, release of nonpolar ligands, and association with other cellular components

Discrete intermediates in the folding process have been isolated for several proteins. Due to the extremely fast time scales of protein folding, partially folded states are often studied at equilibrium where they are stably populated at high or low temperatures, in the presence of chemical denaturants, at pH extremes, or at high pressures. Often these partially constrained intermediates have well-formed local structures but fluctuating or absent long-range

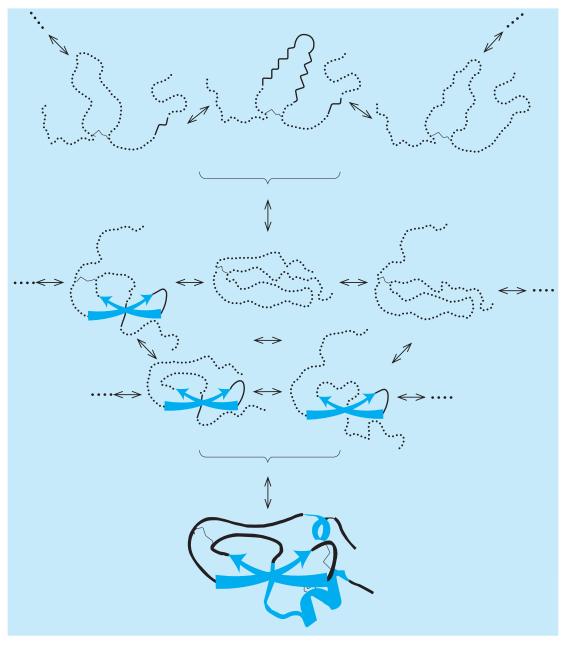


Fig. 2. Schematic of protein bovine pancreatic trypsin inhibitor during the folding process. Arrowhead ribbons are β -strands, and coiled ribbons are α -helices. (Structures provided by Elisar Barbar and Clare Woodward)

tertiary interactions (Fig. 2). Moreover, local structures may not be evenly dispersed throughout the folding intermediate; some regions may remain unfolded. Equilibrium intermediates are frequently detected by spectroscopic methods that monitor formation of local structural elements, or by the differential fluorescence of amino acids buried within the protein. Structural data (determined by nuclear magnetic resonance or x-ray crystallography) have even been obtained for some folding intermediates. Equilibrium intermediates are thought to be relevant to the kinetic folding process for several reasons. Conditions that favor the formation of the intermediates also speed up overall folding rates. Also, kinetic intermediates (partly folded forms that occur transiently during protein folding) often have structures similar

to equilibrium intermediates for those few proteins for which kinetic structural data are available. *See* NUCLEAR MAGNETIC RESONANCE (NMR); X-RAY CRYS-TALLOGRAPHY.

Models. Several theoretical models have been proposed to describe the sequence of intermediates assumed by a protein in the course of its folding reaction. In the hydrophobic collapse model, hydrophobic residues are excluded from the aqueous solvent, driving an initial collapse of the polypeptide chain. The native structure is then achieved by reorganizing and forming noncovalent interactions within the collapsed state. In the nucleation-condensation model, the most stable local structures form and propagate, followed by fusion of these various folded regions to form the native protein. Finally, in the

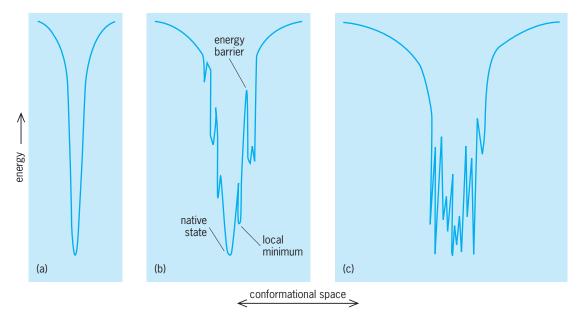


Fig. 3. Two-dimensional cross section through representative folding funnels. (a) A protein that folds without intermediates. (b) A protein that folds through discrete intermediates, located at the local minima. (c) A random heteropolymer with no unique structure.

diffusion-collision model, elements of local structure form and encounter each other by the process of diffusion. Experimental evidence supports each of these models for different proteins, demonstrating that the folding pathway for discrete proteins can be substantially different. Thus, the model that best describes the folding process for any single protein may also vary significantly.

A generalized description of protein folding has been developed that statistically characterizes the energy landscape of folding as a funnel in which the wide top represents the ensemble of high-energy unfolded configurations. The deepest valley represents the native state, the most stable-and therefore lowest energy-conformation (Fig. 3b). Intermediates along the pathway are described as local minima along the sides of the funnel and are separated from the native state by energy barrier peaks. A funnel with smooth sides describes a single, cooperative phase transition (the transition proceeds from starting structure to final structure) with no intermediates (Fig. 3a). The opposite extreme is a funnel with very rough sides and a large number of low-energy structures separated by tall barriers. Such a funnel characterizes the folding of a heteropolymer (a compound comprising two or more molecules that are different from one another) that does not adopt a discrete native state (Fig. 3c).

In vivo, a protein can be trapped in a structure corresponding to a local minimum in its energy funnel. Therefore, cells have evolved chaperone systems that lower the energy barriers and thereby speed folding and prevent aggregation. A commonly studied bacterial chaperone system is the GroEL/GroES protein complex. This complex sequesters unfolded protein in its hydrophobic center, which is cylindrical in shape, allowing it to fold without exposure to other unfolded proteins which would promote aggregation.

Much progress has been made toward understanding folding phenomena in individual proteins. The computer revolution has enabled modeling of the process, but simulating all atoms in the folding of even a small protein still requires a great deal of time, despite using the largest supercomputers. Modeling is limited to a large degree by incomplete descriptions of the force fields describing the noncovalent interactions that drive folding. With current tools and understanding, structure prediction is possible for peptides and very small proteins, and simple proteins have been designed and created. The current challenge is to extend this knowledge to the large, multidomain and multisubunit proteins that perform the processes crucial to living organisms. The ability to predict folding patterns is critical for interpreting data from the Human Genome Project, which can be used to generate the amino acid sequences, but not the structural and consequently functional information, of all proteins in the human body. See HUMAN GENOME.

Exceptions to the folding paradigm. Because protein folding is a balance between a large number of weak intraprotein and protein-solvent interactions, this process is strongly dependent on environmental conditions such as the solvent involved, salinity, temperature, and pressure. In vitro, partially folded or misfolded proteins form under nonoptimal conditions. These incorrect folding states can irreversibly coalesce to form various types of aggregates: amorphous aggregates, films, or amyloid fibrils. Such aggregation is frequently an unwelcome side product during research and for production of protein-based pharmaceuticals.

In vivo, misfolded proteins in a healthy cell are either refolded by chaperones or degraded,

preventing a buildup of aggregates. However, nonoptimal conditions such as inflammation, overproduction of protein, aging/oxidation, and mechanical or chemical stresses may increase the amount of misfolded protein in vivo. These misfolded proteins may then form amyloid fibrils, which are associated with a large number of diseases, among them Alzheimer's disease, Down syndrome, Parkinson's disease, and cataracts. A comprehensive understanding of protein folding and structure will contribute toward the logical design of therapies for these diseases.

Despite the stringent limits imposed upon protein structure by function and the potential effects of aggregation in cells, amazingly regions within many proteins are unfolded or disordered in their native state. Many of these proteins fold upon binding other cellular components—for example, other proteins, nucleic acids, or small molecules—allowing regulation of protein function by a structural oscillation between active and inactive forms. Indeed, many intrinsically disordered proteins participate in inherently transient and tightly controlled processes such as transcription regulation and cell signaling.

Continuing challenges and potential benefits. Protein folding is the least well characterized step in the process by which genetic information is stored, transmitted, and implemented (Fig. 1). Several significant challenges remain, including generating better algorithms or sufficient computational power for modeling large proteins in silico (via computer), monitoring kinetic folding events on very fast time scales, describing the unfolded state, identifying all possible folds (ternary structures), understanding the forces that drive proteins to misfolded states or aggregates and how these forms affect living cells, discovering general principles of protein folding despite varying folding pathways and diverse final structures, and elucidating how and why motion and disorder, like structure, can be encoded by amino acid sequence without eliciting protein destruction. Addressing these issues not only will illuminate a basic and critical aspect of biology but also will expand our ability to dissect uncharacterized biological pathways by predicting the structure and function of novel genes identified by genome sequencing projects. Finally, a thorough description of protein folding would allow exploitation of their functional specificity and efficiency by de novo design of artificial proteins as research reagents, industrial catalysts, and pharmaceutical products.

Sarah Bondos; Kathleen Matthews Bibliography. Y. Bai, Hidden intermediates and Levinthal paradox in the folding of small proteins, *Biochem. Biophys. Res. Commun.*, 305:785-788, 2003; T. E. Creighton, *Proteins: Structures and Molecular Properties*, 2d ed., W. H. Freeman, 1993; H. J. Dyson and P. E. Wright, Intrinsically unstructured proteins and their functions, *Nat. Rev. Mol. Cell Biol.*, 6:197-208, 2005; K. H. Mayo, Recent advances in the design and construction of synthetic peptides: For the love of basics or just for the technology of it, *Trends Biotechnol.*, 18:212-217, 2000; E. R. McCarney, J. E. Kohn, and K. W. Plaxco, Is there or isn't there? The case for (and against) residual structure in chemically denatured proteins, *Crit. Rev. Biochem. Mol. Biol.*, 40:181-189, 2005; P. G. Wolynes, Energy landscapes and solved proteinfolding problems, *Philos. Trans. A Math. Phys. Eng. Sci.*, 363:453-464, 2005.

Protein kinase

One of a family of enzymes that exert regulatory effects on a variety of cellular functions, as well as malignant transformation, by adding a phosphate group to proteins according to the equation

ATP + protein—OH
$$\xrightarrow{\text{protein kinase}}$$
 protein—OPO₃²⁻ + ADP

(ATP represents adenosine triphosphate, ADP represents adenosine diphosphate, and -OH is a hydroxyl group attached to an amino acid residue.) Based upon the nature of the phosphorylated -OH group, these enzymes are classified as serine/ threonine protein kinases and tyrosine protein kinases, where serine, threonine, and tyrosine are amino acid residues found in proteins. Furthermore, there is a small group of dual-specificity kinases, which closely resemble serine/threonine kinases, that catalyze the phosphorylation of both threonine and tyrosine on target proteins. The ratio of phosphoserine/phosphothreonine/phosphotyrosine in proteins from animal cells is about 3000/300/1. Despite the scarcity of tyrosine protein phosphate, it plays a paramount role in cell physiology.

Actions. Protein kinases are enzymes that play a role in nearly every aspect of cell biology. Work in the 1960s showed that protein kinases play a regulatory role in carbohydrate metabolism. Subsequent work indicates that protein kinases participate in nearly all cellular activities including regulation of gene expression, cell division, apoptosis (programmed cell death), differentiation, development, and antibody production. The brain, moreover, is an especially rich site of protein kinase activity. *See* CELL (BIOLOGY); NEUROBIOLOGY.

Protein phosphorylation represents the chief regulatory mechanism in animal cells. In some cases, phosphorylation leads to activation, and in other cases phosphorylation leads to inactivation of the phosphorylated protein. It is estimated that perhaps a quarter of all animal proteins can be phosphorylated by protein kinases. Moreover, many proteins can be phosphorylated at more than one residue by a given protein kinase, and many proteins can be phosphorylated by several protein kinases. *See* EN-ZYME; PROTEIN.

Protein kinase genes. There are more genes encoding protein kinases than there are genes encoding any other family of human enzymes except for proteases. Workers have identified 518 protein kinase genes in humans (478 typical and 40 atypical), about 2% of the total number of human genes (see **table**). All protein kinases have a similar overall three-dimensional structure and chemical mechanism. *See* GENE; GENE ACTION.

Class	Number		Examples ^a
Protein-serine/threonine	428		
Receptor ^b		12	Transforming growth factor- β
Nonreceptor ^c		369	ERK, cyclin-dependent protein kinases, protein kinase A, protein kinase C, raf protein kinase
Dual specificity		47	MEK
Protein-tyrosine	90		
Receptor ^b		58	Insulin receptor, HER
Nonreceptor ^c		32	Src, Abl
TOTAL	518		

^cLacks an extracellular domain and is regulated by second messengers and by phosphorylation-dephosphorylation.

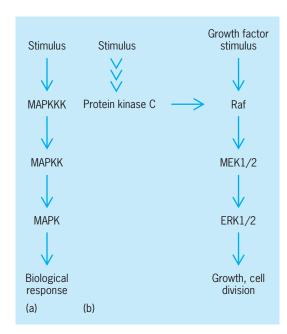
Serine/threonine protein kinases. There are many serine/threonine protein kinases, and three general classes are described here.

Protein kinase A. Protein kinase A, or cyclic adenosine monophosphate (AMP)-dependent protein kinase, was one of the first protein kinases to be described. Many hormones (first messengers) interact with their receptors on the cell membrane, and this interaction leads to the activation of adenylyl cyclase. This enzyme catalyzes the intracellular production of cyclic AMP (a second messenger) from ATP. Cyclic AMP interacts with protein kinase A and activates it. Protein kinase A can lead to the phosphorylation of many different proteins, with many different effects. For example, protein kinase A-mediated phosphorylation of the enzyme that promotes triglyceride degradation leads to its activation, whereas protein kinase A-mediated phosphorylation of the enzyme that promotes glycogen synthesis is inhibitory.

Protein kinase C. Protein kinase C is another serine/ threonine kinase that can lead to the phosphorylation of many different proteins. (The C refers to its requirement for ionic calcium.) It also requires phospholipids and diglyceride for full expression of activity. As a result of the interaction of a hormone or growth factor (first messengers) with a receptor on the exterior of a cell membrane, intracellular second messenger molecules such as diglyceride and calcium are liberated, and these activate protein kinase C.

Many protein kinases function in a series, or a cascade, where one enzyme catalyzes the phosphorylation of a second enzyme that in turn mediates the phosphorylation of a third enzyme. A generic form of this cascade is given in the illustration. A mitogen-activated protein kinase (MAPK) brings about a physiological response. This enzyme is phosphorylated by a kinase acting on a kinase (MAPKK), and this is then activated by a kinase (MAP kinase kinase kinase, or MAPKKK). A stimulus initiates the process by activating the mitogenactivated protein kinase kinase kinase (MAPKKK). There are several examples of cascades that follow this paradigm, one of which (the Raf-MEK-ERK cascade) is shown in illustration b. Protein kinase C-mediated phosphorylation of a serine/threonine protein kinase called Raf leads to its activation, and the activated Raf participates in a cascade that leads to growth and cell division (see illustration). Activation of Raf leads to the phosphorylation and activation of two different protein kinases, MEK1 and MEK2. [MEK stands for mitogenactivated extracellular regulated kinase (ERK)activating kinase.] MEK1 and MEK2 are examples of dual-specificity protein kinases. These lead to the phosphorylation of both a threonine and a tyrosine residue in each of the protein kinases ERK1 and ERK2, and this dual phosphorylation leads to their activation. ERK1 and ERK2 are serine/threonine protein kinases that mediate the phosphorylation of other protein kinases and regulators of gene transcription. As a result of this cascade of events, cell growth and cell division can result. See CELL DIVI-SION; GROWTH FACTOR; HORMONE.

Cyclin-dependent protein kinases. Cyclin-dependent protein kinases are another family of serine/threonine



Protein kinase cascades. (a) A general protein kinase cascade and (b) the Raf-MEK-ERK protein kinase cascade. Protein kinase cascades resulting in an inflammatory response and in apoptosis resemble the general protein kinase cascade. MAPKKK, mitogen-activated protein kinase kinase; MAPKK, mitogen-activated protein kinase kinase; MAPK, mitogen-activated protein kinase. protein kinases that are essential for cell division. Cyclins are proteins that are synthesized and activated during specific phases of the cell cycle, and they are degraded after they have acted. While the level of cyclin varies during the cell cycle, the level of their cognate protein kinases remains constant throughout the cell cycle. *See* CELL CYCLE.

Tyrosine protein kinases. Receptor tyrosine protein kinases possess an extracellular domain, a transmembrane domain, and an intracellular kinase domain. Binding of a growth factor to the extracellular domain results in receptor dimerization (a chemical reaction in which two identical molecular entities react to form a single molecule, a dimer) and protein kinase activation. The paired kinases phosphorylate each other and other substrate proteins. The phosphorylated tyrosines provide binding, or docking, sites for specific proteins. Binding of a target protein to a docking site results in its activation, and the activated protein initiates a series, or cascade, of downstream phosphorylation reactions. The binding of human epidermal growth factor to members of a family of four receptor tyrosine protein kinases attracts docking proteins that lead to the activation of the Raf-MEK-ERK cascade (see illustration) and to growth and cell division. Variations of this cascade occur in yeast and insects in addition to higher animals. See SECOND MESSENGERS.

Protein phosphatases. Protein phosphorylation is a reversible process. Protein kinases add phosphoryl groups to proteins, and protein phosphatases remove these groups during a hydrolysis reaction:

Protein— $OPO_3^{2-} + H_2O \xrightarrow{\text{protein phosphatase}} protein \longrightarrow OH + HOPO_3^{2-}$

To achieve coordinate regulation of cellular processes, both protein kinase activity and protein phosphatase activity are carefully regulated. Much more is known about the regulation of kinases than phosphatases.

Human cells contain genes corresponding to 120 protein phosphatases: 32 are serine/threonine protein phosphatases, 42 are tyrosine protein phosphatases, and 46 are dual-specificity (threonine/ tyrosine) phosphatases. There are three subclasses of tyrosine protein phosphatases. One group contains an extracellular receptor domain, a transmembrane segment, and an intracellular catalytic domain. A second class is tethered to the interior of the cell membrane; this class has the potential to be regulated by transmembrane receptors. A third class of tyrosine protein phosphatases is located within the cell, and this class functions in both the cytosol and the cell nucleus.

Serine/threonine protein phosphatases are generally located in the cell cytoplasm. Phosphoprotein phosphatase-1 is inhibited by a protein called phosphoprotein phosphatase inhibitor-1. Phosphoprotein phosphatases-2A, -2B, and -2C are not inhibited by this protein. Phosphoprotein phosphatase-2B is activated by calcium, a regulatory ion that also activates some protein kinases. There is a balance in the regulation of kinases and phosphatases that leads to physiological responses. Substrates for dualspecificity phosphatases include ERK1, ERK2, and other MAP kinases; these are enzymes that require phosphorylation of both threonine and tyrosine residues for activation.

Drug targets. Owing to the numerous functions of protein kinases and phosphatases, these enzymes are recognized as important drug targets. The HER2 (human epidermal growth factor receptor) protein is overexpressed in about 30% of human breast cancers, and these can be treated effectively with trastuzumab (Herceptin), an antibody directed to the exterior of this receptor tyrosine protein kinase. Imatinib (Gleevec) is an inhibitor of the Bcr-Abl tyrosine protein kinase that occurs in 95% of people with chronic myelogenous leukemia. Both of these medicinal agents are therapeutically effective. A large amount of research on identifying drug targets for kinases and phosphatases is being performed owing to the success of trastuzumab and imatinib. See CANCER (MEDICINE). Robert Roskoski, Jr.

Bibliography. A. Alonso et al., Protein tyrosine phosphatases in the human genome, *Cell*, 117:699-711, 2004; J. M. Berg, J. L. Tymoczko, and L. Stryer, *Biochemistry*, 5th ed., 2001; B. D. Gomperts, I. M. Kramer, and P. E. R. Tatham, *Signal Transduction*, 2002; G. Manning et al., The protein kinase complement of the human genome, *Science*, 298:1912-1934, 2002; D. L. Nelson and M. M. Cox, *Lebninger Principles of Biochemistry*, 4th ed., 2004; R. Roskoski, *Biochemistry*, 1996.

Protein metabolism

The transformation and fate of food proteins from their ingestion and assimilation to the elimination of their excretion products. Proteins are polymers of α -amino acids that are connected by peptide bonds. An average polypeptide chain in a protein contains about 500 amino acid residues. Insulin is a small protein with 51 amino acid residues. Titin, which is the largest known human protein, contains 34,350 residues. A dipeptide contains two amino acid residues, an oligopeptide contains several, and a polypeptide contains many amino acid residues. Proteins are the chief structural and functional components of all living organisms. Their importance was recognized by scientists in the midnineteenth century who coined the name from the Greek proteios, meaning first or primary. Proteins are the main building blocks of the cells, tissues, organs, and systems of the body. Proteins of one species differ from those of another species and, within a single animal, proteins of muscle differ from those of the brain, kidney, liver, and other organs. See AMINO ACIDS; METABOLISM; PEPTIDE; PROTEIN.

Functions of protein. A 70-kg (154-lb) person contains 10–12 kg (22–26 lb) of protein. After water, proteins are the most abundant component of humans, other animals, and bacteria. Muscle accounts for about half of the mass of human protein. Collagen is the most abundant protein in animals and is responsible for the strength of the connective tissue found in muscle and tendons. Actin is a protein that occurs in nearly all animal cells and is responsible for cell motility. Myosin is the chief contractile protein of muscle. Hemoglobin is the main protein of red blood cells and is responsible for oxygen transport from the lungs to the tissues. Many hormones, including insulin, are protein in nature. Enzymes that catalyze the myriad reactions in the breakdown of glucose and fatty acids in intermediary metabolism are proteins. Moreover, the adenosine triphosphate synthase that mediates the conversion of adenosine diphosphate and phosphate into ATP, the common currency of energy exchange in all living organisms, is a protein. See ADENOSINE TRIPHOSPHATE (ATP); COLLAGEN; EN-ERGY METABOLISM; ENZYME; HEMOGLOBIN; MUSCLE; MUSCLE PROTEINS.

Protein turnover. Body proteins are continually broken down and resynthesized, a process called turnover. There are two main pathways that participate in intracellular protein degradation: the ubiquitin-proteasome pathway and the lysosomal pathway. The first step in the former pathway involves the joining of ubiquitin molecules with lysine residues in the target protein in a process that requires ATP as an energy source. Ubiquitin is a small protein containing 76 amino acid residues. Following the attachment of several ubiquitin molecules to the target protein, the complex interacts with the proteasome, a large multicomponent complex that degrades the target protein into peptides while liberating ubiquitin, which is then recycled. Proteasomes are located in the soluble portion of the cell, or the cytosol. The proteasomal pathway is responsible for degrading proteins that are damaged because of oxidative stress or are no longer needed owing to changing metabolic needs. The lysosome is a membrane-enclosed vesicle inside the cell that contains a variety of proteolytic enzymes and operates under acidic conditions. It degrades proteins that combine with receptors found on the plasma membrane of the cell and that are subsequently taken up by the cell where they fuse with lysosomes. See CELL (BIOLOGY); LYSOSOME.

Normal adults are in nitrogen balance; nitrogen intake equals the amount of excreted nitrogen. Proteins, which are made up of amino acid residues, are the main nitrogen-containing compounds in animals. Growing animals are in positive nitrogen balance; nitrogen intake exceeds nitrogen excretion. During starvation, animals are in negative nitrogen balance; nitrogen excretion exceeds nitrogen intake. People with infections, traumatic injuries (such as from automobile accidents), and major burns also exhibit negative nitrogen balance. Under these conditions, tissue and protein breakdown occurs more rapidly than protein synthesis. Even in the absence of trauma, surgery itself results in accelerated protein breakdown, possibly as a result of increased adrenal steroid (glucocorticoid) production and action. When these conditions are accompanied by reduced food intake, which typically occurs in people with these maladies, this imbalance results in wasting and an impaired immune response. During the healing phase of these illnesses, people exhibit positive nitrogen balance.

Besides the breakdown of endogenous proteins, ingested proteins contribute to the amino acid pool. Protein is digested to amino acids in the stomach and small intestine. These amino acids are absorbed and undergo a degree of metabolic interconversion prior to release into the portal vein, which takes them to the liver. Many of the amino acids are taken up by the liver, and the remainder circulates to the other tissues and organs in the body. The newly absorbed amino acids equilibrate with amino acids in cells that result from protein breakdown.

The amino acid pool serves as the source of precursors for new protein synthesis and for the conversion to other metabolites. Protein synthesis occurs on the ribosome in a messenger ribonucleic acid-dependent process. About 50 grams of muscle protein, 12 g of albumin, 2 g of fibrinogen, 3 g of γ -globulin (antibodies), 8 g of hemoglobin, 30 g of gastrointestinal epithelial protein, and 25 g of enzymes and mucins in the salivary glands, stomach, intestine, and pancreas are synthesized and degraded in the body daily. For normal human adults, about 250 g of protein is degraded and synthesized each day. To the 250 g of amino acids that result from protein degradation, about 100 g per day are obtained from the diet.

Nutritive value of dietary proteins. Plants can synthesize all of the amino acids found in proteins; animals cannot, even though the amino acids are essential for life. Animals ingest protein primarily to supply amino acids for the formation of new and depleted body protein and as a source of various other body constituents derived from amino acids. Twenty amino acids are required for protein synthesis. The amino acids of proteins fall into two nutritional categories: essential or indispensable and nonessential or dispensable. The essential amino acids cannot be synthesized by the body or cannot be synthesized in adequate quantities and therefore must be taken in the diet, while the nonessential amino acids can be synthesized by the body in sufficient quantities to sustain health. The 20 genetically encoded amino acids required for protein synthesis are as follows:

Essential	Nonessential
or indispensable	or dispensable
1. Histidine	1. Alanine
2. Isoleucine	2. Arginine
3. Leucine	3. Asparagine
4. Lysine	4. Aspartate
5. Methionine	5. Cysteine
6. Phenylalanine	6. Glutamate
7. Threonine	7. Glutamine
8. Tryptophan	8. Glycine
9. Valine	9. Proline
	10. Serine
	11. Tyrosine

The relative proportions of the 20 amino acids vary in different foods. Foods of animal origin, including meat, fish, eggs, milk, and dairy products, are sources of high-quality protein. Their essential amino acid content meets the human need for protein synthesis. Gelatin, an animal protein that lacks the essential amino acid tryptophan, is not a high-quality protein. Foods of plant origin can be more difficult to digest and are of lower quality. Plant proteins contain less lysine, methionine, and tryptophan than animal proteins, and these amino acids thus limit the synthesis of body proteins. However, a varied vegetarian diet can readily fulfill human protein requirements if the protein-containing foods are balanced such that their essential amino acids complement each other. For example, legumes such as beans are high in lysine and low in methionine, while rice has a higher percentage of methionine. Thus, if beans and rice are eaten together, their joint amino acid composition will complement each other and provide a higherquality diet than would either food alone. Traditional food patterns in various cultures have made good use of protein complementarity.

The recommended daily protein intake is 0.8 g/kg body weight for adults, for example, 56 g for a 70-kg (154-lb) person. This amount of protein is easily obtained in a judicious diet: for example, one cup of dry cereal with a cup of low-fat milk for breakfast, a peanut butter sandwich and an 8 oz glass of milk for lunch, and 4 oz of fish, a cup of rice, and cup of split peas amount to 70 g of protein, enough for the average adult. A daily 2000 calorie diet consisting of 6 oz of grains, $2^{1}/_{2}$ cups of vegetables, 2 cups of fruit, 3 cups of milk, and $5^{1}/_{2}$ oz of meat and beans contains about 90 g of protein. The recommended daily protein intake is greater for infants, for growing children, and for women during pregnancy and during nursing. *See* NUTRITION.

Inadequate protein intake affects health. Kwashiorkor, also called protein malnutrition, is a condition in young children with adequate calorie but inadequate protein intake. Kwashiorkor is most often encountered in developing countries in which the diet is high in starch and low in protein. Signs and symptoms include reddish skin, abdominal swelling, weakness, and nervous irritability. Marasmus results from both insufficient protein and calorie intake in infants under conditions of famine. Marasmus is characterized by growth retardation (in weight more than in height), emaciation, and generalized wasting of muscle and subcutaneous fat. Ribs, joints, and facial bones are prominent owing to the loss of muscle and fat. *See* MALNUTRITION.

Digestion of protein. Protein digestion is initiated in the stomach and completed in the duodenum of the small intestine. The main proteolytic enzyme of the stomach is pepsin, which is secreted by the chief cells in an inactive form called pepsinogen. The hydrogen ion concentration in the stomach is thousands of times greater than that of the blood or in cells; stomach acid is secreted by parietal cells. The conversion of precursor pepsinogen to active pepsin is accelerated by stomach acid and by active pepsin. The activation process involves the hydrolytic removal of an inhibitory 44-aminoacid peptide from pepsinogen. Pepsin preferentially hydrolyzes peptide bonds adjacent to an aromatic amino acid (phenylalanine, tyrosine, tryptophan), and pepsin requires an acid medium in which to function. Stomach acid also unfolds, or denatures, ingested proteins and increases their susceptibility to digestion. *See* DIGESTIVE SYSTEM; PEPSIN.

The acidic stomach contents, which contain partially degraded proteins, are discharged as chyme (semifluid, partially digested food mass) into the slightly alkaline fluid in the duodenum of the small intestine. The pancreatic juice released into the duodenum contains several digestive enzymes that are secreted as zymogens (inactive enzyme precursors) including trypsinogen, chymotrypsinogen, procarboxypeptidase, and proelastase. The activation of trypsinogen involves its proteolytic cleavage by yet another enzyme secreted by the pancreas, namely enteropeptidase, to produce active trypsin. Trypsin, in turn, catalyzes the formation of chymotrypsin, carboxypeptidase, and elastase from their inactive zymogen precursors. In each of these processes, certain peptide bonds are broken to yield the active enzymes. The cells lining the small intestine contain various aminopeptidases that are not liberated into the gut, but act at the cell surface to further proteolytic degradation and to participate in the absorption of the degraded proteins.

Trypsin, chymotrypsin, and elastase are endopeptidases; that is, they cleave peptide bonds that occur in the interior of proteins. Trypsin catalyzes the cleavage of peptide and protein substrates adjacent to positively charged lysine and arginine residues, while chymotrypsin catalyzes peptide bond cleavage next to hydrophobic residues such as methionine, leucine, and the aromatic amino acids. Elastase mediates peptide bond cleavage next to small amino acids such as glycine, alanine, and serine. These three proteases catalyze the cleavage of proteins into small peptides. Aminopeptidases, which are exopeptidases that cleave amino acid residues from the amino-terminal ends of peptides, help to complete protein hydrolysis. Carboxypeptidases, which are also exopeptidases, catalyze the removal of amino acids at the C-terminus of substrate peptides or proteins. Amino acids, dipeptides, and tripeptides are absorbed by the intestinal cells. The hydrolysis of peptides to amino acids is completed intracellularly.

The amino acid digestion products are absorbed as rapidly as they are liberated. The absorption is confined chiefly to the small intestine and is a process that involves the metabolic participation of the cells of the intestinal lining. Amino acids are absorbed from the gut and liberated into the hepatic (liver) portal vein. The outflow of amino acids into the portal vein does not reflect the amino acid composition of dietary protein owing to metabolic transformation by intestinal cells. Most amino acids are taken up and metabolized by the liver except for valine, leucine, and isoleucine—the branched-chain amino acids. These three amino acids may account for 60% of the amino acids entering the circulation from the gut despite their 20% contribution to the total amino acids of dietary protein. The branched-chain amino acids are taken up chiefly by skeletal muscle. Since they contribute to only 20% of skeletal muscle protein, excess branched-chain keto acid metabolites, derived from these amino acids following transamination (transfer of the amino group of an amino acid to a keto acid to form another amino acid), can be oxidized by muscle or transported to other cells. The liver amino acids are converted into other amino acids (the nonessential amino acids) and into cellular and plasma proteins.

Utilization of absorbed amino acids. Protein homeostasis differs from that of carbohydrate and lipid because protein lacks a storage form. Carbohydrate is stored as glycogen, and lipids are stored as triacylglycerol or triglyceride. The absorbed amino acids become part of the amino acid pool of the body. From these amino acids, new tissue proteins are synthesized to meet body needs. To serve the needs for protein synthesis, all of the 20 genetically encoded amino acids must be present. The absence of any amino acid retards the rate of protein synthesis. *See* CARBOHYDRATE; CARBOHYDRATE METABOLISM; LIPID; LIPID METABOLISM.

Tissue protein synthesis. Turnover of tissue proteins occurs in the adult animal in nitrogen balance, with no net gain of body nitrogen. Body proteins are continually undergoing synthesis and degradation, but the total body protein remains relatively constant in quantity. The rate of replacement varies greatly for different tissues. In humans, it is estimated that the average half-life of total body protein is 80 days; that of lung, brain, bone, and muscle together is about 160 days, while that of liver and serum proteins is only 10 days.

Plasma protein synthesis. The liver is the major organ for plasma protein synthesis. It synthesizes albumin, the predominant blood plasma protein. The liver also synthesizes fibrinogen, one of the main components of blood clots. The plasma proteins offer the most readily available test material in determining the protein nutritional status of an individual. A blood sample is easily drawn, and estimation of the different plasma proteins is now a standard procedure. Plasma proteins exhibit fluctuations in conditions associated with inadequate intake or abnormal protein metabolism. The serum albumin level is the traditional standard for nutritional assessment. Protein deprivation diminishes the albumin content to levels below 30 g/liter, whereas the normal levels are 35-70 g/liter. See ALBUMIN; FIBRINOGEN; LIVER.

Blood proteins have numerous physiological functions. Albumin is a major factor in the regulation of blood volume by counteracting fluid expulsion into tissues as a result of hydrostatic blood pressure. Fibrinogen is the chief protein that results in blood clot formation. The globulins include proteins that transport iron, copper, cholesterol, cholesteryl esters, and other lipids. Antibodies, which are synthesized by cells of the immune system, form part of the globulin fraction of plasma proteins. Antibodies defend the body against various pathogenic bacteria and viruses. Protein hormones such as insulin and glucagon, which form a tiny but important fraction of plasma protein, are synthesized in endocrine glands. *See* BLOOD; HORMONE.

Amino acid derivatives. Selected components of the general amino acid pool serve as precursors for numerous nonprotein organic compounds. For example, glycine is converted into heme, an organic component found in hemoglobin. Glycine and aspartate serve as building blocks for the nucleotide bases that occur in deoxyribonucleic acid and ribonucleic acid. Both histamine, which is derived from histidine, and serotonin, which is derived from tryptophan, function as signaling molecules. Tyrosine is converted to norepinephrine, a neurotransmitter, in selected nerve cells and to the melanin pigment in skin cells. *See* DEOXYRIBONUCLEIC ACID (DNA); RI-BONUCLEIC ACID (RNA).

Nitrogen excretion. Excess amino acids in cells, tissues, and organs undergo metabolic degradation. The amino group is lost by transamination or deamination (hydrolysis of amino compounds, removing the amino group). About 85% of excess nitrogen is excreted as urea in the urine. Urea $[C(=O)(NH_2)_2]$ is produced in the liver by the Krebs-Henseleit cycle. The overall process is

 $\begin{array}{l} \mbox{Ammonia} + \mbox{carbon dioxide} + \mbox{aspartate} + 3 \mbox{ATP} + 2 \mbox{H}_2 \mbox{O} \rightarrow \mbox{urea} + \mbox{fumarate} + 2 \mbox{ADP} + 2 \mbox{ phosphate} \end{array}$

+ adenosine monophosphate + pyrophosphate

The carbon and oxygen are derived from carbon dioxide, one nitrogen is derived from ammonia, and the second nitrogen is derived from the amino acid aspartate. The process requires the expenditure of three ATP molecules. Ornithine, a nonprotein amino acid, is required for urea biosynthesis; it is continuously regenerated and recycled for additional urea synthesis:

Arginine $+ H_2 O \rightarrow urea + ornithine$

About 2% of nitrogen in the urine is excreted as uric acid, which is derived from DNA, RNA, and nucleotides. A small amount of nitrogen is eliminated in the feces. The remainder is secreted in approximately equal amounts as ammonia, amino acids, and creatinine. Creatinine is a breakdown product of muscle creatine phosphate, an energyrich storage compound. The residual carbon portions of the amino acids are further metabolized by pathways common to the other major foodstuffs carbohydrates and fats. Proteins account for about 15% of the calories consumed in an average diet. *See* NITROGEN EXCRETION; UREA.

Robert Roskoski, Jr.; David M. Greenberg Bibliography. D. H. Baker, Comparative nutrition and metabolism: Explication of open questions with emphasis on protein and amino acids, *Proc. Nat. Acad. Sci. USA*, 102:17897-17902, 2005; J. M. Berg, J. L. Tymoczko, and L. Stryer, *Biochemistry*, 6th ed., W. H. Freeman, New York, 2006; P. C. Champe, R. A. Harvey, and D. R. Ferrier, *Lippincott's* *Illustrated Reviews: Biochemistry*, Lippincott Williams & Wilkins, Philadelphia, 2005; P. J. Reeds and P. J. Garlick, Protein and amino acid requirements and the composition of complementary foods, *J. Nutr.*, 133:29538-29618, 2003; G. M. Wardlaw, *Contemporary Nutrition*, 6th ed., McGraw-Hill, New York, 2005.

Proteins, evolution of

Proteins are large organic molecules that are involved in all aspects of cell structure and function. Proteins are made up of polypeptide chains, each constructed from a basic set of 20 amino acids, covalently linked in specific sequences. Each amino acid is coded by three successive nucleotide residues in deoxyribonucleic acid (DNA); the sequence of amino acids in a polypeptide chain, which determines the structure and function of the protein molecule, is thus specified by a sequence of nucleotide residues in DNA. *See* GENE; PROTEIN.

Sequence analyses of polypeptides which are shared by diverse taxonomic groups have provided considerable information regarding the genetic events that have accompanied speciation. Interspecies comparison of the amino acid sequences of functionally similar proteins has been used to estimate the amount of genetic similarity between species; species that are genetically more similar to each other are considered to be evolutionarily more closely related than those that are genetically less similar.

The study of functionally related proteins from different animal species has suggested that single amino acid substitutions are the predominant type of change during evolution of such proteins. Insertions or deletions of one or more amino acids have also been reported. The function of a protein is determined by its shape, and the shape is determined by the particular sequence of amino acids; in proteins that serve the same function in dissimilar species small differences in the amino acid sequence will often not affect overall functioning of the protein molecule.

In taxonomic protein sequence analysis, the amino acid sequence of a protein from one species is compared with the amino acid sequence of the protein from another species, and the minimum number of nucleotide replacements (in DNA) required to shift from one amino acid to another is calculated. Peptide "genealogies" can be constructed from many such comparisons in a related group of organisms.

Classical versus protein-derived phylogenies. It has been recognized for a long time that the amino acid sequences of a protein are species-specific. Protein sequencing has been used widely since the mid-1960s to examine taxonomic relationships. Results indicate that, in general, genealogical relationships (phylogenies) based on sequence analyses correspond fairly well with the phylogeny of organisms as deduced from more classical methods involving morphological and paleontological data.

Evolutionary biologists are turning increasingly to the new nucleic acid sequencing technology as an alternative to determining the amino acid sequence of proteins. Knowing the actual nucleotide sequences of genes rather than having to infer them from protein sequence data allows more accurate data to be used in determining genealogical relationships of organisms. For example, silent nucleotide substitutions (that is, base changes in DNA codons that do not result in amino acid changes) can be detected. Nucleotide sequence data has also resulted in increased knowledge of introns (sections of DNA that are not translated into protein), which are usually less conserved than the nucleotide sequences of the coding regions and unexpressed genes, or pseudogenes. Pseudogenes closely resemble a known gene, yet differ from it due to the presence of additions or deletions in their structure, which prevent their normal transcription and translation. Pseudogenes are generally considered to be the diverged products of duplicated genes. See INTRON.

Types of proteins studied. The reconstruction of phylogenies from analysis of protein sequences is based on the assumption that the genes coding for the proteins are homologous, that is, descendants of a common ancestor. Those sequences whose evolutionary history reflects that of the species in which they are found are referred to as orthologous. The cytochrome c molecules (present in all eukaryotes) are an example of an orthologous gene family. Organisms as diverse as humans and yeast have a large proportion of the amino acids in these molecules in common; they derive from a single ancestral gene present in a species ancestral to both these organisms and to numerous others (see below).

Sequences which are descendants of an ancestral gene that has duplicated are referred to as paralogous. Paralogous genes evolve independently within each species. The genes coding for the human α , β , γ , δ , ϵ , and ζ hemoglobin chains are paralogous. Their evolution reflects the changes that have accumulated since these genes duplicated. Analysis of paralogous genes in a species serves to construct gene phylogenies, that is, the evolutionary history of duplicated genes within a given lineage.

Rate of evolution. Sequence data from numerous proteins have shown that different proteins evolve at different rates. Some proteins show fewer amino acid substitutions, or more conservation, than others (see table). Proteins such as immunoglobulins, snake venom toxins, and albumins have changed extensively. Their function apparently requires relatively less specificity of structure and therefore has relatively greater tolerance for variance. By contrast, certain proteins, such as various histones, have changed relatively little over long periods of time. Histone H4 shows extreme conservation; it has essentially the same sequence in all eukaryotes examined. Such extensive sequence conservation is generally interpreted to indicate that the functions of H4 are extremely dependent on its entire structure; thus little or no change is tolerated in its structure. The

Protein	Unit evolutionar period [†]
Histones	
H4	400
H3	330
H2A	60
H2B	60
H1	8
Fibrous proteins	
Collagen (α-1)	36
Crystallin (αA)	22
Intracellular enzymes	
Glutamate dehydrogenase	55
Triosephosphate dehydrogenase	20
Triosephosphate isomerase	19
Lactate dehydrogenase H ₄	19 13
Lactate dehydrogenase M ₄ Carbonic anhydrase B	
Carbonic anhydrase B Carbonic anhydrase C	4 2.1
Electron carriers	2.1
Cytochrome c	15
Cytochrome b_5	13
Plastocyanin	7
Ferredoxin	6
Hormones	0
Glucagon	43
Corticotropin	24
Insulin	14
Thyrotropin (β)	9
Lipotropin (β)	8
Lutropin (a)	7
Proparathyrin	7
Prolactin	5
Growth hormone	4
Lutropin (β)	3
Insulin C peptide	1.9
Oxygen-binding proteins	0
Myoglobin	6
Hemoglobin (α)	3.7 3.3
Hemoglobin (β)	3.3
Secreted enzymes Trypsinogen	6
Lysozyme	2.5
Ribonuclease	2.3
Immunoglobulins	2.0
γ chains (C)	1.7
λ chains (C)	1.7
κ chains (V)	1.0
κ chains (C)	0.9
λ chains (V)	0.8
γ chains (V)	0.7
Snake venom toxins	
Long neurotoxins	0.9
Cytotoxins	0.9
Short neurotoxins	0.8
Other proteins	
Parvalbumin	5
Albumin	3
Lactalbumin	2.3
Fibrinopeptide A	1.7
Casein (κ)	1.4
Fibrinopeptide B	1.1

^{*}After A. C. Wilson et al., Biochemical evolution, *Annu. Rev. Biochem.*, 46:573–639, 1977.

[†]Unit evolutionary period is the average time, in millions of years, required for a 1% difference in amino acid sequence to arise between two lineages; it is thus an inverse measure of the rate of protein evolution.

rates at which different proteins evolve are therefore thought to be due to different functional constraints on the structure of the proteins—the more stringent the conditions that determine the function of a protein molecule, the smaller the chance that a random change will be tolerated in its structure.

Each protein generally has a nearly constant evolutionary rate (the rate of acceptance of mutations) in each line of descent. Exceptions to this rule have been reported, however, and much effort has been spent on determining whether these anomalies are genuine. Some anomalies have been shown to be due to comparison of nonhomologous proteins, and others due to sequencing errors. Other deviations from constant rate of sequence evolution remain to be explained; once uncovered, these may provide useful information about the mechanisms of evolution at the molecular level.

Mitochondrial cytochrome c. Cytochrome *c* is a protein involved in cellular respiration and is found in the mitochondria of plant and animal cells. Although this protein is located in the mitochondria, it is encoded by a nuclear gene. Cytochrome *c* gene sequences have been studied from a wide range of organisms, including protozoa, fungi, various vertebrate and invertebrate groups, and higher plants.

The cytochrome c molecules have generally been considered to be an excellent example of a set of orthologous proteins. However, studies have shown that for various mammalian species, cyctochrome cpseudogenes may be present, and that there may be cytochrome c molecules that are tissue-specific or that occur only during certain developmental stages.

The phylogenetic relationships between various organisms, based on the minimum numbers of nucleotide substitutions necessary to account for the amino acid differences in the cytochrome c molecules of these organisms, correspond fairly well with the phylogeny of the organisms, as determined from fossil evidence and comparative morphology (**Fig. 1**).

Cytochrome c molecules have evolved slowly. Organisms spanning the range from humans to Neurospora have a large proportion of the amino acids in their cytochrome c molecules in common. While the evolutionary conservation of such a molecule makes possible the study of genomic differences between remotely related species, this same property prevents determination of evolutionary change among closely related organisms since they may have cytochrome c molecules that are completely or nearly identical. For example, sequences from mammals within the same order differ by only 1 or 2 residues out of approximately 100. The primary structure of cytochrome c in humans and chimpanzees (which according to paleontological evidence diverged about 10 million years ago) is identical, and in humans and rhesus monkeys (which diverged about 40 to 50 million years ago) it differs by only one amino acid.

Since different proteins evolve at different rates, however, evolutionary relationships among closely related organisms such as humans and chimpanzees can be determined by examining the amino acid sequences of rapidly evolving proteins, such as fibrinopeptides in mammals. Other methods can also be employed to study genetic differences among closely

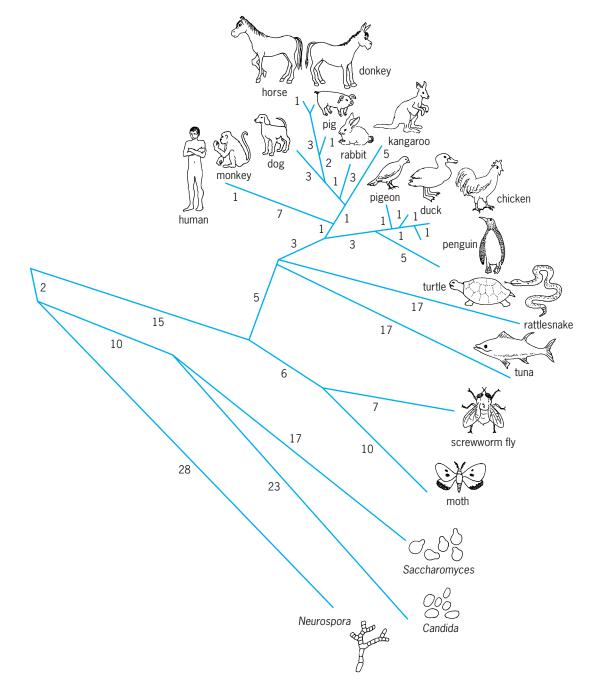


Fig. 1. Phylogeny of 20 organisms, based on differences in the amino acid sequence of cytochrome c. The minimum number of nucleotide substitutions required is shown. (After W. M. Fitch and E. Margoliash, Construction of phylogenetic trees, Science, 155:279–284, 1967)

related species. These include measurement of immunological cross-reaction between proteins from different species; heteroduplex formation between DNAs from different sources and measurement of their thermostability; examination of protein migration in polyacrylamide gels; use of x-ray crystallography to define the three-dimensional structure of proteins; and direct DNA sequencing of genes.

Globin genes. The genes that code for globin, the protein component of hemoglobin, have been extensively studied. Analysis of the globin gene family has provided detailed information regarding its evolutionary history over the past 500 million years.

Hemoglobin is a tetramer, composed of two α -like and two β -like globin molecules. Humans have five different β -like globins (ϵ , γ^{G} , γ^{A} , δ , and β) and three α -like globins (ζ and 2α). These α - and β -globins are synthesized at different times during human development and assemble around a heme group in specific combinations to form functional hemoglobin tetramers.

The human α - and β -globin gene clusters are located on chromosomes 16 and 11, respectively. In both clusters, genes are arranged along the chromosome in the order in which they are expressed during development.

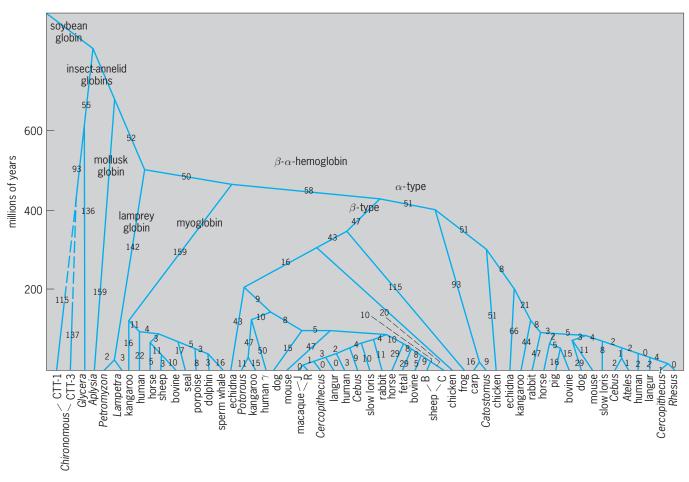


Fig. 2. Globin genealogical tree. Link lengths represent the number of nucleotide replacements between ancestral and descendant sequences. The ordinate scale, in millions of years, is based on paleontological views concerning the ancestral separations of organisms from which the globins came. (*After U. Goodenough, Genetics, 3d ed., Saunders, 1984*)

The similarity in nucleotide sequence of α - and β -globin genes indicates that they arose from a single ancestral gene, presumably by gene duplication. Phylogenetic trees constructed on the basis of amino acid sequence data from vertebrate and invertebrate globin genes suggest that this duplication occurred approximately 425 to 500 million years ago (Fig. 2). From such phylogenetic trees, in which the separation of organisms based on paleontological views is fitted against genealogical trees based on nucleotide replacement rates, it is possible to estimate the rate of molecular evolution for proteins. During the early vertebrate period, when monomeric hemoglobin evolved into a tetramer, an accelerated rate of nucleotide substitutions is observed. The sequences directly affected during this period appear to be those directly involved with heme contacts and others that facilitated oxygen binding. Similarly, following the α - β gene duplication, the sequences in the hemoglobin molecule which were directly involved with efficient oxygen delivery evolved much more rapidly than other positions.

Rates of rapid nucleotide substitutions are correlated with periods in which natural selection was presumably acting to produce beneficial adaptations in the α and β chains. Once optimal versions of these chains were produced, stabilizing selection presumably acted to conserve the perfected adaptions at the functionally important sites in the hemoglobin molecules.

How did the α - and β -gene clusters become dispersed on different chromosomes? Recent studies have shown that in *Xenopus laevis* the α - and β globin genes are closely linked and that the linked genes are found on two different chromosomes ($\alpha^1\beta^1$ and $\alpha^2 \beta^2$). In a related species, X. tropicalis, however, the α - and β -globin genes are found on only one chromosome. It is speculated that this type of duplication resulted from chromosome duplication by tetraploidization. Duplication of whole chromosome sets in an ancient ancestor followed by switching off of α genes in one location and β genes in another may also have taken place in vertebrate evolution. Since the α and β gene clusters are found on different chromosomes in birds and mammals, the dispersal of these genes presumably occurred 300 to 500 million years ago (before the divergence of birds and mammals from their tetrapod ancestor).

Analysis of α - and β -globin gene structure shows that each gene contains two introns, one of which is generally longer than the other. This is consistent with the idea that these genes had a common origin. Indeed, in all genes examined so far from mammals, birds, and amphibians, these introns, although

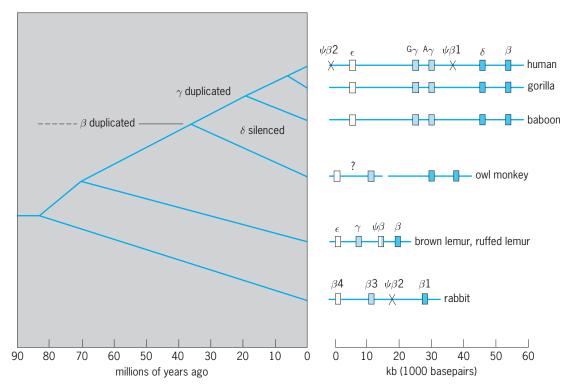


Fig. 3. Phylogeny of the β -globin clusters. The maps of β -globin clusters reveal the sequence in which certain genes changed through evolutionary time. The positions of pseudogenes are shown by crosses. (After R. Lewin, Evolutionary history written in globin genes, Science, 214:427, 1981)

varying in length, have remained in the same homologous positions.

The soybean leghemoglobin gene contains three, rather than two, introns. The first and third introns of the leghemoglobin genes are homologous to the first and second introns of the vertebrate genes, and are located at the same homologous positions found in the vertebrate gene. It is possible that the ancestral hemoglobin genes contained three introns, and the central intron was subsequently deleted, before the ancestral gene duplicated to establish the α and β groups. This explanation is supported by the observation that globin's amino acids are assembled in four distinct regions. The first region corresponds with the first exon, and the last with the third exon; the middle two regions are coded by the single middle exon. Leghemoglobin's "extra" intron, however, falls precisely at the division between the vertebrate globin's middle exon and its two heme-binding regions.

In contrast to the coding regions of the genes, the intervening sequences of genes generally accumulate mutations. Nucleotide sequence analysis of globin genes has shown that the coding regions have remained much more conservative than the noncoding regions, which include the 5' leader and 3' terminal regions, and introns. Some nucleotide sequences in the noncoding regions, however, such as the sequences involved in the RNA splicing mechanism, and those involved in poly-A initiation, have been preserved. The sequences in the α - and β -globin gene introns have nonetheless diverged considerably, although the sizes and positions of these sequences

have remained relatively stable. In contrast to the α and β -globin genes, most introns examined in other genes do not display such constraint. It is inferred from these properties that the hemoglobin introns may be involved in some unknown function.

The globin gene cluster is thought to have arisen as a gene duplication, followed by changes in gene sequences to produce a family of related although distinct genes that function at different points in development. The human β -globin gene cluster, which contains five closely related functional genes, plus two pseudogenes, has been examined extensively.

A phylogeny of the β -globin gene clusters shows that this gene cluster has remained remarkably stable with respect to the size and organization of the gene cluster for 20 to 40 million years of evolution (**Fig. 3**). Since 95% of the β -globin cluster contains DNA that does not code for proteins, it is inferred that this DNA might perform some as yet undefined function.

The complexity of the β -gene cluster is further illustrated by a number of phenomena. For example, although the γ -gene duplicated 20 to 40 million years ago, the high degree of similarity in the sequences of the duplicated genes suggests a very recent separation. Such similarity can be explained, however, by concerted evolution, whereby repeated genes exchange sequences by certain mechanisms, and maintain similar structures over long periods of evolutionary time.

Another observation in the β -globin cluster concerns the δ -globin gene, which is expressed in humans, apes, and New World monkeys, but not in

Old World monkeys. This gene, however, has been shown to be present in the β gene cluster of Old World monkeys as a pseudogene. It has been speculated that the few defects in this gene might be restored (for example, by back mutation) and the gene expressed in some future Old World monkey lineage. Pseudogenes can therefore be seen not only as evolutionary relics but also as potential new genes. Pamela K. Mulligan

Bibliography. A. Efstratiadis et al., The structure and evolution of the human beta-globin gene family, *Cell*, 21:653-668, 1980; M. Goodman, *Macromolecular Sequences in Systematic and Evolutionary Biology*, 1982; D. Jones, M. Goodfellow, and F. G. Priest (eds.), *Twenty-Five Years of Numerical Taxonomy*, 1983; R. Lewin, Evolutionary history written in globin genes, *Science*, 214:426-429, 1981; A. C. Wilson, S. S. Carlson, and T. J. White, Biochemical evolution, *Annu. Rev. Biochem.*, 46:573-639, 1977.

Proteocephaloidea

An order of tapeworms of the subclass Cestoda. With one exception, these worms are intestinal parasites of fresh-water fishes, amphibians, and reptiles. The holdfast organ bears four suckers and, frequently, an apical organ which may be suckerlike. The segmental anatomy is very similar to that of the Tetraphyllidea. Most authorities recognize two families, the Proteocephalidae, in which the reproductive organs are within the central mesenchyme of the segment, and the Monticellidae, in which some or all of the organs are in the cortical mesenchyme and which parasitize catfishes. The life histories of several proteocephalids have been studied. After leaving the host's intestine, the embryos must be ingested by an arthropod, typically a copepod of the genus Cyclops, in the body cavity of which the young worms grow, developing into procercoids and finally plerocercoids. The vertebrate host becomes infected by eating Cyclops containing plerocercoid larvae.

Proteocephalid-like forms in the ancient freshwater fishes probably were ancestral to the other orders of Cestoda. *See* CESTODA; TETRAPHYLLIDEA. Clark P. Read

Proterozoic

A major division of geologic time spanning from 2500 to 543 million years before present (Ma). The beginning of Proterozoic time is an arbitrary boundary that roughly coincides with the transition from a tectonic style dominated by extensive recycling of the Earth's continental crust to a style characterized by preservation of the crust as stable continental platforms. The end of the Proterozoic coincides with the Precambrian-Cambrian boundary, which is formally defined on the basis of the first appearance of diverse coelomate invertebrate animals. Proterozoic Earth history testifies to several remarkable biogeochemical events, including the formation and dispersal of the first supercontinent, the maturation of life and evolution of animals, the rise of atmospheric oxygen, and the decline of oceanic carbonate saturation (see **table**). Tremendous iron and lead-zinc mineral deposits occur in Proterozoic rocks, as do the first preserved accumulations of oil and gas. *See* CAMBRIAN; PRECAMBRIAN.

Tectonics. Many of the Earth's Archean cratons are blanketed by little-deformed sequences of Proterozoic sedimentary rocks, which indicate that vigorous recycling of the Earth's crust, characteristic of Archean time, had slowed markedly by the beginning of Proterozoic time. This decrease in crustal recycling is attributed to the development of thick continental roots, which stabilized the cratons, and the decrease in heat that was escaping from the Earth's interior, believed to drive thermal convection in the Earth's mantle and recycling of the crust. The sedimentary record that is present atop the major cratons is sufficiently continuous and well enough preserved that a rather detailed accounting of major events in Proterozoic Earth history is possible.

Most of the Earth's Archean cratons appear to have participated in the formation of a supercontinent in Mesoproterozoic time, about 1200 Ma. This supercontinent, called Rodinia, was the first for which there is now substantial evidence. Rodinia seems to have assembled with the North American craton (Laurentia) at its center. Rodinia persisted until the latest part of the Neoproterozoic, about 600 Ma, when it fragmented and ejected Laurentia in the process. Almost immediately, the other cratons reassembled to form the supercontinent of Gondwanaland, whose assembly was complete by about 450 Ma. *See* ARCHEAN; CONTINENTS, EVOLUTION OF; EARTH, CON-VECTION IN; EARTH, HEAT FLOW IN; EARTH CRUST; EARTH INTERIOR; PLATE TECTONICS.

Environment. Giant iron oxide deposits were formed by precipitation from seawater about 2000 Ma. These iron formations coincide with the first appearance of red, oxidized sedimentary rocks formed in terrestrial river environments and with the retention of ferric iron in soil horizons. During most of Paleoproterozoic time the oceans and atmosphere were reducing and ferrous iron was abundant in seawater. When photosynthetic microbes produced oxygen in shallow waters, the ferrous iron became oxidized to ferric iron, which is highly insoluble; and the ferric iron precipitated and settled to the sea floor, making the giant iron formations. The iron formations were produced until the iron sink was depleted about 2000 Ma, whereupon oxygen was free to accumulate in the atmosphere and shallow ocean. See ATMOSPHERE, EVOLUTION OF.

The partial pressure of carbon dioxide on the early Earth was very high. During Proterozoic time, much of the mass of carbon shifted from the ocean and atmosphere to the solid Earth. Enormous volumes of limestone [CaCO₃] and dolostone [CaMg(CO₃)₂] were deposited and testify to this shift. These carbonate deposites show significant textural variations, indicating that in the Paleoproterozoic supersaturation with respect to calcium carbonate was very

Proterozoic subdivisions and the major events				
Eon	Era	Historical benchmarks		
Proterozoic	Neoproterozoic 1 billion years Mesoproterozoic 1.6 billion years Paleoproterozoic	Evolution of animals Further rise of atmospheric oxygen Dispersal of Rodinia Assembly of Rodinia supercontinent Evolution of multicellular eukaryotes		
		Deposition of iron formations and rise of atmospheric oxygen Precipitation of limestones Evolution of eukaryotes		

high but declined markedly during the Mesoproterozoic. Neoproterozoic levels of calcium carbonate saturation were similar to those documented for Phanerozoic time. *See* DOLOMITE ROCK; LIMESTONE; SEDIMENTARY ROCKS.

Glaciers covered significant parts of the Earth during two widely separated times in Proterozoic history. The first episode occurred about 2200 Ma, and glacial deposits of that age cover various parts of North America and Scandinavia. The second episode consisted of at least two different pulses spanning from 750 to 600 Ma during Neoproterozoic time. Glaciers formed at that time were of almost global extent, and this was possibly the most frigid time in Earth history. Glaciers are thought to have extended from the poles to the Equator, according to the snowball Earth hypothesis.

Carbon and strontium isotopes present in limestones and dolostones suggest that several other environmentally significant events occurred toward the end of Neoproterozoic time. The presence of strontium isotopes suggests that high mountain ranges developed, probably in association with the earliest stages of Gondwanaland assembly. Rapid erosion of those mountains would have caused an increase in sedimentation rates in the deep ocean. Carbon isotopes suggest that much organic matter, normally a major oxygen sink, was buried because of the high sedimentation rates. As a consequence, oxygen was free to build up to unprecedented levels, perhaps triggering the evolution of animals.

Life. A number of significant events in the evolution of life occurred during Proterozoic time. The record of biological activity is rich, consisting of actual body fossils, in addition to organism traces and impressions, and complex chemical biomarkers. The geologic record of evolution has been independently tested by comparisons with the phylogenetic trees of living organisms constructed from information sequences contained within macromolecules such as ribonucleic acid (RNA) and various proteins. *See* RI-BONUCLEIC ACID (RNA).

Eukaryotic microbes appear to have evolved by about 1900 Ma, when they became major players in ecosystems present at that time. The fossil record is independently confirmed by the presence of steranes, known only from the membranes of eukaryotes, in 1700-Ma-old oil deposits. By the beginning of Neoproterozoic time, about 1000 Ma, multicellular eukaryotic algae are present in numerous sedimentary basins around the world. These eukaryotes, in addition to their reproductive cysts, evolved to fill distinct niches in ecologically complex ancient seas, as shown by onshore-offshore patterns of distribution. The eukaryotic fossil record shows evidence of a major adaptative radiation and subsequent extinction near the end of Neoproterozoic time. *See* EUKARYOTAE.

The evolution of animals did not take place until the close of Neoproterozoic time. The first direct evidence consists of relatively large impressions of soft-bodied organisms whose origin is still broadly debated. The consensus view is that the impressions were formed by unskeletonized animals such as cnidarians and bilaterians. The broader question of why these organisms evolved at this particular time in Earth history remains unanswered. General opinion proposes that it was likely the result of the confluence of a number of environmental factors, such as the rise in oxygen related to increased rates of carbon burial, on top of the previous 3 billion years of molecular and evolutionary preconditioning. Whatever the cause of their origin, these existed until at least 543 Ma, when another major evolutionary adaptive radiation began. This was the most impressive in Earth history, in which most of the extant phyla and classes of marine invertebrates evolved over a remarkably short time interval of about 10 million years. The beginning of this explosive diversification (545 Ma) marks the onset of Cambrian time and the end of the Proterozoic Eon. See ANIMAL EVOLUTION; EXTINCTION (BIOLOGY); GEOLOGIC TIME SCALE. John P. Grotzinger

Bibliography. S. Bengtson, *Early Life on Earth*, Nobel Symp. 84, 1994; J. P. Grotzinger et al., Biostratigraphic and geochronologic constraints on early animal evolution, *Science*, 270:598-604, 1995; P. F. Hoffman, Did the breakout of Laurentia turn Gondwanaland inside-out?, *Science*, 252:1409-1412, 1991; A. H. Knoll, End of the Proterozoic Eon, *Sci. Amer.*, 265:64-73, 1991; J. W. Schopf and C. Klein, *The Proterozoic Biosphere*, 1992.

Protista

The kingdom comprising all single-celled forms of living organisms in both the five-kingdom and sixkingdom systems of classification. Kingdom Protista encompasses both Protozoa and Protophyta, allowing considerable integration in the classification of both these animallike and plantlike organisms, all of whose living functions as individuals are carried out within a single cell membrane. Among the kingdoms of cellular organisms, this definition can be used to distinguish the Protista from the Metazoa (sometimes named Animalia) for many-celled animals, or from the Fungi and from the Metaphyta (or Plantae) for many-celled green plants. This system simplifies the lower-level classification (orders and superfamilies) where closely allied forms of flagellate protistans may possess chlorophyll and, therefore, be plantlike in their metabolism, or lack chlorophyll and feed as animals. However, there are other difficulties, not only of semantics but also of fundamental biology, involved in attempting diagnostic separation of the Protista from the other major groups of living organisms. See METAZOA.

The most significant biological distinction is that which separates the bacteria and certain other simply organized organisms, including blue-green algae (collectively, often designated Kingdom Monera), from both Protista and all many-celled organisms. Distinctions between certain plantlike Protista and the lower many-celled plants are very difficult to make. A distinction between the animallike Protista (or Protozoa) and many-celled animals is possible by a moderately complicated definition. Separation of the bacteria (and other Monera) is also easier semantically. The bacteria are described as prokaryotic; both the Protista and the cells of higher plants and animals are eukaryotic. Structurally, a distinguishing feature is the presence of a membrane, closely similar to the bounding cell membrane, surrounding the nuclear material in eukaryotic cells, but not in prokaryotic ones. This diagnostic feature is paralleled by other features of functional morphology; for example, eukaryotic cells can be considerably more compartmentalized with membrane-bound structures (organelles) responsible for particular cell functions. There are also fundamental differences between prokaryotic and eukaryotic organisms in the mechanics of cell multiplication, particularly as concerns division of nuclear material. See EUKARYOTAE; PROKARYOTAE; PROTO-ZOA

It is important to state that individual protistans are both whole organisms and individual cells. Evidence from studies by electron microscopy has settled earlier controversies. Semantically, protistans are unicellular rather than acellular. In the majority of its ultrastructures, the individual protistan cell appears to be strictly comparable to an individual generalized cell of a metazoan. The definition that can separate the Protista from many-celled animals is that the protistan body never has any specialized parts of the cytoplasm under the sole control of a nucleus. In some protozoa, there can be two, a few, or even many nuclei, rather than one, but no single nucleus ever has separate control over any part of the protistan cytoplasm which is specialized for a particular function. In contrast, in metazoans there are always many cases of nuclei, each in control of cells of specialized function.

Most authorities would agree that the higher plants, the Metazoa, and the Parazoa (or sponges) almost certainly evolved (each independently) from certain flagellate stocks of protistans. This evolutionary significance of the Protista remains valid, but work on the ultrastructure of and physiology in protozoa provides much evidence of considerable levels of adaptive complexity. Thus, the early hope that study of the apparent simplicity of protistans would more readily provide understanding of certain life processes than similar studies in higher forms has proven largely illusory. *See* CLASSIFICATION, BIOLOG-ICAL. W. D. Russell-Hunter

Protobranchia

A subclass of bivalve mollusks characterized by a foot with a sole that is divided sagittally and longitudinally and has papillate margins. The foot is without a true byssus gland, although a nonhomologous byssal gland is frequently present in the heel. The byssal gland does not produce threads and has a structure totally different from the byssus gland of the subclass Lamellibranchia. The gill filaments, which are not reflected, are usually platelike with abfrontal cilia. Thus the gills (etenidia) resemble those of aspidobranch gastropods rather than those of the lamellibranch bivalves. Digestion is entirely extracellular.

The Protobranchia are divided into two orders, Solemyoida and Nuculoida. The Solemyoida have an edentulous shell, and if teeth are present, they are not chevron-shaped. The palps are small, triangular, and without palp proboscides. This is in contrast to the Nuculoida, which have well-developed, chevronshaped hinge teeth and large palps with palp proboscides.

The Nuculoida are subdivided into two superfamilies, Nuculacea and Nuculanacea. The Nuculacea are asiphonate, without fusion or particular specialization of the mantle edge. For example, there is no anteroventral mantle sense organ. The Nuculanacea are siphonate, and ventral to the siphons the mantle is modified to form a feeding aperture. An anterior sense organ and a siphonal tentacle are attached to the siphonal embayment.

The Protobranchia are found throughout the seas of the world, and they are particularly common in the deep sea, where they may form up to 10% of the invertebrate infauna. The subclass has one of the longest geological records within the animal kingdom, dating from the Early Ordovician if not the Late Cambrian. During the course of evolution, they have changed little in their shell form. They probably owe their success in the deep sea to their deposit-feeding habit, which-coupled with extracellular digestion, low metabolic rate, and slow growth-has enabled them to exploit the low organic content of deep-sea soft sediments. Some deep-sea species are extremely long-lived: one species (Tindaria callistiformis) is known to live for 120 years, yet grows to little more than 0.4 in. (1 cm) total length. See BIVALVIA; MOL-John A. Allen LUSCA.

Bibliography. J. A. Allen and F. J. Hannah, A reclassification of the Recent genera of the subclass Protobranchia (Mollusca: Bivalvia), *J. Conchol.*, 32:225-249, 1986; E. R. Trueman (ed.), *The Mollusca*, vol. 10: *Evolution*, 1985.

Protogyny

A condition in hermaphroditic or dioecious animals in which the female reproductive structures mature before the male structures. It is of rare occurrence. Botanically, protogyny occurs in some plant species in which the stigma develops, withers, and dies before the anthers mature. *See* PROTANDRY.

Tracy I. Storer

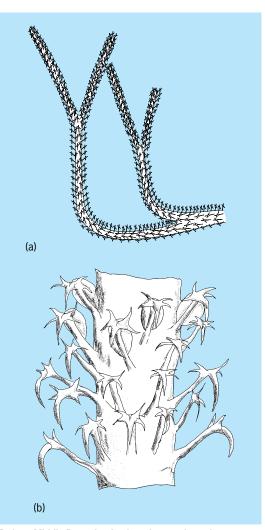
Protolepidodendrales

An extinct order of the class Lycopsida (clubmosses) of the Devonian Period. Most of these clubmosses are poorly preserved and not well understood; the notable exception is the widespread *Leclercqia complexa*, which has been largely reconstructed from fragmentary fossils.

Leclercqia (Early to Middle Devonian) was a medium-sized herb; its horizontal rhizomes generated vertical axes up to 0.4 in. (1 cm) in diameter (see illus.). The vascular tissue formed a starshaped actinostele, with inward maturation of the metaxylem tracheids. Additional structural support was provided in the cortical cylinder by thickened hypodermal fibers. The microphyllous leaves were not routinely shed. Most branched into five unequal, needlelike lobes, each supplied by a lateral branch of the median vascular trace, and arranged threedimensionally. Zones of sterile leaves alternated with zones of fertile sporophylls; both sterile and fertile leaves bore a small ligule on the upper surface. The stomata-bearing sporangia were located on the same surface of the sporophyll, somewhat closer to the stem than the ligule. The sporangia were elliptical and released spores of the same size; thus, Leclercqia is primitively homosporous.

Other putative protolepidodendraleans together span the Devonian Period. *Protolepidodendrom* (Early Devonian) had bifid leaves and sporophylls, whereas the leaves of *Colpodexylon* (Middle to Late Devonian) were trifid; their stems reached 0.8 in. (2 cm) in diameter. Although traditionally included, *Estinnophyton* (Early Devonian) is a doubtful member of the group, as it bore two pairs of spindleshaped sporangia per sporophyll rather than the single sporangium characteristic of the Lycopsida. *Haskinsia* and *Archaeosigillaria* (both Middle to Late Devonian) possessed relatively large axes that bore spear- or arrow-shaped microphyllous leaves.

Complex leaf morphologies are important characteristics in assigning these genera to the Protolepidodendrales, even though it may be that they represent multiple independent evolutionary events. The presence of a ligule has been demonstrated only in



Early to Middle Devonian *Leclercqia complexa*, the best-known protolepidodendralean. (a) General growth habit. (b) Pentafid sporophylls bearing sporangia in the fertile region of the axis. (After P. A. Bonamo et al., *Leclercqia*, Haskinsia, and the role of leaves in delineation of Devonian lycopod genera, Bot. Gaz., 149:222–239, 1988)

Leclercqia; yet it is the combination of homosporous reproduction and the presence of a ligule that delimits the order and testifies to its evolutionary intermediacy between the eligulate homosporous Lycopodiales and the ligulate heterosporous Selaginellales. *See* LYCOPHYTA; LYCOPODIALES; LYCOPSIDA; SELAGINEL-LALES. Richard M. Bateman; William A. DiMichele

Bibliography. P. G. Gensel, Phylogenetic relationships of the zosterophylls and lycopsids, *Ann. Mo. Bot. Gard.*, 79:450-473, 1992; F. M. Hueber, Thoughts on the early lycopsids and zosterophylls, *Ann. Mo. Bot. Gard.*, 79:474-499, 1992; W. N. Stewart and G. W. Rothwell, *Paleobotany and the Evolution of Plants*, 2d ed., 1993.

Proton

A positively charged particle that is the nucleus of the lightest chemical element, hydrogen. The hydrogen atom consists of a proton as the nucleus, to which a single negatively charged electron is bound by an attractive electrical force (since opposite charges attract). The proton is about 1836 times heavier than the electron, so that the proton constitutes almost the entire mass of the hydrogen atom. Most of the interior of the atom is empty space, since the sizes of the proton and the electron are very small compared to the size of the atom. *See* ATOMIC STRUCTURE AND SPECTRA; ELECTRIC CHARGE; HYDROGEN.

For chemical elements heavier than hydrogen, the nucleus can be thought of as a tightly bound system of Z protons and N neutrons. An electrically neutral atom will then have Z electrons bound comparatively loosely in orbits outside the nucleus. *See* NEUTRON; NUCLEAR STRUCTURE.

Most of the overall properties of the proton can be experimentally determined in straightforward ways. Free protons can be easily obtained for experimental purposes by ionization (electron removal) of hydrogen atoms in an electrical gas discharge. In some cases, measurements on the proton are best carried out by using neutral hydrogen atoms. In magnetic resonance experiments, the magnetic dipole moment of the proton has been measured with high accuracy. The numerical values of some overall properties of the proton can be summarized as follows: charge, 1.602×10^{-19} coulomb; mass, 1.673×10^{-27} kg; spin, $(1/_2)\hbar$ (where \hbar is Planck's constant *b* divided by 2π); magnetic dipole moment, 1.411×10^{-26} joule/tesla; radius, about 10^{-15} m. See FUNDAMENTAL CONSTANTS; ION SOURCES; MAG-NETIC RESONANCE; NUCLEAR MOMENTS; SPIN (QUAN-TUM MECHANICS).

Size and structure. In order to understand basic questions concerning the size and structure of the proton, it is instructive to contrast its properties with those of the electron. The behavior of the electron has also been studied in many precision experiments at both high and low energies, and all of the electron's properties have been found to be those expected of a spin-1/2 particle which is described by the Dirac equation of quantum mechanics. Such a Dirac particle has no internal size or structure. *See* ELECTRON; RELATIVISTIC QUANTUM THEORY.

By contrast, although it also has a spin of 1/2, the proton definitely does not behave as a geometrical point Dirac particle. Its magnetic moment, which is different from that for a Dirac particle, and its binding with neutrons into nuclei strongly suggest that the proton has some kind of internal structure, rather than being a point particle. That this is actually the case has been shown in many high-energy physics experiments. Experiments to explore the structure of the proton need to be carried out by using highenergy particles for two reasons. First, the uncertainty relationship between momentum and position, which is a fundamental principle of quantum mechanics, requires the use of high-momentum particles to observe small regions of space. Second, Einstein's equation of special relativity, $E = mc^2$, where E is energy, m is mass, and c is the speed of light, shows that to produce the additional mass which is needed to form excited states of the proton (described below) requires high-energy collisions. *See* UNCERTAINTY PRINCIPLE.

Correspondingly, two different kinds of highenergy physics experiments have been used to study the internal structure of the proton. An example of the first type of experiment is the scattering of highenergy electrons, above say 1 GeV, from a target of protons. The angular pattern and energy distribution of the scattered electrons give direct information about the size and structure of the proton. The second type of high-energy experiment involves the production and study of excited states of the proton, often called baryonic resonances. It has been found that the spectrum of higher-mass states which are produced in high-energy collisions follows a definite pattern. *See* BARYON.

In 1963, M. Gell-Mann and, independently, G. Zweig pointed out that this pattern is what would be expected if the proton were composed of three spin-1/2 particles, quarks, with two of the quarks (labeled *u*) each having a positive electric charge of magnitude equal to 2/3 of the electron's charge (*e*), and the other quark (labeled *d*) having a negative charge of magnitude of 1/3e. Subsequently, the fractionally charged quark concept was developed much further, and has become central to understanding every aspect of the behavior and structure of the proton.

Quark structure. As a result of experimental and theoretical research in high-energy physics since the 1960s, a broad understanding of the internal structure and constituents of the proton has been achieved. The behavior of the proton is essentially that of a system of three quarks, with the charges specified above. The excited, higher-mass states of the proton have been shown to correspond in some detail to the theoretical expectations for such a three-quark system. This correspondence is not yet on a precise basis, because the equations of the quantum-mechanical theory of interactions between quarks, quantum chromodynamics (QCD), have not yet been solved rigorously for a bound system of three quarks.

Quantum chromodynamics has led to other radically new insights into the behavior of the quarks. Among the most fundamental and yet paradoxical properties of the quarks, according to quantum chromodynamics, is the existence of a new quantum number called color, which acts in such a way as to forbid the stable existence of a single, isolated quark. The requirement that isolated physical systems have zero net color in fact forbids the existence of any free particles which have fractional charge. This conclusion has been tested by many experiments which have searched unsuccessfully for such particles. *See* COLOR (QUANTUM MECHANICS).

For the proton, the same principle leads to the conclusion that the proton is the lowest allowed energy state of three quarks. The proton cannot dissociate or decay into separated quarks, for that would lead to the existence of free quarks. In other words, the proton is absolutely stable with respect to the quantum chromodynamics interactions of quarks. *See* QUAN-TUM CHROMODYNAMICS; QUARKS. **Search for decay.** For many years, it has been recognized not only that the stability of the proton is of great practical importance, but also that this stability is probably due to some fundamental conservation law in physics. The proton and the neutron are the building blocks of which all nuclei are formed, and the observed stability of the atoms in the universe is a direct consequence of the stability of the proton itself.

As described above, quantum chromodynamics gives a clear description of why the proton, as a threequark system, cannot break apart into free quarks, but this leaves open the question as to whether the quarks themselves might decay. Such a question goes beyond the scope of quantum chromodynamics, but it can be answered by the unified electroweak theory of S. Weinberg, A. Salam, and S. Glashow. This theory gives a description of all known electromagnetic and weak (radioactive decay) phenomena, and has so far correctly predicted the results of all experimental measurements of these phenomena. The theory also gives a definite prediction that the proton is stable. *See* WEAK NUCLEAR INTERACTIONS.

However, an important class of fundamental theories, called grand unification theories (GUTs), makes the opposite prediction—that the proton will decay. The predicted lifetime of the proton is very long, about 10^{30} years or more—which is some 10^{20} times longer than the age of the universe—but this predicted rate of proton decay may be detectable in practical experiments. *See* GRAND UNIFICATION THE-ORIES.

The fundamental physics issue at stake in the search for proton decay is whether quarks, whose behavior is described in quantum chromodynamics, will ultimately be transformed into electrons or muons, which are described by a completely separate (electroweak) theory. If this process does occur, it will be unambiguous evidence for a single mechanism being at the root of all known forces in nature, excepting gravity. It will also permit the unification of all known particle types into one family.

To search for proton decay, experiments are being carried out which can detect and identify one or two proton decay events per year in a sample whose mass is of the order of 1000 tons (900 metric tons). To achieve such unprecedented sensitivity, the entire experiment is carried out in a deep underground laboratory. This shields the apparatus from most of the background events which would otherwise be caused by cosmic rays. The illustration shows a proton decay detector being constructed in a laboratory beneath the Alps. The detector is constructed of 900 metric tons (992 short tons) of thin iron plates, each 20 ft \times 20 ft \times 0.12 in. (6 m \times 6 m \times 3 mm). Particles resulting from proton decay within an iron nucleus will be detected as they traverse Geiger tubes and flash chambers which are sandwiched between the thin iron plates. Even in an underground laboratory, there is an unavoidable background caused by neutrinos, which are secondary products of cosmic rays. This neutrino background will set an ultimate limit on the sensitivity of proton decay experiments.



Proton decay experiment being assembled in an underground laboratory in the Frejus tunnel under the French Alps.

See GEIGER-MÜLLER COUNTER; NEUTRINO; PARTICLE DETECTOR.

If the proton is observed to decay, this new interaction will also have profound consequences for understanding of cosmology. The very early times of the big bang (about 10^{-30} s) are characterized by energies so high that the same grand unified interaction which would allow proton decay would also completely determine the subsequent evolution of the universe. This could then explain the remarkable astrophysical observation that the universe appears to contain only matter and not an equal amount of antimatter. The present standard theories of fundamental particle interactions, quantum chromodynamics and the electroweak theory, allow only processes in which an equal number of particles and antiparticles is created. These fundamental theories are therefore inconsistent with the apparent absence of antimatter in the universe. See COSMOLOGY; EL-EMENTARY PARTICLE; INFLATIONARY UNIVERSE COS-Thomas H. Fields MOLOGY.

Bibliography. M. Gaillard, Toward a unified picture of elementary particle interactions, *Amer. Sci.*, 70:506-514, 1982; H. Georgi, A unified theory of elementary particles and forces, *Sci. Amer.*, 244(4):48-63, 1981; R. G. Roberts, *The Structure of the Proton: Deep Inelastic Scattering*, 1993; W. B. Rolnick, *The Fundamental Particles and Their Interactions*, 1994; L. Sulak, Waiting for the proton to decay, *Amer. Sci.*, 70:616-625, 1982; S. Weinberg, The decay of the proton, *Sci. Amer.*, 244(6):64-75, 1981; V. F. Weisskopf, The origin of the universe, *Amer. Sci.*, 71:473-480, 1983.

Proton-induced x-ray emission (PIXE)

A highly sensitive analytic technique for determining the composition of elements in small samples. Proton-induced x-ray emission (PIXE) is a nondestructive method capable of analyzing many elements simultaneously at concentrations of parts per million in samples as small as nanograms. PIXE has gained acceptance in many disciplines; for example, it is the preferred technique for surveying the environment for trace quantities of such toxic elements as lead and arsenic. There has also been a rapid development in the use of focused proton beams for PIXE studies in order to produce two-dimensional maps of the elements at spatial resolutions of micrometers.

Apparatus. The typical PIXE apparatus uses a small Van de Graaff machine to accelerate the protons which are then guided to the sample. Nominal proton energies are between 1 and 4 MeV; too low an energy gives too little signal while too high an energy produces too high a background. The energetic protons ionize some of the atoms in the sample, and the subsequent filling of empty inner orbits results in the characteristic x-rays. These monoenergetic x-rays emitted by the sample are then efficiently counted in a high-resolution silicon (lithium) detector which is sensitive to the x-rays of all elements heavier than sodium. *See* JUNCTION DETECTOR; X-RAYS.

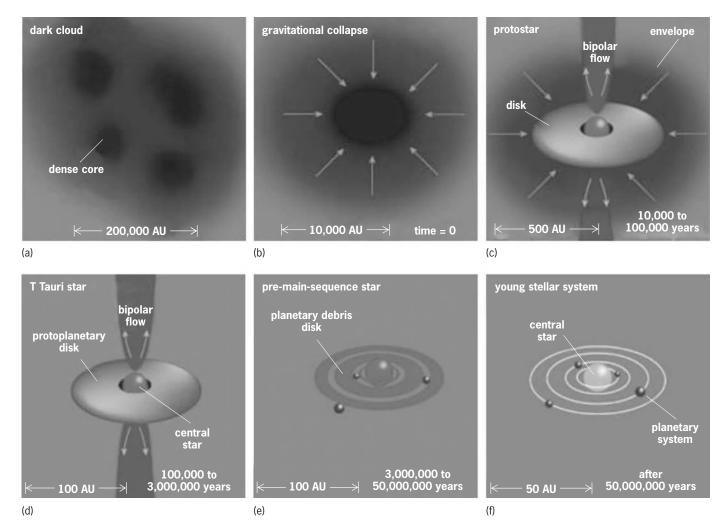
Advantages. The advantages of PIXE over electroninduced x-ray techniques derive from the heaviness of the proton which permits it to move through matter with little deflection. The absence of scattering results in negligible continuous radiation (bremsstrahlung) so that the backgrounds under the characteristic x-ray lines come mainly from secondary effects. As a result, proton-induced x-ray techniques are two to three orders of magnitude more sensitive to trace elements than are techniques based on electron beams. Another important consequence is that the proton beam can be extracted from the vacuum so that samples, in particular of biological material, can be studied in controlled atmospheric environments. *See* BREMSSTRAHLUNG.

Application. PIXE was developed in European laboratories in the early 1970s. At the University of Lund, Sweden, and at the Niels Bohr Institute in Copenhagen, the emphasis was on using broad beams to obtain great sensitivity to trace elements in bulk samples; at the Harwell National Laboratory in England the emphasis was on focused beams to obtain spatial resolution at some sacrifice in analytic sensitivity. Some 40 to 50 PIXE laboratories now carry out investigations in such diverse fields as materials science, earth and planetary sciences, biomedicine, environment, and archeology. A number of these laboratories have sophisticated facilities in which samples are scanned under computer control by micrometer-size beams. Previously the high cost of the proton accelerator confined PIXE to laboratories which used existing Van de Graaffs, which were constructed mainly in the 1950s and 1960s for nuclear studies. Several companies now market small accelerators which are suited to PIXE, and commercial laboratories offer analytic PIXE services to the technology community. See ACTIVATION ANALYSIS. Lee Grodzins

Protostar

A hydrostatic sphere of gas in the process of becoming a star. A protostar forms by the gravitational collapse of a dense core within a giant cloud of dust and molecular gas (mostly H₂). As the core collapses from the inside out, it surrounds a central protostar with a cocoon of accreting dust and gas that hides it from view at optical wavelengths. Observations at longer wavelengths penetrate this material and reveal that the protostar is radiating due to the impact of infalling gas and dust. Most of the envelope material accretes first onto a circumstellar disk, from which it is then conveyed to the protostellar surface. As the envelope dissipates around a low-mass star such as the Sun, the protostar becomes visible as a pre-main-sequence T Tauri star that continues for a time to add to its mass from the accretion disk. It becomes a full-fledged star when the core temperature reaches the level required for nuclear fusion of hydrogen (temperature $T \sim 10^7$ K). For stars at least 10 times more massive than the Sun, nuclear burning begins before infalling material is dissipated, and the resulting high-intensity radiation quickly clears away the remaining envelope. See T TAURI STAR.

Observations. Dark voids in the Milky Way were believed early on to be the result of intervening cold dust clouds that might serve as sites for star formation. Pioneering infrared astronomical surveys in the 1960s and 1970s revealed the first candidate protostars embedded deep within these clouds. Groundwork for a modern understanding of star formation was laid in the following decade. Improved surveys with ground-based telescopes led to an empirical classification of young stars according to their distribution of radiation with wavelength; those with radiation that peaked at the longest wavelengths were interpreted as youngest. Surveys of molecular clouds with submillimeter and millimeter-wave techniques provided the first maps of individual cores in molecular clouds; the Doppler shift of their molecular lines revealed the unexpected presence of bipolar molecular outflows and established their rotation rates. Finally, the Infrared Astronomical Satellite (IRAS) detected infrared protostars in the center of many of these cores and provided a better measure of their radiation at wavelengths of 12 to 100 micronmeters. Cores without IRAS detections were believed to reveal conditions in pre-protostellar cores. However, subsequent observations with the Spitzer Space Telescope extended infrared studies to much fainter levels and showed that at least some of these cores have infrared sources that are either proto-brown dwarfs or protostars in their earliest phases of growth. Since the mid-1990s, high-resolution images of the disks and jets from protostars and T Tauri stars have also been obtained with millimeter-wave interferometers, large ground-based telescopes, and the Hubble Space Telescope. See DOPPLER EFFECT; HUBBLE



Early evolution of a protostar. (a) Dense cores appear within in a larger molecular cloud. AU = astronomical unit. (b) Some cores become unstable to gravitational collapse. (c) Mass rapidly accumulates on a circumstellar disk accreting onto the star and fueling a bipolar outflow. An opaque dusty envelope hides the protostar from view at optical wavelengths. (d) The dusty envelope is dissipated, and the protostar becomes visible as T Tauri star. (e) The disk disperses, perhaps through planet formation, while the star is still in a pre-main-sequence phase. (f) A true main-sequence star emerges when core temperature is hot enough for nuclear fusion. (After T. Greene, Protostars, Amer. Scientist, 89:316–325, 2001)

SPACE TELESCOPE; INFRARED ASTRONOMY; MOLECU-LAR CLOUD; SPITZER SPACE TELESCOPE; SUBMILLIME-TER ASTRONOMY.

Evolution. On the basis of observations and theoretical studies, astronomers now recognize distinct evolutionary phases in the life of a protostar (see illustration). Cold condensations in molecular clouds appear initially as the result of turbulent conditions in the molecular gas with sizes of a fraction of a parsec (which is approximately 200,000 astronomical units), temperatures of 10 to 50 K, and densities of 10⁴ to 10⁵ atoms per cubic centimeter, significantly more than the surrounding medium (illus. *a*). James Jeans showed in the early twentieth century that such a core would collapse under its own gravity if it exceeded a critical mass for its size. During this earliest collapse phase, a protostar may be undetectable even at infrared wavelengths due to the opacity of the surrounding envelope (illus. b). Modern theory modifies Jeans' calculations by taking into account core rotation and magnetic fields in order to explain the observed formation of a circumstellar disk with bipolar outflow (illus. c). Studies of the number of objects in each phase within a single molecular cloud yield estimates of their evolutionary time scales, from 10^4 years for earliest collapse phase (illus. b), to 10^5 years for the optically obscured phase (illus.c), to a few times 10⁶ years for the T Tauri phase (illus. d). Stars such as the Sun reach the main sequence in several times 10^7 years when their cores reach the temperatures of the order of 10^7 K that are needed to begin nuclear hydrogen burning. The disk disperses in the pre-main-sequence phase (illus. e). However, upon reaching the main sequence, some stars may still have residual disk material that is no longer accreting onto the star but is probably related to the formation of a young planetary system (illus. f).

The above description must be revised for stars much more massive than the Sun. These evolve rapidly to a highly luminous phase in which intense ultraviolet radiation ionizes the surrounding material to form an expanding H II region. The infall envelope dissipates in about 10,000 years, but toroidal material around one or more high-mass stars may continue to accrete and fuel powerful bipolar outflows. The exact nature of the accretion process in this environment is not yet well understood. Observations with instruments now under development such as the Atacama Large Millimeter Array and both the *Herschel* and the *James Webb* space telescopes will undoubtedly shed more light on this problem. *See* INTERSTELLAR MATTER; NEBULA; STELLAR EVOLU-TION. David Koerner

Bibliography. P. André, D. Ward-Thompson, and M. Barsony, From prestellar cores to protostars: The initial conditions of star formation, in V. Mannings, A. P. Boss, and S. S. Russell (eds.), *Protostars and Planets IV*, pp. 59-96, University of Arizona Press, 2000; T. Greene, Protostars, *Amer. Sci.*, 89:316-325, 2001; C. J. Lada and F. H. Shu, The formation of sunlike stars, *Science*, 248:564-572, 1990; F. H. Shu, F. C. Adams, and S. Lizano, Star formation in molecular clouds: Observations and theory, *Annu. Rev. Astron. Astrophys.*, 25:23-84, 1987; S. W. Stahler, F. Palla, and P. T. P. Ho, The Formation of massive stars, in V. Mannings, A. P. Boss, and S. S. Russell (eds.), *Protostars and Planets IV*, pp. 327-351, University of Arizona Press, 2000.

Prototheria

One of the four subclasses of the class Mammalia. Prototheria contains a single order, the Monotremata. No ancestral genera of fossil monotremes are known, and the structure of the living monotremes is so specialized that the affinities of the Prototheria are largely conjectural. Most mammalogists believe that the prototheres arose from a different stock of therapsid reptiles than the one that gave rise to the Theria.

This origin from a different stock would mean that the history of the monotremes has been separate from the history of other mammals for at least 1.75×10^8 years.

No fossils earlier than the Pleistocene are known, and these come from Australia. The duckbill platypus (*Ornithorbynchus anatinus*) and several species of the echidnas, *Tachyglossus* and *Zaglossus*, are living representatives of this group. They are found in the Australian region. Everything indicates that the Prototheria represent a very small and relatively unsuccessful group that has miraculously survived in an isolated corner of the Earth. *See* MAMMALIA; MONOTREMATA; PLATYPUS; THERAPSIDA. D. Dwight Davis; Frederick S. Szalay

Bibliography. R. M. Nowak, *Walker's Mammals of the World*, Johns Hopkins University Press, 1999; J. M. Taylor, *Oxford Guide to Mammals of Australia*, 1984.

Prototype

A first or original model of hardware or software. Prototyping involves the production of functionally useful and trustworthy systems through experimentation with evolving systems. Generally, this experimentation is conducted with much user involvement in the evaluation of the prototype.

Primary uses. A primary use for prototyping is the acquisition of information that affects early product development. For example, if requirements for human-computer interfaces are ambiguous or inadequate, prototyping is frequently used to define an acceptable functional solution. It is a method for increasing the utility of user knowledge for purposes of continuing development to a final product. Information obtained through prototyping is important to designers, managers, and users in identifying issues and problems. Prototyping conserves time and resources prior to the commitment of effort to construct a final product. *See* HUMAN-COMPUTER INTERACTION.

Prototyping is an important way to obtain user input related to a product that is being developed. The major ways in which the user participates in development of a product include involvement during requirements identification, review of operational functionality and environmental changes that influence installation and maintenance, and work with designers to gain a better understanding of product application. In large-scale complex systems, such as real-time embedded avionics systems, it is often necessary to construct full-scale models of portions of the system, both hardware and software, to demonstrate the feasibility of the proposed solution. *See* EMBEDDED SYSTEMS.

Prototyping increases worker productivity through modeling of the development process. It provides support for requirements identification and the transformation of requirements into product specifications, as well as for understanding the operational environment. Prototyping assists in understanding the intended and desired systemlevel functionality in terms of what the product should accomplish. For example, in the development of a database to support a complex system, all the operations involving database structures, database query or report features, and data storage are prototyped to assure performance, size, and access of the database. See DATABASE MANAGEMENT SYSTEM

Classification. Several schemes are used to classify prototypes. Examples include functional, structural, purposeful, operational, rapid, experimental, and exploratory. Functional prototyping helps demonstrate that the behavior of the product is what the user intended. Structural prototyping aids in determining that the particular design approach used is feasible and is evolving toward production. Purposeful prototyping is used to ascertain that user requirements and product performance are consistent. Operational prototyping is done in the field as a temporary add-on to an existing system. Rapid prototyping supports early definition and validation of requirements. Experimental prototyping is used for the investigation of alternative approaches to a solution. Exploratory prototyping is used to assist in the formulation of a problem or issue.

Prototypes are also classified according to the uses intended. Examples include architectural, organizational, and simulation or animation. Architectural prototyping assists in visualization of the allocation of functions to system structure. Organizational prototyping assists in the visualization of functional operations along enterprise lines. Simulation and animation prototyping serve to test the viability of proposed solutions and provide graphical representations of the solutions. *See* SIM-ULATION.

A prototype is generally a functionally immature model of a proposed product that is built to explore requirements, investigate alternative approaches, or demonstrate model feasibility. The model may or may not evolve into a mature, functionally useful product.

Methods of use. Prototyping generally is applied in stages. Methods used in these stages may be classified into three groups. The first, concept prototyping, is used to encourage dialogue between the user and the developer to reach mutual agreement that a proposed concept will satisfy user requirements. The concept prototype can vary from a totally static demonstration model to one involving system dynamics. A structural, functional, or purposeful prototype can be built as a concept prototype. Rapid prototyping, involving interface designs, animation of the activity, or multimedia representations, is most often used for concept development.

The next stage, laboratory prototyping, involves a working model and begins as soon as the concept prototype is accepted. Advantage is taken of a friendlier development environment and development tools than might be available in the actual operating environment. Structural (incremental or evolutionary) or functional (experimental) prototyping would be utilized at this stage.

Field prototyping begins after the laboratory prototype is accepted and becomes the first unit to be installed in the field. It is possible to test and evaluate a field prototype under actual operating conditions. This is still regarded as a prototype since user feedback may result in modifications. A field prototype would usually be a structural prototype.

Application phases. There are seven phases in the application of prototyping. These move from the initial request, through selection of an approach, to selection of the methods, to construction and validation.

The first phase involves identifying and evaluating the request for software development services. Preliminary information requirements concerning user needs are determined. The most appropriate approach to product development is identified; in particular, whether prototyping is appropriate is determined in this phase. Influential factors include application area and complexity, user experiential familiarity with the application area, and developer experience with prototyping.

The remaining phases involve identifying the type of prototyping that is most appropriate, which will be some form of structural, functional, or purposeful prototyping; developing a set of user and product requirements appropriate for the prototyping approach selected and the resulting prototype; identifying a set of design specifications for the prototype; creating, testing, and refining the prototype; involving the client or user in use of the prototype in an appropriate way relative to accomplishment of prescribed prototyping tasks; and determining whether the prototype developed is acceptable. If the prototype is acceptable, the next phase in the product development life cycle can be undertaken. Otherwise, the designer must iterate back to identifying the type of prototyping that is most appropriate, if it is believed that the initial judgment concerning prototyping type is flawed; or to refining the prototype if the existing prototype is appropriate.

Characteristics and uses. In many hardware and software development projects, the first prototype product built is barely usable. It is usually too slow, too big, too awkward in use. Hence, the term throw-away prototype is generally applied to describe this early use of prototyping. Usually this is due to lack of understanding of user requirements. There is no alternative but to start again and build a redesigned version in which these problems are solved.

A developmental prototyping approach for incremental design of subsystems is often used to reduce the risk involved in building a system-level prototype. In this prototyping environment an incremental approach to rapid prototyping of subsystems development is used. This provides for management oversight of the entire process to assure that resource usage is effective and efficient. Product assurance is implemented throughout the process to make certain that the prototype operation contains the necessary components to satisfy subsystem requirements. Requirements analysis is performed and reviewed, then incremental specifications are developed and reviewed, followed by design of the approved specifications, and completed by implementation of the product.

The decision concerning whether a developmental prototype should be refined into the complete product or be thrown away should be based on the costs and benefits of each alternative. Two key factors to be addressed are the degree of functionality in the prototype and whether or not the prototype design supports a maintainable product. Response to these concerns determines whether the prototype is worth the investment of more effort and money.

The following example illustrates the use of prototypes: A firm is experiencing difficulties in its correspondence operation and is unable to provide accurate and timely responses and reports to customers. Resolution is required immediately, but system requirements are not clearly stated. A description of the flow shows correspondence is received, reviewed, filed, and then manually distributed to executive offices for review and response generation. The process involves the selection of offices that are to generate a response, called action offices, and selection of reviewers to comment on the proposed response. There is no automated means of maintaining a list of action offices and reviewers appropriate for different subject areas. A packet, containing the correspondence and any background material, is manually circulated throughout the operation, starting and ending with an executive secretariat. Correspondence may come from any number of sources and may need to be responded to in a timely manner. Corporate reports and other materials need to be distributed to grantees, supplier organizations, and other interested parties, both inside and outside the company. Currently, copies of reports are generated on paper and mailed to outside individuals. The listings of groups to which reports are sent are not maintained in the corporate management information system.

Solutions were proposed using rapid prototyping together with animation of current and proposed document flows. Current document flows showed multiple handling of the same document by the same desk and no automatic tracking of the location and status. No reporting capabilities were provided, and no database of reports was available. The rapid prototype animation allowed executive secretariat personnel to view alternative approaches and finally settle on a single approach that satisfied correspondence management needs. The rapid prototype animation was performed on a personal computer using commercially available software.

Advantages. The major potential advantages to the use of prototyping relate to the ability to rapidly construct a product that is representative of the final product at relatively low cost and effort. Nearly all products are developed by using a life cycle, and prototyping enhances the descriptive power of the product life cycle development process. Prototyping allows the usual product evolution to take place in a modeling environment that simulates the natural environment without the expenses associated with replication of that environment. It provides control over the product development process by incorporating aspects of resource management, configuration management, verification, and validation in early phases as well as later phases of the life cycle. Prototyping enables the construction of larger and more complex products with smaller development teams and promotes user feedback early in the development process. See MODEL THEORY; SOFTWARE ENGINEERING; SYSTEMS ENGINEERING.

James D. Palmer Bibliography. W. W. Agresti (ed.), *New Paradigms for Software Development*, 1986; V. E. Church et al., An approach for assessing software prototypes, *ACM Software Eng. Notes*, 11(3):65–76, July 1986; S. Hekmatopour and D. Ince, *Software Prototyping, Formal Methods and VDM*, 1988; H. Hunke (ed.), *Software Engineering Environments*, 1981; J. D. Palmer, J. O. Berkey, and V. Kountouris, Software development environments for reusability and prototyping, *Large Scale Sys.*, 12:201–216, 1987; A. P. Sage and J. D. Palmer, *Software Systems Engineering*, 1990; I. Sommerville, *Software Engineering*, 4th ed., 1992.

Protozoa

A group of eukaryotic microorganisms traditionally classified as one of the five animal kingdoms. Although the name signifies primitive animals, some protozoa (phytoflagellates and slime molds) show enough plantlike characteristics to justify claims that they are plants. This apparent conflict may be reconciled to some extent by considering protozoa as descendants of microorganisms which preceded plants and animals. Such an assumption might account for Euglena gracilis, a green flagellate which can fix carbon dioxide (CO₂) in light as algae do but, unlike algae, can also carry on a plantlike fixation in darkness. In addition, E. gracilis can fix carbon dioxide by pathways characteristic of animals. Inclusion of such metabolically unspecialized organisms, along with certain photoautotrophic chlorophyll-bearing species, might imply that the phylum is partly an arrangement for taxonomic convenience rather than a phylogenetically homogeneous group.

Protozoa are almost as widely distributed as bacteria. Free-living types occur in soil, wet sand, and in fresh, brackish, and salt waters. Protozoa of the soil and sand live in films of moisture on the particles. Habitats of endoparasites vary. Some are intracellular, such as malarial parasites in vertebrates, which are typical Coccidia in most of the cycle. Other parasites, such as *Entamoeba histolytica*, invade tissues but not individual cells. Most trypanosomes live in the blood plasma of vertebrate hosts. Many other parasites live in the lumen of the digestive tract or sometimes in coelomic cavities of invertebrates, as do certain gregarines. *See* COCCIDIA; GREGARINIA; TRYPANOSOMATIDAE.

Many protozoa are uninucleate, others binucleate or multinucleate, and the number also may vary at different stages in a life cycle. Size ranges from less than 3 micrometers to a number of centimeters, as in various Mycetozoia. Colonies are known in flagellates, ciliates, and Sarcodina. Although marked differentiation of reproductive and somatic zooids characterizes certain colonies, such as *Volvox*, protozoa have not developed tissues and organs.

Taxonomy

Various systems of classification have been proposed, and none is yet accepted by all protozoologists. One of those sponsored by the Society of Protozoologists follows:

Phylum Protozoa Subphylum 1: Sarcomastigophora Superclass 1: Mastigophora Class 1: Phytamastigophorea Order 1: Chrysomonadida Order 2: Silicoflagellida Order 3: Coccolithophorida Order 4: Heterochlorida* Order 5: Cryptomonadida Order 6: Dinoflagellida Order 7: Ebriida

Order 8: Euglenida Order 9: Chloromonadida Order 10: Volvocida (Phytomonadida) Class 2: Zoomastigophorea Order 1: Choanoflagellida Order 2: Bicosoecida Order 3: Rhizomastigida Order 4: Kinetoplastida Order 5: Retortomonadida (Protomastigida) Order 5: Retortomonadida Order 6: Diplomonadida Order 7: Oxymonadida Order 8: Trichomonadida Order 9: Hypermastigida Superclass 2: Sarcodina Class 1: Rhizopodea Subclass 1: Lobosia Order 1: Amoebida Order 2: Arcellinida Subclass 2: Filosia Order 1: Aconchulinida Order 2: Gromiida Subclass 3: Granuloreticulosia Order 1: Athalamida Order 2: Foraminiferida Order 3: Xenophyophorida Subclass 4: Mycetozoia Order 1: Acrasida Order 2: Eumycetozoida Order 3: Plasmodiophorida Subclass 5: Labyrinthulia Class 2: Piroplasmea Class 3: Actinopodea Subclass 1: Radiolaria Order 1: Porulosida Order 2: Oculosida Subclass 2: Acantharia Order 1: Acanthometrida Order 2: Acanthophractida Subclass 3: Heliozoia Order 1: Actinophryida Order 2: Centrohelida Order 3: Desmothoracida Subclass 4: Proteomyxidia Subphylum 2: Sporozoa Class 1: Telosporea Subclass 1: Gregarinia Order 1: Archigregarinida Order 2: Eugregarinida Order 3: Neogregarinida Subclass 2: Coccidia Order 1: Protococida Order 2: Eucoccida Suborder 1: Adeleina* Suborder 2: Eimeriina* Suborder 3: Haemosporina Class 2: Toxoplasmea Order 1: Toxoplasmida Class 3: Haplosporea Order 1: Haplosporida Order 2: Sarcosporida

Order 4: Entodiniomorphida Subphylum 3: Cnidospora Class 1: Myxosporidea Order 1: Myxosporida Order 2: Actinomyxida Order 3: Helicosporida Class 2: Microsporidea Class 1: Ciliatea Subclass 1: Holotrichia Order 1: Gymnostomatida Order 2: Trichostomatida Order 3: Chonotrichida Order 4: Apostomatida Order 5: Astomatida Order 6: Hymenostomatida Order 7: Thigmotrichida Subclass 2: Peritrichia Subclass 3: Suctoria Order 1: Suctorida* Subclass 4: Spirotrichia Order 1: Heterotrichida Order 2: Oligotrichida Order 3: Tintinnida Order 4: Entodiniomorphida Order 5: Odontostomatida Order 6: Hypotrichida

See separate articles on each taxonomic group except those marked with an asterisk.

Morphology

A protozoon may be a plastic organism (ameboid type) but changes in form are often restricted by the pellicle. In *Trypanosoma lewisi*, parallel micro-tubules beneath the pellicle may support this layer and may also, as probably contractile structures, cause the typical undulations of these flagellates. Although protozoan form varies considerably, there is a tendency toward universal symmetry in floating species and radial symmetry in sessile types. Bilateral symmetry is rare; most swimmers show appreciable spiral torsion.

A protective layer is often secreted outside the pellicle, although the pellicle itself may be strengthened by incorporation of minerals. Secreted coverings may fit closely, for example, the cellulose-containing theca of Phytomonadida and Dinoflagellida, analogous to the cell wall in higher plants. The dinoflagellate theca (Fig. 1a) may be composed of plates arranged in a specific pattern. Tests, as seen in Rhizopodea (Arcellinida, Gromida, Foraminiferida), may be composed mostly of inorganic material, although organic (chitinous) tests occur in certain species. Siliceous skeletons, often elaborate, characterize the Radiolaria (Fig. 1d and Fig. 2a-e). A vase-shaped lorica, from which the anterior part of the organism or its appendages may be extended, occurs in certain flagellates (Fig. 1b) and ciliates (Fig. 1c). Certain marine ciliates (Tintinnida) are actively swimming loricate forms.

Colonies and aggregates. An important feature of colonies is the material which holds the component zooids together. In spheroid colonies (**Fig. 3**) a

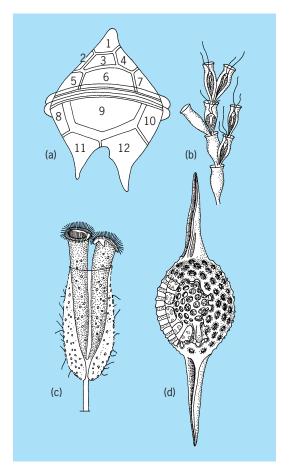


Fig. 1. External coverings of protozoa. (a) Theca of dinoflagellate (*Peridinium*), showing separate plates. (b) Lorica of a colonial chrysomonad, *Dinobryon*. (c) Two zooids within a lorica of a peritrich, *Cothurnia*. (d) A radiolarian skeleton, siliceous type. (After L. H. Hyman, The Invertebrates, vol. 1, McGraw-Hill, 1940)

matrix is secreted during development of the colony. Spheroid colonies of a few Radiolaria (Colozoum), unusually large examples, may measure 4-6 cm across. In arboroid colonies (Fig. 4b and j) a branching pattern is produced by a branching stalk (Fig. 4b) or by attachment of loricae to one another (Fig. 1b). Dispersal of such colonial species sometimes involves migratory stages. In certain spheroid colonies of Pleodorina (Fig. 3b and c) and Volvox, reproductive and somatic zooids are differentiated. More often members of a colony are similar. In addition to colonial types, certain protozoa may form aggregates by not separating promptly after fission. Palmella stages of Phytomonadida, analogous to spheroid colonies, are aggregates of nonflagellated organisms. Chains of Gonyaulax catenella and certain other Dinoflagellida are another type of aggregate.

Flagella. These organelles occur in active stages of Mastigophora and flagellated stages of certain Sarcodina and Sporozoa. A flagellum consists of a sheath enclosing a matrix in which an axoneme extends from the cytoplasm to the flagellar tip. In certain groups the sheath shows lateral fibrils (mastigonemes) which increase the surface area and also may modify direction of the thrust effective lo-

comotion. In electron micrographs axonemes show two central and nine pairs of peripheral microtubules. The base (blepharoplast) in certain species may contain only nine peripheral triplets. The blepharoplast is considered a self-replicating structure from which new flagella arise in fission. Flagellar number ranges from one or two for most free-living species to many as seen on certain symbionts of termites and wood roaches. Most flagella extend forward in swimming; others extend posteriorly, sometimes driving the body forward, sometimes serving as a ski in gliding, or forming the margin of an undulating membrane. Certain organelles, as the kinetoplast (Kinetoplastida), axostyle, parabasal body (Trichomonadida), or sometimes other structures, may be attached to or adjoin the blepharoplasts. See CILIA AND FLAGELLA; KINETOPLASTIDA.

Flagellar movements include planar sinusoidal waves, helical waves, and lateroposterior strokes. When a planar wave travels from base to tip, the force is directed basally, pushing the organism (Ceratium and others). However, a base-to-tip beat pulls Ochromonas, a chrysomonad, the mastigonemes changing a potential push into a pull. The flagellate is pulled (trypanosomes) when a planar wave travels from tip to base. Helical waves from tip to base (trichomonad undulating membrane) push the flagellate. A similar wave, from tip to base, pulls in Peranema. Lateroposterior beating (Monocercomonas neosepsorum and certain other trichomonads) more or less resembles a breast stroke in human swimming. An unusual jet propulsion observed for Medusochloris sp., a phytomonad, involves contractions of a medusa-shaped body. Another substitute for flagellar drive is seen in Mixotricha paradoxa, a parasite of an Australian termite. Most of the body is covered with "brackets," each of which bears a posterior pit into which the tip of a spirochete fits. Coordinated undulations of the spirochetes in metachronal waves, beginning at the anterior end of the flagellate, drive the organism smoothly forward. See MASTIGOPHORA.

Cilia. Although typically shorter than flagella, cilia are similar in structure, even to the fibrillar triplets in the base (basal granule and kinetosome). Multiplicity of cilia entails complex fibrillar systems. It has been suggested that several sets of fibrils help support the cortex and ciliary system. A more delicate fibril, or tubule (Fig. 5), running along each row of basal granules (Tetrahymena) seemingly offers little support and may, as suggested, function in coordination (Fig. 5). Such ciliates as hypotrichs, with compound organelles, may have quite complex fibrillar systems. Compound ciliary organelles represent fused bundles or rows of cilia, with the basal granules detectable in basal plates of the organelles. Membranes (longitudinal rows) occur along the right margin of the buccal cavity, if they are present. Membranelles (transverse rows) often form an adoral zone (Fig. 4a) along the left margin of the buccal cavity. Cirri (Fig. 4a), typically ventral, are fused tufts of cilia, several to 50 or more in large cirri, important in locomotion.

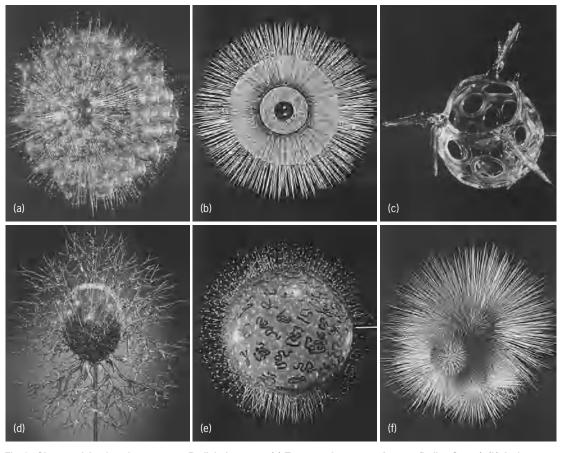


Fig. 2. Glass models of marine protozoa. Radiolarian types: (a) *Trypanosphaera transformata* (Indian Ocean); (b) *Actissa princeps* (Indian and Pacific oceans); (c) *Peridium spinipes* (Pacific Ocean); (d) *Lithocircus magnificus* (Atlantic Ocean); (e) *Collozoum serpentinum* (Atlantic Ocean). (f) Foraminiferan type: the pelagic *Globigerina bulloides* (which is found in all seas). (*American Museum of Natural History*)

Cortical myonemes (**Fig. 6**) are well developed in certain large ciliates (*Spirostomum* and *Stentor*). Similar cortical myonemes, composed of several to a dozen or more fibrils, occur in various gregarines. Ectoplasmic myonemes of certain ciliates also show longitudinal fibrils. Analogous stalk "muscles" occur in the stalks of *Vorticella* and certain colonial ciliates (*Carchesium* and *Zoothamnium*). *See* CILIOPHORA.

Pseudopodia. Two major types have been described, the contraction-hydraulic and the two-way flow types. The first are lobopodia with rounded tips and ectoplasm denser than endoplasm. The larger ones commonly contain granular endoplasm and clear ectoplasm.

According to the contraction-hydraulic theory, pseudopodia of *Amoeba proteus* arise as a result of pressure from the contracting ectoplasm on the endoplasm, forcing the latter forward, away from the posterior contraction. Local weakening of the ectoplasm results in a bulge, or developing pseudopodium. Endoplasm reaching the tip of the growing pseudopodium is diverted to the ectoplasm and converted into gel, thus becoming part of the ectoplasm. In the meantime, ectoplasm at the other end is being converted into endoplasm, which flows forward again under pressure from the ectoplasm.

Two-way flow pseudopodia include reticulopodia of Foraminiferida and related types, filoreticulopodia

of Radiolaria, and axopodia of certain Heliozoia. Axopodia of some genera show an axoneme composed of microtubules extending from the centroplast; in others, of a cylinder surrounded by microtubules. In addition, similar filoreticulopodia (no axonemes) occur in Radiolaria and certain other groups. In all these, cytoplasmic flow in opposite directions occurs on opposite sides of a pseudopodium. *See* HE-LIOZOIA; RADIOLARIA; SARCODINA.

Trichocysts. At maturity these lie in the cortex of certain ciliates and flagellates. Mucocysts (Fig. 5) seem homogeneous in some ciliates in electron micrographs; in others, there is a dense inner zone. The mucocysts of Tetrahymena pyriformis show inclusions forming a "houndstooth" pattern. Analogous mucocysts occur in Euglenida and Dinoflagellida. These inclusions probably serve in secretion of cyst walls and similar materials. Filamentous trichocysts (Paramecium and Frontonia) in the cortex are homogeneous bodies attached to a differentiated tip. During discharge there is formed a cross-striated shaft, ending in a dense tip. Discharged trichocysts are replaced by endoplasmic formation of new ones which, as they mature, are moved to the usual cortical location. Similar trichocysts occur in Oxyrrhis, a dinoflagellate. Cross striations of the shaft show a particular periodicity (55 nanometers) in various ciliates. Fibrillar trichocysts of

Paramecium may be extruded partially to anchor a feeding ciliate. Toxicysts of *Dileptus* show, after discharge, a nonstriated tubule with a rodlike tip. Before discharge a capsule contains telescoping tubules and is closed apically by a cap. New toxicysts arise from endoplasmic vesicles which migrate to an adoral position during development. Toxicysts may, upon contact, induce paralysis of lysis of various Protozoa and sometimes paralysis of rotifers. Such organelles aid carnivores in capture of prey.

Inclusions. In addition to nuclei, food vacuoles (gastrioles) in phagotrophs, chromatophores and stigma in many phytoflagellates, water-elimination vesicles in many Protozoa, and sometimes other organelles, the cytoplasm may contain mitochondria, Golgi material, pinocytotic vacuoles, stored food materials, endoplasmic reticulum, and sometimes pigments of various kinds.

Chromatophores. These contain chlorophylls and commonly additional pigments such as carotenes and xanthophylls. Accessory pigments, varying with the group, may be abundant enough to produce yellowish-green, brown, and even blue or red chromatophores. In different species, chromatophores range from a large one (Fig. 7e, f, and i) to a number of flattened bodies (Fig. 7g and k). A chromatophore contains, within a double membrane, electronopaque lamellae separated by transparent material in which granules may be embedded. Deoxyribonucleic acid (DNA) and ribonucleic acid (RNA), chemically different from nuclear DNA and cytoplasmic RNA, occur in chromatophores of Euglena gracilis. DNA has been reported also from chromatophores of Chlamydomonas moewusii. Certain strains of E. gracilis have been bleached by exposure to high temperature (about 97°F or 36°C), maintenance in darkness, or treatment with streptomycin, erythromycin, pyribenzamine, and a nitrosoguanidine, among other drugs. Bleaching has been temporary or permanent with different treatments. Certain antibiotics damage chloroplast DNA (preventable by supplementary hydroxylamine). Plastids, containing starch but no chlorophyll, lie below the cortical mitochondria in Polytoma uvella. See CHROMATO-PHORE.

Pyrenoids, ranging from a single body to a cluster, are associated with the chromatophore in certain phytoflagellates but not in others (Fig. 7*b*, *c*, and *j*). Taxonomic distribution may be erratic; for example, certain species of *Euglena* have pyrenoids but others do not. In *Chlamydomonas* the pyrenoid, lying within a chloroplast, is surrounded by starch granules. Electron micrographs show delicate tubules extending from the pyrenoid to the lamellae of the chloroplast.

Stigma. A stigma (Fig. 7*d*) occurs in many green flagellates and a few of their colorless relatives. Observations on behavior have suggested that the stigma functions in orientation with relation to a light source. Also, stigmaless mutants (*Chlamydomonas*) respond to light more erratically than the green parent stock. Whether the stigma is a photoreceptor or merely a shield for a photosensitive organelle is de-

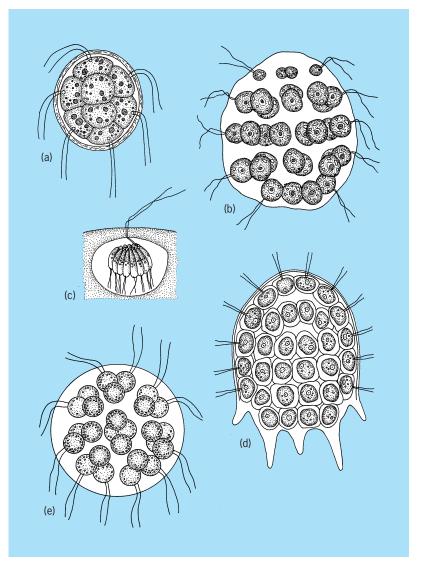


Fig. 3. Colonies of Phytomonadida. (a) Pandorina. (b) Pleodorina illinoisiensis. (c) Sperm packet, Pleodorina. (d) Platydorina caudata. (e) Eudorina. (After L. H. Hyman, The Invertebrates, vol. 1, McGraw-Hill, 1940)

batable. A typical stigma shows globules containing carotenoid pigment. The organelle lies on the wall of the reservoir, as in *Euglena*, or may be adherent to or lie within a chromatophore, as in the phytomonads. The pigment globules (100 nm or larger) are packed into one or more plates forming the stigma. In experimentally bleached *Euglena*, the fate of the stigma is not closely correlated with that of the chromatophores. A more complex ocellus of certain dinoflagellates (Fig. 7*b* and *m*) shows a lenslike body in a pigmented cup. *See* PHOTORECEPTION.

Cytoplasmic pigments. These occur in many phytoflagellates and certain other protozoa, such as *Blepharisma* and *Stentor*, ranging from violet to red. Reddish pigments are common in Euglenida and Phytomonadida. In *Haematococcus pluvialis* production of such pigments (largely astaxanthin) in abundance results in a "hematocyst" in which the green color is masked. In *Euglena rubra* a similar pigment, normally endoplasmic, is moved to the periphery upon exposure to bright light or

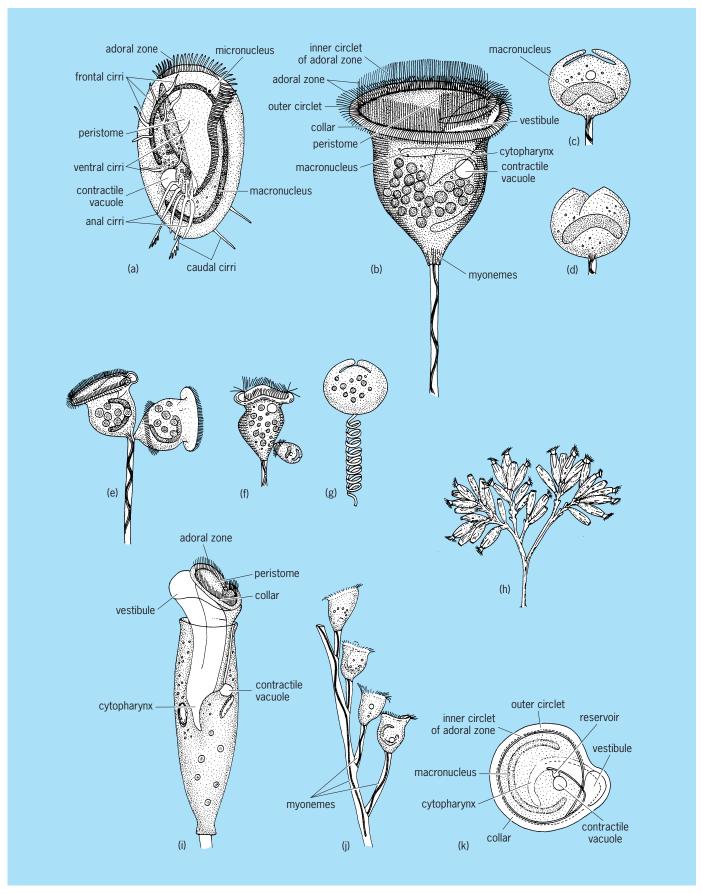


Fig. 4. Hypotricha and Peritricha. (a) Euplotes. (b) Vorticella. (c-e) Stages of fission in Vorticella. (f) Macro- and microconjugants in Vorticella. (g) Vorticella, stalk contracted. (h) Arboroid colony, Opercularia. (i) Zooid from Opercularia colony. (j) Portion of colony (Carchesium) showing stalk muscle. (k) Peristome of Vorticella, seen from above. (After L. H. Hyman, The Invertebrates, vol. 1, McGraw-Hill, 1940)

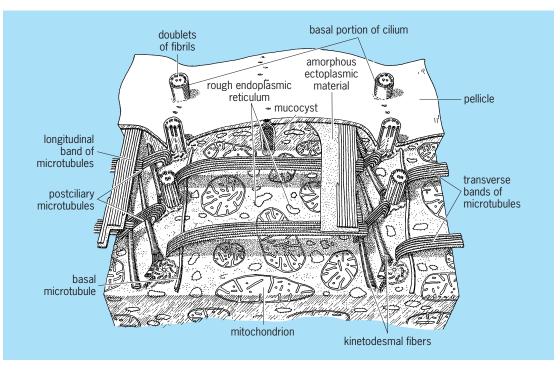


Fig. 5. Schematic of cortex of Tetrahymena pyriformis. (After Allen, J. Protozool., 14:553-565, 1967)

temperatures near 104° F (40° C). Red blooms may develop in ponds during hot summers.

Certain ciliates also produce cytoplasmic pigments. Of these the pink pigments, similar to hypericin of plants, of *Blepharisma undulans* are converted by bright light into products toxic to microorganisms, including *Blepharisma*. Pigment extracts, tested in light, show similar toxicity. An analogous material, ichthyotoxin, produced by a flagellate (*Prymnesium parvum*), is inactivated in bright light.

Mitochondria. Although apparently absent in certain anaerobes, these normally occur in aerobic protozoa **Fig.** 8*a*–*c*). These inclusions are usually elongated and may fuse into nets in certain species (Fig. 8*a*). Morphology may be influenced by the condition of the organism. When reproduction slows down in a culture of *Tetrabymena pyriformis*, the previous ovoid and peripheral mitochondria round up and become more widely distributed. Also, ribosomes which are normally clumped along mitochondrial villi tend to disappear. DNA has been detected in mitochondria of several species. Oxidative phosphorylation seems to be associated with protozoan mitochondria (*Euglena*), as would be expected. *See* MITOCHONDRIA.

Electron micrographs indicate two general kinds of protozoan mitochondria. In one, folds of the inner membrane form cristae extending into the central lumen; this variety, characteristic of higher animals, occurs in certain phytoflagellates (*Euglena* and chrysomonads) and apparently certain trypanosomes. The kinetoplast may be continuous with such a mitochondrion, and metamorphosis from a leishmanial to a leptomonad stage may involve an increase in the size and complexity of trypanosomid mitochondria. In the other type (amebas, ciliates, and certain parasitic flagellates), cristae are replaced by microvilli.

Lysosomes. Small vacuoles $(0.5-1.0 \ \mu\text{m})$ in *Tetraby*mena pyriformis (**Fig. 9**) are considered lysosomes containing acid hydrolases (not yet involved in digestion) and probably arising from the rough endoplasmic reticulum. *See* ENDOPLASMIC RETICULUM; LYSO-SOME.

Golgi material. In electron micrographs a typical Golgi body is a stack of flattened vesicles, sometimes up to 30 (*Peranema*). Although the identity of protozoan Golgi material is confused in older literature, some of the described Golgi elements should prove to be such. Although functional significance in Protozoa is not yet certain, some workers favor a secretory activity of these inclusions. *See* GOLGI APPARATUS.

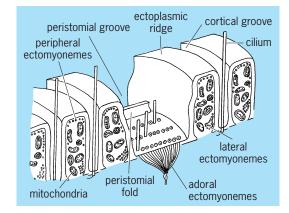


Fig. 6. Diagrammatic reconstruction of the cortex of Spirostomum ambiguum. (After Finley, Brown, and Daniel, J. Protozool., 11:264, 1964)

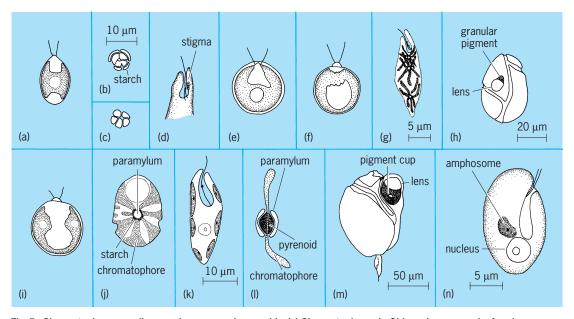


Fig. 7. Chromatophores, ocellus, amphosome, and pyrenoids. (a) Chromatophores in Chlamydomonas agloeformis.
(b, c) Compound pyrenoids of Pyramidomonas, with adherent starch. (d) Stigma of Euglena. (e) Chromatophores in Chlamydomonas umbonata. (f) Chromatophores in Chlamydomonas inversa. (g) Chromatophores in Euglena geniculata.
(h) Ocellus in Protopsis. (i) Chromatophores in Chlamydomonas bicocca. (j) Chromatophore in Peridinium umbonatum.
(k) Chromatophores in Colacium. (l) Chromatophore having a pyrenoid and a paramylum. (m) Ocellus in Erythropsis.
(n) Amphosome (stained) in Cryptomonas. (After R. P. Hall, Protozoology, Prentice-Hall, 1953)

Food reserves. Polysaccharides are often stored as visible inclusions. These include glucose polymers: paramylum of Euglenida, leucosin of Chrysomonadida, starches of Cryptomonadida, Phytomonadida, and Dinoflagellida; and glycogens of Sarcodina, Zoomastigophorea, and ciliates. Protozoan glycogens resemble liver and muscle glycogen but differ in molecular weight and number of glucose residues in unit chains. Lipids may be stored as small to large globules which are characteristic of old rather than young laboratory cultures. A few flagellates, like certain diatoms, accumulate oils responsible for undesirable flavors and odors of drinking water. Nitrogenous reserves include chromatoid bodies, described in axenic cutures of trophozoites of Entamoeba invadens and cysts of certain other species. These inclusions arise before encystment by aggregation of small vesicles into bodies of characteristic form. Such inclusions aid in differentiating encysted E. bistolytica and E. coli. Protein and RNA have been reported from chromatoid bodies, each of which contains small (about 20 nm) globules, interpreted by some workers as ribosomes. In electron micrographs (E. invadens) these inclusions show fibrils in a closely packed helical arrangement. Chromatoid bodies gradually disappear as the cyst matures. See AMOEBIDA.

Water expulsion vesicle. This organelle (**Figs. 10** and **11**), commonly termed a contractile vacuole, is characteristic of fresh-water protozoa and occurs also in certain marine and parasitic phagotrophs. Such vesicles show cyclic increase in volume, followed by collapse during which the contents are discharged. Motion pictures show no contraction of the "contractile" vacuole in *Amoeba proteus* (Fig. 11). Instead, cytoplasmic pressure increases with accumu-

lating fluid until the vesicle is forced to the surface and its proximal surface is thrown into folds as the cortex ruptures to release the contents through a pore. The result is collapse of the vesicle. Electron micrographs of *Tetrahymena pyriformis* suggest a

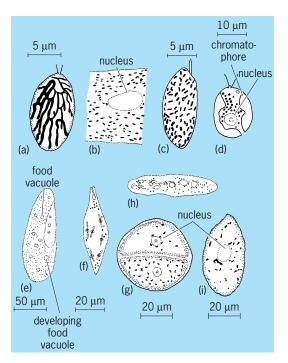


Fig. 8. Mitochondria, granules, and osmophilic inclusions. Mitochondria (a) in *Polytoma*; (b) in *Protoopalina*; and (c) in *Chilomonas*. Granules (d) in *Chlamydomonas*; (e) in *Paramecium*; and (f) in *Euglena*. Osmiophilic inclusions (g) in *Gregarina*, gametocytes in cyst; (h) in *Paramecium*; (i) in *Protoopalina*. (After R. P. Hall, *Protozoology*, *Prentice-Hall*, 1953)

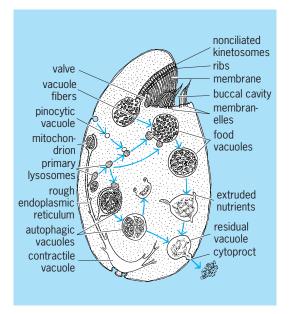


Fig. 9. Digestive pathways in *Tetrahymena pyriformis*. (After Elliot, J. Protozool., 13:311, 1966)

similar collapse of the vesicle as its contents are discharged through the pores. Accumulation of fluid in the enlarging vesicle is aided by collecting tubules (Fig. 10*f* and *g*) which may, as in *T. pyriformis*, extend through much of the cytoplasm as a "nephridial" system apparently continuous with the endoplasmic reticulum. The major function of the vesicle is elimination of excess water, which may be traced to ingestion (phagotrophs), endosmosis, or metabolic origin. Cyclic activity of the vesicle may be accelerated by adding ATP to the medium. Since there is no evidence that only pure water is secreted into the collecting tubules, it is possible that soluble materials also are eliminated by the vesicle.

Sensory vacuoles. These occur in certain parasitic and free-living ciliates (Fig. 10c-e). The similarity to statocysts of certain invertebrates inspired the assumption implied by the name.

Nuclei. Dimorphic nuclear equipment (micronucleus and macronucleus) is typical of ciliates. Analogous dimorphism is known in a few Foraminiferida but has not been observed in various others. Except for possibly similar cases yet to be investigated carefully, protozoa outside the ciliates contain one kind of nucleus, although the number ranges from one to many. The polyploid macronucleus of ciliates develops from a diploid derivative of the synkaryon (conjugation and autogamy), presumably by repeated replication of chromosomes in a macronuclear anlage. Filamentous "chromosomes" containing DNA have been described in macronuclei of Blepharisma. Later development involves increase in volume and sometimes elongation (Euplotes and Vorticella) or constriction into nodes (Stentor and some strains of Blepharisma). Micronuclear differentiation may involve a slight decrease in size. Both kinds of nuclei may contain about twice as much RNA as DNA, while the macronuclear content may be about 40 times that of the micronucleus (Paramecium).

In protozoa other than ciliates, nuclei vary in size and content of chromatin and nucleoli (or endosomes). Nuclear division is mitotic except in the ciliate macronucleus. It is not certain that the polyploid "somatic" nuclei of various Foraminiferida ever divide normally.

Nutrition

In protozoan feeding, either phagotrophic (holozoic) or saprozoic (osmotrophic) methods predominate in particular species. In addition, chlorophyllbearing flagellates profit from photosynthesis; in fact, certain species have not been grown in darkness and may be obligate phototrophs.

Phagotrophic. Ingestion of food, followed by digestion in vacuoles, is characteristic of Sarcodina, ciliates, and many flagellates. Digestion follows synthesis of appropriate enzymes and their transportation to the food vacuole. Ingestion of nonnutritive latex particles also leads to acid-phosphatase activity of the "food" vacuoles. Primary lysosomes, which can fuse with the food vacuole, may transport hydrolases from the rough endoplasmic reticulum, as suggested for *Tetrabymena* (Fig. 9). Even a few

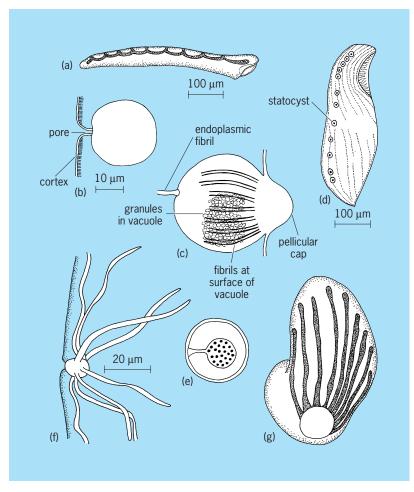


Fig. 10. Water expulsion vesicles and sensory vacuoles of protozoa. (a) Contractile tube in *Haptophrya*, an intestinal ciliate from a salamander. (b) Water expulsion vesicle and pore in *Eudiplodinium*, a parasitic ciliate from ruminants. (*c*–e) Sensory vacuoles of *Blepharoprosthium* and *Loxodes*. (f, g) Water expulsion vesicle and contributory canals in *Paramecium* and *Tillina*. (After R. P. Hall, Protozoology, Prentice-Hall, 1953)

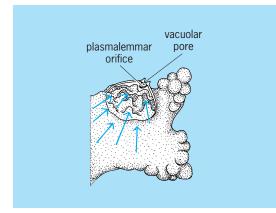


Fig. 11. Water elimination vesicle in *Amoeba proteus*, based on tracings of photomicrographs. Arrows indicate movement of granules. Endoplasm gets pushed against vesicle membrane, producing wrinkles. (*After Wigg, Bovee*, and Jahn, J. Protozool., 14:104, 1967)

Sporozoa, the malarial parasites and certain gregarines, are known to ingest host proteins or sometimes tissue cells or portions of them. Details of ingestion vary. Formation of food cups, or gulletlike invaginations, to enclose prey is common in more or less ameboid organisms, such as various Sarcodina, many flagellates, and at least a few Sporozoa. Entrapment in a sticky reticulopodial net occurs in Foraminiferida and certain other Sarcodina. A persistent cytostome and gullet are involved in phagotrophic ciliates and a few flagellates. A small cytostome and gullet in Trypanosoma mega suggest ability to ingest food, although importance of this activity in trypanosomes is still uncertain. Many ciliates (Fig. 9) have buccal organelles (membranes, membranelles, and closely set rows of cilia) arranged to drive particles to the cytostome. Particles pass through the cytostome into the cytopharynx (gullet), at the base of which food vacuoles (gastrioles) are formed. Digestion occurs in such vacuoles. An additional feature, reported in Paramecium and Tetrahymena, is entanglement of particles in mucus secreted by the peristomial cortex or buccal cavity. The effect is a clumping of particles into larger aggregates. In many less specialized ciliates the cytostome opens at the surface of the body. Such organisms (Didinium and Perispira) usually ingest large prey, such as other Protozoa. In some cases rodlike trichites surrounding the cytopharynx appear analogous to toxicysts of certain other ciliates.

Saprozoic. By definition this involves passage of dissolved foods through the cortex. It is uncertain to what extent diffusion is responsible, but enzymatic activities presumably are involved in uptake of various simple sugars, acetate and butyrate in *Polytoma* and even potassium and phosphate ions. In addition, external factors, for example, the pH of the medium, may strongly influence uptake of fatty acids and phosphate.

Pinocytosis. This process resembles phagocytosis in miniature but with ingestion apparently limited to materials in solution. First recognized in *Amoeba proteus*, the process has since been described in other amebas, certain ciliates, and *Eimeria* (Sporozoa). Pinocytosis is stimulated by "inducers," such as proteins, certain salts, and toluidine blue. In the process small vacuoles are formed at the base of a tubular invagination from a surface, either cortex or the wall of cytopharynx. The vacuoles drift into the endoplasm and mingle with other inclusions. In certain ciliates, a similar process yields secondary vacuoles from a food vacuole. In certain malarial parasites, but not in others, such pinocytosis precedes completion of digestion. *See* ENDOCYTOSIS.

Nutritional requirements. Establishment of a strain in axenic cultures in defined media make possible a determination of specific food requirements. In nature, these requirements are met by the available diet which includes other microorganisms and materials dissolved in the external medium. *See* AXENIC CUL-TURE.

Mineral requirements. Certain materials, including calcium, silicon, and strontium salts, are needed for tests and other skeletal structures, and metal ions are needed in traces for enzyme systems. Trace requirements of one species or another include calcium, cobalt, copper, iron, potassium, magnesium, manganese, phosphorus, sulfur, and zinc. Although indispensable status is uncertain, aluminum, barium, boron, iodine, silicon, sodium, and vanadium have each stimulated growth of one or more species.

Nitrogenous foods. Nitrogen requirements of many phytoflagellates are met by an ammonium salt or by either a nitrate or an ammonium salt. Other protozoa need specific amino acids which they cannot synthesize, the number ranging from 1 (*Perenema*) to 7 for an ameba (*Hartmannella rhysodes*), 13 for *Paramecium multimicronucleatum*, and at least 15 for *Leishmania tarentolae*. In a few cases, radioactive amino acids have been used in tracing their fate in metabolism. In analogous fashion, labeled food organisms have been used to identify species ingested by Foraminiferida in laboratory aquariums. In addition to amino acids, certain protozoa need a purine, or sometimes both a purine and a pyrimidine. *See* AMINO ACIDS; PURINE; PYRIMIDINE.

Carbon requirements. The only carbon source needed by certain chlorophyll-bearing flagellates is CO_2 . In fact, *Gymnodinium breve*, a dinoflagellate occurring in "red tides," apparently cannot use any organic compound which has been tested and seems to be an obligate phototroph. *Euglena gracilis*, in contrast, can grow in darkness in a good medium. Requirements of heterotrophs, as a group, may be met by a polysaccharide, by a fatty acid, by an acid of the tricarboxylic acid (TCA) cycle, or by an alcohol. Suitability varies with the species.

Growth factors. Organic substances, such as vitamins, are needed in traces. A few phytoflagellates, namely, certain species of *Brachiomonas, Chlamydomonas*, and *Chlorogonium*, are known to synthesize all required vitamins. Other protozoa need one or more vitamins and sometimes other growth factors. Loss of an ability to synthesize a particular growth factor (niacin or thioctic acid) has occurred in *Tetrabymena pyriformis* after prolonged

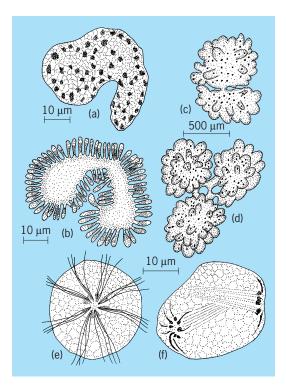


Fig. 12. Schizogony and plasmotomy. (a) Schizogony in Ovivora thalassemae. (b) Schizont of O. thalassemae. (c, d) Plasmotomy in Pelomyxa carolinensis. (e) Coronympha octonaria, vegetative stage showing nuclei and flagellar groups. (f) Nuclear groups at end of telophase, as they appear just before plasmotomy in C. octonaria. (After R. P. Hall, Protozoology, Prentice-Hall, 1953)

maintenance of strains in laboratory cultures. Certain amebas (Acanthamoeba and Hartmannella) need only three vitamins (B12, biotin, and thiamine); similar requirements are known for certain phytoflagellates. A ciliate, Glaucoma chattoni, needs biotin, folic acid, nicotinic acid, pantothenic acid, pyridoxine, riboflavin, thiamine, thioctic acid, plus certain lipids. Less common requirements include ethanol or methanol for Tetrabymena setifera, chlorine for certain European strains of T. pyriformis, and a pteridine for Crithidia fasciculata. Identification of vitamin requirements was followed by use of certain protozoa in biological assays-vitamin B₁₂ (Ochromonas malhamensis, Euglena gracilis), thiamine (O. malhamensis), pyridoxine (Tetrahymena pyriformis), biotin (Ochromonas danica), nicotinic acid (T. pyriformis), and pantothenic acid (T. pyriformis). See VITAMIN.

Reproduction

Reproduction occurs after a period of growth which ranges, in different species, from less than half a day to several months (certain Foraminiferida). General methods include binary fission, budding, plasmotomy, and schizogony. Fission, involving nuclear division and replication of organelles, yields two organisms similar in size. Budding produces two organisms, one smaller than the other. In internal budding (certain Suctoria) a bud is formed within a brood pouch, an almost closed invagination in the "parental" body. The bud, after it develops cilia, is expelled through a "birth pore" as a sort of larva which, after a few hours of activity, undergoes metamorphosis (losing cilia, developing tentacles) into a typical suctorian. In external budding, the bud is formed at the surface of the "parental" body. In plasmotomy (**Fig. 12***c* and *d*), a multinucleate organism divides into several, each containing a number of nuclei. Schizogony, characteristic of Sporozoa, follows repeated nuclear division, yielding many uninucleate buds (Fig. 12*a* and *b*). In some cases, as in *Eimeria bovis*, the schizont first divides into several multinucleate forms, in each of which the nuclei approach the surface to take part in schizogony (merogony).

Reproduction includes formation of new organelles, either by division of old structures or by neoformation, sometimes preceded by resorption of an old organelle. Blepharoplasts and basal granules, in particular, seem to be self-replicating. Formation of a new mouth (stomatogenesis) in ciliates often involves origin of an anlage from basal granules of one or more ciliary rows and its differentiation into new buccal organelles. Even the formation of a new pellicle in Euglenida may involve formation of new pellicular strips between adjacent strips of the original pellicle.

Life cycles. Simple life cycles include a cyst and an active (trophic) stage undergoing growth and reproduction. In certain free-living and parasitic species, no cyst is developed. Dimorphic cycles show two active stages; polymorphic show several. The former include adult and larva (Suctoria); flagellates and ameba (certain Mastigophora and Sarcodina); flagellate and palmella (nonflagellated; certain Phytomonadida); and ameba and plasmodium (Mycetozoia especially).

Encystment. This process involves resorption of locomotor and feeding organelles, loss of water, often rounding up of the body, and secretion of one or more cyst membranes. The wall usually is composed

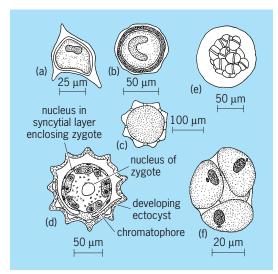


Fig. 13. Protective and reproductive cysts. (a) Protective cyst of Ceratium; (b) of Didinium; (c) of Bursaria. (d) Encysted zygote of Volvox. (e) Reproductive cyst of Gyrodinium, a dinoflagellate and (f) of Colpoda, a ciliate. (After R. P. Hall, Protozoology, Prentice-Hall, 1953)

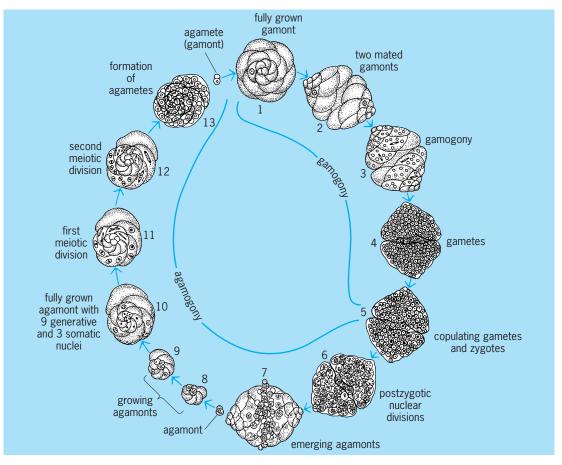


Fig. 14. Life cycle of *Glabratella sulcata*, showing alternation of haploid phase (gamogony) and diploid phase (agamogony). Two meiotic divisions (11 and 12) precede the production of young haploid organisms (13). (*After Grell, Research in Protozoology, vol. 2, Pergamon Press, 1968*)

of secreted materials (proteins, carbohydrates, keratin, and others), although foreign particles may be cemented into a wall. Precystic activities may include storage of reserves. Protective cysts (**Fig. 13***a*–*d*), with fairly thick walls, often resist desiccation. Dried cysts of *Colpoda cucullus* have remained viable for 5 years; those of certain amebas and flagellates, for 49 years in dry soil samples. Within reproductive cysts (Fig. 13*e* and *f*) fission, budding or schizogony, sometimes gametogenesis, and syngamy occur in various species. The usually thin-walled reproductive cysts afford relatively little protection.

Excystment involves absorption of water leading to rupture of cyst membranes and, when necessary, development of organelles for feeding and locomotion.

Sexual activities. These involve production of haploid (gametic) nuclei and their fusion into a synkaryon (zygotic nucleus). Meiosis occurs at different stages in life cycles. In haploid organisms (Phytomonadida, certain flagellates of termites and wood roaches, Coccidia, malarial parasites, and gregarines) the zygote is the only diploid stage. Meiosis occurs as the zygotic nucleus divides. In these haploid cycles a heterozygous condition can exist only in the zygote. Diploid cycles, in which meiosis occurs during gametogenesis, are found in ciliates, Heliozoia (*Actinophryx* and *Actinosphaerium*), and

in certain Hypermastigida. In several flagellates of the wood roach, sexual activity is stimulated by ecdysone (molting hormone) of the host. A few Foraminiferida show an alternation of generations involving a haploid gamete-producting phase and a diploid phase with ordinary reproduction (**Fig. 14**). *See* ECDYSONE.

Sexual activities include syngamy, autogamy, and conjugation. Syngamy involves fusion of two gametes which may be morphologically similar (isogamy) or different (anisogamy). Autogamy occurs within a single organism; haploid nuclei are formed and subsequently fuse into a synkaryon. Conjugation involves pairing of ciliates (Fig. 15), apparently following appearance of specific proteincontaining substances at the surface, including cilia. Pairing in Tetrahymena includes fusion of the two pellicles and formation of channels joining the conjugants. In tests with radioactive foods, cytoplasmic materials (RNA and proteins) passed from one conjugant to the other. Pairing stimulates nuclear activities. Old macronuclei gradually degenerate as a rule, although fragments may be exchanged occasionally in Stylonychia. The micronucleus undergoes three divisions, the second being reductional. The third produces gametic nuclei which, in Tetrahymena, arise from one haploid nucleus. A migratory gametic nucleus, whose status is apparently determined by

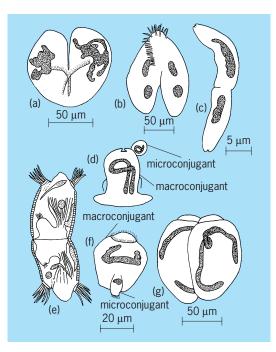


Fig. 15. Early conjugation, just after fusion of conjugants. (a) Nyctotherus. (b) Pleurotricha. (c) Ancistrocoma. (d) Cycloposthium. (e) Scyphidia. (f) Vorticella. (g) Euplotes. (After R. P. Hall, Protozoology, Prentice-Hall, 1953)

position, passes from each conjugant to its mate, where fusion with the stationary nucleus occurs. The resulting synkaryon, in different species, divides one or more times and the products become macroand micronuclei. Reorganization may involve one or more fissions which restore nuclear numbers to normal. In this sense, a pair of conjugants produces two or more ciliates. In the peritrichs, with micro- and macroconjugants, only the latter survives.

Autogamy occurs in certain ciliates and a few parasitic flagellates, *Barbulanympha*, for example. Since the gametic nuclei arise from one nucleus, the resulting synkaryon is homozygous.

Normal conjugation involves ciliates of complementary mating types, two or more belonging to a syngen, originally termed "variety." Each species contains a number of syngens. Typically, mating types within one syngen can conjugate normally. Abortive conjugation, sometimes fatal, follows pairing of ciliates from different syngens.

Syngamy in Phytomonadida typically requires two kinds of gametes (+ and –). Syngens analogous to those of ciliates have been reported in certain Phytomonadida. In *Astrephomene* at least, different syngens may have different chromosome counts. In dioecious species (*Chlorogonium* and some species of *Chlamydomonas*) the two gametes come from different clones; each zygote produces two clones of each "sex." In monoecious types (several species of *Chlamydomonas* and *Haematococcus pluvialis*) a single clone produces both types of gametes. In colonial phytomonads gametes may be isogamous (*Gonium* and *Pandorina*) or anisogamous (*Eudorina* and *Volvox*). Differentiation of male colonies in *Volvox aureus* is induced by a rather thermostable protein-containing substance. In less specialized genera (*Gonium*) all zooids become gametes; at the other extreme (*Volvox* and *Eudorina*) gametes arise in the posterior half of the colony.

According to the concept of a "physiological life cycle," laboratory strains of ciliates pass through phases of youth; maturity, when conjugation is possible; senesence, when conjugation is impossible; and death. Senescence supposedly is averted by opportune occurrence of conjugation. Although certain species apparently can live indefinitely without conjugation, it is possible that autogamy replaces conjugation in rejuvenating such ciliates.

Parasitic Protozoa

Parasites occur in all major groups. Sporozoa are exclusively parasitic, as are some flagellate orders (Trichomonadida, Hyperamastigida, and Oxymonadida), the Opalinata, Piroplasmea, and several ciliate orders (Apostomatida, Astomatida, and Entodiniomorphida). Various other groups contain both parasitic and free-living types. Protozoa also serve as hosts of other Protozoa, certain bacteria, fungi, and algae. An unusual addition is Leptomonas karyophilus, a trypanosomid flagellate which invades the macronucleus of Paramecium trichium. Such symbionts as Chlorella paramecii in Paramecium bursaria are considered symbiotes. A similar example is a bacterial symbiont of Crithidia oncopelti, which apparently contributes products to nitrogen metabolism of the host.

Many protozoan symbionts, called commensals, are more or less harmless. Cellulose-digesting flagellates of termites and ciliates of ruminants are symbiotic in the sense that they benefit their hosts. In addition to helpful digestive activities, rumen ciliates are potential food for the herbivores.

Pathogens. Relatively few parasites are distinctly pathogenic, causing amebiasis, visceral leishmaniasis (kala azar), African sleeping sickness, Chagas' disease, malaria, tick fever of cattle, dourine of horses, and other diseases. *See* MALARIA.

Harmful effects may be produced in different ways. Tissue cells are invaded and destroyed in Chagas' disease, leishmaniasis, coccidiosis, and malaria. Tissues may be destroyed, as in abscesses and ulcers involving *Entamoeba histolytica*. Although a potent "toxotoxin" is reported for *Toxoplasma gondii* and a Chagas toxin from *Trypanosoma cruzi*, little is yet known about possible toxins of protozoan parasites.

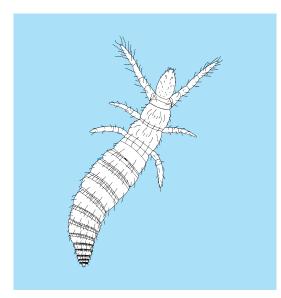
Epidemiology. Transfer may involve direct migration of parasites living in aquatic hosts; direct contact (osculatory, veneral); ingestion with contaminated food or drink; and vector transfer, as in malaria or Chagas' disease. Variations include placental transfer from mother to fetus, known in malaria and Chagas' disease; ovarian transfer, to eggs in ovary, as in ticks infected with *Babesia bigemina*; and lacteal transfer, from female to suckling young, reported in laboratory trypanosomiasis. *See* MEDICAL PARASITOL-OGY. Richard P. Hall

Bibliography. A. C. Chandler and C. P. Read, Introduction to Parasitology: With Special References to the Parasites of Man, 10th ed., 1961; J. O. Corliss, The Ciliated Protozoa: Characterization, Classification, and Guide to the Literature, 2d ed., 1979; J. Kreier, Parasitic Protozoa, 2d ed., 1997; R. R. Kudo, Protozoology, 5th ed., 1977; Progress in Protistology, 2 vols., 1986–1987; M. Sleigh, Protozoa and Other Protists, 3d ed., 2002; J. J. Wolken, Euglena: An Experimental Organism for Biochemical and Biophysical Studies, 1967.

Protura

An order of ancestrally wingless insects (subclass Apterygota). The order is usually classified into four families in two suborders (Eosentomoidea and Acerentomoidea), and about 500 species are recognized.

The insects are under 0.08 in (2 mm) in length, having an elongated body with a small head, relatively large abdomen, and three pairs of functional legs. The first pair of legs is normally held out in front of the head, antennalike. They serve to replace the sensory role of true antennae, which, like eyes, are absent. Paired pseudoculi (probably chemosensory) occupy eyelike positions on the head. The mouthparts include an elongate and styliform mandible, a maxilla, and a two-part labium. All parts are enclosed within a pocket in the head, so that only their tips are exposed. The thorax bears five-segmented legs, each with an apical claw. In the adult, the abdomen has 12 segments; the anus is terminal, and cerci (segmented sensory appendages on the last abdominal segment) are lacking. The sterna of the first through third segments bear short styli, probably representing vestigial limbs. A large gland opens at the rear of the eighth dorsal plate, which sometimes has a comblike lid. The genitalia are enclosed within a ventral pouch between the eleventh and twelfth segments, and have paired protrusible stylets (slender, elongated appendages).



A proturan, Acerentulus barberi (dorsal view).

Protura live in forest soil, leaf litter, and similar places, and are thought to feed on mycorrhizal fungi. They are distributed globally but unevenly in temperate and tropical climates, and they sometimes occur in enormous numbers, as in some Oregon fir forests in the United States. *See* INSECTA. William L. Brown, Jr.

Bibliography. G. Imadate, *The Insects of Australia*, 2d ed., 1991; S. L. Tuxen, *Protura: A Revision of the Species of the World*, 1964.

Proustite

A mineral having composition Ag₃AsS₃ and crystallizing in the hexagonal system. It occurs in prismatic crystals terminated by steep ditrigonal pyramids, but is more commonly massive or in disseminated grains. There is good rhombohedral cleavage. Hardness is 2–2.5 (Mohs scale) and specific gravity is 5.55. The luster is adamantine and the color ruby red. It is called light ruby silver in contrast to pyrargyrite, dark ruby silver. Proustite is less common than pyrargyrite but the two minerals are found together in silver veins. Noted localities are at Chañarcillo, Chile; Freiberg, Germany; Guanajuato, Mexico; and Cobalt, Ontario, Canada. *See* PYRAR-GYRITE; SILVER.

Provenance (geology)

In sedimentary geology, all characteristics of the source area from which clastic sediments and sedimentary rocks are derived, including relief, weathering, and source rocks. *See* WEATHERING PROCESSES.

The goal of most provenance studies of sedimentary rocks is the determination of source characteristics of the mountains or hills from which the constituent sediment was derived. Such determinations are difficult to make because sediment composition and texture are continually modified during erosion, transport, deposition, and diagenesis (postdepositional modification). These modifications must be accounted for in order to determine source characteristics (provenance). It is most straightforward to determine provenance in situations in which these modifying effects are minimal; provenance may be indeterminate or ambiguous in situations involving extensive modification of sediment composition and texture. The former situation is most common in tectonically active areas, resulting in rapid uplift and erosion of mountains, rapid transport and deposition, and slight diagenetic modification after deposition. In contrast, stable continental areas (for example, cratons) provide ample opportunity for intense weathering so that chemical, mineralogical, and textural characteristics of sediment are intensely modified, thus rendering determination of source characteristics difficult. See DIAGENESIS; EROSION; SEDIMENTARY ROCKS.

Clastic (or detrital) sediment is commonly recycled during multiple episodes of mountain building, erosion, sedimentation, lithification, and renewed mountain building. This process constitutes the rock cycle, within which igneous, sedimentary, and metamorphic rocks are created and modified. During this process, it is common for older sedimentary rocks to be uplifted and eroded, so that individual sedimentary particles (clasts) are recycled to form new sediment, which may be lithified to form new sedimentary rock. Provenance studies must determine the proportion of a sedimentary rock derived directly from indicated source rocks versus the proportion derived directly from another sedimentary rock (that is, rocks exposed during previous cycles of sedimentation). This determination is essential, but commonly difficult to accomplish because the multicyclic nature of sediment may be difficult to recognize. See METAMORPHIC ROCKS.

Fundamentally different methods of study are utilized, depending on what aspect of provenance is emphasized, what type of sediment is studied, and what scale of sampling and study is attempted. Grain size of detrital sediment is a dominant control over what methods may be employed.

Coarse-grained sediment (gravel and conglomerate) provides the most direct determination of source rocks because individual clasts may be studied by any method available for the study of source rocks themselves. Statistical counts of clasts may be compared directly to proportions of rock types in potential source areas. However, gravel and conglomerate is unavailable for study in many locations, and many rock types do not survive weathering or transport in this grain size. *See* CONGLOMERATE; GRAVEL.

Intermediate grain sizes (sand and sandstone) are the most commonly studied because they are abundant and preserve many source-rock characteristics. Statistical point-counting of thin sections of sand and sandstone mounted on glass slides using polarizing microscopes is the most common method of study. Details of point-counting methods must be carefully chosen in order to maximize the usefulness of the data for different types of studies. *See* SAND; SAND-STONE.

Fine-grained sediment (mud and mudrock) owes much of its history to weathering and alteration of sediment during transport, deposition, and burial. Therefore, composition of fine clastic particles (primarily clay minerals) reflects these factors more than it reflects provenance. For this reason, provenance is more difficult to decipher based on fine-grained rocks, although whole-rock chemical characteristics may persist through various stages of sediment modification.

The first-order control on rock composition at Earth's surface is plate tectonics. Mountain building (orogeny) and rifting (taphrogeny) are the direct results of plate tectonic processes; most studies of provenance, therefore, account for modern or ancient plate tectonic processes. Provenance studies must also account for scale of study; scale of sampling directly reflects the scale of variation of rock composition in source areas. For example, local drainages produce sediment derived from local rock types, with possible great variability, whereas continentalscale sedimentary systems (e.g., big rivers and turbidite fans) produce homogeneous sediment reflecting plate-tectonic setting. *See* OROGENY; PLATE TEC-TONICS.

Methods of determining provenance include direct determination of rock types (primarily used for coarse to medium grains); direct determination of mineralogy (used for all grain sizes); whole-rock geochemistry (used for medium to fine grains); geochemistry of individual mineral species (used for all grain sizes, but especially for medium grains); and radiometric dating of individual mineral species (primarily for medium grains). *See* GEOCHEMISTRY; MIN-ERALOGY; ROCK AGE DETERMINATION.

A guiding principle in provenance studies is actualism (the concept that the present is the key to the past). Studies of modern environments of uplift, erosion, transport, deposition, and diagenesis provide key insights into how sedimentary systems work. Fundamental controls on composition of gravel, sand, and mud can be determined from the study of modern systems, and compositional models may be developed based on these controls. The models in turn may be applied to the study of ancient conglomerate, sandstone, and mudrock; significant progress in the reconstruction of paleogeography and paleotectonics inevitably results. *See* PALEO-GEOGRAPHY. Raymond V. Ingersoll

Bibliography. W. R. Dickinson, Provenance and sediment dispersal in relation to paleotectonics and paleogeography of sedimentary basins, in K. L. Kleinspehn and C. Paola (eds.), New Perspectives in Basin Analysis, pp. 3-25, Springer-Verlag, New York, 1988; H. Ibbeken and R. Schleyer, Source and Sediment: A Case Study of Provenance and Mass Balance at an Active Plate Margin (Calabria, Southern Italy), Springer-Verlag, New York, 1991; R. V. Ingersoll, Actualistic sandstone petrofacies: Discriminating modern and ancient source rocks, Geology, 18:733-736, 1990; R. V. Ingersoll et al., The effect of grain size on detrital modes: A test of the Gazzi-Dickinson point-counting method, J. Sed. Petrol., 54:103-116, 1984 (see also Discussions and Replies, 55:616-621); R. V. Ingersoll, A. G. Kretchmer, and P. K. Valles, The effect of sampling scale on actualistic sandstone petrofacies, Sedimentology, 40:937-953, 1993; M. J. Johnsson and A. Basu (eds.), Processes controlling the composition of clastic sediments, Geol. Soc. Amer. Spec. Pap., no. 284, 1993; L. J. Suttner, Sedimentary petrographic provinces: An evaluation, Soc. Econ. Paleontol. Mineralog. Spec. Publ., no. 21, pp. 75-84, 1974.

Prussian blue

The common name for a material prepared by either the addition of trivalent iron (Fe^{III}) to ferrocyanide [Fe(CN)₆]^{4–} or Fe^{II} to ferricyanide [Fe(CN)₆]^{3–} with the composition of Fe^{III}₄[Fe^{II}(CN)₆]₃ · *x*H₂O, where $x \sim 15$. Prussian blue, also known as Turnbull's blue or ferric hexacyanoferrate, was discovered

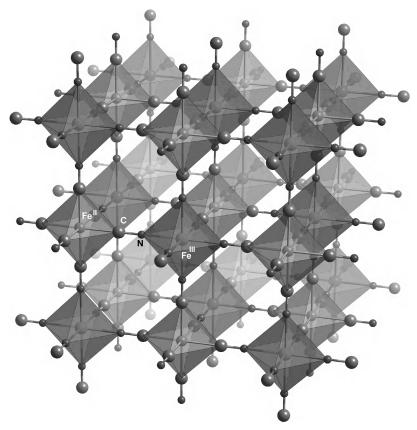


Diagram of cubic structure of Prussian blue.

accidently 300 years ago by Diesbach of Berlin, and its structure has been determined by several diffraction methods. Prussian blue has a structure with each Fe^{II} bound to six C-bounded cyanides (CN⁻), while the nitrogens (N) are bonded to Fe^{III}; thus, it has a cubic structure with Fe^{II}CNFe^{III} linkages along each of the edges of the cube (see **illustration**). Owing to the presence of more Fe^{III} than Fe^{II}, some Fe^{II} sites are vacant and filled with water. The presence of iron in two different oxidation or valence states in Prussian blue makes it a mixed-valence compound. This gives Prussian blue its intense blue color and its application as a pigment. The reduction of Prussian blue results in the formation of Everitt's salt, K₂Fe^{II}[Fe^{II}(CN₆)].

Uses of Prussian blue include artists' colors, carbon paper, typewriter ribbons, blueprinting, and printing inks. It is prescribed for the treatment of radioactive ¹³⁷Cs⁺ and thallium poisoning. Substitution of the Fe^{II} and Fe^{III} ions with other metal ions leads to Prussian blue-structured materials with different metal ions. These materials typically are magnets. Substitution of the Fe^{II} and Fe^{III} ions with Cr^{III} and V^{II} ions, respectively, results in a magnet with an ordering temperature around 100°C (212°F). *See* COORDINATION CHEMISTRY; COORDINATION COM-PLEXES; CURIE TEMPERATURE; CYANIDE; FERRICYA-NIDE AND FERROCYANIDE; IRON. Joel S. Miller

Bibliography. K. R. Dunbar and R. A. Heintz, Chemistry of transition metal cyanide compounds: Modern perspectives, *Prog. Inorg. Chem.*, 45:283–392, 1997; J. S. Miller, 3-D-network structured cyanidebased magnets, *MRS Bull.*, 25(11):60-64, 2000; A. G. Sharp, *The Chemistry of Cyano Complexes of the Transition Metals*, Academic Press, New York, 1976.

Prymnesiophyceae

A class of algae (also known as Haptophyceae) in the chlorophyll *a-c* phyletic line (Chromophycota). In protozoological classification these organisms constitute an order, Prymnesiida or Haptomonadida, in the class Phytamastigophora. Most of the approximately 300 species of prymnesiophytes are biflagellate monads.

This class has been segregated from the Chrysophyceae, with which it shares many biochemical and ultrastructural characters. Prymnesiophytes differ from chrysophytes, however, in several significant characters: (1) the typical monad bears a filiform organelle, the haptonema, between the two flagella; (2) except in the order Pavlovales, the flagella are of equal length and smooth; (3) organic scales, which may be calcified, cover most motile cells. (By contrast, motile cells of Chrysophyceae lack haptonemata, have one smooth and one pleuronematic flagellum, and have siliceous scales, if any.) Depending on the species, the haptonema is longer than the flagella and slender, short and bulbous, or vestigial. It consists of a bundle of longitudinal microtubules surrounded by a fenestrated sheath of endoplasmic reticulum, the whole being enclosed in the plasmalemma. At its base, which may be interconnected with the basal bodies of the flagella, the microtubular bundle consists of nine microtubules in a closely packed hexagonal arrangement. The emergent structure has the form of a shallow trough (an arc in cross section), enclosing seven microtubules. No function for haptonemata is known. Although they are capable of bending or flickering and, if long, can coil rapidly and uncoil slowly in response to shocks, haptonemata apparently play no role in either locomotion or capture of prey.

Calcified scales (cellulosic scales impregnated with calcite) are characteristic of many prymnesiophytes. These scales, which are given the general term coccoliths, were discovered in marine sediment before they were observed on living cells. Coccoliths are classified into several morphological types (such as rhabdoliths, discoliths, zygoliths, and ceratoliths), which are of great diagnostic value in the taxonomy of the prymnesiophytes that bear them (coccolithophorids).

Cells of prymnesiophytes contain one or two chloroplasts, with ultrastructure and pigmentation similar to that in Chrysophyceae. A pyrenoid is usually present, and an eyespot usually absent. The single Golgi apparatus with characteristically distended central cisternae is the site of scale production. The metabolic storage product is chrysolaminaran. All species are photosynthetic, although some are facultatively osmotrophic or phagotrophic.

Members of the order Pavlovales, which resemble other prymnesiophytes in having a haptonema, differ by lacking platelike scales, by having eyespots, and by having unequal flagella that are covered with fine hairs.

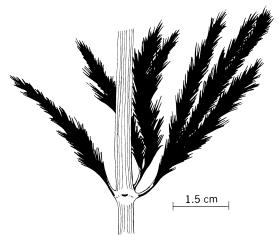
In most prymnesiophytes a nonmotile phase alternates with a motile phase. The nonmotile phase is a free-living unicell or a palmelloid or pseudofilamentous colony. In some cases the alternation is mediated by sexual reproduction. Usually, however, reproduction is effected by binary fission or the production of zoospores.

Prymnesiophytes are primarily marine, with coccolithophorids constituting one of the three major components of phytoplankton (the others being diatoms and dinoflagellates). Their importance in past geologic periods is documented by the vast amounts of coccolith deposits (for example, the white cliffs of Dover) dating back at least to the Jurassic (140 million years ago). More than 350 genera of fossil coccolithophorids have been described, so that the living representatives must be considered relict. Some prymnesiophytes are restricted to brackish or fresh water or are euryhaline. A few species are grown as food sources for invertebrates in mariculture projects, while other (Prymnesium parvum) develop explosively (bloom) as weeds in brackishwater fish ponds and poison the fishes. Phaeocystis pouchetii, which is characteristic of North Atlantic and Antarctic waters, can bloom so densely in the open sea that fishing nets are clogged by the gelatinous colonies and patterns of herring migration are affected. See ALGAE; COCCOLITHOPHORIDA; PHYTO-PLANKTON; PROTOZOA. Paul C. Silva; Richard L. Moe

Bibliography. E. R. Cox (ed.), *Phytoflagellates*, pp. 273-317, 1980; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982.

Pseudoborniales

An order of fossil plants found in Middle and Upper Devonian rocks. The group is related to Sphenophyllales and includes a single family and two monotypic genera. *Pseudobornia ursina* is known from Bear



Pseudobornia node with a whorl of three leaves.

Island (north of Norway), Alaska, and Germany. *Prosseria grandis* is found in New York State. Sphenopsid characters are more firmly established in this order than in Hyeniales.

Pseudobornia (see **illus.**), the better known of the two, has rhizomes and stems up to 2.4 in. (6 cm) wide and over 3 ft (1 m) long. The axes are jointed. Larger axes bear two branches at a node and smaller ones have whorls of four leaves. The leaves are 1.4– 2.4 in. (3.5–6 cm) long, short-stalked, and palmately divided, and they have laciniate margins. Fertile shoots bear reduced leaves with sporangia on their lower sides. *See* HYENIALES; PALEOBOTANY; SPHENO-PHYLLALES. Harlan P. Banks

Pseudomonas

A genus of gram-negative, non-spore-forming, rodshaped bacteria. *Pseudomonas* is a heterogeneous taxon; division at the genus level may be revised with some species being removed to *Xanthomonas*, *Comamonas*, *Alteromonas*, and *Acetobacter*. The type species is *P. aeruginosa*.

Motile species possess polar flagella. They are strictly aerobic, but some members do respire anaerobically in the presence of nitrate. Some species produce acids oxidatively from carbohydrates; none is fermentative and none photosynthetic. Pseudomonads are catalase-positive and usually indophenol oxidase-positive. Yellow-green fluorescent pigments are produced by some species, and blue, yellow, red, or brown nonfluorescent pigments are also characteristic of certain species.

Members of the genus Pseudomonas cause a variety of infective diseases; some species cause disease of plants. One species, P. mallei, is a mammalian parasite, and is the causative agent of glanders, an infectious disease of horses that occasionally is transmitted to humans by direct contact. Pseudomonas pseudomallei is the causative agent of melioidosis, an endemic glanderslike disease of humans and animals, in Southeast Asia. Pseudomonads such as P. maltophilia, P. cepacia, and P. stutzeri have been implicated in nosocomial human infections. Pseudomonas aeruginosa is the most significant cause of hospital-acquired infections, particularly in predisposed patients with metabolic, hematologic, and malignant diseases. The spectrum of clinical disease ranges from urinary tract infections to septicemia, pneumonia, meningitis, and infections of postsurgical and posttraumatic wounds.

Such toxic factors as lipase, esterase, lecithinase, elastase, endotoxin, and exotoxin are produced by *P. aeruginosa*, some of which may contribute to its pathogenesis. Serological, bacteriophage, and pyocin typing systems are used as markers to trace sources of infections to the epidemic foci. Serious *P. aeruginosa* infections require combination drug therapy with an aminoglycoside and a beta-lactam penicillin. *See* GLANDERS; HOSPI-TAL INFECTIONS; MENINGITIS; PASTEURELLOSIS; PNEU-MONIA. Gerald L. Gilardi Bibliography. M. Bendinelli, M. Campu, and H. Friedman (eds.), *Pseudomonas Aeruginosa as an Opportunist Pathogen*, 1993; R. Hugh and G. L. Gilardi, *Pseudomonas: Manual of Clinical Microbiology*, 6th ed., 1995; R. Y. Stanier, N. J. Palleroni, and M. Doudoroff, The aerobic pseudomonads: A taxonomic study, *J. Gen. Microbiol.*, 43:159-271, 1966.

Pseudophyllidea

An order of tapeworms of the subclass Cestoda, parasitic in the intestine of all classes of vertebrates. Typically, the head is simple in structure with two groovelike attachment organs (**Fig.** 1*a*), the bothria.

Most pseudophyllideans are segmented and polyzoic with replication of the reproductive systems, although there are a number which do not show such replication and are monozoic. The genital openings are typically in the midline rather than lateral, and there is usually a uterine pore in the midline from which embryos are discharged (Fig. 1*b*). *Dibothriocephalus latus*, the broad or fish tapeworm of humans and certain piscivorous mammals, is a pseudophyllidean. In humans, this worm sometimes pre-

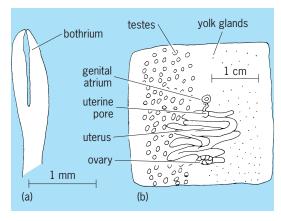


Fig. 1. Pseudophyllidea morphology. (a) Scolex of *Dibothriocephalus*. (b) A body segment.

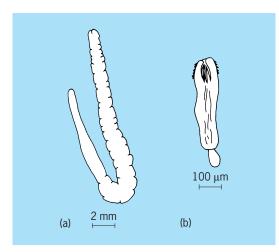


Fig. 2. Larval forms of *Dibothriocephalus*. (a) Procercoid larva. (b) Plerocercoid larva.

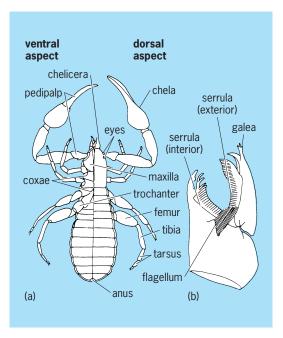
cipitates a pernicious anemia by competing with the host for vitamin B_{12} .

After leaving the intestine of the mammalian host, the ciliated embryo, or coracidium, of D. latus escapes from its shell and must be eaten by an arthropod, a copepod of the genus Cyclops or Diaptomus. In the copepod, the embryo develops into a procercoid (Fig. 2a). For development to continue, the copepod must be eaten by a fish in which the worm develops into a plerocercoid larva (Fig. 2b). If this plerocercoid is eaten by the mammalian host, the worm rapidly grows to sexual maturity, sometimes attaining a length of more than 30 ft (9 m). Larval pseudophyllideans are occasionally found as parasites in the extraintestinal tissues of humans, producing a condition known as sparganosis. See EU-CESTODA; SPARGANOSIS. Clark P. Read

Bibliography. L. H. Hyman, *The Invertebrates*, vol. 2, 1951; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982; R. A. Wardle et al., *Advances in the Zoology of Tapeworms*, 1950-1970, 1974.

Pseudoscorpionida

An order of terrestrial Arachnida having the general appearance of miniature scorpions without the postabdomen and sting (**illus.** *a*). These animals



Pseudoscorpionida. (a) Adult, in two views. (b) Chelicera of male pseudoscorpion. (After H. S. Pratt, A Manual of the Common Invertebrate Animals, rev. ed., McGraw-Hill, 1951)

are also known as the Chelonethida. The body length is seldom greater than 0.2 in. (5.0 mm). Typically, each finger of the anterior appendages, or chelicerae, has a serrula composed of a row of ligulate plates (illus. *b*). Ducts of silk glands open near the end of

the movable finger, often in connection with a simple or branched spinneret, or galea. The second pair of appendages, or palpi, are large and conspicuous, usually with glands that discharge venom through a terminal tooth on one or both of the chelal fingers. The four pairs of legs are ambulatory. Eggs and larval young are frequently carried on the underside of the abdomen of the female. Pseudoscorpions feed chiefly on small arthropods and, although frequently found on birds, mammals, and insects, are considered nonparasitic. Pseudoscorpions are common in the nests of mammals, birds, and social insects, in woody debris and forest litter, under stones, and in crevices in the bark of trees. A few species are found in barns and chicken houses and the virtually cosmopolitan species Chelifer cancroides is frequently found in libraries and dwellings. Tertiary fossil pseudoscorpions from Baltic amber are very similar to modern forms. About 2000 species of pseudoscorpions have been described. See ARACHNIDA. C. Clayton Hoff

Bibliography. M. Beier, *Pseudoscorpionidea*, in F. E. Schulze and W. Kükenthal (eds.), *Das Tierreich*, vols. 57–58, 1932; J. C. Chamberlain, The arachnid order Chelonethida, *Stanford Univ. Publ., Biol. Sci.*, 7(1):1-284, 1931; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982; P. Weygoldt, *The Biology of Pseudoscorpions*, 1969.

Pseudosphaeriales (lichenized)

An order of the class Ascolichenes, shared by the class Ascomycetes. The order is also called the Pleosporales. The genera now assigned to this order were formerly classified in the Pyrenulales. They resemble the typical pyrenomycetous lichens except for the structure of the ascocarp, which is not a true perithecium. It is flask-shaped and lined with a layer of interwoven, branched pseudoparaphyses. The asci, with bitunicate walls, are located in scattered locules. Little is known about the structure and development of these ascocarps.

Except for the presence of symbiotic algae, these lichens are very close to the nonlichenized Pseudosphaeriales, but none are pathogenic. The usual habitat is tree bark, and the species are common in temperate and tropical regions. There are two major families. The larger one, Arthopyreniaceae, is a widespread family with at least 5 genera, the largest of which, *Arthopyrenia*, has more than 50 species. The Mycoporaceae is a small family with two well-known genera, *Dermatina* and *Mycoporellum*. All of the species in this order are crustose and many lack a well-defined thallus. Mason E. Hale

Psilophytales

A group long recognized as an order of fossil plants (subdivision Psilopsida) collected in rocks of Late Silurian and Devonian age. It has been subdivided, on the basis of many data, into three categories whose descriptions include the chief kinds of plants formerly included in Psilophytales. Some authors will treat the three groups as classes under the division Rhyniophyta; others may prefer to consider them as three separate divisions of the subkingdom Embryobionta. They are given in this Encyclopedia as three classes of the division Rhyniophyta: Rhyniopsida, Zosterophyllopsida, and Trimerophytopsida. *See* EMBRYOBIONTA; PALEOBOTANY; RHYNIOPHYTA; RHYNIOPSIDA; TRIMEROPHYTOPSIDA; ZOSTEROPHYL-LOPSIDA. Harlan P. Banks

Psilotophyta

A division of the plant kingdom consisting of only two genera with three living species, *Psilotum nudum*, *P. complanatum*, and *Tmesipteris tannensis*. *Psilotum* is widespread in tropical and subtropical regions of both hemispheres, but *Tmesipteris* is confined to Australia and some of the Pacific islands.

The Psilotophyta have the typical life cycle of vascular cryptogams, with an alternation of sporophyte and gametophyte generations, the sporophyte being much the larger and more complex. The sporophyte consists of a slender, evergreen, more or less dichotomously branched stem with small, alternate leaves, which in *Psilotum* are tiny, simple or bifurcate scales without vascular tissue. The sporangia are compound, trilocular, and terminal on very short branches axillary to, but sometimes grown together with (adnate), some leaves. The underground part of the stem is slightly differentiated from the aerial part, and the plant has no true roots. The gametophyte is tiny, subterranean, and mycorhizal. *See* REPRODUC-TION (PLANT).

The Psilotophyta have no economic importance, but they are interesting as possible remnants of an ancient (Silurian and Devonian) group of plants, the Rhyniophyta, which is regarded as ancestral to all other vascular plants. The dichotomously branched stems, morphologically terminal sporangia, and absence of roots are considered particularly significant in linking the two groups, and indeed some systems of classification put them into a single division under the name Psilophyta. Other botanists have suggested a possible relationship of the Psilotophyta to the Polypodiophyta or the Lycopodiophyta. *See* LYCO-PHYTA; PLANT KINGDOM; POLYPODIOPHYTA; PSILO-PHYTALES; RHYNIOPHYTA. Arthur Cronquist

Psittaciformes

The parrots, a large order of land birds found worldwide but with most species concentrated on the landmasses of the Southern Hemisphere, particularly Australasia, the Neotropics, Africa, and southern Asia. Only a few species are found in the northern temperate regions. The parrots likely are closely allied to pigeons, from which they have probably evolved. *See* COLUMBIFORMES. **Classification.** Although the parrots are generally placed in a single family, the Psittacidae, they are divided into a number of distinct subfamilies, which are listed below:

Order Psittaciformes Family Psittacidae Subfamily: Psittacinae (parrots; 247 species; worldwide, but few in the Palearctic) Cacatuinae (cockatoos; 18 species; Australasia to Indonesia) Micropsittinae (pygmy parrots; 6 species; New Guinea) Strigopinae (owl parrot; 1 species; New Zealand) Nestorinae (keas; 3 species; New Zealand) Loriinae (lories; 54 species; Australia to Indonesia) Psittrichadinae (Pesquet's parrot; 1 species; New Guinea) Loriculinae (hanging parrots; 10 species; southeast Asia)

Some workers consider the cockatoos (see **illustration**) more distinctive and have placed them in a separate family, the Cacatuidae. Among other changes, their tongue is considerably reduced in size; although it is still used to manipulate food, cockatoos can no longer drink by lapping water with their tongues. Rather, they have to scoop up water with their broad bill, which restricts them to areas with standing water. Since occupation of Australia by Europeans and the spread of ranching with water troughs for livestock, several species of cockatoos have expanded their range greatly on that continent.

Fossil record. Parrots have a very poor fossil record, as would be expected of a group of forest and grassland birds. The earliest form is from the lower Eocene of England, followed by a late Eocene fossil from France. Only one fossil parrot is known from the North American Tertiary and was found in the late Miocene deposits of Nebraska. Most of the few fossil parrots that have been found are from the Holarctic, outside the present geographic range of the order.

Characteristics. Parrots have a characteristic strong hooked bill. The tongue is large and fleshy; it is used for manipulating food in the bill, for obtaining nectar (many papillae exist on the tongue tip in some groups such as the lories), and for drinking. A short neck connects the large head to the stocky body. The wings are of medium length, varying from pointed to rounded. The legs are short and stout with strong, clawed toes arranged in a zygodactyl pattern, two pointing forward and two backward. The tail varies from short to long. Pigmy parrots climb on the vertical trunks of trees and are said to eat lichens on the bark. Parrots fly well and can attain high speeds, and the larger cockatoos can soar, but even those few parrots that are migratory journey only short distances. One species, the



Sulphur-crested cockatoo (Cacatua galerita). (Photo by Lloyd Glenn Ingles; © California Academy of Sciences)

owl parrot, is flightless. Parrots can walk well but not rapidly, and the feet and the bill are used for climbing about in trees as well as for eating. Most parrots are arboreal, but some are terrestrial. A few are nocturnal. Plumage of parrots is variable; most often it is green, but red, orange, yellow, blue, black, and white in bold bright patterns are also common. Parrots are mainly vegetarians, eating seeds, nuts, fruit, nectar, or pollen; a few consume animal food.

Parrots are social birds and usually live in flocks. Almost all species have a strong pair bond, remaining together year-round for life. The nest is built in a cavity, generally in a tree, but sometimes in the ground, except for the monk parakeet (*Myiopsitta monachus*), which builds a large globular nest of sticks. Usually both sexes incubate the clutch of three or four eggs. The young are unfeathered and remain in the nest to be cared for by both parents. An exception is the owl parrot (Kakapo, *Strigops habroptilus*) of New Zealand. The males of this species establish courting stations in a loose lek, or courtship area, to which the females come. No pair bond is formed, and only the female incubates and cares for the young.

Parrots are found mainly in the tropics and in the southern continents, and the center of parrot radiation is the Australasian-New Zealand region. A large monophyletic radiation of psittacine parrots is found in the Neotropics, and a few species inhabit areas of Africa and tropical Asia.

Many parrots in captivity have an excellent ability to mimic words, and some can be taught a large vocabulary. However, parrots are not known to mimic the calls of other birds in the wild. Their vocalizing ability is probably important for the constant communication between mates.

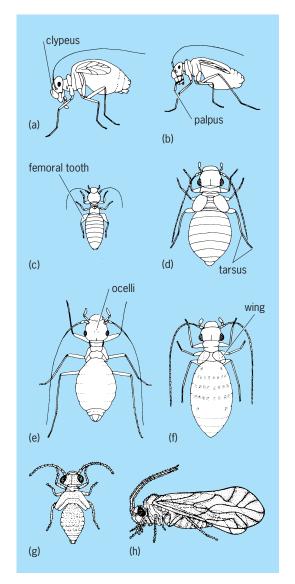
Economic significance. Parrots are the most important birds in the avicultural and pet trade, with many species kept in captivity. A number of species have escaped, and feral populations exist in the United States mainly in California and Florida. The monk parakeet is established in more northerly parts of the United States, such as the New York City area. Because the demand for parrots is so great, many countries have imposed exportation bans on native birds or importation restrictions on wild-trapped parrots. Two species, the cockatiel (Nymphicus hollandicus) and the budgerigar (Melopsittacus undulatus; the parakeet of the pet trade), have been domesticated, and several distinct breeds of each have been developed. Some species are seriously endangered because of habitat destruction, introduction of mammalian predators (such as weasels into New Zealand causing great reduction of the flightless owl parrot), and excessive capture of the pet trade. The Carolina parakeet (Conuropsis carolinensis) of North America is now extinct, as are several other species. Some psittaciforms are agricultural pests, consuming tree fruits. Parrots can transmit an especially virulent form of ornithosis to humans. See AVES; ENDANGERED SPECIES. Walter J. Bock

Bibliography. J. M. Forshaw, *Parrots of the World*, 3d ed., Landsdowne Editions, 1989; J. M. Forshaw, *Parrots of the World: An Identification Guide*, Princeton University Press, 2005; T. Juniper and M. Parr, *Parrots: A Guide to the Parrots of the World*, Yale University Press, 1997.

Psocoptera

An order of insects frequently referred to as the Corrodentia, or Copeognatha. Common names for members of this order are book lice, bark lice, and psocids, the latter a general term for all members of the order. They are usually less than 0.25 in. (0.6 cm) long, though rarely some may reach about 0.5 in. (1.2 cm). Wings may be absent, and when present are of differing distinctive venational types. Tarsi are two- or three-segmented, cerci are absent (see **illus.**), and metamorphosis is gradual. Chewing mouthparts usually have a much enlarged clypeus; the lacinia of the maxilla is usually elongate and chisellike, and the antennae have 13 or more segments.

Book lice are most common among old papers on dusty shelves, in cereals, or other domestic situations. They are usually pale, wingless types of insects. Many bark lice, the majority winged, occur on the bark or foliage of trees, and some are found under dead bark or beneath stones. Nymphs of a few species occur on tree trunks as clusters of gregarious individuals, but disperse when mature. Food consists of microscopic molds, cereals, fragments of dead insects, and other organic debris. Sometimes psocids infest granaries, houses, straw packing, or museum collections of insects to the extent that injury, or a



Some genera of Psocoptera. (a) Psocathropos; (b) Dorypteryx; (c) Liposcelis; (d) Lepinotus; (e) Psyllipsocus; (f) Trogium (after A. B. Gurney, Pest Control Technology, National Pest Control Association); (g) Lachesilla, nymph; (h) Lachesilla, adult.

nuisance, is involved, but control measures are seldom necessary.

Psocoptera are worldwide, especially in warm coun-tries, and some 1300 species are known. Current classification now lists about 27 families for this group. About 150 species, in 11 families, have been found in the United States. *See* INSECTA.

Ashley B. Gurney

Bibliography. D. J. Borror and D. M. DeLong, An Introduction to the Study of Insects, rev. ed., 1964; P. J. Chapman, Corrodentia of the United States of America, I: Suborder Isotecnomera, J. N. Y. Entomol. Soc., 38:219–290, 319–403, 1930; E. L. Mockford and A. B. Gurney, A review of the psocids or booklice and barklice of Texas, J. Wash. Acad. Sci., 46:353–368, 1956; S. P. Parker (ed.), Synopsis and Classification of Living Organisms, 2 vols., 1982; C. N. Smithers, A bibliography of the Psocoptera, Aust. Zool., 13:137–209, 1965.

Psychoacoustics

All of the psychological interactions between humans (and animals) and the world of sound. It encompasses all studies of the perception of sound, as well as the production of speech.

Hearing. The simplest sounds-that is, pure sine waves-can be described in terms of two physical characteristics, frequency and intensity. The primary perceptual attributes of sound are loudness (perceived intensity from soft to loud) and pitch (perceived quality from low to high). However, there is no one-to-one correspondence between the physical characteristics and the perceptual attributes. Although loudness depends largely on a sound's physical intensity, it also depends on frequency. A basic measurement is the audibility function, the absolute threshold for detecting sound as a function of sound frequency. Threshold is far from uniform across the audible spectrum. The physical intensity needed to detect a sound with a frequency of 100 Hz must be a million times greater than the intensity needed to detect a sound with a frequency of 1000 Hz. See LOUDNESS.

Similarly, though the pitch of a pure tone depends primarily on its sound frequency, it also depends to some extent on intensity. As intensity increases, low-frequency tones decrease in pitch, and highfrequency tones increase in pitch. A complex sound, composed of many frequencies, has a pitch that usually corresponds to the pitch of the fundamental frequency. *See* PITCH.

Musical sounds also vary in the attribute known as timbre, which characterizes the sounds of individual instruments. The same note played on an oboe and a piano may have the same fundamental frequency and therefore the same basic pitch; however, the notes differ in timbre (quality) because of differences in their patterns of higher-frequency harmonics and because of differences in the temporal characteristics of onset of the notes. *See* MUSICAL ACOUSTICS.

The localization of a source of sound is made possible by the spatial separation of the two ears. A sound coming from the left arrives at the left ear before the right ear, and this difference in time of arrival makes it possible for the auditory system to determine the location of the sounds. Intensity differences at the two ears provide additional information about location. With high-frequency sound waves, the head acts as a barrier, reducing the sound intensity at the opposite ear. Low-frequency sound waves tend to bend around the head, so the intensity difference between the ears is much smaller. Movements of the head are particularly important in enabling humans and animals to locate the source of a sound.

The capacity to detect sounds in the environment and to discriminate among sounds depends on the frequency spectra, intensities, and temporal properties of the sounds. This capacity also depends on the presence of other acoustic energy in the environment, usually noises, which may mask the sounds. In general, masking agents exert their most deleterious effects on sounds with the same or similar frequen-

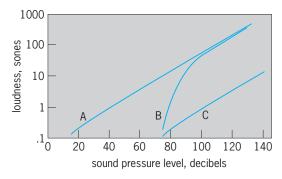


Fig. 1. Relationship between perceived loudness in sones and sound pressure level in decibels for a normal ear (curve A), an ear with a sensory-neural hearing loss (B), and an ear with a conductive hearing loss (C).

cies. In order for a person to detect a target sound above the level of background masking noise, the masking noise cannot be more than 10 dB greater in intensity than the target. Even though a sound can be detected in noise, it may not be possible to discern what the sound is. For speech to be fully intelligible, it must have an average intensity level 20–30 dB above the background noise. *See* ACOUSTIC NOISE; HEARING (HUMAN).

Hearing loss and deafness. Hearing deficits may be congenital or acquired. It is especially important to detect cases of congenital hearing loss early in life— before the age of 2 or 3 years—because the acquisition of normal speech and language can be permanently impaired by profound deafness. Modern audiometric techniques permit the evaluation of hearing in infancy, and infants and young children can be fitted with hearing aids tailored to their individual needs in the same way adults can. *See* AU-DIOMETRY; HEARING AID.

There are basically two kinds of hearing loss, conductive and sensory-neural. In conductive hearing loss, the deficit stems from the failure of sound to be transmitted properly to the receptors in the inner ear. Often, this results from blockage in the ear canal or from fusion of the bones in the middle ear, the latter condition known as otosclerosis. Because conductive loss consists of a deficiency in sound transmission, its effect appears as a constant relative loss in loudness, as shown in **Fig. 1**. A conductive loss of 60 decibels means that every sound is reduced in physical intensity by a factor of a million. In terms of perceived loudness in sones, an ear with a conductive loss of 60 dB hears sounds 1/64 as loud as a normal ear hears them.

Sensory-neural hearing loss typically results from damage to the receptive elements in the ear or to the auditory nerve itself. This is the kind of hearing loss that in many individuals encroaches with advancing age. Typically, such losses are most evident at high sound frequencies (above about 1000 Hz). Unlike conductive disorders, which affect intense and weak sounds equally, sensory-neural losses show up mainly at low sound levels, which are inaudible or much attenuated perceptually. High sound levels, however, may be perceived to be just as loud by a person with a sensory-neural loss as by a person with normal hearing. This is shown in Fig. 1. The rapid rise in loudness perception just above threshold is called recruitment. *See* HEARING IMPAIRMENT.

Speech production. Speech sounds are acoustic patterns imposed on air expelled by the lungs. The specific properties of speech—sound intensity, frequency spectrum, and time course—depend on the force with which the air is expelled, by the vibrations of the vocal cords in the larynx, by the shapes of the cavities in the mouth and throat, and by the movements of the tongue, lips, and jaw. It is the position of the tongue that is most important in determining individual vowel sounds, while individual consonants are determined by place and manner of articulation, and by whether they are voiced or unvoiced. *See* PHONETICS; SPEECH.

Speech perception. Most of the energy in the sound waves of speech falls in the frequency region from 50 to 10,000 Hz. Of this, the region from about 200 to 7000 Hz carries most of the information for speech perception. Vowel sounds are characterized by bands of energy concentrated at certain frequencies; these energy bands are known as formants. The spectral locations of the formants correspond to resonances in the vocal tract. The formants are numbered in order or increasing sound frequency: first, second, third, and so on. In general, the first three formants—especially the first and second—are most important in identifying vowels.

Bursts of noise, pauses, and transitions to the formants mark locations in the speech signal where consonants are heard. It is not possible, however, to identify every consonant by means of a single cue in the acoustic spectrum: The very same burst of noise may be heard as *k* when it precedes the vowel *ab* but as *p* when it precedes *oo* or *ee*. Similarly, very different acoustic events can be perceived as the same consonant when they appear in the context of different vowels. The perception of consonants, therefore, depends not only on the acoustic signal at that moment but also on the pattern of acoustic information surrounding it.

The intelligibility of speech signals transmitted over a communication channel depends on such factors as the intensity of the sound, the fidelity of the communication system, and the presence of masking agents such as noise. It is common for communication systems to truncate the sound waves at high amplitudes; this type of distortion is known as peak clipping. Fortunately, intelligibility is affected little by removing high-amplitude peaks in a sound wave, for it is the central region of the speech wave, where information about consonants is represented, that is most important.

Although most of the energy in speech comes from vowels, most of the information comes from consonants. For this reason, as much information is lost if all frequencies above 1900 Hz are removed as if all frequencies below 1900 Hz are removed, despite the fact that the frequencies above 1900 Hz contain only about 2% of the average energy in a speech signal: sound frequencies above 1900 Hz are largely devoted

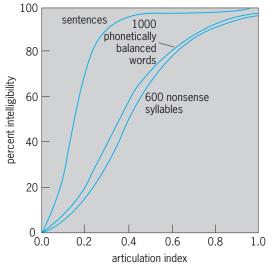


Fig. 2. Relationship between intelligibility and the articulation index for various types of speech.

to consonants, frequencies below 1900 Hz to both consonants and vowels.

It is possible to predict the intelligibility of speech by the following scheme: The frequency range from 200 to 6100 Hz is divided into 20 bands, each of which has the potential to contribute equally to intelligibility. Each band's actual contribution is determined by the intensity of the speech signal relative to any masking noise in that band. The actual contributions of all 20 bands sum to yield an articulation index, a variable that ranges from 0.0 to 1.0. From the articulation index one can determine the intelligibility of various kinds of speech, whether simple syllables, single words, or connected discourse (**Fig. 2**). *See* PERCEPTION; PSYCHOLINGUIS-TICS. Lawrence E. Marks

Bibliography. R. B. Batkin, *Hearing and Hearing Disorders*, 1988; R. J. Dooling and S. H. Hulse (eds.), *The Comparative Psychology of Audition: Perceiving Complex Sounds*, 1990; N. J. Lass, *Speech, Language and Hearing*, 1989; R. D. Luce, *Sound and Hearing*, 1992.

Psychoanalysis

Psychoanalysis is a psychological theory; a form of psychotherapy, especially for the treatment of neurotic and character or personality disorders; and a method for investigating psychological phenomena. Psychoanalysis was created and developed by Sigmund Freud, who presented his method, clinical observations, and theory in *Interpretation of Dreams* and other major works, including *The Psychopathology of Everyday Life* and *Three Essays on the Theory of Sexuality*, as well as in many of his case studies.

Psychoanalytic theory. Generally, psychoanalysis is concerned with the causal role of wishes and beliefs in human life. More specifically, it attempts to explain mental or behavioral phenomena that do not appear to make sense as the effects of unconscious wishes and beliefs. Such phenomena include dreams, disturbances in functioning such as slips of the tongue or pen and transient forgetting, and neurotic symptoms.

Typically, unconscious wishes and beliefs are constituents of conflicts. Most simply, the conflict consists of a wish, a belief that there is danger in gratifying the wish, and a solution of sorts. The solution can take the form of a sequence of imagined events, a fantasy, in which the wish is fulfilled in such a way that the danger is evaded or mitigated. Such wishes, beliefs, and fantasies constitute the person's psychic reality, which influences conscious perceptions, feeling states, memories, beliefs, thought processes, and actions.

The term unconscious in psychoanalysis does not mean simply that mental contents are out of awareness. Its psychodynamic meaning is that the person does not want to be aware of these contents, and takes active steps to avoid being aware of them. A fundamental hypothesis of psychoanalysis is that because a mental entity is dynamically unconscious it has the causal power to produce the phenomena that are of interest to psychoanalysis.

At first, the dynamic unconscious was thought to consist of traumatic memories. Later, it was believed to consist of impulses or wishes-especially sexual (and aggressive) impulses or wishes-having some causal link or, at least, some similarity to impulses or wishes in childhood. The significance of memories of actual events is believed to lie in their connection to impulses or wishes. Psychoanalysis now emphasizes that the dynamic unconscious consists of fantasies, which have a history reaching back to childhood. These fantasies are internal scenarios in which sexual (and aggressive) wishes are imagined as fulfilled. Wish fulfillment occurs in association with imagined threats. Therefore, anxiety signals danger when an unconscious fantasy is activated and the person is then motivated to modify the internal scenario in such a way as to mitigate the threat and thereby avoid any further anxiety. A person elaborates the fantasy in attempting to discover a way to fulfill the wish while more or less successfully evading the imagined threat

The construction of an internal scenario throughout a lifetime, then, involves layers of revisions resulting from different responses to the same threat, or to different or multiple threats, imagined during different developmental epochs. These revisions are essentially symbolic operations (referred to as defense mechanisms, dream work, or primary processes) on the contents of mental states. Defense mechanisms, such as repression, expel or delete contents from consciousness. In addition, they may serve to prevent these contents from returning to consciousness-as reaction formation does by substituting for an expelled wish or impulse its opposite (for example, love for hate, kindness for cruelty). Other defense mechanisms dispose of rejected wishes or impulses by altering the contents of a selfrepresentation or object-representation, or by shifting attention from one to the other. The dream work or primary processes, which include condensation and displacement, transform latent mental contents into manifest contents that may not make sense. Condensation is the use of one manifest content to stand for or allude to a whole set of latent ideas. Displacement is a shift of attention or emphasis from a significant latent content to an indifferent manifest content. Images, which are referentially vague or ambiguous, may be used instead of words to stand for or allude to latent ideas. Iconic (metaphoric) symbols, also known as Freudian symbols, may be used to stand for or allude to latent ideas that go unrecognized in this guise.

Psychoanalysis is distinct in attributing causal powers to unconscious sexual wishes. Such attribution depends on extending the meaning of sexual to encompass the quest for sensual pleasure in childhood (so-called infantile sexuality) and choices of objects and aims other than those exemplified by two adults of the opposite sex engaging in coitus. Unconscious fantasies have themes that correspond to postulated psychosexual stages of development. These stages are characterized and dominated by experiences, thoughts, feelings, kinds of relations with others, and kinds of sensual pleasure associated with a particular zone of the body, including the skin, mouth, teeth, anus, and genitals. One theme that is thought to have particular importance is the Oedipus complex, in which the child rivals one parent in seeking sensual gratifications of various kinds from the other parent.

When an unconscious fantasy is activated, it manifests itself in conscious mental states or in actionsimportantly, in emotions; in interpretations of the significance of events or states of affairs; in attributions of motives to others; and in daydreams, dreams, and neurotic symptoms. An unconscious fantasy may be activated periodically and spontaneously, as it is part of its nature to so manifest itself, independently of situational vicissitudes. It may be activated by something in an external situation that reminds the person of the fantasy. For example, a fantasy may come to be manifested by a mood that seems either out of place or not to make sense in the situation that evokes it but that belongs to and makes sense in the unconscious internal scenario. An unconscious fantasy may also be activated by the willful mental activity of a person who wants to experience one of its manifestations, such as a conscious daydream or fantasy, a particular mood or emotion, or pleasure or relief from anxiety.

Unconscious fantasies, as distinct from both conscious reality-oriented imagining and conscious daydreaming, are constructed when imagination functions under very special conditions. The person, immersed in imagining, ignores the fact that he or she is merely imagining: the truth of beliefs is not permitted to matter. Connections are made that have nothing to do with what is true by means of such nonlogical mental operations as condensation, displacement, translation of verbal symbols into imagery, and iconic, or metaphoric, symbolization. The mind is imagined to be both spatial and corporeal: spatial because it can be inhabited by actors in internal scenarios; and corporeal since it is imagined to be a body or a part of a body and is therefore imagined to have the causal powers of a body or part of a body. Because people can identify with an actor in the internal scenario, what is imagined as giving pleasure can be actually experienced as pleasurable.

This emphasis on fantasy underscores the fact that psychoanalysis gives priority to the relation between wishes (including wishes a person knows could not conceivably be gratified in reality) and imagination (functioning under very special conditions). In particular, psychoanalysis focuses on how sexual wishes and imagination are implicated in the vicissitudes of unconscious wish fulfillments, their construction, and their effects. This is in contrast to other branches of the psychology that, for example, might give priority instead to relations among conscious mental states, and in particular to the causal relation between desire and belief acting together, on the one hand, and rational or irrational action, on the other, where irrational action arises because beliefs are erroneous or relevant beliefs are missing.

Variants of psychoanalytic theory have emphasized the role of destructive object-relations in determining inner reality; how attempts to avoid anxiety or blows to self-esteem lead to distortions in interpersonal relations; the effects of actual early relations, especially with the mother, on the development and maintenance of a sense of identity and self-esteem; and the effects of deficits in early life on psychological capacities and personality. All four of these variants share two characteristics: They tend to assign a major (if not exclusive) causal role to experiences in external reality, especially during the preverbal period of infancy and in relations with the mother. They also tend to deemphasize the causal role of sexual wishes and to reject the primacy of the quest for sensual pleasure.

Psychotherapy. In psychoanalytic psychotherapy, 45- or 50-min sessions are scheduled three to five times a week over a number of years. The patient lies on a couch rather than sitting facing the psychoanalyst, so that the maximum amount of attention is directed to what is within the person rather than to external cues.

Free association is the method of psychoanalysis, both as a form of psychotherapy and as a method of investigation. Patients are encouraged not to talk about some particular problem or aspect of their lives but rather to suspend any conscious purposive organization of what they say, speaking freely whatever comes to mind, and avoiding as much as possible censoring of what may seem trivial, embarrassing, shameful, bad, offensive, or senseless. Both psychoanalyst and patient follow the patient's productions: conscious purposes are replaced by unconscious purposes, which, under these conditions, can determine the direction of the patient's mental processes with less interference. Information that is not easily if ever acquired in other situations can be obtained in the psychoanalytic situation because the participation of the psychoanalyst is disciplined, and the influence of external cues and internal inhibitions on patients and what they say is minimized.

Psychoanalysts do not take what their patients talk about at face value. Interventions are predominantly interpretative; psychoanalysts do not seek primarily to tell their patients what to do, to educate them about the world, to influence their values, or to reassure them in one way or another that everything is or will be all right. Psychoanalysts look for patterns in what each patient says and for signs of feelings of which the patient is more or less unaware. They then engage their patients (who are increasingly aware of these patterns and able to experience and articulate these feelings) in an inquiry about the reasons for them or motives behind them. The focus is on what the patients do not know-and do not want to know-about themselves and their inner life, including strategies for avoiding such knowledge and the consequences of these strategies. Much of the work of psychoanalysis involves coping with the ways in which patients, more or less wittingly, oppose what they regard as efforts to interfere with these strategies. Interpretation of resistance is a distinctive feature of psychoanalytic psychotherapy.

The goal of psychoanalytic psychotherapy is to extend the realm of what patients permit themselves to experience. That does not mean simply to know intellectually about a mental state, but to know firsthand what it feels like to be in such a state. This change occurs especially as patients begin to attribute to their psychoanalysts what are actually their own wishes, impulses, and attitudes or to identify their psychoanalysts with early significant figures who were the object of these wishes, impulses, feelings, or attitudes, or who were seen or imagined as responding to these wishes, impulses, feelings, or attitudes in certain ways. Interpretation of transference is another distinctive feature of psychoanalytic psychotherapy.

Any person can have multiple wishes. Since some of these conflict with each other and some compete for inevitably limited resources for their gratification, they cannot all be gratified. The inevitable normal misery that this as well as unfavorable conditions in external reality causes patients is not what psychoanalytic psychotherapy seeks to alleviate. Rather, it tries to mitigate the misery that patients with a neurotic, character, or personality disorder inflict on themselves by their own more or less unwitting attitudes toward their own wishes and feelings, or by the more or less unwitting strategies that they use to avoid becoming aware of these wishes or feelings or anything associated with them, including their own attitudes toward them.

Method of investigation. The case-study method is characteristic of psychoanalytic research. The arguments that can be used in case studies are not much different from those used by Charles Darwin in his arguments for the credibility of the theory of evolution: namely, analogy (the use of familiar or homely models in which postulated causes and mechanisms can be shown to exist); consilience (the convergence of inferences from different kinds of information on a common cause); and abduction (inference to the best explanation). *See* PSYCHOTHERAPY. Marshall Edelson

Bibliography. M. Edelson, *Psychoanalysis: A Theory in Crisis*, 1990; S. Freud, *Interpretation of Dreams*, 1900; S. Freud, *Introductory Lectures on Psychoanalysis*, 1966, reprint 1989; S. Freud, *Question of Lay Analysis*, 1964, reprint 1969; E. Jones, *The Life and Work of Sigmund Freud*, 3 vols., 1953-1957; L. A. Salome, *Freud*, 1994; R. Waelder, *Basic Theory of Psychoanalysis*, 1960.

Psycholinguistics

An area of study which draws from linguistics and psychology and focuses upon the comprehension and production of language in its spoken, written, and signed forms. Although psychologists have long been interested in language, and the field of linguistics is an older science than psychology, historically scientists in the two fields have had little contact. The two fields were thrown together, however, by the publications of Noam Chomsky in the late 1950s. Chomsky's writing had the effect of making psychologists acutely aware of their lack of knowledge about the structure of language, and the futility of focusing theoretical and research attention upon the surface structure of language while arguing that linguistics is as much concerned with the mind as psychologists should be. Psychologists, in turn, made linguists more aware of the larger cognitive context into which language fits as well as a broader set of methodological approaches available to scientifically test hypotheses about language. As a result, psycholinguists, who have a background of training in both linguistics and psychology, have been attempting since the early 1960s to gain a better understanding of the abstract rules which underlie human language and how they are acquired, used to communicate meaningful messages, and influenced by the biological and psychological context in which they occur. Research has been directed to the evolutionary development of language, the biological bases of language, the nature of the sound system, the rules of syntax, the nature of meaning, and the process of language acquisition. See LINGUISTICS.

Evolution and biological bases of language. While animals do convey messages to each other, it is clear that there are species-specific, finite limitations to those messages. Lower animals have limited capacities for generating new and appropriate messages when the occasion arises. Some animals can be taught to expand their communicative capabilities. The achievements of nonhuman primates which have been taught to communicate with humans and other primates via sign language, symbolic objects, and computer terminals are well beyond initial expectations. While it has been hard work for both the animal learners and the human teachers in these experiments, the achievements of the primates have been greater than expected, although hardly comparable to the human use of the symbolic language system so easily and naturally acquired by the human child. *See* ANIMAL COMMUNICATION.

Efforts to reconstruct the evolutionary development of language suggest that early hominid forms were structurally incapable of language. Their supralaryngeal vocal tracts, for example, were similar to those of present-day nonhuman primates. The modification in the supralaryngeal tract, along with other physiological and neurological changes which allowed language to develop, probably occurred long after the evolutionary branching which separated humans from other primates. Thus, although the other primates are the closest phylogenetic relatives of humans, there is little likelihood that they and humans are closely related with regard to symbolic capability.

The neuroanatomical bases of language have been studied by examining the language functioning of individuals with known brain injuries and those whose corpus callosum has had to be severed, by electrical evoked potentials taken from the scalps of normal persons, and by dichotic listening studies. Research in this area supports the early Greeks' hypothesis that there is a lateralization of brain function such that the left hemisphere appears to be much more intimately involved in most language functions than the right hemisphere. There is now evidence, however, that both hemispheres are involved to some extent in most language functioning, especially those aspects relevant to meaning. Differentiation of brain functioning seems to be present from birth, although developmental changes in function, plasticity of function, and cerebral hemispheric support for the functioning continue well beyond birth. The theoretical accounts of the relation between brain function and brain organization remain, however, a matter of considerable controversy. See BRAIN; HEMISPHERIC LATERALITY.

Sound system of language. An examination of the acoustic properties of speech reveals that the physical properties of speech sounds are unusually complicated insofar as their relations to the phonological components of the language system are concerned. The fundamental frequency set in motion by the air forced by the lungs through the vocal folds is modulated by the shape and length of the supralaryngeal vocal tract. The constantly changing configuration of the vocal tract during speech produces noise, defining some consonants and transitions to and from relatively steady-state formants, or bands of acoustic energy, defining vowels, to create syllables composed of consonants and vowels. While the vowels have fairly constant acoustic properties modified by the contexts in which they occur, the consonants seem to have no apparent one-to-one relation to the acoustic signal. See PHONETICS.

The phonological system is, however, abstracted in terms of discrete segments, phonemes, or syllables from the continuous sound wave. It appears that humans have a phonetic mode, special mechanisms that are engaged to process speech sounds differently than other auditory signals. These special mechanisms involve an integration of the motor gestural system, usually associated only with production, and the perceptual system, usually associated only with reception, such that neither system operates without engaging the other. This system accounts for the fact that speakers of a language will perceive radically different acoustic signals produced in different contexts as the same phoneme or syllable.

Before acquiring speech, the very young infant appears to be specifically attuned to linguistically relevant acoustic characteristics of the signal and capable of making distinctions among phonemes and syllables of speech as well as clausal units in natural speech. When speech production begins at about 1 year of age, the child begins to acquire the phonology of the language, which includes the distinctive features that define the phonemes of the language, the suprasegmental features such as pitch and stress that extend over syllables or words, and the phonological rules that apply to phonemic strings to alter them in various ways to derive their phonetic pronunciation. *See* SPEECH.

Morphology. Early research ignored the nature of words, or morphology, because in Chomsky's linguistics emphasis had been focused upon the sentence, but some studies have been exploring the manner in which words are stored in the lexicon, or mental dictionary, of the individual, the acquisition of rules that apply to morphemes derived from, for example, Latin as opposed to English, and the influences of context upon access to lexical items. Results suggest that morphemes may be stored independently of the various inflectional forms that may modify them; differential rules for inflecting morphemes derived from different sources are known to the adult speaker and are acquired at a relatively early age; and content effects appear to be important to the lexical processing of words

Syntax of language. The early research on comprehension and production of syntax began with the assumption that linguistic theories of syntax were psychologically based. That assumption was quickly challenged as it became clear that, in addition to the rules of syntax, processing of syntax involves memory, meaning, context, conversational rules, and the general knowledge of the persons involved in the communication. Much of the early research focused upon the processing of single sentences in isolation. It was demonstrated that the underlying structure of sentences which relates parts of sentences to each other affects, for example, speech production and perception; that ambiguous sentences sometimes take longer to process; that syntactically related sentences are more easily confused than unrelated sentences; and that negative sentences take longer to process than affirmative sentences. Some of these results are attributable to the finding that ambiguous words that have multiple meanings are processed more slowly because all of the meanings are called to mind in the sentence context prior to selecting the appropriate meaning for the particular context. Processing negative sentences appears to involve establishing the affirmative version of the sentence and then negating it. It was also discovered that when sentences are presented in memory tasks, people tend to forget the syntactic form of the sentences but retain the meaning unless they are specifically required to remember the syntax. In addition, it was found that sentences cannot be analyzed in isolation, because each sentence involves presuppositions and implications not explicitly stated but taken into account more or less automatically by a person processing those sentences. Furthermore, it has become clear that many of the influences of syntax on processing of sentences in isolation do not seem to be important when sentences are a part of a larger meaningful conversation or story. It is clear that syntax is required for the comprehension and production of meaningful sentences, but it has become progressively clearer that, from a psychological point of view, syntax cannot be treated in isolation. However, the question of whether the processing of syntactic as well as other sentence components occurs in isolated modules or as an interactive system is a controversial issue. The search for a psychological theory of syntax has led to a host of hypotheses and a number of techniques to test them. The simplest hypotheses have been rejected, and the complexity of sentence production, perception, and memory has become increasingly apparent. See INFORMATION PROCESSING (PSYCHOLOGY); MEMORY.

Meaning of language. Obviously language is used to convey meaningful messages from one person to another, and yet meaning, as such, has been the least studied component of language. The reason for this lack of attention has been the lack of a comprehensive theory to stimulate research. Three theoretical approaches, two of which have not proved totally satisfactory and a third too new to evaluate, have dominated thinking in this area. The first, which comes directly from linguistics, takes the word as the unit of analysis and focuses upon a componential analysis of words into feature elements of which meaning is composed. Synonymy, antonymy, and other relations among words are accounted for in terms of the presence or absence of the feature elements making up the meaning of the words involved. Analysis of a whole word into its component elements requires a knowledge of the whole of which the elements are the parts; recognition of this has led other theorists to propose alternative theories which are more holistic in nature. The latter conceptualizations of meaning take the abstract conceptual idea as a unit, and attempt to infer the structure of the concept based upon the relations among exemplars of that concept as used by speakers. The structure of the concept of fruit, for example, may be inferred from the relations among the exemplars apple, pear, guava. The apple exemplar is closer to the core meaning, or is more prototypic, of the concept fruit than is, for example, guava. Based upon this conceptualization of meaning, borrowed from L. Wittgenstein, it has been found that prototypic exemplars can be classified more rapidly, learned more quickly, and remembered more easily than peripheral exemplars. Neither of these theoretical approaches takes into consideration linguistic and extralinguistic contextual effects upon meaning, nor the fact that the broader knowledge system of the communicator influences the meaning of language.

A theoretical approach called connectionism is based upon the transmission of activation levels in large networks of densely interconnected simple units rather than the manipulation of structured symbolic elements. The tie to the computer and the analogy to neural networks make this an appealing but as yet untested theoretical approach. An important test may come with efforts to account for figurative language, where other theories have had difficulties. In any case, some inroads are being made into the understanding of meaning, but the problem remains relatively intractable. *See* INTELLIGENCE; NEURAL NET-WORK.

Language acquisition. Actual production of language is not evidenced until about the first birthday (although the signing child may start sooner), when repeated syllables such as "mama" and "papa" are likely to be the earliest utterances. These early utterances appear to be articulations of the mental and emotional states of the child. When a single word is the productive limitation of the child, extralinguistic context is used effectively to convey the message intended. In effect, the context is used by the child (and the listener) to expand the syntax of the message which the child is unable to produce linguistically. Gradually, the child frees himself or herself from the immediate context by adding to the linguistic portion of the message as syntactic and morphophonemic rules are acquired. Syntactic and morphophonemic complexities are usually expressed initially in forms simpler than those used by adults. The child seems to formulate increasingly complex hypotheses or strategies for coping with the problems of encoding and decoding messages.

As for the complexity of the messages themselves, research has led investigators away from language as such and into the broader realm of cognitive development. The nature of what children discuss is directly related to the level of their cognitive development. Children express their affective and cognitive mental states. However, there seems to be little relation between the acquisition process and what is ordinarily thought of as efforts to teach language by reward and imitation. Clearly children take advantage of exemplary input of various forms when they are ready to acquire those forms, but they are neither imitating nor learning by reward as adults decide to teach. It is apparent that adults and children do adapt their language input to the level of the child acquiring the language, but there is little evidence that the child attempts to imitate their language. Rather, the child formulates broad rules, and the processes of acquisition appear to be a matter of refining those rules to better express his or her mental contents in terms of the specifics of individual structures in the language. While the strategies have been shown to vary among individual children, and the structures for some messages are more complicated in some languages than in others, the construction of utterances that follow the rules for comprehending and producing specific constructions within a language are remarkably regular between children, both within a language and across languages.

In trying to understand the acquisition process, however, developmental psycholinguists have been driven at every turn from the study of phonology and syntax as independent modularly separate units to the interrelations among them and, in particular, to the study of semantics as well. Each component affects the others in the process of acquisition. While sound and syntax provide the medium for the message, it is meaning that is the message. The acquisition of that meaning has been shown to relate to biological givens, the broader cognitive system, and the relation of those biological and cognitive factors to the context in which the child develops. Analytic feature theories, synthetic prototype theories, and connectionist theories have been related to the general cognitive development of the child in an effort to understand the child's semantic development. In addition, learnability theory has been introduced in an effort to specify a class of grammars that include the grammars of all languages and to identify the language-learning procedures that a child could use to learn those languages, given the data available to the child. The child is certainly constrained in some ways in the acquisition process. For example, children overextend some words and underextend others, but they seldom form bizarre relations between words, and concepts. It seems clear that children have some sort of theoretical framework for conceptualizing the world with which they interact, and they search for the proper language tools for expressing their concepts and the relations among those concepts. See COGNITION; PSYCHOACOUSTICS. David S. Palermo

Bibliography. D. W. Carroll, *Psychology of Language*, 2d ed., 1994; V. A. Fromkin and R. Rodman, *An Introduction to Language*, 6th ed., 1998; J. B. Gleason and N. B. Ratner (eds.), *Psycholinguistics*, 2d ed., 1997; R. Hasan and J. R. Martin, *Language Development: Learning Language*, *Learning Culture*, 1989; C. K. Ogden and I. A. Richards, *The Meaning of Meaning*, 1989; M. Singer (ed.), *Psychology of Language: An Introduction to Sentence and Discourse Processing*, 1990.

Psychology

The study of human behavior and mental processes. Psychology is distinguished from similar fields of research and applied science by the unit of analysis, which is the individual. However, some psychological studies stretch this definition. For example, a study of group problem-solving may compare groups of different types, in which case the group is the unit of analysis. However, a psychological theory will still explain the effects by processes that apply at the level of individuals in a group.

Applied and experimental aspects. Psychology is sharply divided into applied and experimental areas. Although the field of psychology is devoted to and determined by clinical treatment of psychological problems, basic research in the field, which includes analysis of many problems, such as thought and perception, has little (if any) relation to clinical applications. However, many fields are represented in both research and applied psychology. For example, clinical psychology exists as an area of application in which individuals are treated for many types of disorders; however, on the research side, there is extensive work on the causes, diagnosis, and treatment outcome of psychological disorders. In many cases, a practitioner or researcher contributes to both aspects of the area. Despite such crossovers, research and application often remain distinct in methodology, purpose, institutional venue, training, professional affiliation, and publication outlets.

Researchers in psychology study a wide range of areas. Cognition includes such areas as human memory, use of language, perception, and attention. Cognitive research is often included as part of subdiscipline called cognitive science. This area examines central issues such as how mental process work, the relation between mind and brain, and the way in which biological transducing systems can convert physical regularities into perceptions of the world. Cognitive science is carved from the common ground shared by computer science [especially simulations of perception and thought (artificial intelligence and machine perception)], cognitive psychology, philosophy of mind, linguistics, neuropsychology, and cognitive anthropology. The study of human attention is a cognitive area that is central in the field. See COGNITION.

The study of consciousness involves such basic questions as the physiological basis of mental activity, the freedom of will, and the conscious and unconscious uses of memory. The latter topic can be classified under the rubric of implicit memory, in which people (sometimes also neurological cases with various forms of amnesia) can be shown to dissociate memory use from awareness of memory. That is, they might be unable to retrieve material learned, or may even be unaware that they learned it, while demonstrating in their behavior that they are using the information. *See* INSTINCTIVE BEHAVIOR; MEMORY; PSY-CHOLINGUISTICS; SENSATION.

Social psychology. Social psychology includes the study of interactions between individuals and groups, as well as the effects of groups (whether present or imagined) on the attitudes, opinions, and behavior of individuals. The field covers such topics as persuasion, conformity, obedience to authority, stereotyping, prejudice, and decision making in social contexts. Although much interest in the area has been stimulated by questions that arise from pressing social issues, the overall motivation has been to develop general theoretical accounts of human social behavior. *See* MOTIVATION; PERSONALITY THEORY.

Developmental psychology. This discipline has three subfields: life-span development, child devel-

opment, and aging. Most research in the area concentrates on child development, which examines the development of abilities, personality, social relations, and, essentially, every attribute and ability seen in adults. Life-span development and aging are growing fields of interest. *See* AGING; INTELLIGENCE.

Practitioners. The public awareness of psychology is dominated by the field of clinical psychology, which represents a fraction of theory and research in the field but which is the only point of contact for most of the population. A clinical psychologist is usually known by the term psychologist, which in some states is a term that can be used only by a registered practitioner. A psychiatrist is a physician with a specialty in psychiatric treatment and, in most states, with certification as a psychiatrist by a board of medical examiners. A psychoanalyst is typically trained by a psychoanalytic institute in a version of the Freudian method of psychoanalysis. To be accepted by a psychoanalytic institute for training, an applicant normally has an advanced degree in some form of psychological therapy. A large number of practitioners qualify both as psychoanalysts and psychiatrists, but these terms represent distinct areas of training. See PSYCHOANALYSIS.

Neuropsychologists are usually psychologists, who may come from an experimental or a clinical background but who must go through certification as psychologists if they are to render therapy. They treat individuals who have psychological disorders with a clear neurological etiology, such as stroke, brain trauma, and Tourette's syndrome.

Clinical practice includes individual consultation with clients, group therapy, and work in clinics or with teams of health professionals. Psychological therapists work in many settings and on problems ranging from short-term crises and substance abuse, to psychosis and major disorders such as schizophrenia and clinical depression.

Also cutting across the categories of practitioners is the theoretical orientation and therapeutic methodology of choice. While there are definite biases within each field, it is possible for a practitioner with any background to prefer behavior therapy, a humanistic approach, a Freudian (dynamic) approach, or an eclectic approach derived from these and other areas.

Nonclinical professional work in psychology includes the human-factors element, which traditionally is applied to the design of the interface between a machine and its human operator. Cognitive engineering is a branch of applied psychology that deals mainly with software and hardware computer design and other types of human-machine interfaces that are cognitively intensive. Industrial psychology also includes personnel selection and management and organizational planning and consulting.

The use of psychology in forensic matters is a natural result of the fact that much of law is based on psychology. Psychologists have been involved in jury selection, organization of evidence, evaluation of eyewitness testimony, and presentation of material in court cases. Psychiatrists and psychologists are also called on to diagnose potential defendants for mental disorders and the ability to stand trial. William P. Banks

Bibliography. H. Gardner, *The Mind's New Science: A History of a Cognitive Revolution*, 1987; H. Gleitman, *Psychology*, 5th ed., 2000; M. I. Posner and M. Raichle, *Images of Mind*, 1997; D. L. Schacter, *Memory: The Brain, the Mind, and the Past*, 1996.

Psychoneuroimmunology

The study of the interactions among behavioral, neural and endocrine, and immune functions. It is perhaps the most recent convergence of disciplines that has evolved to achieve a more complete understanding of adaptive processes. At one time, the immune system was considered an independent agency of defense that protected the organism against foreign material (that is, proteins that were not part of one's "self"). Indeed, the immune system is capable of considerable self-regulation. However, converging data from the behavioral and brain sciences indicate that the brain plays a critical role in the regulation or modulation of immunity. This research indicates that the nervous and immune systems, the two most complex systems that have evolved for the maintenance of homeostasis, represent an integrated mechanism for the adaptation of the individual and the species. Thus, psychoneuroimmunology emphasizes the study of the functional significance of the relationship between these systems-not in place of, but in addition to, the more traditional analysis of the mechanisms governing the functions within a single system-and the significance of these interactions for health and disease. See NEUROIMMUNOLOGY.

Brain-immune system interactions. Evidence for nervous system-immune system interactions exists at several different biological levels. Primary (thymus, bone marrow) and secondary (spleen, lymph nodes, gut-associated lymphoid tissues) lymphoid organs are innervated by the sympathetic nervous system, and lymphoid cells bear receptors for many hormones and neurotransmitters. These substances, secreted by the pituitary gland, are thus able to influence lymphocyte function. Moreover, lymphocytes themselves can produce neuropeptide substances. Cytokines produced by macrophages and activated lymphocytes (and by cells of the central nervous system) are critical elements in the cascade of immune responses to antigenic stimulation and also serve to energize the hypothalmic-pituitary-adrenal axis.

Thus, there are anatomical and neurochemical channels of communication that provide a structural foundation for the several observations of functional relationships between the nervous and immune systems. Lesions or electrical stimulation of the hypothalamus, for example, can alter antibodyand cell-mediated immune responses, and the eliciting of an immune response results in an increase in the firing rate of of neurons within the ventromedial hypothalamus at the time of peak antibody production. Changes in hormonal states can influence immunologic reactivity; and conversely, the immune response to antigenic challenges includes the release of cytokines which influence the neural regulation of psychophysiological processes, and is also associated with changes in circulating levels of hormones and neurotransmitter substances.

Stress and immunity. The link between behavior and immune function is suggested by experimental and clinical observations of a relationship between psychosocial factors, including stress, and susceptibility to or progression of disease processes that involve immunologic mechanisms. There are abundant data documenting an association between stressful life experiences and changes in immunologic reactivity. The death of a family member, for example, is rated highly on scales of stressful life events and, depending on gender and age, is associated with depression and an increased morbidity and mortality. Bereavement and depression are also associated with changes in some features of immunologic reactivity such as reduced lymphoproliferative responses [a general measure of the physiological status of T and B (bone marrow-derived) lymphocytes] and impaired activity of natural killer cells, which are lymphocytes capable of destroying cancer and virally infected cells without having had prior contact with the foreign material. Changes in immunity are also associated with the affective responses to other traumas such as marital separation and divorce. Other, less severe, naturally occurring stressful experiences (such as taking examinations) result in transient impairments in several parameters of immune function. See DISEASE; STRESS (PSYCHOLOGY).

It should be emphasized, however, that the association between stressful life experiences and disease and the association between stressful life experiences and changes in immune function do not in themselves establish a causal link between stress, immune function, and disease. In animals, a variety of stressors can, under appropriate experimental circumstances, influence a variety of immune responses in a variety of species in a variety of ways. Stressful circumstances can also alter the host's defense mechanisms, allowing an otherwise inconsequential exposure to a pathogen to develop into clinical disease. Understanding of the interactions between neuroendocrine and immune function under normal and stressful conditions, however, is incomplete. Glucocorticoids secreted by the adrenal cortex, a common endocrine feature of the stress response, are, in general, immunosuppressive, and there are numerous examples of stress-induced, adrenocortically mediated changes in immunity. However, there are numerous other observations of stressinduced changes in immunity that are independent of adrenocortical activation. It is evident that the immunologic consequences of stressful experiences involve complex neural, endocrine, and immune response interactions. Since immune responses are themselves capable of altering levels of circulating hormones and neurotransmitters, these interactions probably include complex feedback and feedforward mechanisms as well. See ENDOCRINE MECHANISMS.

The direction, magnitude, and duration of stressinduced alterations of immunity are influenced by a number of factors: (1) the quality and quantity of stressful stimulation; (2) the capacity of the individual to cope effectively with stressful events; (3) the quality and quantity of immunogenic stimulation; (4) the temporal relationship between stressful stimulation and immunogenic stimulation; (5) sampling times and the particular aspect of immune function chosen for measurement; (6) the experiential history of the individual and the existing social and environmental conditions upon which stressful and immunogenic stimulation are superimposed; (7) a variety of host factors such as species, strain, age, sex, and nutritional state; and (8) interactions among these several variables.

Conditioning. Central nervous system involvement in the modulation of immunity is dramatically illustrated by the classical (Pavlovian) conditioning of the acquisition and extinction of suppressed and enhanced antibody- and cell-mediated immune responses. Using a one-trial taste-aversion conditioning situation, a distinctively flavored drinking solution (the conditioned stimulus) was paired with an injection of the immunosuppressive drug cyclophosphamide (the unconditioned stimulus). When subsequently immunized with sheep red blood cells, conditioned animals reexposed to the conditioned stimulus showed a reduced antibody response compared to nonconditioned animals and conditioned animals that were not reexposed to the conditioned stimulus. See CONDITIONED REFLEX.

The acquisition and the extinction (elimination of the conditioned response by exposures to the conditioned stimulus without the unconditioned stimulus) of the conditioned enhancement and suppression of both antibody- and cell-mediated immune responses-and nonimmunologically specific host defense responses as well-have been demonstrated under a variety of experimental conditions. For example, the immunological effects of stress have been conditioned, and conditioning effects using antigen itself as the unconditioned stimulus have been demonstrated. The hypothesis that conditioned alterations of immunity are merely a reflection of stress responses, notably, adrenocortical secretions, is not supported by research results. In keeping with the bidirectional nature of nervous and immune system interactions, it is also possible to condition the physiological effects elicited by the products of an activated immune system.

The biological impact of conditioned alterations in immunity is illustrated by experiments in which conditioning operations were applied in the pharmacotherapy of spontaneously developing systemic lupus erythematosus in New Zealand mice. In conditioned animals, substituting conditioned stimuli for active drug on some of the scheduled treatment days delays the onset of autoimmune disease using a cumulative amount of immunosuppressive drug that is ineffective by itself in altering the progression of disease. Similarly, reexposure to a conditioned stimulus previously paired with immunosuppressive drug treatment prolongs the survival of foreign tissue grafted onto mice. These dramatic results address the clinical implications of the behavioral component of research in psychoneuroimmunology, but they have yet to be experimentally verified in humans. *See* IM-MUNOSUPPRESSION.

Again, in keeping with the reciprocal nature of the relationship between neural and endocrine and immune responses, there are indications that immune status influences behavior. For example, emotional and cognitive changes are associated with lupus in humans, and there are changes in the behavior of animals that accompany the progression of their autoimmune disease. *See* AUTOIMMUNITY.

Prospects. The mechanisms underlying the functional relationships between the nervous system and the immune system illustrated by conditioned and stressor-induced modulations of immune functions cannot yet be fully described. It is assumed that different conditioning and stressful experiences induce different patterns of neuroendocrine changes that define the milieu within which immunologic reactions occur. This milieu is influenced by neural and endocrine signals to the immune system and signals from the immune system that initiate further neural and endocrine changes-and by regulatory feedback loops between as well as within these "systems." An elaboration of the integrative nature of neural, endocrine, and immune processes and the mechanisms underlying behaviorally induced alterations of immune function is likely to have important clinical and therapeutic implications that will not be fully appreciated until more is known about the extent of these interrelationships in normal and pathophysiological states. See ENDOCRINE SYSTEM (VERTEBRATE); IMMUNOLOGY; NERVOUS SYSTEM (VER-Robert Ader TEBRATE)

Bibliography. R. Ader and N. Cohen, Psychoneuroimmunology: Conditioning and stress, *Annu. Rev. Psychol.*, 44:42-85, 1993; R. Ader, D. L. Felten, and N. Cohen, Interactions between the brain and the immune system. *Annu. Rev. Pharmacol. Toxicol.*, 30:561-602, 1990; R. Ader, D. L. Felten, and N. Cohen (eds.), *Psychoneuroimmunology*, 3d ed., 2000.

Psychopharmacology

A discipline that merges the subject matter of psychology, which studies cognition, emotion, and behavior, and pharmacology, which characterizes different drugs. Thus, psychopharmacology focuses on characterizing drugs that affect thinking, feeling, and action. In addition to this general concern, psychopharmacology places particular emphasis on those drugs that affect abnormalities in thought, affect, and behavior, and thus has a relationship to psychiatry.

Psychopharmacology is predominantly, but not exclusively, concerned with four major classes of drugs that are of clinical significance in controlling four major categories of psychiatric disorder: anxiety, depression, mania, and schizophrenia.

Anxiety. Anxiety is an emotional state that can range in intensity from mild apprehension and nervousness to intense fear and even terror. It has been estimated that 2-4% of the general population suffer from an anxiety disorder at some time. Although anxiety in some form is a common experience, it can become so intense and pervasive as to be debilitating; it may therefore require psychiatric attention and treatment with an anxiolytic drug. There are three major groups of anxiolytics. Members of the first group are called propanediols; meprobamate is the most widely used. The second group is the barbiturates, of which phenobarbital is the most generally prescribed. The third group, most frequently prescribed, is the benzodiazepines, the best known of which is diazepam.

All anxiolytics have, in varying degrees, anticonvulsant and muscle relaxant properties. However, the most clinically relevant features of these drugs are two side effects characteristic of them all-sedation and toxicity. Sedation refers to the drowsiness and minor motor incoordination that all anxiolytics can induce. The propanediols and barbiturates are relatively more sedating than the benzodiazepines. The benzodiazepines are also less toxic in that fatal overdoses are much less likely to occur with them alone. However, it is important to recognize that although suicide attempts with a benzodiazepine alone do occur, such attempts much more often involve a benzodiazepine and another drug such as alcohol or a barbiturate. Benzodiazepines can significantly increase the suicide potential because they potentiate these other drugs.

A major advance in understanding the benzodiazepines was the identification of the cellular sites at which these drugs act (so-called benzodiazepine receptors). The distribution of these receptors in the brain has also been found to have a striking parallel to the distribution of the receptors for a naturally occurring substance called gamma-amino butyric acid (GABA). Furthermore, it is known that GABA has a ubiquitous inhibitory role in modulating brain function. Most importantly, it is now clear that benzodiazepines share a biochemical property in that all augment the activity of GABA.

These findings imply that the augmentation of GABA activity could be the common denominator for all forms of anxiety reduction. Some drugs, such as buspirone, however, do not augment GABA activity but do reduce anxiety, while other drugs, including valproic acid, augment GABA activity but do not reduce anxiety. Thus, although it is likely that augmented GABA activity is the mechanism that mediates the muscle relaxant and anticonvulsant actions of the benzodiazepines themselves, it is not clear that such augmentation is critical in the reduction of anxiety that benzodiazepines also produce. Indeed, other lines of evidence favor a critical role for another endogenous brain substance (5-hydroxytryptophan or its decarboxylated derivative, serotonin) in mediating the effects of the benzodiazepines on anxiety.

Attempts at elucidating the mechanisms underlying anxiety reduction have also been complicated by the finding that the extreme forms of anxiety that manifest themselves as panic and phobic states can be controlled by members of another, completely separate group of drugs called antidepressants. *See* ANX-IETY DISORDERS; SEROTONIN; TRANQUILIZER.

Depression. The symptoms of depression can include a sense of sadness, hopelessness, despair, and irritability, as well as suicidal thoughts and attempts, which are sometimes successful. In addition, physical symptoms such as loss of appetite, sleep disturbances, and psychomotor agitation (restlessness, hand wringing, pacing, and so on) are often associated with depression. When depression becomes so pervasive and intense that normal functioning is impaired, antidepressant medication may be indicated. It has been estimated that as much as 6% of the population will require antidepressant medication at some time in their lives.

There are two major groups of antidepressant drugs. Members of the first group are called heterocyclics because of their characteristic chemical structures. Members of the second group, which are less often prescribed, are called monoamine oxidase inhibitors because one of their biochemical effects is to inhibit an enzyme called monoamine oxidase. This action does not, however, directly account for the clinical efficacy of these drugs.

The antidepressants typically require at least several weeks of chronic administration before they become effective in alleviating depression. This contrasts with the anxiolytics, which are effective in reducing anxiety in hours and even minutes. Another difference between these two classes of drugs is that the anxiolytics are more likely to be efficacious: anxiolytics are effective in the vast majority of nonphobic, anxious patients, whereas the antidepressants are effective in only about 65–70% of depressed patients. *See* AFFECTIVE DISORDERS.

Historically, the activity of the antidepressants has been linked to the activity of norepinephrine, among other naturally occurring substances. Detailed analyses of this hypothetical linkage were made possible because of two developments: first, the cellular sites at which norepinephrine acts have been isolated; and second, because these receptors can be identified, it has been possible to map out an anatomically defined brain system dependent on norepinephrine for its activity. Furthermore, this system can be supposed to be involved in mediating the signs of depression. Chronic administration of most antidepressants decreases the sensitivity of some norepinephrinedependent cells because of a reduction in the number of one type of norepinephrine receptors. However, not all efficacious antidepressants reduce the number of these receptors, and other drugs reduce norepinephrine activity but are not antidepressants. See NORADRENERGIC SYSTEM.

Mania. Manic episodes are characterized by hyperactivity, grandiosity, flight of ideas, and belligerence; affected patients appear to be euphoric, have racing thoughts, delusions of grandeur, and poor if not self-destructive judgment. Periods of depression follow these episodes of mania in the majority of patients. The cycles of this bipolar disorder are typically interspersed among periods of normality that are, in most cases, relatively protracted.

Mania can usually be managed by chronic treatment with lithium and can be expected to be effective in 70–80% of the individuals treated. Furthermore, the period of depression that typically follows the manic episode can usually be prevented, or at least attenuated, if lithium treatment is maintained after the manic phase has subsided. Thus, lithium not only controls the period of mania itself but has a prophylactic effect on the subsequent depression. Any periods of depression that do occur can be managed by antidepressant drugs. Lithium is no longer the only drug used in the management of mania. Carbamazepine, an anticonvulsant that is used in the treatment of epilepsy, is also useful in the treatment of periods of mania.

Schizophrenia. Schizophrenia is a form of psychosis; it incorporates a broad range of symptoms that can include bizarre delusions, hallucinations, incoherence of thought processes, inappropriate affect, and grossly disorganized movements. It affects 1–2% of the population.

The symptoms of schizophrenia can be controlled, in varying degrees, by a large group of drugs called antipsychotics. Symptom management requires chronic medication and can be expected in about 80% or more of the schizophrenics treated. However, management is only partially successful in that normal functioning is not completely restored in most patients.

The antipsychotics have a broad range of side effects among which are disturbances of movement that fall into two general classes. The first class includes a loss of muscle tone, motor restlessness, and an array of symptoms very like those characteristic of Parkinson's disease. These disturbances can generally be managed by adjunctive medication with other drugs. Such control is not yet possible for the second class of movement disorder called tardive dyskinesia. Signs of this disturbance typically include involuntary movements that most often affect the tongue and facial and neck muscles but can also include the digits and trunk.

Tardive dyskinesia can develop in the course of chronic treatment with antipsychotics; as a rough rule of thumb, about 25% of those treated will show signs of the disorder. Tardive dyskinesia is relatively permanent once developed. This feature, as well as its refractoriness to treatment, places serious limits on the use of the antipsychotic drugs.

Although different antipsychotic drugs have different kinds and degrees of side effects, all share a single biochemical action: they all attenuate the activity of dopamine, a naturally occurring substance in the brain. The reduction in dopamine activity produced by the antipsychotics directly accounts for their effects on motor behavior. That is, dopamine is known to be present in high concentrations in the nigrostriatal system, an anatomically defined system of the brain that is intimately involved in the regulation of movement. It is to be expected, therefore, that disrupted dopamine activity in this system would produce disturbances of movement as, in fact, the antipsychotic drugs do. It is less clear, however, whether reduced dopamine function is also a factor in the process by which these drugs control psychotic (including schizophrenic) symptoms. *See* SCHIZOPHRENIA.

Laboratory analysis of physiological mechanisms. The mechanisms by which different drugs exert their effects on clinically significant abnormalities in behavior remain to be determined. Four major factors have impeded the kinds of analyses required: (1) The multiplicity of variables that control behavior typically cannot be simply measured. (2) Each of these variables can interact with others so that the behavioral consequences of one interaction are not the same as the consequences of another interaction. (3) It may be that different mechanisms control similar behavior so that patients, all labeled as schizophrenic, for example, may share many behaviors that are actually mediated by different underlying processes. (4) Much of the research cannot be undertaken in humans, for ethical reasons, and therefore laboratory models of various clinical entities must be developed in infrahuman species.

The development of laboratory models is itself a formidable tasks, because the disorders that must be mimicked are, in contrast to most other disorders, uniquely human. Unlike a disorder such as diabetes, there are no direct analogs of mania, depression, or schizophrenia in laboratory animals. While anxiety can be produced in the laboratory, the extent to which such anxiety truly mirrors anxiety in humans is uncertain, because the measurement of anxiety in humans is based on their verbal report.

These problems can be partially circumvented by showing that drug-induced behavioral change in the laboratory has properties parallel to those seen in the clinic. For instance, if chronic, but not acute, administration of a drug produces clinical change, then chronic, but not acute, administration of that drug should produce behavioral change in the laboratory. In addition, if clinical change is selectively induced by one group of drugs, then the laboratory behavior should also be selectively sensitive to that one group but not to other groups of drugs. For example, behavior in a laboratory animal that changes only after chronic, not acute, treatment with both lithium and carbamazepine-but not other drugs-could have promise as a model of mania. As more and more of the characteristics of clinical change are paralleled in laboratory behavior, the greater are the chances that such laboratory behavior will serve as a useful baseline for the elucidation of underlying physiological mechanism. See NEUROBIOLOGY. Peter L. Carlton

Bibliography. R. J. Baldessarini, *Chemotherapy in Psychiatry: Principles and Practice*, rev. ed., 1990; P. L. Carlton, *A Primer of Behavioral Pharmacology: Concepts and Principles in the Behavioral Analysis of Drug Action*, 1983; S. G. Dahl and L. F. Gram (eds.), *Clinical Pharmacology in Psychiatry*, 1989; L. F. Hollister and J. Csernansky, *Clinical* *Pharmacology of Psychotherapeutic Drugs*, 3d ed., 1990; P. G. Janicak, *Principles and Practices of Clinical Psychopharmacotherapy*, 1993; R. Spiegel, *Psychopharmacology: An Introduction*, 3d ed., 1996.

Psychophysical methods

Methods for the quantitative study of the relations between physical stimulus magnitudes and the corresponding magnitudes of sensation, for example, between the physical intensity of a light and its perceived brightness or the concentration of a sugar solution and its observed sweetness. To establish these relations measurement scales are needed, not only for physical magnitudes but also for subjective magnitudes. Subjective scales are, of course, not obtained directly from observation. They are theoretical models which summarize observed relations between stimuli and responses. *See* SENSATION.

The original three psychophysical methods, first systematically worked out and published by G. Fechner in 1860, were designed to yield the data from which to construct subjective scales. Other techniques, more direct than Fechner's, are superseding his approach to psychophysical problems; his methods and variants of them are still in use, however.

The term psychophysical methods is sometimes extended to include certain scaling techniques which are most often used with, although not limited to, subjective dimensions to which there correspond no simple physical dimensions, for example, food preferences. These techniques have in common the fact that the rationale for scaling involves the amount of dispersion of repeated or multiple judgments. *See* MULTIDIMENSIONAL SCALING.

This article surveys the principal types of methods employed in relating subjective magnitudes to physical stimulus magnitudes. The first group of methods are for measuring the least stimulus, or least stimulus change, that is perceptible; the second group of methods are for establishing quantitative relationships between readily perceptible stimuli and their corresponding sensations.

Threshold methods. Two of the three psychophysical methods developed by Fechner were designed to measure the absolute threshold, defined as the minimum stimulus energy that an organism can detect, and the differential threshold, defined as the minimum detectable change in a stimulus. Both quantities had to be defined as statistical averages, because their repeated measurement under externally constant conditions yielded moment-to-moment fluctuations in the values obtained. To obtain reliable measurements for these averages, Fechner devised the method of limits (also called the method of minimal changes) and the method of constant stimuli. It is necessary in scaling to establish the units of the scale and the zero point. Fechner used the difference threshold as the unit and the absolute threshold as the zero point in scaling the magnitude of subjective sensations.

Method of limits. To measure the absolute threshold by this method, the experimenter begins with a stimulus which is too weak for the subject to detect. In successive presentations, the stimulus intensity is increased in small, equal steps, the subject reporting after each presentation whether the stimulus was perceived. The end of this ascending series is reached when the subject reports that the stimulus has been detected. The descending series is then begun, the stimulus intensity beginning at an abovethreshold value and decreasing in steps until the subject signals the disappearance of the stimulus. Many such series are given. To help the subject avoid basing judgments on the number of steps occurring in a series, the initial stimulus intensity of each series is varied on successive repetitions. Frequently, the average threshold value that is obtained from ascending series differs from that obtained from descending series; the usual practice is to average the two.

In measuring the difference threshold, essentially the same procedure is involved, except that the subject now signals the relation of a comparison stimulus to a standard stimulus. The comparison stimulus begins at a value clearly less than that of the standard, and increases in steps until the subject reports the apparent equality of the two. This gives a lower difference threshold. The comparison stimulus is further increased until the subject first reports it as greater than the standard. That is the upper difference threshold. The comparable procedure is followed in a descending series, which yields a descending upper and lower threshold. After a large number of such trials, the average of each of these four threshold values is computed. The differences of each of these four averages from the value of the standard stimulus gives four values for the difference threshold. Usually, these four values are averaged.

The most common modifications of the method of limits involve the omission of descending series in order to minimize sensory adaptation, and the use of continuous rather than discrete stimulus variation. Another procedural modification is the up-and-down method. The stimulus intensity (or intensity difference) is decreased on the following trial if the subject has detected it, increased on the following trial if the subject has not. It is a more efficient method because it concentrates the testing in the region of the threshold, although there is the disadvantage that the level of the next following stimulation is predictable by the subject. It has proved notably useful in its application with the Békésy audiometer for the testing of human hearing and in the study of the visual sensitivity of animals.

Method of constant stimuli. To measure the absolute threshold, the experimenter selects a small number of stimulus values in the neighborhood of the absolute threshold (previously roughly located by informal use of the method of limits) and presents them to the subject a large number of times each, in an irregular order unknown to the subject. Each time a stimulus is presented, the subject reports the presence or absence of sensation. At the end of the experiment the data provide the proportion of times that each stimulus resulted in a report of sensation by the subject. By interpolation one can estimate the stimulus value that has a probability of .50 of producing sensation, this value being defined as the absolute threshold.

An analogous procedure is followed in obtaining difference thresholds. One finds, for each of a set of comparison stimulus values, the proportion of stimulus presentations in which each of the values is reported to be greater than that of a standard stimulus. By interpolation one can estimate the stimulus value that has a .75 probability of being reported to be greater than the standard stimulus value. The difference between these two stimulus values has conventionally been defined as the difference threshold.

The method of constant stimuli, although a much lengthier procedure than the method of limits, has the advantage that the sequence of stimulus levels is not predictable by the subject. The lengthiness of the procedures is roughly comparable to that required by methods based on the theory of signal detection.

Theory of signal detection. Both the method of limits and the method of constant stimuli have a serious limitation. The threshold values obtained by their use are unavoidably affected by the criterion that a subject adopts in deciding when to report that a stimulus or a stimulus change has been perceived. The discriminations the subject is asked to make are difficult, and a subject who decides not to report that a stimulus or a stimulus change has been detected, until he or she is quite sure of it, will appear to have larger thresholds than would be obtained by being less conservative. Furthermore, as long as the subject is fairly consistent in criterion, there is no way to measure the extent to which it has affected the threshold values obtained. This difficulty is surmounted by the theory of signal detection, an approach to methods of measuring sensory discrimination that dispenses entirely with the concept of threshold.

From the following example of this theory it can be seen how it is possible to assess separately the detectability of a stimulus and the subject's criterion for reporting that it is detected.

Imagine a series of trials, on each of which a subject is to listen for a tone so weak that he or she can never be sure that it has been heard. The subject knows that on some trials the tone will not be sounded and that on others it will be, and after each trial he or she must say whether or not it may have been heard. On some trials the subject will be correct when saying that the tone was or was not present, but there will be some trials on which he or she says the tone was present when it was not, and some trials on which he or she says that the tone was not present when it was. Where there have been enough trials to yield reliable proportions of these kinds of trial outcome, it is possible to infer independently the detectability of the tone and the strictness of the subject's criterion in deciding whether or not the tone has been sounded.

For example, if the tone is of equal detectability for two subjects, but one employs a stricter criterion than the other, the subject with the stricter criterion will less often say that the tone was present when it was not, but will more often say that the tone was not present when it was. Thus, differences in criterion produce an inverse relationship between these two kinds of error. Differences in detectability, on the other hand, produce differences in the total proportion of correct responses. Because detectability and criterion have different effects, it is possible to use the proportions of the subject's responses on trials when a stimulus was present and on trials when a stimulus was absent to assess independently the detectability of the stimulus and the strictness of the subject's criterion. For a more detailed explanation *See* SIGNAL DETECTION THEORY.

It should be noted that the concept of a sensory threshold for a stimulus is not involved, only the detectability of a stimulus as measured by the proportion of correct reports of the presence or absence of the stimulus. This has led a number of investigators to deny that there exists such a thing as a sensory threshold, since this theory has no need of such a concept. Because both detectability and sensory threshold are statistical definitions, the question is not which definition is more valid, but which is more useful. The less laborious method of limits, with its associated concept of the sensory threshold, is often used when a high degree of precision is not required; the methods of the theory of signal detection have been coming into more and more use for accurate work when lengthier procedures are not a handicap

Subjective metric methods. Fechner proposed to use the results of threshold measurement in developing a subjective metric or scale. He defined the difference threshold, or just noticeable difference (jnd), as the subjective unit and the absolute threshold as the zero point of the subjective scale. Thus the subjective intensity of a particular brightness of light, for example, would be specified when it was given as 100 jnds above threshold. The subjective scale so defined is not a linear function of the physical stimulus scale since jnds, though defined as subjectively equal units, are not of physically equal magnitude throughout the intensity scale. The size of the jnd is approximately proportional to physical stimulus intensity. To the extent that this relation holds, Fechner deduced that subjective intensity should be proportional to the logarithm of the stimulus intensity.

There has been considerable debate over the merits of this formulation and the assumptions involved. The chief modern objection is an empirical one, that the results obtained by Fechner's methods are different from those resulting from other, more direct techniques. A tone 40 jnds above the threshold does not sound twice as loud as a tone 20 jnds above threshold. Furthermore, the employment of the direct ratio discrimination methods that is outlined below suggests that subjective magnitude is a power function of stimulus intensity, rather than the logarithmic function that is required by Fechner's postulates. However, these same methods indicate that on at least some nonintensive dimensions, such as length, jnds may be subjectively equal.

Direct-ratio discrimination methods. Rather than requiring of the subject merely either yes-no or ordinal judgments, these methods require the subject to make direct-ratio discriminations. For instance, he or she may be presented with a moderately loud tone, and then required, by turning a knob, to adjust the loudness of a comparison tone until it is half as loud, or twice as loud, as the first. The first case illustrates the method of fractionation, the second the method of multiplication. In the method of magnitude estimation, the subject is given a stimulus, such as the brightness of a light, to serve as a modulus with a value assigned to it, for example, 10. The task, as other lights of different intensities are presented, is to assign them numbers which shall stand in the same ratio to 10 as their brightness stands to that of the modulus. One twice as bright is given the designation 20; one half as bright is 5. In these and other similar methods, whether the subject's task is to estimate or to produce the prescribed ratio or the prescribed fraction, there are certain common characteristics. Direct-ratio assessments are obtained from the subject; there can be experimental checks on internal consistency of the results, and since the individual judgments are not of high precision, repetition is required if stable averages are to be obtained.

The empirical results obtained by the various methods are in fairly good agreement. They agree in that, to at least a first approximation, subjective magnitudes on a variety of dimensions are found to be power functions of suprathreshold stimulus intensity; that is to say, subjective magnitude is proportional to the suprathreshold stimulus magnitude raised to a power. The powers have a range from 0.3 for auditory loudness to 3.5 for subjective intensity of alternating current that is applied to the skin.

There is not yet general agreement on whether these power functions indicate a fundamental relationship between physical stimulus intensities and corresponding magnitudes of sensation, or whether they also reflect people's response characteristics when they are asked to gage their sensations. Although recordings of electrical activity in sensory nerve fibers sometimes appear to support the former interpretation, until more is learned about how the brain processes these incoming signals, this line of evidence is inconclusive.

Ratio methods are applicable to nonintensitive, or qualitative, attributes, such as auditory pitch and hue of colors, and their application is not necessarily limited to stimuli from simple physical dimensions. For example, the methods have been used to develop a subjective scale of the seriousness of various crimes.

Direct-matching methods. The subject is not required to produce or assess the ratio of one subjective magnitude to another, but only to adjust a comparison stimulus until some attribute appears to match that of a standard stimulus. For example, the subject might be asked to adjust the physical intensities of tones of various frequencies until their loudness matched that of a 1000-Hz tone of fixed intensity. The result would

be an equal-loudness contour, showing the intensities to which tones of various frequencies must be set to produce sensations of equal loudness. These data are of use in acoustics. *See* LOUDNESS.

Method of average error. This method, one of the three methods devised by Fechner, is a special application of direct-matching methods to cases in which the point of interest is in discrepancies between perception and stimulation. The subject adjusts a comparison stimulus to match a standard stimulus; the average of a number of such settings gives the point of subjective equality, and the difference between this point and the standard stimulus is the average error. Two illustrative uses of the method are the measurement of accuracy of distance perception and the measurement of the magnitude of so-called optical illusions. *See* HEARING (HUMAN); PSYCHOLOGY. John F. Hahn

Bibliography. L. R. Aiken, *Psychological Testing* and Assessment, 10th ed., 1999; N. Arthur and R. R. Fay (eds.), *Human Psychophysics*, 1993; J. R. Barclay, *Psychological Assessment: A Theory Sys*tems Approach, 1991.

Psychosis

Any disorder of higher mental processes of such severity that judgments pertaining to the reality of external events are significantly impaired. A wide range of psychiatric, medical, and neurological conditions can bring about a psychotic state. They include schizophrenia, mania, depression, ingestion of drugs, withdrawal from drugs, liver or kidney failure, endocrine disorders, metabolic disorders, and Alzheimer's disease, epilepsy, and other neurologic dysfunctions. Not all psychotic states are caused by disease or drugs. The dreams of normal sleep are a form of psychosis; while dreaming, a reality can be seen and heard which, at the moment, is not known to be false but which is roused from upon awakening.

Psychotic alterations of beliefs are called delusions. For manic patients, a typical delusion is that they possess omnipotent powers; for schizophrenics, it is often a belief that their minds are controlled by external forces or that they can project their thoughts into other people's minds. Patients with schizophrenia, Alzheimer's disease, or drug-induced psychosis may have paranoid delusions, such as a belief that other people or evil forces are attempting to do them harm. Psychotic alterations of perception are referred to as hallucinations. Addicts who are undergoing withdrawal from drugs may suffer from tactile hallucinations, for example, a sensation that bugs are crawling on them. Schizophrenics often hear voices of persons who are not actually present. A depressed person may "see" the silhouette of a lost loved one. A person with epilepsy may hallucinate certain smells. Psychotic states that are due to alcoholism, metabolic diseases, or other medical conditions are frequently accompanied by general mental confusion. On the other hand, psychiatric

illnesses and drugs, such as lysergic acid diethylamide (LSD) and phencyclidine (PCP), can produce hallucinations and delusions in the absence of general confusion. *See* PSYCHOTOMIMETIC DRUGS.

Patients are often very distressed by their psychotic symptoms. Unfortunately, very few of those symptoms are unique to a particular illness, which can make proper diagnosis difficult and challenging. Correct diagnosis, however, is critical so that appropriate treatment can be provided. *See* ADDICTIVE DIS-ORDERS; AFFECTIVE DISORDERS; ALZHEIMER'S DISEASE; PARANOIA; SCHIZOPHRENIA. Ralph E. Hoffman

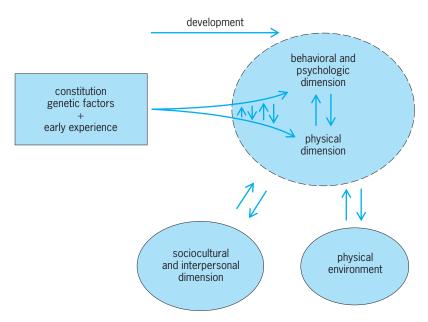
Bibliography. P. Buckley (ed.), Essential Papers on Psychosis, 1988; R. Marcus, Psychosis and Near Psychosis: Ego Function, Symbol, Structure, and Treatment, 1993; R. E. Hoffman and T. H. McGlashan, Speech, thought, and perceptual disorders, in A. Tasman, J. Kay, and J. A. Lieberman (eds.), Psychiatry, vol. II, Saunders, 1996.

Psychosomatic disorders

Disorders characterized by physical symptoms that originate, at least in part, from emotional factors. The *Diagnostic and Statistical Manual of the American Psychiatric Association* (DSM) *IV* now assigns such conditions to two diagnostic categories: somatoform disorders and psychological factors affecting physical condition. Somatoform disorders are characterized by physical symptoms that are suggestive but not fully explained by a diagnosable medical disease. The second category, psychological factors affecting physical condition, refers to the presence of a general medical condition that is negatively affected (for example, prolonged or exacerbated) by psychological factors.

The fact that emotional factors play a role in the pathogenesis of physical disease was known as early as the time of Hippocrates. Sigmund Freud demonstrated that physical symptoms such as paralysis and paresthesia could be caused by emotional conflicts (conversion disorder or hysterical conversion, now known as a type of somatoform disorder). F. Alexander distinguished between conversion disorder and psychophysiologic disorders, now considered to be psychological factors affecting physical condition. His specificity theory also proposed that specific psychological conflicts caused specific medical diseases, such as peptic ulcer and hypertension, that he considered to be "psychosomatic."

Evolution of concept. The explosive developments in the fields of molecular biology, genetics, and computer-assisted large-scale epidemiologic studies fueled an expansion of the concept of psychosomatic medicine during the past several decades. Stress research, in humans and animals, has shown that general stress (for example, common life changes such as marriage and divorce) as well as specific stresses (such as bereavement and combat) can have a profound impact on a number of organ systems (including the immune system, the cardiovascular system, and the brain) and the behavior of the organism, in-



Comprehensive model for all diseases. (After H. Leigh and M. F. Reiser, Major trends in psychosomatic medicine, Ann. Intern. Med., 87:233–239, 1977)

cluding its coping styles. It has been shown that emotional factors, including stress, may play an important role in diseases such as cancer and various infections which previously had not been considered psychosomatic. On the other hand, the role of genetic factors and nonspecific stress is now considered to be as great as that of specific psychological conflicts. The development of a disease state (pathogenesis) must be considered in terms of genetics, physical and psychological development, and the interaction of the individual with the physical and social environment. Thus, the model of psychosomatic medicine is now an interactional systems model that encompasses all disease (see **illus.**).

Developments in medical sociology and psychology have also elucidated the role of psychological and psychosocial factors not only in pathogenesis but also in help-seeking behavior, the course of illness, compliance to a medical regimen, and convalescence from practically all diseases. It has been determined, for example, that stress and anxiety tend to increase help-seeking behavior and that depression tends to worsen the prognosis of any procedure and delay recovery from disease. In modern medicine, the psychosomatic approach involves investigating the interaction of biological, psychological, and social factors in each person, regardless of the diagnosis. This integrative approach is referred to as the biopsychosocial model for the patient and disease.

Psychological factors affecting physical condition. This category encompasses disorders characterized by the existence of a general medical condition that is negatively affected by psychological factors, such as stress. For example, stress can delay recovery from or exacerbate the symptoms of a general medical condition or even cause additional health problems. There is nothing mysterious about the idea that the mind changes body states: voluntary movements are certainly psychosomatic. For example, when one thinks about sour apples, salivation occurs as a classically conditioned response. Psychological states influence body organs through a combination of three interrelated mechanisms: neural, hormonal, and immunologic. Voluntary movements (for example, clenching the teeth) are mediated through the motor neurons by the conscious command of the brain. In stress, clenching of the teeth, mediated by the same motor neurons, may also occur, but the act may not be voluntary and conscious. Stress usually causes an activation of the sympathetic nervous system and the hypothalamo-pituitary-adrenal axis followed by a decrease in immunocompetence. Immune mechanisms may be suppressed in part through corticosteroid activation, but a decrease in T-lymphocyte activity in stress (for example, following bereavement) may not be mediated by hormones.

The voluntary nervous system as well as the autonomic nervous system are capable of learning through operant conditioning (reward and punishment), and the immune system can be classically conditioned. Individual specific, but inadvertent, conditioning of specific conflict or stress to specific bodily malfunction may be an important psychosomatic mechanism. *See* CONDITIONED REFLEX; NEUROIMMUNOLOGY; STRESS (PSYCHOLOGY).

Somatoform disorders. Included in this category are disorders whose symptoms are suggestive of but not fully explained by a physical disease. Somatoform disorders differ from factitious disorders and malingering in that the symptoms are not under intentional (voluntary) control. There are six types of somatoform disorders: conversion disorder, somatization disorder, pain disorder, hypochondriasis, body dysmorphic disorder, and undifferentiated somatoform disorder.

Conversion disorder. Conversion disorders are characterized by physical symptoms referrable to the somatosensory nervous system or special sensory organs that cannot be explained on the basis of a medical or neurologic disease and that are caused by psychological factors. Common symptoms include paralysis, blindness, ataxia (lack of muscular coordination, especially in walking), aphonia (loss of voice and power of speech), and numbness of the feet (stocking anesthesia). The symptoms may represent a psychological conflict or may be a form of body-language communication.

Somatization disorder. In somatization disorder (also known as Briquet's syndrome), the patient recurrently complains of multiple somatic symptoms that are referrable to practically every organ system in the body. Upon medical investigation, however, the symptoms turn out not to be a diagnosable physical disease. This chronic, fluctuating disorder begins before age 30 and has a familial tendency; both genetic and environmental factors seem to play a role. The chronicity and multiplicity of symptoms in somatization disorder distinguish it from conversion disorder. The symptoms do not usually symbolize psychological conflicts but may represent general dysphoria and distorted illness behavior, such as a psychological addiction to the sick role. There is no definitive treatment; patients should be managed by one physician who coordinates all diagnostic and treatment plans and provides ongoing support and follow-up without unnecessary invasive procedures.

Pain disorder. When pain is the prominent clinical symptom and psychological factors are considered to play an important role in the onset, severity, exacerbation, or maintenance of the pain, pain disorder may be diagnosed. If pain is part of somatization disorder, depression, or another major psychiatric disorder, a separate diagnosis is not appropriate. *See* PAIN.

Hypochondriasis. Hypochondriasis is preoccupation with the idea of having a serious medical disease based on a misinterpretation of one or more bodily symptoms or signs. Such preoccupation continues in spite of a negative reasonable medical workup, and causes significant distress or impairment in functioning. There may be an amplification of somatic sensations in hypochondriasis. Hypochondriasis is not diagnosed if another major psychiatric disorder, such as anxiety disorder or major depression, accounts for the preoccupation.

Body dysmorphic disorder. Otherwise known as dysmorphophobia, this condition denotes a preoccupation with a defect in appearance that may be either imagined or greatly exaggerated, and that causes significant distress or impairment in function. If such preoccupation occurs in the presence of another major psychiatric disorder, such as schizophrenia or anorexia nervosa, it is subsumed under that diagnosis.

Therapeutic approaches. The management of psychological factors affecting physical condition should exemplify the biopsychosocial model, that is, effective treatment of the physical condition as well as psychotherapeutic and psychopharmacologic interventions as needed to ameliorate the psychological factors, including stress management and relaxation techniques.

The treatment of choice for conversion disorder is psychotherapy designed to provide understanding and resolution of the psychological issues underlying the condition. Physical or speech therapy may be indicated to facilitate recovery of function. Supportive psychotherapy and judicious psychopharmacologic treatment may be indicated for pain disorder, hypochondriasis, and body dysmorphic disorder. Antidepressants may be effective in treating pain disorder, and supportive psychotherapy, relaxation techniques, and antianxiety medications may be useful for patients with various somatoform disorders. *See* PSYCHOPHARMACOLOGY; PSYCHOTHERAPY. Hoyle Leigh

Bibliography. R. Ader, D. Felton, and N. Cohen (eds.), *Psychoneuroimmunology*, Academic Press, San Diego, 2001; F. Alexander, *Psychosomatic Medicine: Its Principles and Applications*, W. W. Norton, New York, 1987; *Diagnostic and Statistical Manual Of Mental Disorders*, 4th ed. (DSM IV), American Psychiatric Association, 1994; H. Leigh (ed.), *Biopsychosocial Approaches in Primary Care: State of the Art and Challenges for the 21st Century*, Plenum Press, New York, 1997; H. Leigh and M. F. Reiser, Major trends in psychosomatic medicine, *Ann. Intern. Med.*, 87:233-239, 1977; H. Leigh and M. F. Reiser, *The Patient: Biological, Psychological, and Social Dimensions of Medical Practice*, 3d ed., Plenum Medical Book, New York, 1992.

Psychotherapy

Any treatment or therapy that is primarily psychological in nature. In recent years, counseling also has been included in this categorization. The number of different forms of psychotherapy has increased tremendously, particularly in the past three or four decades. Once the main forms of psychotherapy were Freud's psychoanalysis and those of his former disciples. Now there are at least 400 different forms of psychotherapy.

Types. In addition to designating psychotherapies in terms of their theoretical or procedural differences, they can be classified in other ways—for example, individual or group therapy, marital therapy, adult or child therapy, crisis intervention, timelimited or brief therapy, long-term or time-unlimited therapy, and manualized therapy. Thus, one must be cautious when discussing psychotherapy generally.

Psychodynamic therapies. Historically, psychoanalysis has played an important role in the growth and development of psychotherapy. It was created by Sigmund Freud, who developed his views over a period of years. Central to his theories was the importance of unconscious conflicts in producing the symptoms and defenses of the patient. Thus, the goal of therapy is to gradually overcome the patient's defenses and to help the patient attain insight into the repressed conflicts which are the source of difficulty. Since patients resist these attempts to uncover and bring to consciousness the painful repressed material, therapy must proceed slowly. Consequently, psychoanalysis is a long-term therapy requiring several years for completion and almost daily visits.

Since Freud's time, there have been important modifications associated with former disciples such as Alfred Adler and Carl Jung, as well as by a number of others in more recent decades. Self psychology and ego psychology are among more recent emphases. However, the popularity of psychoanalysis has waned for a number of reasons, including cost, length of treatment, selectivity of patients, and evidence of effectiveness. Still, several briefer forms of psychodynamic therapy derived from psychoanalysis have emerged and are clearly in tune with recent emphases on brevity, cost, and health maintenance organizations. *See* PSYCHOANALYSIS.

Experiential therapies. A number of related therapies are included in this group. Probably best known was the patient-centered therapy of Carl Rogers appearing in the 1940s. Like many of the therapies that have appeared since then, it was developed because

of dissatisfaction with analytic therapy. In Rogers' therapy, a major emphasis is placed on the ability of the patient to change when the therapist is empathic and genuine and conveys nonpossessive warmth. The therapist is nondirective in the interaction with the patient and attempts to facilitate the growth potential of the patient. Rogers also was a pioneer in using recorded therapy sessions for training and research.

Other therapeutic approaches considered as experiential include the Gestalt therapy of Perls, existential approaches, and transpersonal approaches. The facilitation of experiencing is emphasized as the basic therapeutic task, and the therapeutic relationship is viewed as a significant potentially curative factor. The uniqueness of the patient and his or her potential for personality growth is also emphasized.

Cognitive, behavioral, and interpersonal therapies. Another group of therapies that have received increased attention in recent years are the behavioral, cognitive-behavioral, and interpersonal therapies. Behavioral therapies, which developed from behaviorism in the field of psychology, emphasized theories and procedures that differ markedly from psychodynamic and experiential approaches. The therapists play a more directive role in therapy. The emphasis is on changing the patient's behavior, using positive reinforcement, and increasing selfefficacy. They also emphasized research. More recently, cognitive therapies such as those of A. T. Beck have tended to be combined with behavioral emphases. The cognitive-behavioral therapies have focused on changing dysfunctional attitudes into more realistic and positive ones and providing new information-processing skills. Cognitive-behavioral therapies have become quite popular in the last two decades or so for several reasons. They tend to be brief forms of therapy, rather detailed manuals for conducting therapies for depression and anxiety disorders have been produced, and an impressive amount of research has been published to support the claims made for the effectiveness of these therapies. See COGNITION.

Interpersonal therapy is somewhat different from the behavioral and cognitive-behavioral therapies, and most of its developments have occurred in work with depressed patients. The goal of interpersonal therapy (a brief form of therapy) is centered on increasing the quality of the patient's interpersonal interactions. Emphasis is placed on enhancing the patient's ability to cope with stresses, improving interpersonal communications, increasing morale, and helping the patient deal with the effects of the depressive disorder. Like the cognitivebehavioral approaches, a detailed manual was prepared by Klerman, Weissman, and colleagues, and interpersonal therapy was compared with Beck's cognitive-behavioral therapy in a large-scale multisite study of depression. At the end of treatment, the results were roughly comparable and positive. See PERSONALITY THEORY.

Eclectic and integrative therapies. Although there are a large number of psychotherapies, several

surveys of different groups of psychotherapists have indicated that the largest number consider themselves to be eclectics. They do not adhere strictly to one theoretical orientation or school but use any procedures that they believe will be helpful for the individual patient. Rather than having the patient adapt to a specific form of therapy, the therapeutic procedures are adapted to the patient. More recently, eclecticism has been linked with the development of a movement for integration in psychotherapy. The emphasis in this new development is on openness to the views of other approaches, a less doctrinaire approach to psychotherapy, and an attempt to integrate two or more different theoretical views or systems of psychotherapy. There is thus a greater emphasis on theoretical integration than is generally true for eclecticism. This also provides more opportunity for research aimed at evaluating therapeutic factors or techniques of different approaches.

Group, family, and marital therapy. Most psychotherapy is conducted on a one-to-one basis-one therapist for one patient-and the confidentiality of these sessions is extremely important. However, there are other instances where more than one patient is involved because of particular goals. Marital therapy, for example, involves both marital partners, as the most efficacious approach to such problems. Therapeutic work with families has also developed from the recognition of the importance of the family when working with individuals who have psychological difficulties. This emphasis has been noted particularly in the case of children and severely disturbed individuals. A variety of group therapies have also been developed and used in different ways. Outpatient groups have been used for smoking cessation, weight loss, binge eating, and similar problems as well as for what were traditionally viewed as psychoneurotic problems. Inpatient group therapy was frequently employed in mental hospital settings along with physical, recreational, and pharmacological therapies.

Therapeutic factors. With so many different forms of psychotherapy available, one may wonder what factors or variables make for positive therapeutic change, particularly when the theoretical emphases appear so different. For example, in psychodynamic therapies the relationship to the therapist and the interpretation of it by the therapist is stressed, as well as the recovery of repressed negative thoughts and feelings. The ability of the patient to secure insight into the causes of the patient's feelings and behavior is also very important. In patient-centered therapy, the therapist's empathy, warmth, and genuineness are the therapeutic variables stressed. In the behavioral therapies, other factors are considered important, including learning to relax, facing the situations that produce anxiety, receiving positive reinforcement and feedback for desired behaviors, and learning self-control. Cognitive therapies emphasize replacing negative cognitions with positive ones, and interpersonal therapists pay particular attention to the patient's interpersonal problems and behaviors in the patient's current life.

Are the therapies really so different? Most of them have not been systematically evaluated, and in a number of the comparative studies the differences in outcome have generally been small. It is important to note that the participants in psychotherapy, the client and the therapist, play a significant role in the types of outcome. There has been an increasing belief that factors common to most psychotherapies may be of some importance. The expectations and motivations of the patient, the type of patient, and the type of therapeutic relationship that develops have been considered possible common factors. Patients who are highly motivated to cooperate with the therapist, whose problems are not severe, and with whom a positive relationship can be developed are most likely to secure positive results regardless of the form of therapy. Furthermore, in most forms of therapy the therapist is likely to provide explanations, information, suggestions, encouragement, and opportunities for emotional expression. See PSY-CHOSIS; PSYCHOSIS.

Effectiveness. As the popularity of psychotherapy has increased, a number of events have led to increased research evaluating its effectiveness. Better means of appraisal and refined research designs have contributed. Most recently, a design has been used in which carefully evaluated research subjects are assigned randomly to the groups being compared. One form of therapy can be compared to a no-treatment control group, to a group receiving medication, or to another form of therapy. Patients with specific diagnoses are compared during therapy, at termination, and at posttherapy intervals. In general, the results have favored psychotherapy over no-treatment controls. The therapies have tended to be brief and time-limited, and few differences have been secured between pharmacotherapy and psychotherapy, although these latter findings are more complex. In some instances, such as depression, an antidepressant works more quickly and appears to act somewhat better with the most severe cases. However, there may be more "drop-outs" from treatment and possible side effects with medication. There has also been research on the combined use of medication and psychotherapy. In general, where two highly successful treatments are combined in cases with depressive or anxiety disorders, there appears to be little gain in effectiveness. However, in several studies of hospitalized patients with schizophrenia where individual psychotherapy has been ineffective, a combination of psychotherapy and medication has produced better results than medication alone. The combination of certain types of family therapy and psychotropic medication for patients diagnosed with schizophrenia has proven more effective than the use of either therapy alone. Thus, increased knowledge of all treatment modalities as well as types of psychopathology may gradually lead Sol L. Garfield to improved outcomes.

Bibliography. D. H. Baucom and N. Epstein, *Cognitive-Behavioral Marital Therapy*, 1990; B. D. Beitman and G. L. Klerman (eds.), *Integrating* Pharmacotherapy and Psychotherapy, 1991; A. E. Bergin and S. L. Garfield (eds.), Handbook of Psychotherapy and Behavior Change, 4th ed., 1994; G. L. Klerman and M. M. Weissman (eds.), New Applications of Interpersonal Psychotherapy, 1993; S. L. Garfield, The Practice of Brief Psychotherapy, 2d ed., 1998; S. L. Garfield, Psychotherapy, An Eclectic-Integrative Approach, 2d ed., 1995.

Psychotomimetic drugs

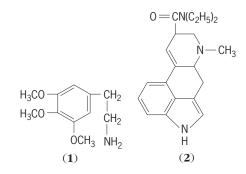
A class of drugs, also called psychedelic or hallucinogenic drugs, that induce transient states of altered perception resembling or mimicking the symptoms of psychosis and are characterized by profound alterations in mood. The term psychotomimetic was originally coined to reflect the similarities between the mind-altering effects of certain psychoactive drugs and some of the symptoms of schizophrenia. *See* PSY-CHOSIS; SCHIZOPHRENIA.

Effects. The net effect of any psychoactive drug is a complex interaction between the personality and expectations of the individual, the pharmacological effects of the drug, and the setting in which the drug is used. Psychotomimetic drugs produce a range of perceptual alterations and distortions. Sensory inputs are heightened or diminished (colors, textures, sounds, tastes, and touch become more or less intense). Sensations may assume compelling significance, to the point that an individual becomes completely immersed in an experience and disregards more salient stimuli. Individuals may experience illusions (distortions in the perceived nature or meaning of real objects), such as seeing a grotesque distortion of a familiar face, or frank hallucinations (perceptions in the absence of any obvious stimuli). Typically the hallucinations produced by psychotomimetic drugs are in the visual domain, and individuals realize that the perceptions are false. Synesthesias may occur, wherein individuals perceive stimuli from one sensory modality in another modality-for example, "seeing" sounds, or "hearing" colors. Often the body or its parts may appear larger, smaller, heavier, or lighter than normal. Time perception may be distorted, with time seeming to pass slower or faster. Concentration and attention can also be impaired, and subjects may sometimes focus on seemingly trivial details of an experience. Inability to "filter" out trivial external or internal stimuli may result in the feeling of being overwhelmed by sensations. Individuals may feel detached from themselves or the immediate environment (depersonalization). This may be described as watching oneself as a spectator or observer. Similarly, individuals may experience their surroundings as unreal or dreamlike (derealization). At times, individuals may describe having "greater understanding" or enlightenment of events, their own existence, or the motivation of others while under the influence of psychotomimetic drugs. Rarely, } individuals may develop false fixed beliefs or inferences that are not consistent with reality (delusions), such as the paranoid belief that people in their environment are trying to harm them. *See* HALLUCINATION; PARANOIA.

These effects on perceptions and thinking are invariably accompanied by changes in the individual's mood. Thus, individuals may experience intense and often rapidly shifting mood states ranging from mild apprehension to panic or severe depression to elation, or they may experience concurrent emotions (such as sadness and joy) which are not ordinarily experienced simultaneously.

Classes of drugs. Psychotomimetic drugs can be classified loosely into several groups based on chemical structure and their profile of effects.

Classical serotonergic hallucinogens. These drugs include mescaline (1), lysergic acid diethylamide (LSD) (2),



and dimethyltryptamine (DMT), as well as the "magic mushroom" (*Psilocybe mexicana*) derivatives psilocybin and psilobin. Mescaline, used in highly structured Native American tribal religious rituals, is derived from peyote buttons, which are the dried tops of a cactus found in the southwestern United States.

Albert Hoffman synthesized LSD-25 (the twentyfifth in a series of synthesized derivatives) in 1938 with compounds derived from ergot (a fungus, *Claviceps purpurea*, which infests such grasses as rye). He discovered its psychotomimetic properties in 1943. *See* ERGOT AND ERGOTISM.

Serotonergic hallucinogens produce their effects by partial stimulation of the 2_A subtype of serotonin receptors in the brain. While these drugs differ in the onset, duration, and intensity of effects, their acute behavioral effects are quite similar. LSD, the most potent drug in this class, initially produces somatic symptoms including paraesthesias (tingling of the extremities), dizziness, weakness, and tremor. These are followed by perceptual alterations including blurring of vision, illusions, visual hallucinations, less discriminant hearing, and altered time perception. Dreamlike imagery may develop when eyes are closed. Synesthesia sometimes occurs. Abstract geometric patterns are frequently seen. These effects are accompanied by changes in emotions: several emotions may occur at once, or emotions may become intensified and change rapidly. Depersonalization (the feeling of being separated from oneself) and difficulty separating self from environmental stimuli may occur. Cognitive disturbances are generally mild and rare, and include impaired memory and difficulty with concentration. Individuals may report a capacity for greater introspection and the revelation of "new insights" about themselves, leading to the reputation of these drugs as "mind revealing" or "psychedelic." A "bad trip" is a panic reaction occurring during the drug state (usually manageable with sensible reassurance, or sedatives). "Flashbacks" can occur in a minority of individuals, and these vivid momentary reactivations of the drug state in the absence of drug can be frightening. *See* SERO-TONIN.

Tolerance to the physical effects of these drugs develops rapidly, and decreases rapidly with cessation of use. However, the drugs are not associated with dependence or withdrawal and have low reinforcement liability relative to stimulants. Unlike other drugs of abuse, most do not reinforce selfadministration in animals.

Amphetamine. Amphetamine and related drugs are stimulants, increasing the release of dopamine, norepinephrine, and serotonin in addition to blocking the reuptake of these neurotransmitters. These drugs can be ingested, smoked, or injected. They produce a "rush" (an orgasmlike state) followed by a state of euphoria, alertness, and a feeling of increased energy and self-confidence. Relative to classical serotonergic hallucinogens, perceptual alterations are not as prominent with amphetamine. Some individuals start to binge (bursts of repeated consumption spanning hours to a few days), and this pattern may be associated with symptoms similar to those of paranoid schizophrenia. Bingeing is followed by a "crash" phase or withdrawal syndrome characterized by depression, agitation, anxiety, and low energy, followed by insomnia and fatigue, and culminating in increased appetite and hypersomnolence (excessive sleepiness).

Amphetamine and other dopaminergic stimulants are associated with tolerance (decreased effect upon repeated use), intense psychological dependence, and withdrawal, and may be associated with sensitization (increased effect upon repeated use). *See* DOPAMINE; NORADRENERGIC SYSTEM.

Substituted amphetamine analogues. These drugs include methylenedioxymethamphetamine (MDMA) or "ecstasy," N-ethyl-3,4-methylenedioxyamphetamine (MDEA), 2,5-dimethoxy-4-methylamphetamine (DOM or "STP"), methylenedioxyethylamphetamine (MDE or "EVE"), and methylenedioxyamphetamine (MDA). They were designed from amphetamine, but share properties of both amphetamine (stimulant) and mescaline (hallucinogen). MDMA was developed in 1914 as a parent compound for the synthesis of other drugs. Though intranasal use has been reported, these drugs are typically ingested. The psychoactive properties of these drugs appear to be related to some combination of effects on both serotonin and dopamine release, and partial agonist effects at 5-HT_{2a} receptors. The psychoactive effects begin with a very mild rush followed by other stimulantlike effects including euphoria, increased self-confidence, and decreased appetite. Like serotonergic hallucinogens, these drugs

produce perceptual alterations (heightened sensory perception, illusions, hallucinations), difficulty in concentrating, and changes in mood (calmness, or anxiety and panic). Other effects include ataxia (instability of gait), insomnia, jaw clenching (bruxism), and muscle rigidity. MDMA reportedly makes people feel more "open" or more connected to others. Its use became illegal in the United States in 1985 after reports of neurotoxicity, and suggestions that MDMA and MDA may cause irreversible damage to brain serotonin neurons.

Cannabinoids. One of the oldest and most widely used psychotomimetic drugs is derived from the flowers and leaves of the Cannabis sativa plant. Delta-9-tetrahydrocannabinol (THC) is the principal active ingredient of cannabis, but there are more than 80 other cannabinoids present in cannabis which may also contribute to its effects. While usually smoked, cannabis is sometimes consumed in brownies, cakes, and sweetened milk. Cannabinoids produce their psychotomimetic effects by actions at brain cannabinoid receptors. The psychotomimetic effects of cannabis appear rapidly (5-10 min) after smoking but slowly (up to 1-2 h) when eaten. The effects of cannabis include changes in perception, cognition, and mood. The cognitive effects include impairment of short-term memory and free recall, difficulty with attention and concentration, difficulty keeping track of thoughts, rapid flow of ideas, and poor judgement. Perceptual alterations may include a heightened sensitivity to all sensory modalities (colors and sounds seem brighter and richer), distortion of time, body image, and space (objects seem or the body feels heavier or longer), depersonalization, derealization, illusions, hallucinations, and synesthesias. Individuals may report a sense of great insight or understanding of life events. Paranoia (irrational suspiciousness or distrustfulness) is not infrequent and frank psychosis, while rare, is possible. The initial effect on mood is euphoria, often accompanied by uncontrollable laughter and a feeling of hilarity. This is followed by feeling calm, relaxed, mellow, passive, apathetic, and drowsy. However, anxiety and panic attacklike symptoms are also possible. "Bad trips" and flashbacks can also occur

There is some evidence suggesting tolerance to some of the effects of cannabis, and increasing evidence suggesting the presence of a mild and subtle (though definite) cannabis withdrawal syndrome.

MMDA antagonists. These agents were originally developed as "dissociative" anesthetics. Rather than complete loss of consciousness, their usual dosage produces an inattentiveness to pain, an amnesia for ongoing events, and a kind of stupor in which the eyes are open but "not seeing." The class includes 1-[1-phenylcyclohexyl]piperidine (phencyclidine, PCP or "angel dust"), ketamine ("special K"), and 1-(1-2-thienyl-cyclohexyl)piperidine (TCP). They produce their effects by noncompetitively blocking the NMDA (N-methyl-D-aspartate) subtype of glutamate receptor. PCP can be snorted, smoked, consumed orally, or injected intravenously. The

effects occur within minutes of smoking the drug and include initial euphoria, followed by perceptual alterations (feelings of detachment, distortion of body image, heightened sensitivity to sounds and images, illusions, and hallucinations) and cognitive deficits (disruption of memory, difficulty in planning, impairment of judgement, difficulty with abstract thinking, and difficulty concentrating and maintaining a train of thought). Other effects include numbness, paresthesias, unsteady gait, and slurred speech. Some of the effects of PCP and ketamine resemble some of the symptoms of schizophrenia, and in rare instances PCP use may be associated with a psychotic syndrome that outlasts intoxication.

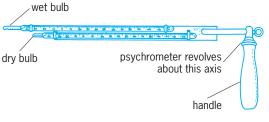
There is little evidence to suggest that the abuse of these drugs by humans is associated with tolerance, dependence, or withdrawal. *See* PSYCHOPHAR-MACOLOGY. Cyril D'Souza

Bibliography. D. J. Balfour, *Psychotropic Drugs of Abuse*, 1990; D. S. Charney, E. Nestler, and B. S. Bunney (eds.), *Neurobiology of Mental Illness*, 1999; J. R. Cooper, F. E. Bloom, and R. H. Roth, *Biochemical Basis of Neuropharmacology*, 2002; K. L. Davis, et al. (eds.), *Neuropsychopharmacology: The Fifth Generation of Progress*, 2002; B. L. Jacobs (ed.), *Hallucinogens: Neurochemical, Behavioral and Clinical Perspectives*, 1984; B. L. Jacobs, How hallucinogenic drugs work, *Amer. Scientist*, 75:386-392, 1987; M. Lader (ed.), *Psychopharmacology of Addiction*, 1994.

Psychrometer

An instrument consisting of two thermometers which is used in the measurement of the moisture content of air or other gases. The bulb or sensing area of one of the thermometers either is covered by a thin piece of clean muslin cloth wetted uniformly with distilled water or is otherwise coated with a film of distilled water. The temperatures of both the bulb and the air contacting the bulb are lowered by the evaporation which takes place when unsaturated air moves past the wetted bulb. An equilibrium temperature, termed the wet-bulb temperature (T_W) , will be reached; it closely approaches the lowest temperature to which air can be cooled by the evaporation of water into that air. The water-vapor content of the air surrounding the wet bulb can be determined from this wet-bulb temperature and from the air temperature measured by the thermometer with the dry bulb (T_D) by using an expression of the form $e = e_{SW} - aP (T_D - T_W)$. Here *e* is the water-vapor pressure of the air, e_{SW} is the saturation water-vapor pressure at the wet-bulb temperature, P is atmospheric pressure, and a is the psychrometric constant, which depends upon properties of air and water, as well as on speed of ventilation of air passing the wet bulb.

Other moisture parameters, such as relative humidity and dew-point temperature, can be conveniently evaluated from the wet- and dry-bulb measurements by means of psychrometric tables.



Sling psychrometer. (After D. M. Considine, ed., Process Instruments and Controls Handbook, McGraw-Hill, 1957)

These tables are accurate for a ventilation speed of 4.6 m/s (15 ft/s) when applied to measurements made using standard U.S. Weather Bureau thermometers. Other sensors may require different ventilation rates.

The common sling psychrometer consists of two mercury-in-glass thermometers which are mounted side by side on a metal or plastic frame (see **illus.**). The frame is attached to a handle by means of a chain or a suitable bearing in such a manner that the frame can be freely rotated about the handle. A clean muslin patch is tightly mounted over the bulb of one of the thermometers and is thoroughly wetted with distilled water prior to an observation. To make a measurement, the frame is rotated about the handle at a rate of approximately four revolutions per second, and this rotation is continued until the wet-bulb temperature is decreased to a constant value.

Another form of the psychrometer consists of two temperature-sensing elements, one of which is wetted, mounted in an air duct. A suction fan is used to aspirate air past the sensing elements at a rate sufficient to assure that the wet-bulb temperature will be reached for the particular thermal sensor employed in the instrument. This form of the instrument is used frequently at permanent weather stations.

The error associated with the measurement of atmospheric moisture by means of a psychrometer originates from a number of sources. These can be classified as errors due to (1) inherent deficiencies in temperature sensing, (2) improper ventilation, (3) improper muslin preparation or bulb coverage, (4) impure water, (5) extraneous heat flow to the bulb from radiative or conductive sources, and (6) an uncertainty in the phase of the water when the wet-bulb temperature is below the freezing point.

In general, the accuracy of the instrument is greatest at temperatures above the freezing point of water and decreases substantially at low temperatures. For example, normal calibration errors in the standard sling psychrometer can lead to a measurement error of 25% in air of nominal 50% relative humidity when the dry-bulb temperature is 5° F (-15° C). The corresponding measurement error at a dry-bulb temperature of 59° F (15° C) is only 5%. *See* MOISTURE-CONTENT MEASUREMENT. Richard M. Schotland

Bibliography. C. F. Marvin, *Psychrometric Tables*, U.S. Department of Commerce, W.B. 235, 1941; A. Wexler (ed.), *Humidity and Moisture*, 1965.

Psychrometrics

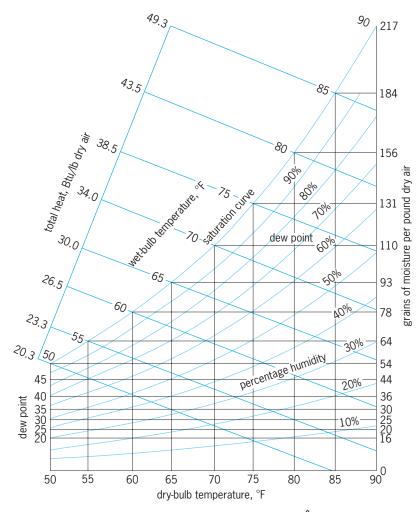
A study of the physical and thermodynamic properties of the atmosphere. The properties of primary concern in air conditioning are (1) dry-bulb temperature, (2) wet-bulb temperature, (3) dew-point temperature, (4) absolute humidity, (5) percent humidity, (6) sensible heat, (7) latent heat, (8) total heat, (9) density, and (10) pressure.

The atmosphere in a clean pure state is a mechanical mixture of dry air and water vapor. Each is independent of the other and follows the laws of physics in accordance with its respective physical properties. Normal atmosphere contains impurities such as carbon dioxide, ozone, dust, and dirt in varying quantities.

Graphic representation. Even though dry air and water vapor are independent entities, there are certain relationships which permit the inclusion of these factors on a common chart, called a psychrometric chart (see **illus.**).

The psychrometric chart is a valuable tool for understanding the relationship of the many variables encountered in the atmosphere and for solving problems in air conditioning.

Dry-bulb temperature. The dry-bulb temperature is the ambient temperature of the air and water vapor



Psychrometric chart for air at a pressure of 29.92 in. of mercury (10^2 kPa). 1 grain/pound = 1/7 g/kg; 1 Btu/lb = 2326 J/kg; °C = (°F-32)/1.8.

as measured by a thermometer or other temperaturemeasuring device in which the thermal element is dry and shielded from radiation. Temperature scales are usually in degrees Fahrenheit or degrees Celsius. The vertical lines on the chart are dry-bulb temperature lines.

Wet-bulb temperature. If the bulb of a dry-bulb thermometer is covered with a silk or cotton wick saturated with distilled water and the air is drawn over it at a velocity not less than 1000 ft/min (5 m/s), the resultant temperature will be the wet-bulb temperature.

If the atmosphere is saturated with water vapor, the water on the wick cannot evaporate and the wet bulb will read the same as the dry bulb. If the atmosphere is not saturated, water will evaporate from the wick at a rate dependent upon the percentage of saturation of the atmosphere. The cooling produced by the evaporation will result in a lowering of the temperature of the bulb and the consequent reading is the wet-bulb tempera- ture.

The lines on the psychrometric chart extending from upper left to lower right are wet-bulb temperature lines. Where the dry-bulb and wet-bulb temperatures are the same, the atmosphere is saturated and a line drawn through these points is called the saturation or 100% relative humidity curve.

Dew-point temperature. The dew-point temperature is the temperature at which the water vapor in the atmosphere begins to condense. This is also the temperature of saturation at which the dry-bulb, wetbulb, and dew-point temperatures are all the same. *See* DEW POINT.

The dew-point temperature is a measure of the actual water-vapor content in the atmosphere. The water-vapor content is constant for any dew-point temperature regardless of the dry-bulb or wet-bulb temperature. The dew-point temperatures are the horizontal lines on the psychrometric chart.

Absolute humidity. The actual quantity of water vapor in the atmosphere is designated as the absolute humidity. It is measured as pounds of water vapor per pound of dry air, grains of water vapor per pound of dry air, grains of water vapor per cubic foot of dry air, or grams of water per kilogram of dry air. One pound of water vapor is equal to 7000 grains. Because the absolute water-vapor content is directly related to dew-point temperature, the vapor content in grains per pound of dry air is shown on the vertical scale on the right-hand side of the psychrometric chart for different dew-point temperatures. *See* HUMIDITY.

Percentage or relative humidity. Percentage or relative humidity is the ratio of the actual water vapor in the atmosphere to the quantity of water vapor the atmosphere could hold if it were saturated at the same temperature. For example, if the atmosphere at 50°F (10°C) is saturated with water vapor, there would be 53 grains of water vapor per pound of dry air (7.57 grams per kilogram of dry air). If the dry-bulb temperature of this atmosphere were then raised to 69°F (21°C), it could hold 106 grams of water vapor per pound or 15.14 grams per kilogram (corresponding to a saturation temperature of 69°F or 21°C). As no moisture was added in changing the temperature

from 53 to 69°F (12 to 21°C), the actual vapor content is still 53 grains. The percentage humidity would then be 53/106 or 50% at 69°F.

Sensible heat. Sensible heat, or enthalpy of dry air, is heat which manifests itself as a change in temperature. It is expressed in British thermal units (Btu) or joules (1 Btu = 1055 J). One Btu will raise the temperature of 1 lb of water $1^{\circ}F$ (or 1 kg of water by 0.252°C). One Btu will also raise the temperature of 1 lb of dry air approximately 4.25°F (or 1 kg by 1.07°C) at atmospheric temperatures and sea-level pressure. *See* ENTHALPY.

Latent heat. Latent heat, or enthalpy of vaporization, is the heat required to change a liquid into a vapor without change in temperature. For example, it would require 1054 Btu to change 1 lb (or 2.45 megajoules to change 1 kg) of water at 70° F (21°C) from a liquid to a dry, saturated vapor at 70° F.

Latent heat is sometimes referred to as the latent heat of vaporization and varies inversely as the pressure. The higher the pressure (or saturation temperature) the lower the Btu required to evaporate water from a liquid to a vapor. **Table 1** shows the thermodynamic properties of water for various temperatures.

Total heat. The total heat, or enthalpy, of the atmosphere is the sum of the sensible heat, latent heat, and superheat of the vapor above the saturation or dew-point temperature. At saturation, the total heat measured in Btu/lb (1 Btu/lb = 2326 J/kg) of dry air is measured at the wet-bulb temperature and includes both the sensible heat of the dry air and the latent heat of the water vapor at the temperature measured. For example, at 60° F (15°C) on the saturation curve the total heat is 26.46 Btu/lb (61.55 kilojoules/kg) of dry air with water vapor to saturate the air at that temperature.

Total heat is relatively constant for a constant wetbulb temperature, deviating only about 1.5–2% low at relative humidities below 30%. **Table 2** shows enthalpy values for saturated air.

Density. The density of the atmosphere varies with both altitude and percentage humidity. The higher the altitude the lower the density, and the higher the moisture content the lower the density. *See* DENSITY.

Temperature, °F (°C)	Pressure, Ib/in. ² absolute	Enthalpy of evap- oration, Btu/lb	
40 (4.44)	0.122	1070.46	
45 (7.22)	0.147	1067.81	
50 (10.00)	0.178	1064.99	
55 (12.78)	0.214	1062.16	
60 (15.56)	0.256	1059.34	
65 (18.33)	0.306	1056.52	
70 (21.11)	0.363	1053.71	
75 (23.89)	0.430	1050.89	
80 (26.67)	0.507	1048.07	
85 (29.44) 0.596		1045.23	
90 (32.22)	0.698	1042.40	
95 (35.00)	0.816	1039.56	
100 (37.77)	0.950	1036.72	

Temperature, F (°C)	Volume, ft ³ /lb of dry air	Enthalpy, Btu/lb of dry air	
40 (4.44)	40 (4.44) 12.70		
45 (7.22)	12.85	17.65 20.30 23.22 27.15	
50 (10.00)	13.00		
55 (12.78)	13.16		
60 (15.56)	13.33		
65 (18.33)	13.50	30.06	
70 (21.11)	13.72	34.09	
75 (23.89)	13.88	38.61	
80 (26.67)	14.09	43.69	
85 (29.44)	14.31	49.43	
90 (32.22)	14.55	55.93	
95 (35.00)	14.80	63.32	
00 (37.77) 15.08		71.73	

TABLE 3. Altitude and pressure (standard atmosphere)*			
Altitude, ft	Pressure, in. Hg		
-1,000	31.02		
-500	30.47		
0 (sea level)	29.92		
500	29.38		
1,000	28.86		
5,000	24.89		
10,000	20.58		

At sea level (29.92 in. Hg or 101.325 kilopascals absolute pressure) and 59°F (15°C) the density is 0.0765 lb/ft³ (1.225 kg/m³). At 5000 ft (1524 m) elevation and 50°F (10°C), the density would be 0.0637 lb/ft³ (1.020 kg/m³).

At sea level and 65° F (18°C) saturated, the density is 0.074 lb/ft³ (1.185 kg/m³). At 65° F (18°C) dry bulb and 30% saturation, the density is 0.0752 lb/ft³ (1.205 kg/m³). The reciprocal of density (cubic feet per pound of dry air) is usually used rather than the density. Table 2 shows temperature-volume relations for saturated air.

Pressure. Atmospheric pressure, usually referred to as barometric pressure, is measured either in inches of mercury (29.92 in. Hg at sea level), in pounds per square inch absolute [14.7 lb/in.² absolute (psia) at sea level], in millibars (1013.25 mb at sea level), or in kilopascals (101.325 kPa at sea level).

Pressure varies inversely as elevation, as temperature, and as percentage saturation. Pressure decreases with elevation as shown in **Table 3**. *See* MOISTURE-CONTENT MEASUREMENT; PSYCHROME-TER. John Everetts, Jr.

Bibliography. D. J. Croome and B. M. Roberts, *Air Conditioning and Ventilation of Buildings*, vol. 1, 2d ed., 1981; D. J. Croome and A. F. Sherratt (eds.), *Condensation in Buildings*, 1972; A. Pande, *Handbook of Moisture Determination and Control: Principles, Techniques, Applications*, vol. 1, 1974, vol. 3, 1975; B. Stamper and R. L. Koral (eds.), *Handbook of Air Conditioning, Heating, and Ventilating*, 3d ed., 1979; A. Wexler (ed.), *Humidity and Moisture Measurement and Control in Science and Industry*, vols. 1, 3, and 4, 1964–1965.

Pteridosperms

A group of extinct seed plants characterized by fernlike leaves that produced naked seeds. The discovery of the seed ferns was a major contribution to the study of plant evolution because it demonstrated the existence of a group of vascular plants that is today extinct. Although the seed ferns probably have their ultimate origin within the progymnosperm order Aneurophytales, the best evidence of the group comes from lower Carboniferous (Mississippian) and younger sediments. Some seed ferns are reconstructed as trees with stout stems, while perhaps the majority were vines or lianas that supported massive fernlike fronds (**Fig. 1**).

The seed ferns consist of six Paleozoic orders (Calamopityales, Buteoxylonales, Lyginopteridales, Medullosales, Callistophytales, Glossopteridales) and three orders (Peltaspermales, Corystospermales, Caytoniales) found in Mesozoic rocks.

Calamopityales. This order is included with the seed ferns because of the nature of the wood and features of the cortex (**Fig. 2**). The stems may be protostelic to eustelic, but all are characterized by a conspicuous pith and petiole bases with numerous vascular bundles.

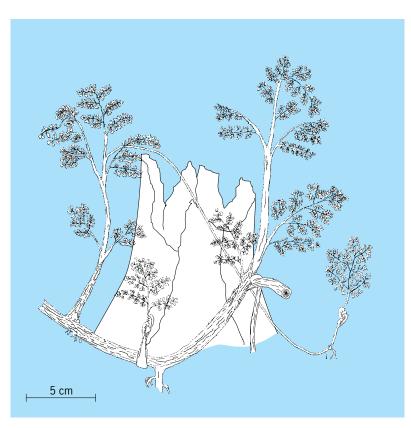


Fig. 1. Suggested reconstruction of Callistophyton. (After G. W. Rothwell, The Callistophytaceae (Pteridospermopsida): I. Vegetative structures, Palaeontographica, 151B:171–196, 1975)

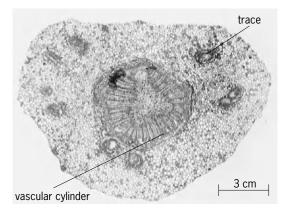


Fig. 2. Transverse section of Stenomyelon. (From C. B. Beck, The appearance of gymnosperm structure, Biol. Rev., 45:379–400, 1970)

Buteoxylonales. This group of lower Carboniferous seed ferns is also based on the anatomy of stems and petioles, but the primary features of the order include protostelic petioles that are trilobed in cross section. The cortex has sclerenchyma bands around the periphery. As with the Calamopityales, nothing is known about the foliage or reproductive organs of this group.

Lyginopteridales. This group of Carboniferous seed ferns has eustelic stems with a limited amount of secondary xylem. The leaves are large and planated, with numerous small pinnules. The seeds are small and characterized by thick integument and distinct micropyle. Lagenostoma is a seed that was approximately 0.2 in. (6 mm) long and was produced in a cup-shaped structure termed a cupule. Extending from the tip of the seed was a small conical projection that trapped wind-borne pollen. One feature that made it possible to unite some of the dispersed organs of the Lyginopteridales was small multicellular glands that extended from the epidermis of the stems, petioles, and seed cupules. Pollen organs consist of elongate pollen sacs that are fused laterally to form synangiate organs that produced spherical, trilete pollen grains. The middle of the pollen organ was hollow, suggesting that the sacs released pollen toward the center of the organ.

Medullosales. These seed ferns extend from the lower Carboniferous into the Permian and are characterized by stems that contain several segments of vascular tissue, each surrounded by secondary xylem. Medullosa is the best-known genus of stems, with some Permian forms being up to 20 in. (50 cm) in diameter and containing more than 40 vascular segments. The leaves were massive, with some specimens being several meters long; the petioles contained numerous vascular bundles in the base. Two of the common medullosan leaf types are Alethopteris and Neuropteris. The seeds of this group were noncupulate and radial in transverse section. Many were large (5 in. or 13 cm long) and ornamented by longitudinally oriented ribs. Pollen organs of the medullosans were large and synangiate, and ranged from simple forms with a few pollen sacs to compound types with more than 80 elongate sacs. The pollen grains of this group are some of the largest produced by either living or fossil plants, with many up to 0.02 in. (0.5 mm) long.

Callistophytales. Although the seed ferns in this order form only a small group of upper Carboniferous pteridosperms, they are by far the best understood. Stems of *Callistophyton* were eustelic, and they produced pinnately compound leaves with axillary buds or branches at each node. The seeds were small and platyspermic. Pollen organs were synangiate and attached to the lower surface of laminar pinnules. Pollen grains possessed two distally inclined air bladders like those of modern pines. Pollen tubes were produced, and the microgametophyte consisted of an axial row of prothallial cells.

Glossopteridales. These plants are recognized as specialized, highly diverse seed ferns that dominated Southern Hemisphere continents (Gondwana) during the Permian. Leaves of *Glossopteris* are characterized by a reticulate pattern of veins, with some species more than 12 in. (30 cm) long. In transverse section, the underground parts consist of radiating wedges of wood with hollow areas between, a type of anatomy like that found in certain swamp-dwelling plants today. Seeds were produced in cupules or on exposed leaves. Pollen in this group is bisaccate.

Caytoniales. These Mesozoic seed ferns produced palmately compound leaves with a reticulate pattern of venation. Bisaccate pollen was produced in elongate synangia attached to shortened lateral branches. The seed-bearing units (*Caytonia*) consisted of helmet-shaped cupules approximately 0.2 in. (5 mm) in diameter. *See* CAYTONIALES.

Corystospermales. This small group of seed ferns occurs in Triassic sediments. One common foliage type was *Dicroidium*, a leaf characterized by a basal dichotomy of the rachis and small laminar pinnules. Some stems were more than 32 in. (80 cm) in diameter, suggesting that certain corystosperms were woody plants. Pollen sacs were borne on the lower surface of modified pinnule segments and produced bisaccate pollen. Seeds were partially covered by cupules, and the tip of the integument was curved and projected from the cupule surface.

Peltaspermales. These seed plants range from the upper Permian into the Triassic. The genus *Callipteris* is used for large bi- or tripinnate leaves bearing alethopterid pinnules. Seeds were produced on the lower surface of umbrella-shaped modified leaves, and pollen grains in the group were believed to have been boat shaped. *See* PALEOBOTANY. Thomas N. Taylor

Bibliography. W. N. Stewart and G. W. Rothwell, *Paleobotany and the Evolution of Plants*, 2d ed., 1993; T. N. Taylor and E. L. Taylor, *The Biology and Evolution of Fossil Plants*, 1993.

Pteriomorphia

An order containing seven superfamilies of marine and brackish bivalves with byssal attachment at some time in their geological history (the Ostreacea are exceptional, with the lower valve attached by byssal cement at settlement, and with no record of ancestral attachment by byssal threads). Byssal attachment is associated with departure from the ancestral isomyarian form; Mytilacea and Pinnacea are heteromyarian, with a small anterior adductor muscle; others are monomyarian, with one central posterior adductor. Byssal attachment is lost in some: *Placuna* is free-living; *Pecten maximus* can right itself if inverted; *Chlamys* can swim to escape from starfish; some *Lima* can swim, while the tropical *Amusium* swims strongly. Pearls are obtained from *Pinctada* (Pteriacea).

The Pteriomorphia possess some primitive features and some secondarily simple features: the hinge is edentulous or with inconspicous teeth (except Spondylus); there are no siphons; minimal points of fusion occur between the left and right mantle lobes; the Mytilacea, Pteriacea, Pectinacea, and Anomiacea have filibranch ctenidia. Specialized features include simultaneous hermaphroditism in some Pectinacea, and sex reversal and incubation of larvae in some Ostreacea, and protandrous hermaphroditism in others. Few features are common to all Pteriomorphia, and the systematic status is debatable. Studies suggest addition of the isomyarian Arcacea and Limopsacea. See BIVALVIA; MOL-Richard D. Purchon LUSCA.

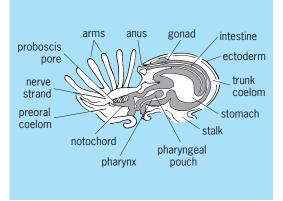
Pterobranchia

A group of small sessile hemichordates that may be colonial (*Rhabdopleura*), pseudocolonial (*Cephalodiscus*), or solitary (*Atubaria*). Each individual, or zooid, lives inside a nonchitinous tube secreted by the protosome, except for *Atubaria*; an aggregation of zooids is called a coenecium and can vary in shape. The protosome, or oral shield, is disciform, closes the mouth ventrally, and secretes the tube. The protocoel has symmetrical pores at the base of the first pair of arms. The mesosome, or collar, has an anteroventral mouth and one to nine pairs of dorsally ciliated tentaculated arms, which are used to collect small organisms for food. The mesocoels extend into both arms and tentacles.

The metasome is divided into a sacciform trunk and a slender ventral stalk that may be free at its end. The trunk contains the U-shaped digestive tract, with the pharynx having a single pair of gills, except in *Rhabdopleura*. The stomach is a sacciform expansion of the gut, and the tubular intestine curves dorsally and opens behind the arms by a middorsal anus, which in *Rhabdopleura* is on the right side.

The nervous system is very simple and lacks a neurochord. The buccal diverticulum is hollow, with an anteroventral heart vesicle and central sinus. The glomerulus is poorly developed, and the circulatory system is simpler that in enteropneusts.

Sexes are separate; the zooids show two anterodorsal sacciform gonads, except for *Rhabdopleura*, which has only one on the right side. The gonads open near the base of the last pair of



Cephalodiscus, of the Pterobranchia. (After H. W. Rand, Comparative Anatomy, Blakiston-McGraw-Hill, 1936)

arms. Development is not completely understood in Cephalodiscus; in Rhabdopleura, the eggs undergo radial holoblastic cleavage and develop into lecithotropic larvae that feed on stored yolk and have a ventral glandular depression that secretes the larval cocoon. After metamorphosis, which takes place in the cocoon, a hole is formed in the cocoon's wall, and the ancestrula, the first zooid in the colony, builds the coenecium by budding. In Rhabdopleura, the ancestrula produces a branching tube that creeps over a substratum and gives off small erect tubes. In the inner side of the creeping tube a syncytial structure, the black stolon, forms and connects all the zooids of the colony by budding. In Cephalodiscus, the coenecium is composed of many unconnected individuals that originated from an ancestrula, and budding activity takes place in the distal zone of the stalk (see illus.). See HEMICHORDATA. Jesús Benito

Pteropsida

A large group of vascular plants characterized by having parenchymatous leaf gaps in the stele (vascular system of the stem) and by having leaves which are thought to have originated in the distant past as branched stem systems. Some botanists regard the Pteropsida as a natural group which they recognize as a class, subdivision, or division. Others regard it as an artificial assemblage of plants that have undergone certain similar changes from a rhyniophyte ancestry. The various components of the Pteropsida are here treated as three separate divisions under the names Magnoliophyta, Pinophyta, and Polypodiophyta. *See* MAGNOLIOPHYTA; PINOPHYTA; PLANT KINGDOM; POLYPODIOPHYTA. Arthur Cronquist

Pterosauria

The "winged reptiles" of the Mesozoic Era, constituting the closest major group to Dinosauria and sharing many features with them. Their common ancestor was a small, bipedal, agile archosaur reptile probably similar to the small *Scleromochlus* of the Late Triassic. This ancestor probably possessed a large, lightly built skull, a short body, long hindlimbs, and digitigrade feet with four long metatarsals. Pterosaurs inherited all these features, and they further evolved the power of flight. Bats have a wing of skin stretched among four fingers and extending from the arms to the legs (and often the tail), and birds have a wing of feathers supported by the arm and the fused wristhand bones. Pterosaurs had a wing of skin that was internally supported by long, fine, possibly keratinous stiffening fibers and braced by the forelimb, including a greatly elongated fourth finger (the first three fingers remained small). Their brains were relatively large and somewhat birdlike; the canals of the ear region show that they were well suited for life in the air. *See* DINOSAUR; REPTILIA.

Noteworthy finds. The first remains of pterosaurs were found in Germany in the 1780s, and it was not clear what sort of animals they were. Regarded as aberrant birds by some and aberrant bats by others, they were established as reptiles by the French comparative anatomist Georges Cuvier in the early 1800s. He noted that they were not at all like typical reptiles: they stood erect like birds, they had hollow bones and probably an insulatory covering, and they flew.

Since that time, hundreds of pterosaur skeletons and isolated bones have been discovered on most continents. The best records originally came from Germany, where the Late Jurassic Solnhofen Limestones, the lithified remains of ancient lagoon sediments, preserved many complete and beautiful specimens of Rhamphorbynchus, Pterodactylus, and other forms. Equally striking were pterosaurs such as Dorygnathus and Campylognathoides from the Early Jurassic of the Holzmaden region of southern Bavaria, and remains of Dimorphodon from southern England (Fig. 1). In the late 1800s, many worn and broken pterosaur bones were recovered from the newly explored badlands of Kansas and Nebraska; they included Pteranodon and Nyctosaurus. The earliest pterosaurs are known from the Late Triassic, about the time that the first dinosaurs appeared: they include Eudimorphodon, Peteinosaurus, and Preondactylus, all from marine rocks of Italy and Greenland. In recent decades, some of the most interesting finds have come from China and from a series of limestone nodules in the Santana region of Brazil, where fossil remains have included three-dimensional skeletons, some preserved soft parts, and bizarre skull crests. Specimens discovered in the lake district of Kazakhstan and Kirgizia preserve a kind of epidermal covering that appears to include both skin and a furlike covering, not identical to mammalian fur. New discoveries from the Early Cretaceous of China have included eggs of pterosaurs with the embryos intact.

Classification. Pterosaurs were traditionally divided into two subgroups. The "Rhamphorhynchoidea" comprised the early pterosaurs with long tails, a long fifth toe, and short metacarpal (palm) bones in the hand; the Pterodactyloidea reduced the tail and fifth toe to stumps and lengthened the metacarpals. The first category has been

abandoned because some rhamphorhynchoids are more closely related to pterodactyloids than to other rhamphorhynchoids. (The term pterodactyl, first used to describe these animals, now informally denotes any pterodactyloid; the term pterosaur describes any member of the Pterosauria.) Pterodactyloids ("wing fingers") were a subgroup of Late Jurassic and Cretaceous pterosaurs that replaced the early pterosaurs. The best-known members of this group include the small Pterodactylus and the much larger Pteranodon with 7-m (23-ft) wing span (Fig. 2). An even larger (11-12 m or 35-40 ft), partially known form, Quetzalcoatlus, was discovered in Texas, and similar large remains are known from the Middle East and elsewhere. These larger pterosaurs we were primarily soaring animals, like large birds.

At present there are over 100 known genera of pterosaurs, and most genera have only one known species. Many genera have historically been based on incomplete material, so it is difficult to compare them to others. Several conflicting schemes of the interrelationships of pterosaurs have been proposed; none seems clearly preferable. *See* ARCHOSAURIA; ORNITHISCHIA; SAURISCHIA.

Most analyses of the relationships of pterosaurs conclude that they shared a common ancestor with dinosaurs. These groups share long legs, a thigh bone with an offset head, a reduced fibula (the splint bone on the side of the shin in humans), a hingelike ankle, long foot bones and toes, and other features seen in no other animals. However, it has also been observed that all these features are related to locomotion and could be independently derived (though there is no evidence for this). Unfortunately, pterosaurs are so modified through the rest of their skeletons that they leave little trace of features that might link them to other reptiles. In order to advance this question, further analyses must include more kinds of pterosaurs and other reptiles, and more features held in common.

The hindlimb features that pterosaurs share with birds and other dinosaurs have suggested that pterosaurs also walked upright, on their toes, with a narrow gait, using the hindlimbs exclusively or nearly so. An alternative view suggests that the hindlimbs were too fragile, that the foot was connected to the wing and so impeded terrestrial progression, or that the joints of the feet did not allow walking on the toes. Recent discoveries in France of tracks with forefoot impressions far outside the body midline were clearly made by small pterosaurs slowly foraging for seaside invertebrates while walking on all fours. The feet seem to have been placed beneath the hip joints, and the heels are impressed. Thus, at least some pterodactyloids, with their elongate metacarpals, were habitual quadrupeds on the ground. However, we still know nothing of the stances and gaits of the first pterosaurs. Other indistinct tracks found elsewhere continue to be attributed to pterosaurs, but these tracks do not have clear features of pterosaur anatomy, and the forefoot prints seem too close to the body, so their trackmakers remain unresolved.

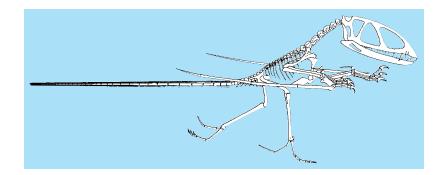


Fig. 1. Skeleton of *Dimorphodon*, an Early Jurassic pterosaur, restored as a biped. Wingspan is about 1.6 m (5.3 ft).

The fifth toe of early pterosaurs, much like the wing finger, was elongated and of variable length. Some reconstructions of pterosaurs have represented this toe as an anchor for the wing, but this idea is controversial because the size of the toe is so variable and its jointed morphology is apparently ill suited for the purpose (compared to the short, stiff, and potentially analogous calcar bone of some bats). Pterodactyloids lost the tail and the fifth toe, which poses difficulties for the explanation of this toe as a wing anchor. *See* SKELETAL SYSTEM.

Flight. Experts agree that pterosaurs were true flying reptiles, powered by wings that flapped vigorously. The flight muscles had extensive attachments to the wing bones and the expanded breastbone, and the shoulder girdle was modified to brace the force of the flight stroke against the skeleton. The wing itself was made of thin skin attached to the bones of the forearm and wing finger. It was reinforced by thin, perhaps keratinous fibers only a few micrometers in diameter but several centimeters long. These radiated over the undersurface of the wing in a pattern similar to the feather shafts of birds, and provided the stiffness and flexibility that made the wing membrane airworthy. All but the largest pterosaurs were active flappers; the largest relied on soaring because muscle power cannot sustain flapping flight for extended periods in such large forms.

Pterosaurs have other modifications for flight. The bones are extremely thin-walled and hollow, and many have openings on the surfaces that are remnants of channels through which the respiratory tissues ran. These tissues opened inside the bones, apparently providing further respiratory



Fig. 2. Late Cretaceous pterodactyloid *Pteranodon*, from western North America (reconstruction).

surface and helping to cool the blood. Birds have similar structures and functions, and their dinosaurian ancestors have at least the beginnings of these structures, which may reflect relatively high metabolic levels. The shoulder girdles of pterosaurs are braced against their large, expanded breastbones that anchored the flight muscles. Many large pterodactyloids have a series of fused vertebrae in the upper back, often showing evidence of sockets in which the ends of the scapulae (shoulder blades) inserted as a further brace. Specialized structures such as internal struts, plywoodlike layers of bone, and torsion-resistant tubular design helped the skeleton resist the forces of flight.

Pterosaurs, like birds and bats, evolved a downand-forward flight stroke that created a donut-shaped vortex wake that provided the forward component of thrust. Anatomical studies demonstrate that pterosaurs produced various flight strokes. The most basic is roughly the shape of a lazy figure-8, at the shoulder joint, and an up-and-down motion was also possible. The elbow is a hinge, the wrist is blocky but slightly flexible, and most of the length of the wing is formed by a single elongated finger. The wing finger dwarfs the first three, which were free and could still manipulate objects. The base of the wing finger has a pronounced pulley joint, modified so as to flex the finger to the side rather than up and down like the other fingers. The rotation of this joint is very important in the flight stroke.

The shape of the wings of pterosaurs is an open question. Some specimens of Rhamphorhynchus suggest a narrow, gull-like wing that tapered toward the waist. Soft tissues preserved in some specimens of Pterodactylus suggest an attachment of the wing to the upper leg. However, without evidence of the supporting fibers, it is difficult to tell which softpart remains represent the true wing and which are merely body remains squashed into two dimensions. Claims have been made that the Kirgiz specimens of Sordes, which preserve both wing fibers and other epidermal structures, demonstrate that the wings were attached to the ankles. This would have given the wing a much broader, more batlike planform (the shape and layout of a wing) than suggested by the Rhamphorbynchus specimens. In aerodynamic terms, the aspect of the wing (roughly the relationship of the length to the mean breadth) would be lower, and less like the aspect of many vigorously flying birds. This might suggest different styles of flight and different ecological roles. But the Sordes specimens have the same problems of flattening and ambiguous interpretation as the Pterodactylus specimens. At least one study has suggested that the specimens are systematically distorted and merely create an illusion of wing membranes that are nothing but skin.

Much more needs to be learned about the soft tissues of pterosaurs. Some filamentous epidermal structures appear not only on and near the wing areas but all over the body. These are apparently different from the structural fibers of the wing membranes, but it is not clear whether they grade into each other or are entirely different. The best-preserved specimens of *Rbamphorhynchus* indicate that the wing's structural fibers occur only on the surface of the underside of the wing, and they are embedded superficially in the epidermis, not free. The ability to differentiate the various kinds of fibers would be important in discerning whether a given preserved tissue in a pterosaur fossil is from the wing or another part of the body. Such information could have broad implications for flight style, terrestrial locomotion, and physiology. *See* FLIGHT. Kevin Padian

Bibliography. S. C. Bennett, Terrestrial locomotion of pterosaurs: A reconstruction based on *Pteraichnus* trackways, *J. Vert. Paleontol.*, 17:104–113, 1997; E. Buffetaut and J.-M. Mazin, *Evolution and Palaeobiology of Pterosaurs*, 2003; K. Padian, A functional analysis of flying and walking in pterosaurs, *Paleobiology*, 9:218–239, 1983; K. Padian, Pterosaur stance and gait, and the interpretation of trackways, *Ichnos*, 10:115–126, 2003; K. Padian and J. M. V. Rayner, The wings of pterosaurs, *Amer. J. Sci.*, 293A:91–166, 1993; D. Unwin, *The Pterosaurs: From Deep Time*, 2005; P. Wellnhofer, *The Illustrated Encyclopedia of Pterosaurs*, 1991; L. M. Witmer et al., Neuroanatomy of flying reptiles and implications for flight, posture and behaviour, *Nature*, 425:950–953, 2003.

Pterygota

A subclass of the class Insecta comprising the winged insects and their secondarily wingless descendants. This subclass, including the great majority of insects, is usually divided into two infraclasses: Paleoptera, or ancient-winged insects (mayflies, dragonflies, and several extinct orders, most of which could not fold the wings over the body), and Neoptera, or foldingwinged insects (the rest of the winged insects and some secondarily wingless lineages), which ancestrally folded the wings over the body by a particular articulatory mechanism.

The acquisition of wings occurred once in the Insecta and was the major general adaptation in their evolution. It is possible that wings arose as the muscled gills of aquatic nymphal Apterygota, which gained propulsive capabilities in the water. On the return of the adults to land for mating and reproduction, the movable gills acquired a new function in sexual and agonistic signaling, which soon became supportive for mating flight (as in many extant mayflies). With the basic mechanism already in place, modifications led to directed flight.

The winged insects, both Paleoptera and Neoptera, first appear in the fossil record in the Early Pennsylvania (upper Carboniferous), suggesting their origin in the Mississippian. *See* ENDOPTERY-GOTA; EXOPTERYGOTA; INSECTA.

William L. Brown, Jr. Bibliography. Commonwealth Scientific and Industrial Research Organization, *The Insects of Australia: A Textbook for Students and Research Workers*, 2d ed., 1991; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, vol. 2, 1982.

Public health

An effort organized by society to protect, promote, and restore the people's health. It is the combination of sciences, skills, and beliefs that is directed to the maintenance and improvement of health through collective or social actions. The programs, services, and institutions of public health emphasize the prevention of disease and the health needs of the population as a whole. Additional goals include the reduction of the amount of disease, premature death, disability, and discomfort in the population.

The sciences and techniques utilized in public health aim to protect, preserve, and restore good health. The basic sciences of public health include epidemiology and vital statistics, which measure health status and assess health trends in the population. Epidemiology is also a powerful research method, used to identify causes and calculate risks of acquiring or dying of many conditions. For example, it has been the principal technique for clarifying the nature of infection by human immunodeficiency virus (HIV), the causative agent of acquired immunodeficiency syndrome (AIDS). Many sciences, including toxicology and microbiology, are applied to detect, monitor, and correct physical, chemical, and biological hazards in the environment. Such applications are being used to address concerns about a deteriorating global environment. The social and behavioral sciences have become more prominent in public health since the recognition that such factors as indolence, loneliness, personality type, and addiction to tobacco contribute to the risk of premature death and chronic disabling diseases. See EPIDEMIOL-OGY.

History and evolution. Some aspects of public health practice have ancient roots in empirical observations dating back to biblical times; for example, lepers carried bells to warn others of potential contagion, and simple sanitary rules were indicated. The Romans also practiced sanitary precautions on military campaigns 2000 years ago, but the nature of their plumbing and water supply systems showed a lack of understanding of the role of polluted water in the transmission of disease. Quarantine began in Europe in the fourteenth century and led to public health laws aimed at protecting populations against infection. Many basic principles of public health law were codified by the Austro-Hungarian physician J. P. Frank in 1779. In 1798, the English physician Edward Jenner discovered a vaccine to prevent smallpox. This find led to the eventual elimination of the disease, declared eradicated by the World Health Organization in 1980-the greatest achievement of coordinated worldwide public health action. See SMALL-POX

Effective sanitary systems, supported by laws and regulations, did not become widespread until the late nineteenth century. Food hygiene followed in the early twentieth century. Medical discoveries in the late nineteenth and early twentieth centuries, beginning with the work of the great microbiologists Robert Koch and Louis Pasteur, led to the development of sera and vaccines to treat and prevent many infectious diseases. Biochemists and physiologists worked out the essential components of a healthy diet. Geneticists identified inherited diseases and conditions due to disorders of genes and chromosomes, laying the foundations for genetic screening and counseling programs that would eventually lead to prevention of some of these conditions. The importance of life-styles in the cause of heart disease and cancer was recognized in the 1960s. All these developments have helped to improve public health. *See* HUMAN GENETICS; NUTRITION; VACCINATION.

Services. It is possible to protect, preserve, and restore good health by making the environment safe, fostering sensible behavior, immunizing against infections, maintaining good nutrition, and providing prudent health care, including prenatal care. All these are part of the work of public health services.

In most industrial nations, public health services are organized nationally, regionally, and locally. National public health services exist primarily to protect the nation's health. They are usually responsible for setting, monitoring, and maintaining health standards, for promoting good health (for example, through national fitness and health education programs), for collecting and compiling national health statistics, and for supporting and performing research on diseases important to public health. Regional (for example, state) public health services deal mainly with major health protection activities such as ensuring safe water and food supplies; they may also operate screening programs for early detection of disease and are responsible for health care of certain groups such as chronic mentally ill persons. Local public health services (in cities, large towns, and some rural communities) conduct a variety of personal public health services, such as immunization programs, health education, health surveillance and advice for mothers and newborn babies, and personal care of vulnerable groups such as the elderly and housebound long-term sick. Local health services also investigate and control epidemics and other communicable conditions such as sexually transmitted diseases.

National public health services communicate with each other in efforts to control diseases of international importance, and they collaborate worldwide under the auspices of the World Health Organization (WHO). While much of the work of WHO has been concentrated in the developing nations, it has also been involved in global efforts to control major epidemic diseases and to set standards for hazardous environmental and occupational exposures. In tropical and developing nations, WHO conducts many programs aimed at controlling and, if possible, eliminating such diseases as malaria, schistosomiasis, infant diarrhea, tetanus, and measles. Premature death and long-term disabling diseases are still serious problems worldwide and in particular in the developing world.

The ultimate purpose of public health is to improve the human condition. Much has been accomplished, but much remains to be done, especially to improve the health of the disadvantaged, minority groups, the poor, and, above all, people in the developing nations. *See* MEDICAL BACTERIOLOGY; MEDICAL MYCOLOGY; MEDICAL PAR-ASITOLOGY. John M. Last

Bibliography. W. W. Holland, R. Detels, and G. Knox (eds.), Oxford Textbook of Public Health, 3d ed., 1997; J. M. Last (ed.), Maxcy-Rosenau Public Health and Preventive Medicine, 14th ed., 1998; J. M. Last, Public Health and Human Ecology, 2d ed., 1998; World Health Organization, Global Strategy for Health for All by the Year 2000, 1981.

Pulley

A wheel with a flat, crowned, or grooved rim (**Fig. 1**) used with a flat belt, V-belt, or a rope to transmit motion and energy. Pulleys for use with V-belt and rope drives have grooved surfaces and are usually called sheaves. A combination of ropes, pulleys, and pulley blocks arranged to gain a mechanical advantage, as for hoisting a load, is referred to as block and tackle. *See* BELT DRIVE; BLOCK AND TACKLE.

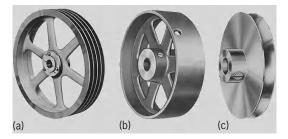


Fig. 1. Some typical examples of pulleys. (a) Multigroove V-belt sheave. (b) Solid-hub crown face pulley. (c) Brass pulley for round belt (*Boston Gear, INCOM International, Inc.*).

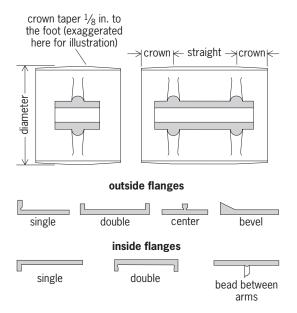


Fig. 2. Diagrams of crowning and flanging on flat-belt pulleys. $1_{/6}$ in. to the foot \cong 1 cm to the meter. (After D. C. Greenwood, Mechanical Power Transmission, McGraw-Hill, 1962)

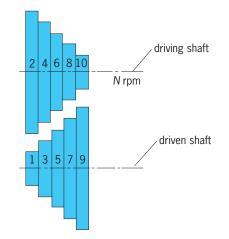


Fig. 3. Equal-stepped pulleys.

Flat-belt pulleys. Pulleys for flat belts are made of cast iron, fabricated steel, wood, and paper. A particular pulley design must be based on such considerations as the ability to resist shock, to conduct heat, and to resist corrosive environments. The face must be smooth enough to minimize belt wear; yet there must be adequate friction between belt and pulley face to carry the load.

Cast-iron pulleys are made in sizes from 3 to 72 in. (7.5 to 183 cm) and are particularly well adapted for service where dampness or acid fumes are present, in contrast to wood or steel pulleys. Because of their greater weight they are useful wherever a flywheel effect is wanted. Available in a wide variety of hub and spoke designs, they can also be furnished in either flat, crowned, or flanged rims. Pulleys are crowned (**Fig. 2**) to keep the belt on the pulley. The amount of crowning seldom exceeds a taper of 1/8 in./ft (1 cm/m); excessive crown is harmful to the belt and may cause the belt to be thrown off the pulley as quickly as if there were no crown at all. Outside flanging (Fig. 2) helps hold the belt on the pulley; inside flanging stiffens the rim.

The arms on cast-iron pulleys are typically elliptical in cross section to achieve both a pleasing appearance and low windage loss, even though an H-section arm can provide equal strength for less material. Commercial cast-iron pulleys are balanced for rim speeds of 3500-4000 ft/min (18-20 m/s). For higher speeds the pulleys have to be specially balanced. Stock pulleys are made to withstand belt tensions of up to 65 lb/in. (115 newtons/cm) of width, though they can be furnished for tensions up to 210 lb/in. (370 newtons/cm).

The majority of pulleys used in flat-belt drives are steel split pulleys. They are low in cost, light in weight, and easily available, and are made with interchangeable bushings which permit their use on any shaft diameter up to the bore size of the pulley itself. Standard sizes range from 3-in. (7.5-cm) diameter by 2-in. (5-cm) width up to 144-in. (366-cm) diameter by 36-in. (91-cm) width.

Where lightness is important and moisture is not present, wood pulleys can be used. The resilience of wood makes these pulleys able to withstand shock

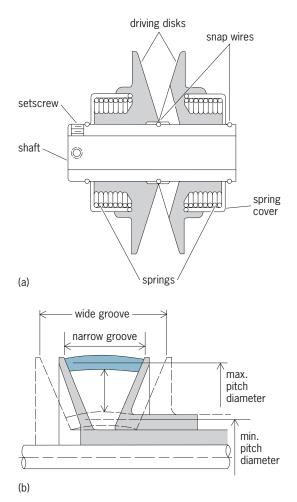


Fig. 4. Typical variable-pitch V-belt pulley (a) for standard-section V-belts and (b) for wide V-belts. Pulley can accommodate up to six belts. (*Lovejoy, Inc.*)

loads very well. Layers of soft wood are glued together and then compressed under high pressure to increase the density. A ribbed cast-iron hub is forced into an undersize hole in the wood block to complete the construction. Belt speeds of up to 6000 ft/min (30 m/s) are common.

Paper pulleys are made of impregnated paper board and are used particularly for motor pulleys. The recommended rim speed is 5000 ft/min (25 m/s).

If a belt drive must be able to produce several output speeds from a single input speed, stepped pulleys (**Fig. 3**) can be used. It is good practice to select the several output speeds so that they form a geometric progression. For a crossed-belt drive, the sum of the pulley diameters must be the same for each step. In practice a pair of stepped pulleys is often designed so that they have the same dimensions (Fig. 3). As shown in the equation below,

$$N = (n_3 n_7)^{1/2}$$

the driving shaft speed *N* is then the geometric mean of the output speeds of the two steps adjacent to the middle one.

V-belt pulleys. The two common types of V-belt pulley are pressed-steel and cast-iron. The pressed-

steel pulleys are suitable for single-belt drives. For multiple-belt drives, or in single-belt drives where pulley mass should be high to get a flywheel effect, cast-iron pulleys are used (Fig. 1*a*). Generally the pressed-steel pulleys have noninterchangeable hubs, but the cast-iron pulleys have interchangeable hubs. Interchangeability of hubs reduces the number of pulleys that need be kept in stock: A single pulley can be made to fit a range of shaft diameters. Small pulleys have a groove angle of 30° to increase the wedging action of the belt in the groove. For pulley diameters of 5–8 in. (12.5–20 cm) a 34° angle is recommended. Over 8 in. (20 cm) a 38° angle is used.

Variable-pitch pulleys (**Fig. 4**) are made for drives where it is necessary to vary the speed ratio of the drive continuously over a range. Made for capacities up to 5 hp (3.7 kW), they can accommodate one to six belts.

Rope sheaves. Rope drives are much less efficient than V-belt drives, and so V-belt drives have virtually replaced rope drives. But in cranes and power shovels wire-rope drives are still in use. If a relatively soft material such as gray cast iron is used for the sheave, the wire rope will leave its imprint in the groove. A cast-steel sheave will stand up satisfactorily under the rope pressures experienced in typical installations, and heat-treated alloy steel, hardened to about 350 Brinell, will stand very high loads.

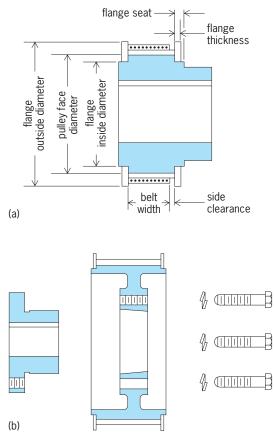


Fig. 5. Timing-belt pulley. (a) External plate-type flanges are used to retain belt. (b) Standard mounting arrangement permits various pulleys to be used on same shaft diameter. (*T. B. Wood's Sons Co.*)

Timing-belt pulleys. A timing-belt pulley (**Fig. 5**) is rather like an uncrowned flat-belt pulley, except that the grooves for the belt's teeth are cut in the pulley's face parallel to the axis, using a milling cutter, a gear shaper, or a hob. Most of the materials used in the manufacture of gears can be used as well in making timing-belt pulleys. Steel and cast iron are the most common materials, particularly for heavy loads. For light loads aluminum alloys, brass, plastics, and zinc or aluminum die castings are used. External flanges are used to keep the belt on the pulley. John R. Zimmerman

Bibliography. E. A. Avallone and T. Baumeister III (eds.), *Marks' Standard Handbook for Mechanical Engineers*, 10th ed., 1996; R. T. Kent, *Mechanical Engineer's Handbook*, pt. 1: *Design and Production*, 12th ed., 1950; W. Patton, *Mechanical Power Transmission*, 1979; Trade and Technical Press Editors, *Handbook of Mechanical Power Drives*, 1977.

Pulmonata

A subclass of the molluscan class Gastropoda, containing about 23,500 species of snails that are grouped into three orders (or superorders): Systellommatophora, Basommatophora, and Stylommatophora. Pulmonates include most of the common snails found on land or in fresh waters, although a few prosobranchs have invaded both habitats. Certain pulmonates, the Onchidiidae, Amphibolidae, Otinidae, Siphonariidae, Trimusculidae, and some members of the Ellobiidae are intertidal to subtidal on rocky shores of tropical and temperate regions. *See* BASOMMATOPHORA; SYSTELLOMMATOPHORA.

Comparatively few characteristics are common to most species, and there are exceptions to almost every statement. Pulmonates usually are hermaphroditic with cross-fertilization, but many inhabitants of temporary ponds, for example, can selffertilize as an aid to colonization of new areas. All pulmonates, except the Amphibolidae, lack an operculum. Instead of an open mantle cavity with a gill, fusion of the mantle to the neck region has resulted in the formation of a pulmonary sac with a heavily vascularized surface and a small pneumostomal opening; the mantle cavity has become a lung. However, pseudobranchs (false gills not homologous with the ctenidia of other mollusks) are secondarily developed in some fresh-water taxa. Normally there is a spiral shell into which the animal can withdraw completely, but reduction of the shell to an external or internal remnant plate, or even complete loss of the shell, is a common evolutionary pattern in certain land mollusks. Such shell-less snails are called slugs. The animal is detorted, but in a different way from that in opisthobranchs. The single auricle lies (except in some slugs) in front of the ventricle. The size of pulmonates ranges from 0.04 to 9 in. (1 to 225 mm) in maximum adult shell size.

Most pulmonates are secretive dwellers in forests

or nonsilted waters. They reach their highest levels of diversity in well-watered limestone hill regions and on relatively isolated islands. A few fresh-water species are of medical importance because they act as intermediate hosts for parasitic worms whose final hosts are humans or domestic animals. A very few land snails and slugs occasionally are pests in gardens or on crops. The vast majority of pulmonate species are beneficial in that they mechanically break up the dead parts of plants into pieces small enough to be quickly reduced by bacteria and fungi into chemicals that can be reused by plants to build new organic matter.

The earliest known pulmonates are stylommatophorans from the Late Carboniferous of eastern North America. All of these species belong to extant families of land snails. The first basommatophorans do not appear in the fossil record until the Jurassic-Cretaceous boundary, almost 1.5×10^8 years after the appearance of the stylommatophorans.

Fresh-water pulmonates belong to the superorder Basommatophora and are adapted to make use of accidental transport on birds or insects from one isolated body of water to another. Mostly they have short life-spans and their numbers vary dramatically with the seasons.

Land pulmonates belong to the superorder Stylommatophora and are marginally terrestrial in that they can be active only when the humidity in their microhabitat is 90% or above. Thus many are cryptic or nocturnal in both feeding and reproduction. They can survive in deserts by estivating for up to 6 years. *See* GASTROPODA. G. Alan Solem

Bibliography. R. C. Brusca and G. J. Brusca, *Invertebrates*, 2nd ed., Sinauer Associates, Sunderland, MA, 2002; R. Chase, *Behavior and Its Neural Control in Gastropod Molluscs*, Oxford Univ. Press, 2002; V. Fretter and J. Peake, *Pulmonates*, vol. 2A, 1978; C. P. Hickman and L. S. Roberts, *Animal Diversity*, Wm. C. Brown, Dubuque, IA, 1994; L. Hyman, *The Invertebrates*, vol. 6: *Mollusca I*, 1967; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982; A. Solem, *The Shell Makers*, 1974.

Pulsar

A celestial radio source producing intense short bursts of radio emission. Since their discovery in 1968, over 1750 pulsars have been found (as of October 2006), and it has become clear that 100,000 pulsars must exist in the Milky Way Galaxy—most of them too distant to be detected with existing radio telescopes. *See* RADIO ASTRONOMY.

Pulsars are distinguished from most other types of celestial radio sources in that their emission, instead of being constant over time scales of years or longer, consists of periodic sequences of brief pulses. The interval between pulses, or pulse period, is nearly constant for a given pulsar, but for different sources ranges from 0.0014 to 8.5 s. The bursts of emission are generally confined to a window whose width is a few percent of the period (**Fig. 1**). Individual pulses

can vary widely in intensity; however, their periodic spacing is accurately maintained.

Neutron stars. The association of pulsars with neutron stars, the collapsed cores left behind when moderate- to high-mass stars become unstable and collapse, is supported by many arguments. Prior to the discovery of pulsars, neutron stars were thought to be unobservable owing to their extremely small size (radius less than 15 km or 10 mi). The standard model for pulsars is a spinning neutron star with an intense dipole magnetic field (surface field of 108 teslas or 10^{12} gauss) misaligned with the rotation axis. The off-axis rotating dipole field develops a huge voltage difference between the neutron star surface and the surrounding matter. Charges accelerate in this voltage and generate an avalanche of electrons and positrons, a relativistic current leaving the polar zones of the star. Highly directive radio emission is formed in this current, which is observed as pulses, one per rotation, just like a rotating searchlight. See ELECTRON; POSITRON.

A fundamental observation that supports the rotating neutron star model is the remarkable stability of the basic pulsation periods, which typically remain constant to a few tens of nanoseconds over a year. This stability is natural to the free rotation of a compact, rigid object like a neutron star, but is extremely difficult to produce by any other known physical process. A small emitting body is also required by the observed rapid variations within pulses. A spinning, magnetized body will gradually slow down because it emits low-frequency electromagnetic radiation at harmonics of its rotation period. This slowing down is observed in all objects where observations of sufficient precision are available.

Further support of the neutron star model comes from the discovery of a 33-millisecond pulsar in the center of the Crab Nebula, the remnant of the supernova observed in A.D 1054. The lengthening of this pulsar's short period on a time scale of 1000 years matches the age of the remnant. The rotating, magnetized neutron star model is further supported by observations of diffuse radiation near the pulsar in the Crab Nebula by the Hubble, XMM-Newton, and Chandra space observatories. Currently, about 20 associations have been found between pulsars and supernova remnants which lend further support to these ideas. A 66-ms pulsar has been found in the supernova remnant 3C58. This is likely associated with the historical supernova of 1181, making this the youngest radio pulsar currently known. See CRAB NEBULA; NEUTRON STAR; SUPERNOVA.

Pulsed emission. The waveform of a pulsar's periodic emission, averaged over several hundred pulses or more, has a distinctive shape characteristic of each individual source (**Fig. 2**). Variations in pulse shape from one pulsar to another are probably the result of differences in detailed structure of the neutron star's magnetic field, or of a different orientation of the rotating beam relative to the Earth, or both. The emission is usually highly polarized, and the angle of the polarization vector (projected onto the plane of the sky) rotates smoothly across the pulse profile.

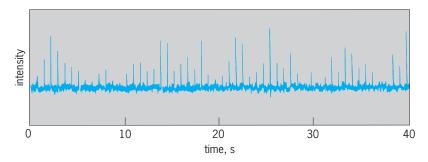


Fig. 1. Chart recording of radio emission from the pulsar PSR B0329+54, observed at a frequency of 400 MHz. This pulsar lies at a distance of about 7500 light-years (7 \times 10¹⁶ km or 4.5 \times 10¹⁶ mi) from Earth.

This behavior is taken to be the straightforward result of the changing projection of the pulsar's magnetic field lines as the star spins relative to the line of sight.

Individual pulses are found to vary rapidly in intensity, shape, and polarization (**Fig. 3**). Individual pulses normally consist of one or more subpulses, and the subpulses may in turn contain microstructure of even shorter duration. In some pulsars, subpulses vary in intensity in an obviously nonrandom way, and sometimes exhibit secondary periodicities incommensurate with the basic pulsar period. Individual pulses are more highly polarized than the average pulse profiles, and frequently undergo abrupt changes of polarization angle of almost exactly 90°. *See* POLARIZATION OF WAVES.

Spectra and links to other wavelengths. With a few notable exceptions, pulsars are observable only at radio frequencies, typically between 50 and 5000 MHz. Over this range the average flux density decreases rapidly, usually in approximate proportion to f^1 or f^2 , where *f* is the observing frequency. At frequencies below a few hundred megahertz, a combination of galactic background radiation and terrestrial interference makes observations very difficult; at frequencies above a few thousand megahertz, most pulsars become too weak to be observed with existing instruments.

When extrapolated to optical or higher frequencies, the steep radio-frequency spectra of pulsars predict flux densities far below detectable limits. However, a few of the youngest pulsars, including the pulsar in the Crab Nebula, have been detected at optical, x-ray, and gamma-ray frequencies. Evidently, another emission mechanism is responsible for the high-frequency radiation from these pulsars. A mysterious gamma-ray source called Geminga has pulsed optical, x-ray, and gamma-ray emission but no convincing radio counterpart. This may be a case where the narrow radio beacon does not intercept the Earth. *See* GEMINGA.

Equally mysterious is the discovery of magnetars, pulsars that emit x-rays and gamma rays but may be radio-quiet owing to extreme values of magnetic field that suppress the electron-positron avalanche. However, radio emission has been detected during the outburst of a transient 5.5-s magnetar J1819-197. This remarkable discovery has provided a firm link between radio pulsars and magnetars, and challenges

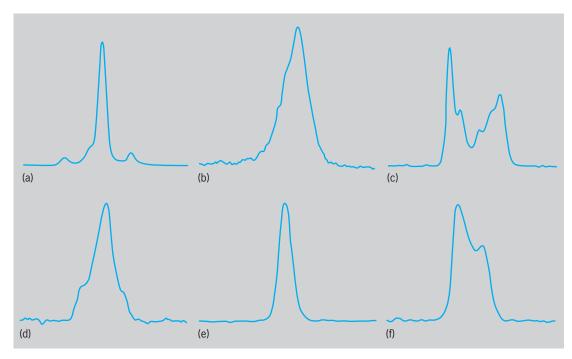


Fig. 2. Average pulse shapes of six pulsars. The data plotted cover a small, roughly 10% fraction of one rotation period. (a) PSR B0329+54. (b) PSR B0950+08. (c) PSR B1237+25. (d) PSR B1508+55. (e) PSR B1642-03. (f) PSR B1919+21.

the details of current theories of radio pulsar emission.

The emissions of most neutron stars are powered by the star's rotation. In contrast, there are accretionpowered neutron stars where matter from the atmosphere of a gravitationally bound companion star heats up to x-ray emission temperatures as it falls into the deep gravitational well of the neutron star. The infalling matter interrupts the process responsible for radio emission. *See* ASTROPHYSICS, HIGH-ENERGY; BINARY STAR; X-RAY ASTRONOMY.

Interstellar medium. Pulsars have provided a unique set of probes for the investigation of the diffuse gas and magnetic fields in interstellar space. Measurement of absorption at 1420 MHz, the frequency of the hyperfine transition in ground-state neutral hydrogen atoms, gives information on the structure of gas clouds, and in many cases provides an estimate of the pulsar distance. The index of refraction for radio waves in the ionized interstellar gas is strongly frequency-dependent, and low-frequency signals propagate more slowly than those at high frequencies. The broadband, pulsed nature of pulsar signals makes them ideal for measurements of this dispersion. Measurements of dispersion when coupled with independent distance estimates lead to both a model of the electron density in the nearby interstellar medium and the pulsar distribution. The average electron density is 0.03 cm⁻³, and most pulsars are within 30,000 light-years of the Earth (1 light-year equals 9.46×10^{12} km or 5.88×10^{12} mi). The interstellar electrons are not distributed uniformly on small length scales, and their turbulent structures disrupt the observed signals with a variety of diffractive and refractive effects that are related to the twinkling of optical stars. See TWINKLING STARS.

Because of the large-scale Galactic magnetic field, the index of refraction in the interstellar medium is different for the two circular components of a linearly polarized wave. For this reason, the plane of polarization rotates along the propagation path, an effect known as Faraday rotation. Measurements of this effect for over 400 pulsars have shown that the magnetic field strength in the solar neighborhood of the Milky Way Galaxy is approximately 2×10^{-10} T (2×10^{-6} G) in a direction along the local spiral arm. The direction of the magnetic field along other spiral arms has also been measured using this technique, which provides strong evidence for field direction reversals in the inner Galaxy. *See* FARADAY EFFECT; INTERSTELLAR MATTER; MILKY WAY GALAXY.

Galactic distribution. The whole sky has been searched for pulsars using a variety of radio telescopes to carry out surveys at different observing frequencies. Figure 4 shows the distribution of all currently known pulsars in galactic coordinates. The concentration of pulsars within $\pm 10^\circ$ of the galactic equator and in the hemisphere toward the galactic center (longitudes $0-90^{\circ}$ and $270-360^{\circ}$) demonstrates that pulsars are members of the Milky Way Galaxy and that their distances are large compared to the thickness of the galactic disk. When distance estimates are combined with the two-dimensional information, a more elaborate analysis shows that pulsars occupy a disk-shaped region of space approximately 2000 light-years thick and 40,000 light-years in radius. The Sun, at a distance of about 30,000 lightyears from the galactic center, lies near the outer edge of the pulsar distribution. Twenty pulsars are found in the Magellanic Clouds, while a further 129 pulsars have been associated with galactic globular clusters. See MAGELLANIC CLOUDS.

The massive, luminous stars that are believed to be the progenitors of pulsars have a galactic distribution similar to that of the pulsars, except that the thickness of their disk is only about 500 light-years. It appears that pulsars are found at much larger distances from the galactic plane because most pulsars acquire a substantial velocity (300 km s⁻¹ or 200 mi s⁻¹ or more) at birth, as a result of the supernova explosion. During a typical pulsar's active lifetime, roughly 10⁷ years, the pulsar has time to move as much as 10,000 light-years away from its birthplace.

If pulsars typically remain active for 10^7 years, and if the present inferred galactic population of 100,000 pulsars is to be maintained, then a new pulsar must be formed somewhere in the Galaxy approximately every 100 years. This birth rate is reasonably close to the estimated rate of occurrence of galactic supernovas, if allowance is made for those that are optically obscured from view by interstellar dust.

Millisecond pulsars. The view that few if any pulsars spin faster than the 30-ms pulsar in the Crab Nebula was shattered with the discovery in 1982 of PSR B1937+21 with the remarkably short period of 1.6 ms, the first millisecond pulsar. This period is within a factor of 2 or so of the stability limit of neutron stars of 1 solar mass. The period is lengthening, but on time scales of 108 years, not 106 years or less. Evidently, this object is old and has a low magnetic field. About 150 millisecond-period pulsars are now known. Despite intensive searches over the past 20 years, PSR B1937+21 remains the most rapidly rotating neutron star known in the galactic disk. In 2005, a pulsar was found in the globular cluster Terzan 5 with a spin period of 1.4 ms, that is, a spin rate of 716 Hz. It is an open question as to whether a significant number of pulsars spinning even more rapidly exist.

Millisecond pulsars are extremely accurate clocks which exhibit few or none of the rotation irregularities seen in most pulsars. The observed stabilities rival the best terrestrial atomic time scale, which is derived from cesium clocks maintained around the world. Timing an array of millisecond pulsars will provide the best time scale for durations exceeding 1 year. The data may also improve knowledge of the orbit of the Earth in the solar system, and be used to set limits on, or even detect, a background of gravitational radiation left over from chaotic events in the early universe. *See* ATOMIC CLOCK; ATOMIC TIME; GRAVITATIONAL RADIATION.

Binary pulsars. Over 120 pulsars with orbiting companions are now known. Orbital periods range from hours to years. Timing observations of the binary pulsars, which provide indications of the companion masses, together with optical observations which often identify or constrain the nature of the companion star, show that there are several different types of orbiting companion: planets, white dwarfs, other neutron stars, and main-sequence stars.

The origin of most binary pulsars can be understood by following the evolution of a pair of mainsequence stars. After the supernova explosion of

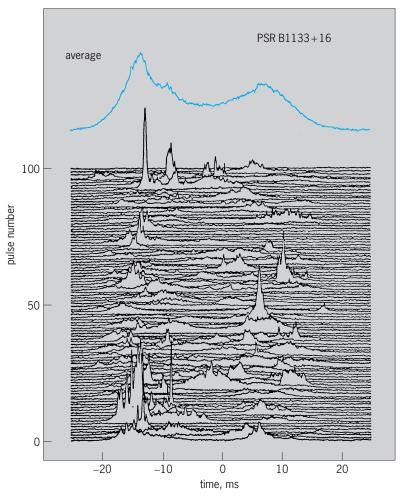


Fig. 3. Individual pulses from pulsar PSR B1133+16, illustrating the variations which take place within pulses and from one pulse to the next.

the initially more massive (primary) star, the resulting neutron star initially behaves normally—spinning down to longer periods. The spin evolution changes, however, when the initially less massive (secondary) star in the system reaches the end of its mainsequence lifetime and enters a red giant phase. For favorable orbital configurations, the strong gravitational field of the neutron star can draw material from the evolved secondary. Accretion of mass and angular momentum from the secondary can spin-up the neutron star to millisecond periods. This process is known as recycling, and the resulting rapidly spinning neutron stars are often termed recycled pulsars. *See* GIANT STAR; STELLAR EVOLUTION.

If the secondary star is massive enough to also explode as a supernova and the binary system survives this explosion, a double neutron star system may be formed. For less massive secondary stars, a white dwarf is formed at the end of the red giant phase. Tidal circularization of the orbit during the recycling phase means that the resulting system is observed as a millisecond pulsar in a circular orbit around a white dwarf. *See* WHITE DWARF STAR.

While the spin-up scenario explains the currently observed spin parameters of most binary and millisecond pulsars, PSR B1937+21 and a few others

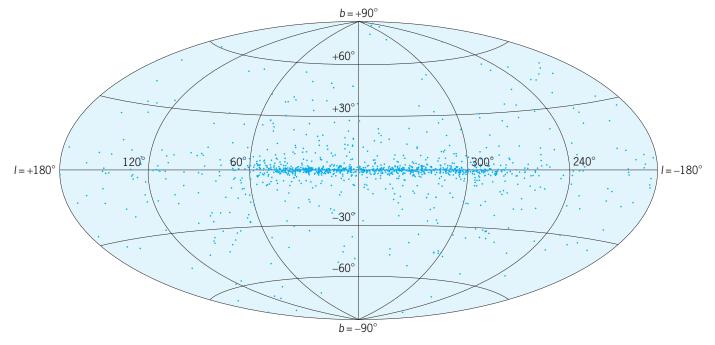


Fig. 4. Distribution of pulsars in Galactic coordinates, showing all pulsars known in late 2006.

are still a puzzle since they have no companion. Several binary millisecond pulsars show eclipses by matter surrounding the companion, which suggests that a pulsar can ablate its companion out of existence. These systems may be the missing links between binaries and solitary millisecond pulsars.

Double pulsar system. Currently (2006), eight examples of double neutron star binary systems are known. In all but one of these systems, only one of the neutron stars has its emission beam passing over the Earth. In 2003, the first system where both pulsar beams are visible was found. This double pulsar binary, PSR J0737-3039, consists of a 22-ms pulsar in a highly relativistic 2.4-h orbit with a 2.7-s pulsar companion. This compact orbital system is viewed almost edge-on to the line of sight and has allowed observers to study the interaction of two radio pulsar beams for the first time. Two effects have been seen so far: the eclipse of the 22-ms, pulsar by the 2.7-s pulsar once per orbit and the modulation of the 2.7-s pulsar by the radiation beam of the 22-ms pulsar at specific orbital phases. Further studies of these interactions may help provide a better understanding about the pulsar radiation mechanism in general.

The pulsars in this unique binary have orbital speeds about one-thousandth the speed of light. The large speeds and intense gravitational fields of the two neutron stars, combined with the accurate clock mechanism provided by both components, make this system an ideal testing ground for relativistic gravity. In addition to the five classical (Keplerian) orbital parameters determined for this system and the mass ratio of the two stars, a further five relativistic (so-called post-Keplerian) parameters have been measured so far. Among the many effects seen in this system is the precession of periastron which occurs at a rate of almost 17° per year—a factor of 10^{5} times

larger than the general relativistic perihelion advance observed for Mercury.

Within 3 years of its discovery, the double pulsar system surpassed the previous two best double neutron star systems as a precision test of general relativity. Currently (2006) the comparison between observed and predicted post-Keplerian parameters shows that general relativity is correct at the 99.95% confidence level. Ongoing timing measurements should provide even more stringent constraints and new tests of Einstein's theory.

One especially important prediction of the general theory of relativity is that a close binary star system should gradually lose energy by the radiation of gravitational waves, and consequently the two stars should slowly spiral closer together. Observations of double neutron star binary systems show just such an effect, and provide the most stringent experimental evidence available in support of the existence of gravitational waves.

The fate of close double neutron star binaries is a furious coalescence event that will produce a burst of gravitational radiation detectable by Earth-based gravity-wave telescopes now in operation. Although double neutron star mergers in the local Universe are not thought to be frequent enough to be readily observable by current detectors, planned upgrades in sensitivity are expected to allow the direct detection of gravitational waves within the next decade. *See* GRAVITATION; RELATIVITY.

Globular-cluster pulsars. Stars in the centers of globular clusters are only a few light-weeks apart so that frequent stellar encounters occur. Neutron stars in clusters have a 10% chance of capturing a companion during the life of the cluster. This constant production of binaries can lead to an abundance of x-ray objects as well as millisecond pulsars.

Currently, there are 129 pulsars known in 44 globular clusters. Studies of this population exploit the clocklike stability of pulsars to probe the physical properties of the pulsars and the cluster itself. For example, in the cluster 47 Tucanae, in which there are currently 22 pulsars known, timing measurements which map out the pulsar positions in three dimensions have provided the first evidence for the presence of ionized gas in any globular cluster.

The richest cluster, Terzan 5, has 33 pulsars with rotation periods between 1.4 and 80.3 ms. Studies of these pulsars are being used to probe the pulsar properties and overall mass distribution of Terzan 5. Of particular interest is a so-far unidentified radio source in the cluster. This enigmatic source may be a bright millisecond pulsar that has evaded detection by virtue of being in an extremely compact orbit, perhaps around a black hole. The identification of such a system is a high-priority objective in pulsar astronomy, and Terzan 5 may be the place where the system can be found. *See* STAR CLUSTERS. Duncan R. Lorimer; Donald C. Backer

Bibliography. M. Bailes, D. J. Nice and S. E. Thorsett (eds.), Radio Pulsars, Astron. Soc. Pacific Conf. Ser. 302, 2003; D. R. Lorimer, Binary and millisecond pulsars, Living Rev. Relativity, vol. 8, no. 7, 2005; D. R. Lorimer and M. Kramer, Handbook of Pulsar Astronomy, 2005; A. G. Lyne and M. Kramer, Gravitational labs in the sky, Phys. World, 18(3):29-32, March 2005; A. G. Lyne and F. G. Smith, Pulsar Astronomy, 2006; T. Piran, Binary neutron stars, Sci. Amer., Spec. Ed., pp. 84-91, 2004; F. A. Rasio and I. H. Stairs (eds.), Binary Radio Pulsars, Astron. Soc. Pacific Conf. Ser. 328, 2004; F. G. Smith and M. A. McLaughlin, A magnetopause in the double-pulsar binary system, Astron. Geophys., 46:23-26, 2005; E. P. J. van den Heuvel and J. van Paradijs, X-ray binaries, Sci. Amer., Spec. Ed., pp. 58-67, 2004.

Pulse demodulator

A device that extracts the analog information from a received pulse or from a set of pulses.

Constant-amplitude pulse-width-modulated (PWM) and pulse-position-modulated (PPM) signals carry their information in the time interval that the pulse is "on" in the first case, and in the time interval between the pulse of interest and the framing or timing pulse in the second case. PPM signals can be converted to binary PWM signals (amplitudes of "V" volts when on, amplitude zero when off) by using the signal pulses to latch a gated V-volt signal "on" and the timing signal to turn the gate "off." The source information is in the average value of the resulting PWM signal; therefore, the following source-information-recovery circuit can be a simple low-pass filter whose cutoff frequency must lie above the bandwidth of the source signal to preserve the information content, but below $1/_2$ the frequency of the framing or timing signal to reject the timing-pulse noise. The source analog information behind ternary PWM signals of amplitudes +V, 0, and -V is recovered in the same manner. *See* ELECTRIC FILTER; LOGIC CIRCUITS; PULSE MODULATOR.

Pulse-amplitude-modulated (PAM) signals are pulses of constant width, and usually the product of the source signal and a sampling-pulse train of frequency f_s . The sampling pulses may be regarded as windows through which the source signal is observed. The information is carried in the average amplitude of the pulses. Recovery again is achieved by simple low-pass filtering. The filter cutoff frequency must be above the information bandwidth of the source signal and less than $1/_2$ the pulse-repetition frequency f_s of the sampling-pulse train in order to avoid foldover distortion.

Pulse-density-modulated (PDM) signals consist of trains of constant-amplitude and constant-duration pulses with variable spacing; therefore, those pulse rates are proportional to the amplitudes of the source signals. The analog source information is again recovered by a simple low-pass filter.

Pulse-code-modulated (PCM) signals are digital representations of their source signals. A conventional digital-to-analog converter (DAC) recovers the analog source signal. Variations on PCM such as delta Modulation (DM), adaptive delta modulation (ADM), and differential PCM (DPCM), all of which have been introduced to improve channel efficiency, require specialized and often complicated recovery circuits. These specialized circuits may be considered as performing a two-step process: recovery of the basic PCM signal followed by a conventional digital-to-analog converter. *See* DIGITAL-TO-ANALOG CONVERTER.

Phase-modulated (PM) signals, sometimes called PSK, or phase-shift keyed, are constant-amplitude sinusoidal segments shifted in phase from a reference sinusoid according to the amplitude of the source signal. The source signal may be recovered by synchronous detection (multiplication) with a reference sinusoid. *See* PHASE MODULATION.

Frequency-modulated (FM) signals are constantamplitude sinusoids whose frequencies are varied in accordance with the amplitude of a reference sinusoid. A pulse containing a burst of FM data at known discrete frequencies can be demodulated by applying that pulse to a bank of filters tuned to the known discrete frequencies. The power appearing at the filter outputs signifies which frequencies are present in the pulse. FM signals can also be recovered by a phase-locked loop (PLL), but not as quickly. Integration of the recovered FM signal yields PM recovery, an alternative to the method described above. *See* FREQUENCY MODULATION; MODULATION; PHASE-LOCKED LOOPS.

Quadrature-amplitude-modulated (QAM) signals appearing as a burst within a pulse must be synchronously demodulated in two parallel channels by reference signals in quadrature with one another. Subsequent processing of each of the two channels involves integration, digitizing, bit formatting, and data-stream interleaving in order to recover the PCM-formatted data. Finally, a pair of digital-to-analog converters is required to recover the analog source signals. *See* DEMODULATOR; MODULA-TOR. Stanley A. White

Bibliography. G. A. Breed (ed.), Wireless Communications Handbook, Cardiff Publishing, Englewood, CO, 1992; S. Gibilisco, Handbook of Radio and Wireless Technology, McGraw-Hill, New York, 1999; J. G. Proakis, Digital Communications, 4th ed., McGraw-Hill, New York, 2001; T. S. Rappaport, Wireless Communications, 2d ed., Prentice Hall, Upper Saddle River, NJ, 2002; M. S. Roden, Analog and Digital Communications Systems, 4th ed., Prentice Hall, Englewood Cliffs, NJ, 1995; M. E. Van Valkenburg and W. M. Middleton (eds.), Reference Data for Engineers: Radio, Electronics, Computer and Communications, 9th ed., Butterworth-Heinemann, Woburn, MA, 2002.

Pulse generator

An electronic circuit capable of producing a waveform that rises abruptly, maintains a relatively flat top for an extremely short interval, and then rapidly falls to zero. A relaxation oscillator, such as a multivibrator, may be adjusted to generate a rectangular waveform having an extremely short duration, and as such it is referred to as a pulse generator. However, there is a class of circuits whose exclusive function is generating short-duration, rectangular waveforms. These circuits are usually specifically identified as pulse generators. An example of such a pulse generator is the triggered blocking oscillator, which is a single relaxation oscillator having transformer-coupled feedback from output to input. *See* MULTIVIBRATOR.

Pulse generators sometimes include, but are usually distinguished from, trigger circuits. Trigger circuits, by means of *RC* (resistance-capacitance) differentiating, gated *RLC* (resistance-inductancecapacitance) peaking circuits, or blocking oscillators, generate a short-duration, fast-rising waveform for initiating or triggering an event or a series of events in other circuits, such as monostable or bistable multivibrators. In the pulse generator, the pulse duration and shape are of equal importance to the rise and fall times. In this sense the blocking oscillator is a circuit which can be made to perform well in both respects. *See* TRIGGER CIRCUIT.

Digitally controlled pulse generator. The term pulse generator is often applied not only to an electronic circuit generating prescribed pulse sequences but to an electronic instrument designed to generate sequences of pulses with variable delays, pulse widths, and pulse train combinations, programmable in a predetermined manner, often microprocessorcontrolled.

For example, in the circuit shown in **Fig. 1**, the basic pulse generator is a bistable *R-S* or *J-K* flipflop under control of a sequence of clock pulses with a digital logic block to determine the appropriate combinations of set (*S*), reset (*R*), and clock pulses. An output pulse can be programmed to start with the next clock pulse following R = (1) and stop with the clock pulse following S = (1). Thus, the pulse width which can be pro-

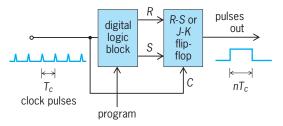


Fig. 1. Digitally controlled pulse generator.

grammed is nT_c , where T_c is the clock pulse interval and n is any number equal to or greater than unity.

Pulse-forming networks. A network, formed in such a way as to simulate the delay characteristics of a lossless transmission line, and appropriate switching elements to control the duration of a pulse form the basis for a variety of types of pulse generators. *See* DELAY LINE.

A lossless transmission line has a characteristic impedance, given by Eq. (1), where *L* is the series

$$R_0 = \sqrt{\frac{L}{C}} \tag{1}$$

inductance and *C* the shunt capacitance per unit length. Such a line may be approximated by a network consisting of a number *n* of cascaded *LC* elements. A pulse applied to the input of such a line reaching the output offers a time delay T_d , given by Eq. (2).

$$T_d = n\sqrt{LC} \tag{2}$$

An idealized circuit showing how such a network is used in a pulse generator is shown in **Fig. 2**. A source is connected by a switch to a simulated, unterminated transmission line in series with a resistance R_0 equal to the characteristic impedance of the line. At the time voltage is applied, and until all the capacitors in the line become fully charged, the input impedance of the line is equal to R_0 , and the current *I* is given by Eq. (3). The step

$$I = \frac{V_i}{2R_0} \tag{3}$$

function progresses along the line, charging each C in succession. When it reaches the end, it is reflected back with no change in phase and

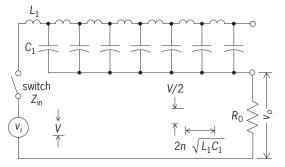


Fig. 2. Principle of line-controlled pulse generator.

returns to the source in a time T, in Eq. (4).

$$T = 2n\sqrt{LC} \tag{4}$$

At this time the line is fully charged, the impedance becomes infinite, and current ceases to flow. The pulse appearing across R_0 is suddenly terminated as shown. For discussion of switching circuits suitable for supplying the line-charging current *see* GATE CIR-CUIT

Similar results can be obtained from the use of a current generator as a source and a short-circuited line as the controlling circuit element.

Various forms of delay-line-controlled pulse generators can be found, and some are capable of generating pulses containing considerable amounts of power for such applications as modulators in radar transmitters. Glenn M. Glasford

Bibliography. D. A. Bell, *Solid State Pulse Circuits*, 4th ed., 1997; R. L. Castellucis, *Pulse and Logic Circuits*, 1976; R. J. Tocci, *Fundamentals of Pulse and Digital Circuits*, 3d ed., 1983.

Pulse jet

A type of jet engine characterized by periodic surges of thrust. The pulse jet engine was widely known for its use during World War II on the German V-1 missile (Fig. 1). The basic engine cycle was invented in 1908. The inlet end of the engine is provided with a grid to which are attached flap valves. These valves are normally held by spring tension against the grid face and block the flow of air back out of the front of the engine. They can be sucked inward by a negative differential pressure to allow air to flow into the engine. Downstream from the flap valves is the combustion chamber. A fuel injection system is located at the entrance to the combustion chamber. The chamber is also fitted with a spark plug. Following the combustion chamber is a long exhaust duct which provides an inertial gas column.

When the combustion chamber is filled with a mixture of fresh air and fuel, a spark is discharged; it ignites the fuel-air mixture, producing a pressure surge that advances upstream to slam shut the inlet valves and to block off the entrance. Simultaneously, a pressure pulse goes downstream to produce a surge of combustion products out the exhaust duct. Thrust results from the rearward discharge of this gas at high velocity. With the discharge of gas from the combustion chamber, its pressure tends to drop. Inertia causes the column of gas in the exhaust duct to continue to flow rearward even after the explosion

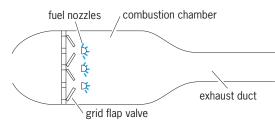


Fig. 1. Diagram of a pulse jet.

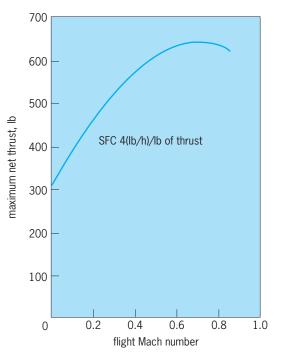


Fig. 2. Effect of flight Mach number on thrust of a German pulse jet; length 137.2 in. (3.49 m), diameter 21.6 in. (0.55 m). 1 lb = 4.448 N. SFC = specific fuel consumption. 4 (lb/h)/lb of thrust = 0.41 (kg/h)/N of thrust.

pressure in the combustion chamber has been dissipated, and this drops the combustion chamber pressure below atmospheric. As a result, the flat valves open and a fresh charge of air enters the combustion chamber. As this air flows past the fuel nozzles, it receives an injection of fuel and the mixture is then ignited by contact with the hot gas residue from the previous cycle. This causes the mixture to explode and the cycle repeats. Thrust increases with engine speed up to a maximum dependent on design (**Fig. 2**). *See* TURBINE ENGINE SUBSYSTEMS.

Unlike the ramjet, the pulse jet has an appreciable thrust at zero flight speed. However, as the flight speed is increased, the resistance to the flow of air imposed by the flap valves eventually causes substantial loss in performance and the pulse jet becomes less efficient than the ramjet. Failure of flap valves and valve seats by fatigue was found to be a problem.

In addition to their use on the German V-1 buzzbomb, pulse jets have been used to propel radio-controlled target drones and experimental helicopters. In the latter case, they were mounted on the blade tips for directly driving the rotor. The high fuel consumption, noise, and vibrations generated by the pulse jet limit its scope of applications. *See* PROPULSION. Benjamin Pinkel

Pulse modulation

A set of techniques whereby a sequence of information-carrying quantities occurring at discrete instances of time is encoded into a corresponding regular sequence of electromagnetic carrier pulses. Varying the amplitude, polarity, presence or absence, duration, or occurrence in time of the pulses gives rise to the four basic forms of pulse modulation: pulse-amplitude modulation (PAM), pulse-code modulation (PCM), pulse-width modulation (PWM, also known as pulse-duration modulation, PDM), and pulse-position modulation (PPM).

Analog-to-digital conversion. An important concept in pulse modulation is analog-to-digital (A/D) conversion, in which an original analog (timeand amplitude-continuous) information signal s(t)is changed at the transmitter into a series of regularly occurring discrete pulses whose amplitudes are restricted to a fixed and finite number of values. An inverse digital-to-analog (D/A) process is used at the receiver to reconstruct an approximation of the original form of *s*(*t*). Conceptually, analog-to-digital conversion involves two steps. First, the range of amplitudes of s(t) is divided or quantized into a finite number of predetermined levels, and each such level is represented by a pulse of fixed amplitude. Second, the amplitude of s(t) is periodically measured or sampled and replaced by the pulse representing the level that corresponds to the measurement. See ANALOG-TO-DIGITAL CONVERTER; DIGITAL-TO-ANALOG CONVERTER.

Figure 1 shows an example of an analog signal and its corresponding pulse sequence resulting from the sampling and quantizing process. Here the dynamic range of the amplitude of the signal is assumed to extend from a minimum of 0 volt to a maximum of 7 volts, and the seven quantization levels are uniformly spaced and of magnitude 1 volt. Sampling occurs at the rate of one sample per second, and each sample value is represented by a pulse corresponding to the nearest quantization level and lasting the entire sampling interval of 1 second.

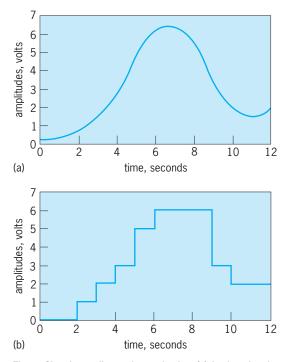


Fig. 1. Signal sampling and quantization. (a) Analog signal. (b) Corresponding pulse sequence.

According to the Nyquist sampling theorem, if sampling occurs at a rate at least twice that of the bandwidth of s(t), the latter can be unambiguously reconstructed from its amplitude values at the sampling instants by applying them to an ideal low-pass filter whose bandwidth matches that of s(t).

Quantization, however, introduces an irreversible error, the so-called quantization error, since the pulse representing a sample measurement determines only the quantization level in which the measurement falls and not its exact value. Consequently, the process of reconstructing s(t) from the sequence of pulses yields only an approximate version of s(t). Generally speaking, the size of the quantization error, and hence the distortion in the reconstructed version of s(t), is related to the coarseness of the quantization scheme. It can thus be controlled by appropriately selecting the size of the quantization levels or the number of levels covering the dynamic range of the signal. In linear quantization the levels are chosen to be of constant size, whereas nonlinear quantization uses levels nonuniform in size, with the use of smaller levels for signal amplitudes that are particularly sensitive to distortion. But any increase in the number of levels covering a given dynamic range implies that the digital-to-analog conversion process must be able to distinguish between levels whose amplitudes differ by smaller amounts. This may result in an increase in the probability of error in determining the correct quantization level when the signals representing the levels have been corrupted by noise and distortion. See ELECTRIC FILTER; INFORMATION THEORY.

Pulse-amplitude modulation. In PAM the successive sample values of the analog signal *s*(*t*) are used to effect the amplitudes of a corresponding sequence of pulses of constant duration occurring at the sampling rate. No quantization of the samples normally occurs (**Fig.** *2a*, *b*). In principle the pulses may occupy the entire time between samples, but in most practical systems the pulse duration, known as the duty cycle, is limited to a fraction of the sampling interval. Such a restriction creates the possibility of interleaving during one sample interval one or more pulses derived from other PAM systems in a process known as time-division multiplexing (TDM). *See* MULTIPLEXING AND MULTIPLE ACCESS.

Pulse-width modulation. In PWM the pulses representing successive sample values of s(t) have constant amplitudes but vary in time duration in direct proportion to the sample value. The pulse duration can be changed relative to fixed leading or trailing time edges or a fixed pulse center. To allow for time-division multiplexing, the maximum pulse duration may be limited to a fraction of the time between samples (Fig. 2*c*).

Pulse-position modulation. PPM encodes the sample values of s(t) by varying the position of a pulse of constant duration relative to its nominal time of occurrence. As in PAM and PWM, the duration of the pulses is typically a fraction of the sampling interval. In addition, the maximum time excursion of the pulses may be limited (Fig. 2*d*).

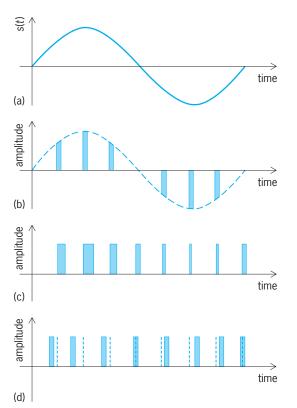


Fig. 2. Forms of pulse modulation for the case where the analog signal, s(t), is a sine wave. (a) Analog signal, s(t). (b) Pulse-amplitude modulation. (c) Pulse-width modulation. (c) Pulse-position modulation.

Pulse-code modulation. Many modern communication systems are designed to transmit and receive only pulses of two distinct amplitudes. In these socalled binary digital systems, the analog-to-digital conversion process is extended by the additional step of coding, in which the amplitude of each pulse representing a quantized sample of s(t) is converted into a unique sequence of one or more pulses with just two possible amplitudes. The complete conversion process is known as pulse-code modulation.

Figure 3*a* shows the example of three successive quantized samples of an analog signal s(t), in which sampling occurs every *T* seconds and the pulse representing the sample is limited to *T*/2 seconds. Assuming that the number of quantization levels is limited to 8, each level can be represented by a unique sequence of three two-valued pulses. In Fig. 3*b* these pulses are of amplitude *V* or 0, whereas in Fig. 3*c* the amplitudes are *V* and -V.

In general, if the original analog signal s(t) has a bandwidth of *B* hertz, the number of samples per second must be at least 2*B* and the time between successive samples can be no greater than T = 1/(2B) seconds. If the amplitudes of the samples are quantized into *L* levels, each level can be uniquely represented by a sequence of $\log_2 L$ binary pulses. Consequently the PCM process generates at least 2*B* $\log_2 L$ binary pulses per second. Increasing the number of levels generally reduces the quantization error and therefore improves the fidelity of the reproduced analog signal, but also increases the number of pulses per

second required to represent the signal.

PCM enjoys many important advantages over other forms of pulse modulation due to the fact that information is represented by a two-state variable. First, the design parameters of a PCM transmission system depend critically on the bandwidth of the original signal s(t) and the degree of fidelity required at the point of reconstruction, but are otherwise largely independent of the information content of s(t). This fact creates the possibility of deploying generic transmission systems suitable for many types of information. Second, the detection of the state of a two-state variable in a noisy environment is inherently simpler than the precise measurement of the amplitude, duration, or position of a pulse in which these quantities are not constrained. Third, the binary pulses propagating along a medium can be intercepted and decoded at a point where the accumulated distortion and attenuation are sufficiently low to assure high detection accuracy. New pulses can then be generated and transmitted to the next such decoding point. This so-called process of repeatering significantly reduces the propagation of distortion and leads to a quality of transmission that is largely independent of distance. In contrast, repeated amplification of analog signals along a transmission path results in the accumulation of the signal's distortion.

Baseband transmission. The modulated carrier pulses derived from any of the pulse modulation schemes may be transmitted directly, provided that the passband of the transmission medium includes the part of the electromagnetic spectrum that contains the energy of the pulses. In such a case the transmission system is referred to as baseband.

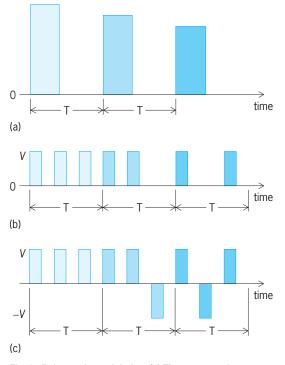


Fig. 3. Pulse-code modulation. (a) Three successive quantized samples of an analog signal. (b) With pulses of amplitude V or 0. (c) With pulses of amplitude V or -V.

Otherwise the modulated carrier pulses may in turn be used to modulate the amplitude, frequency, or phase of an analog carrier signal such as a sine wave in order to shift its energy to a higher frequency range appropriate for the medium.

PCM baseband transmission systems employ several different pulse patterns to represent the binary pulses obtained from the coding process. They are generally distinguished by their ability to provide timing content and error detection capability at the receiver, and by the amount of channel bandwidth required for their reliable transmission. Among the most important of these baseband codes are the nonreturn-to-zero (NRZ), biphase, and bipolar pulse codes. NRZ employs two pulses, typically of equal amplitude but opposite polarity, +V and -V, each pulse lasting for the entire time between samples. Biphase, also known as Manchester coding, uses a two-pulse sequence for each sampling interval, with the polarity of the pulse changing from +V to -Vor from -V to +V at the midpoint of the sampling interval. Bipolar, also known as alternate mark invert (AMI), employs two pulse amplitudes of equal and opposite polarity +V and -V, and a third pulse of zero amplitude. Here +V and -V are used in an alternating manner to represent successive occurrences of one of the binary pulses, with the zero pulse always representing the other binary pulse.

The primary objective in any of these coding schemes is the provision of timing at the receiver by the observation of changes in the polarity of the pulses. In this respect, NRZ is inferior to the other forms due to the fact that polarity changes from one sampling interval to the other occur only when the underlying code pulse changes. The alternating polarity of the nonzero pulses of a bipolar code also provides a certain error detection capability, since the reception of two successive pulses of the same polarity must be the result of an error occurring during transmission. The occurrence of two successive pulses of the same polarity can also be used to signal the presence of a special condition in the information stream.

Time-division multiplexing. An advantage inherent in all pulse modulation systems is their ability to transmit signals from multiple sources over a common transmission system through the process of time-division multiplexing. By restricting the time duration of a pulse representing a sample value from a particular analog signal to a fraction of the time between successive samples, pulses derived from other sampled analog signals can be accommodated on the transmission system. For example, if five analog signals, each of bandwidth 100 hertz, are sampled at the Nyquist rate of 200 samples per second, then the time between successive samples from each signal is 5 milliseconds. By restricting the duration of a pulse to less than 1 millisecond, and by appropriate coordination of the timing of the samples, the pulses from all five signals can be transmitted during one sampling interval. Assigning the pulses from each signal to fixed positions in successive sampling intervals allows the demultiplexing of the composite pulse stream into its individual components by a simple counting process.

One important application of this principle occurs in the transmission of PCM telephone voice signals over a digital transmission system known as a T1 carrier. In standard T1 coding, an original analog voice signal is band-limited to 4000 hertz by passing it through a low-pass filter, and is then sampled at the Nyquist rate of 8000 samples per second, so that the time between successive samples is 125 microseconds. The samples are quantized to 256 levels, with each of them being represented by a sequence of 8 binary pulses. By limiting the duration of a single pulse to 0.65 microsecond, a total of 193 pulses can be accommodated in the time span of 125 microseconds between samples. One of these serves as a synchronization marker that indicates the beginning of such a sequence of 193 pulses, while the other 192 pulses are the composite of 8 pulses from each of 24 voice signals, with each 8-pulse sequence occupying a specified position (Fig. 4). T1 carriers and similar types of digital carrier systems are in widespread use in the world's telephone networks.

Bandwidth and distortion. An important consideration in any of the pulse modulation systems is the level of distortion experienced by a pulse in its transmission over a communications channel. In this regard a major effect is due to the band-limited nature of the channel. Theoretically, any pulse of finite duration, no matter what its shape, requires a channel of infinite bandwidth for undistorted reception. Since signals of infinite duration can in practice not be generated, and since the bandwidth of real channels is always finite, some distortion is unavoidable. The amount, however, depends on the duration T of a pulse and its amplitude variation, relative to the channel bandwidth B. In general, it can be said that for a given amount of distortion, pulses in which the amplitude changes abruptly or which are of shorter duration require larger channel bandwidths than pulses with continuously changing amplitudes or longer

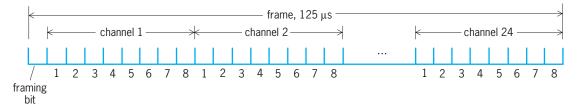


Fig. 4. Frame format for a T1 carrier, showing the arrangement of the pulses for one frame, corresponding to one sampling interval.

duration. As an example, the required bandwidth for a square pulse lasting *T* seconds is usually taken to equal 1/T hertz, whereas for a so-called raised cosine pulse, whose shape approximates that of a cosine wave, the value can be made as low as 1/(2T) hertz.

The required channel bandwidth for the transmission of pulses can be reduced by extending them in time over more than one sampling interval. However, such pulses incur intersymbol interference, in which the trailing part of one pulse overlaps with one or more subsequent pulses. Various pulse-shaping techniques that produce controlled amounts of this type of distortion, while at the same time limiting the amount of required channel bandwidth, have been developed. These are referred to as partial response codes. An important consequence of the distortion incurred by a pulse in a band-limited channel is the loss of timing content, since the spreading of the pulse in time obliterates sharp transitions in the voltage levels.

Pulses may also be corrupted by atmospheric noise produced by rain, sandstorms, galactic emissions, ionospheric fading, and other natural phenomena. Other major sources of distortion and interference external to the transmission channel are artificial, and include electric motors, and transmissions from radio and television stations in the same or adjacent spectral ranges. Both atmospheric noise and artificial interference manifest themselves largely in the form of random variations of the pulse amplitude and therefore affect the accuracy with which the quantization levels or sample values of the original signal can be reproduced at the receiver. *See* DISTORTION (ELECTRONIC CIRCUITS); ELECTRICAL IN-TERFERENCE; ELECTRICAL NOISE.

Bandwidth requirements. Pulse modulation systems may incur a significant bandwidth penalty compared to the transmission of an analog signal in its analog form. An example is the standard PCM transmission of an analog voice signal band-limited to 4000 hertz over a T1 carrier. Since the sampling, quantizing, and coding process produces 8 binary pulses 8000 times per second for a total of 64,000 binary pulses per second, the pulses occur every 15.625 microseconds. Depending on the shape of the pulses and the amount of intersymbol interference, the required transmission bandwidth will fall in the range of 32,000 to 64,000 hertz. This compares to a bandwidth of only 4000 hertz for the transmission of the signal in analog mode. See BANDWIDTH REQUIREMENTS (COMMUNICATIONS).

Pulse timing. An important consideration in the design of transmission systems based on any of the pulse modulation methods is the provision of pulse timing information at the receiver. Specifically, the receiver must have accurate knowledge of the points in time when pulses were generated at the sending side in order to determine properly their amplitude, position, or duration. In PAM, PWM, or PCM systems, the timing information usually identifies the beginning, middle, or end of a pulse, whereas in PPM the center of the unmodulated pulse may be indicated.

Pulse timing can be provided in three ways, re-

ferred to as autonomous, explicit, and implicit timing. In the autonomous method, timing is determined at the sender and receiver by means of local clocks that are periodically aligned from master reference clocks. These master clocks derive highly stable timing signals from the natural frequency of oscillation or resonance of certain materials such as quartz, cesium, and rubidium. In the explicit timing method, timing pulses are transmitted in parallel with the sequence of information-bearing pulses over a separate channel, and identify the moments in time when the information pulses were generated. Implicit timing relies on a local clock and the presence of timing information in the received pulse sequences themselves from which the local clock can be realigned. For example, in a 100% duty cycle NRZ-coded or AMIcoded pulse sequence, the occurrence of a transition in the voltage level indicates the end of one pulse and the beginning of another and provides an opportunity for the adjustment of the receiver clock. See ATOMIC CLOCK; QUARTZ CLOCK.

NRZ or AMI baseband codes suffer from the deficiency that certain information signals may generate long sequences of pulses without any level change, which may cause the receiver clock to drift. Several methods for dealing with this problem are found in operational systems. In so-called pulse stuffing, an extra pulse chosen to interrupt this pattern is inserted at the sending side and extracted at the receiving side after timing extraction. In the technique known as bit sequence transparency coding, a pulse sequence of specified length with no level changes is replaced with a predetermined pattern of pulses that normally does not occur as a sequence of information pulses but contains a high percentage of level changes. In scrambling, the original pattern of information pulses is deterministically altered prior to transmission in a manner that guarantees a sufficiently high incidence of level changes, and is restored at the receiver through an inverse scrambling process. The bipolar code avoids the difficulty altogether in that it guarantees a change in level at the midpoint of every pulse.

Frame synchronization. In TDM systems, besides pulse timing, the additional problem of frame synchronization arises. Here the receiver, in order to demultiplex the pulses in a frame (that is, the pulses received during a sampling interval) must have information relating to the beginning of each such sampling interval. Demultiplexing then can proceed by a simple counting process.

In one method of frame synchronization in widespread use, the sender transmits at the beginning of each frame of information pulses a specially selected pattern of pulses. Synchronization is indicated when a copy of this pattern stored at the receiver matches pulse-by-pulse with a received sequence of pulses. So-called pseudonoise (PN) sequences, whose pulse patterns simulate the random occurrence of binary pulses, and a set of pulse patterns known as Barker codes, find frequent application in frame synchronization.

A somewhat different approach is used in T1

carrier systems, which rely on a single pulse transmitted at the beginning of each frame whose polarity alternates from one frame to the next, that is, every 193 pulses. The receiver derives frame synchronization from the observation of this alternating pattern over several consecutive frames.

In any of the frame synchronization methods, special provision must be made to prevent the appearance of the synchronization pattern in the information portion of the frame, for in such cases false frame synchronization may result. Various techniques have been developed for altering the information sequence when such patterns are detected.

Applications. PAM, PWM, and PPM found significant application early in the development of digital communications, largely in the domain of radio telemetry for remote monitoring and sensing. They have since fallen into disuse in favor of PCM.

Since the early 1960s, many of the world's telephone network providers have gradually, and by now almost completely, converted their transmission facilities to PCM technology. The bulk of these transmission systems use some form of time-division multiplexing, as exemplified by the 24-voice channel T1 carrier structure, and are intended for short-haul transmission. Higher-capacity TDM facilities such as the T1-C, T2, T3, and T4 carriers, which combine 48, 96, 672, and 4032 digital voice channels, respectively, have also been deployed in large quantities to provide connections between large central office switches and toll office switches, and between subscribers and central offices over subscriber lines. Internationally, carrier systems referred to as E1, E2, E3, and E4, which combine 30, 120, 480, and 1920 voice channels, respectively, are in widespread use. These carrier systems are implemented over many types of transmission media, including twisted pairs of telephone wiring, coaxial cables, fiber-optic cables, and microwave. See COAXIAL CABLE; COMMU-NICATIONS CABLE; MICROWAVE; OPTICAL COMMUNI-CATIONS; OPTICAL FIBERS; SWITCHING SYSTEMS (COM-MUNICATIONS); TELEPHONE SERVICE.

The deployment of high-speed networks such as the Integrated Service Digital Network (ISDN) in many parts of the world has also relied heavily on PCM technology. PCM and various modified forms such as delta modulation (DM) and adaptive differential pulse-code modulation (ADPCM) have also found significant application in satellite transmission systems. *See* COMMUNICATIONS SATELLITE; DATA COMMUNICATIONS; ELECTRICAL COMMUNICATIONS; INTEGRATED SERVICES DIGITAL NETWORK (ISDN); MODULATION. Hermann J. Helgert

Bibliography. B. P. Lathi, *Modern Digital and Ana*log Communication Systems, 3d ed., Oxford University Press, 1998; W. D. Reeve, Subscriber Loop Signaling and Transmission Handbook: Digital, IEEE Press, New York, 1995; M. Schwartz, Information Transmission, Modulation, and Noise, 4th ed., McGraw-Hill, New York, 1990; D. R. Smith, Digital Transmission Systems, 2d ed., Van Nostrand Reinhold, New York, 1993.

Pulse modulator

A device that converts an analog information signal to a pulse format. A pulse modulator encodes the amplitude information at a particular time by a set of short- or long-duration pulses using such methods as pulse-amplitude, pulse-width, pulse-position, pulsedensity, or pulse-code modulation. A pulse modulator, therefore, accepts a continuous analog input signal and provides a pulse output that conveys the amplitude of the continuous analog signal as seen at an observation, or sampling, time t_s , often optimistically described as a sampling instant.

Pulse-amplitude modulation. A continuous analog signal, when viewed through equally spaced time windows, provides a pulse-amplitude modulation (PAM) signal. This is analogous to viewing the continuous signal through the vertical slats in a picket fence. The signal is sampled at equally spaced time intervals of T_s , the sampling period. The reciprocal of T_s is the sampling frequency f_s . If the highest frequency in the continuous analog signal is f_{max} , then in order to avoid loss of information through sampling, the sampling theorem dictates that $f_s > 2 \cdot f_{\text{max}}$. The duration of each sampling window, during which time we look at the analog signal, is the aperture time T_a (the spacing between the slats). The duty cycle, T_a/T_s , represents the fraction of the time that the PAM signal is present. The time that the PAM signal is absent (idle time) can be profitably used to multiplex $N = T_s/T_a$ different PAM signals together. Channel efficiency is improved by time-division-multiplexed (TDM) transmission, or the interleaving of N analog signals over a single channel. Multiple analog instrumentation signals are frequently sent to a remote recorder or monitor via a PAM signaling system. See INFORMATION THEORY; MULTIPLEXING AND MULTIPLE ACCESS; TELEMETERING.

A pulse-amplitude modulator is simply a transmission gate, or switch, that opens every T_s seconds for duration of T_a seconds to produce a PAM signal. On the other hand, a sample-and-hold circuit may be used. The value of the signal detected at the sampling instant is clamped or "held" until the next sample instant by a sample-and-hold (S&H) device.

The preceding discussion concerns baseband PAM. If the transmited signal is then multiplied by a sinusoidal carrier, it becomes passband PAM.

Phase modulator. A close relative of the pulseamplitude modulator is the phase-shift-keyed (PSK) modulator. The amplitude of the signal is held constant, but the information is provided in the phase of a carrier. PSK and PAM can be combined to form QAM (quadrature-amplitude-modulated) signals, discussed below. *See* PHASE MODULATION.

Pulse-width modulation. Unlike a PAM signal, the amplitude of each pulse-width modulation (PWM) pulse is constant. The width τ of each output pulse carries the amplitude information of the continuous analog signal during the observation interval. The input signal x(t) may be sampled and "held" or clamped to its value at the sampling instants, thereby generating a stair-step waveform; or it may simply

be used in its continuous form. Most PWM signals are generated by combining the stair-step or continuous analog signal with a triangular or sawtooth waveform in a thresholding device that provides an output of +1 when the sum is positive, and an output of zero otherwise if the input signal is unipolar (binary PWM), that is, if $0 \le x(t) \le 1$. If, for example, the sawtooth function is $-t/T_s$ for $0 \le t \le T_s$, the summed signal $x(t) - t/T_s$ reaches zero at $\tau = T_s$. $x(\tau)$. Since the threshold circuit output is +1 until $t = \tau$ and then drops to zero, the average value of the PWM output is $\tau/T_s = x(\tau)$. If $-1 \le x(t) \le -0$, we have simply flipped the polarity of the input, and we can similarly flip the polarity of the ramp to make a PWM modulator for negative in put signals. If a -1is produced in response to a negative input, the two systems can be tied in parallel to provide a ternary PWM system.

An alternative but less power-efficient form of binary PWM outputs a +1 or a -1 (never a 0), maintaining a correct average value over the pulse period. PWM is attractive because the signal can be clamped to amplitudes of +1, -1, and zero, making it very insensitive to additive amplitude noise.

PWM is especially useful in applications where power, as opposed to information or data, is being controlled because electrical power-switching devices are more power-efficient than proportionalcontrol devices. Switched-device resistance is either very low or near-infinite. In dc-motor speed-control operation, PWM motor-drive signals virtually eliminate the static friction (stiction) of motor bearings and mechanical linkages because the torque is applied abruptly. PWM is also used for voltage regulation in switched power supplies, light dimmers, and class D amplifiers for audio applications. *See* ELEC-TRONIC POWER SUPPLY.

Pulse-position modulation. This is now a seldomused format. The amplitude information of the analog input signal is carried by the location of a constant-amplitude, constant-duration pulse whose location along the time axis carries the input information. Pulse-position modulation (PPM) involves inexpensive electronics, but is quite susceptible to additive noise. A PPM signal can be generated by differentiating, or high-pass-filtering, a PWM signal.

Pulse-density modulation. The pulse-density modulation (PDM) pulse rate during the sampling period is proportional to the amplitude of the analog input signal, as is its average value. A PDM signal may be generated by driving a voltage-controlled oscillator with the analog input signal, clipping the output to generate a square wave, then differentiating to produce a pulse train whose rate, or density, varies proportionally with the analog input voltage. *See* OSCILLATOR.

Pulse-code modulation. A pulse-code modulator produces numbers, rather than the analog values that are produced by all of the pulse-modulation methods discussed so far, which quantize the signal in time only. In those methods, the signal amplitude is specified at sampling instants t_s (quantized in time), but the signal amplitude can take on any values between the signal maximum and minimum values. For PCM,

the signal is quantized also in amplitude steps of size $q = 2^{-B}$, where *B* is the number of bits used to represent the sample of the amplitude of the continuous analog signal x(t), and it is assumed that $0 \le x(t) \le 1$. The parameter *B* determines the size of the granularity of the quantization step as well as the error. In binary notation, the signal quantization is given by Eq. (1), where the b_n are the bits 0 or 1 used

$$x(t_s) \approx \sum_{1}^{B} b_n \cdot 2^{-n} \tag{1}$$

to approximate $x(t_s)$. The peak quantization error ε does not exceed $\varepsilon_{\text{max}} = \frac{1}{2}q$. Since the quantization error is equally likely to fall anywhere between $-\frac{1}{2}q$ and $\frac{1}{2}q$, the mean or average squared quantization error (MSE) is given by Eq. (2). The RMS (root-

$$\frac{1}{q} \int_{-q/2}^{q/2} \varepsilon^2 d\varepsilon = q^2 / 12 = 2^{-2B} / 12$$
(2)

mean-squared) error is, of course, the square root of the MSE. This is an additive quantization noise. As *B* grows, the peak and RMS errors fall. for example, in the quantization of an analog signal sample to 4 bits, that is, B = 4, the quantization step size is $q = 2^{-4} =$ 1/16, the peak quantization error is $\varepsilon_{max} = \frac{1}{2} \cdot 2^{-4} =$ 1/32, and the RMS error is $\varepsilon_{max}/\sqrt{3} = 0.018$. Now, suppose $x(t_s) = 0.6875 = 1 \cdot 2^{-1} + 0 \cdot 2^{-2} + 1 \cdot 2^{-2} +$ $1 \cdot 2^{-4}$. The bits are $b_1 = 1$, $b_2 = 0$, $b_3 = 1$, and $b_4 =$ 1. That string of bits (or data word), 1011, forms the PCM-coded representation of $x(t_s)$.

In the representation of signed data, positive numbers have a zero as a sign bit appended to the left of the most significant data bit. Conversion to a negative number in two's complement format (the most common) is accomplished easily by copying the data, from right to left, through the first "one," then complementing (making ones into zeros and vice versa) all bits to the left. For the example above, with its sign bit, the PCM form of $x(t_s)$ is 01011; its negative is 10101. *See* NUMERICAL REPRESENTATION (COMPUT-ERS).

For additional robustness against noise, the digital format is often not straight binary, but reflected binary or Gray code, in which any two adjacent numerical values can differ by only one bit.

Corresponding to each sampling instant, the PCM produces one data word. A PCM signal is said to be digitized because its quantized amplitude values are in digital format. The PCM device itself is usually known as an analog-to-digital converter. A digitized signal is reasonably noise-resistant and can be modulated, demodulated, filtered, and otherwise processed more accurately than an analog signal. Many PCM data streams may be time-division-multiplexed together for transmission. *See* ANALOG-TO-DIGITAL CONVERTER.

Delta modulation (DM) or delta PCM (DPCM). DPCM conversion, in which the difference, rather than the actual values, in amplitudes of successive samples is encoded and transmitted, is often more efficient than PCM. This is particularly so if the signal is oversampled, because large sample-to-sample changes in

amplitude are unlikely to occur. DPCM attempts to approximate the input signal by a sequence of positive or negative steps of fixed size. This method is simple but is degraded by "slope overload," analogous to saturation, if the DPCM is unable to keep up with the changes of the input signal. Because only the error is encoded, fewer bits are necessary, permitting a lower transmission data rate, or improved signal-tonoise ratio. For example, a reduction in word-length requirement can be translated to a faster transmission rate or a 6-dB reduction in quantization noise.

Adaptive delta modulation. This method avoids the slope overload of DM by adjusting the step size according to the demands of the input signal. One adaptive delta modulation (ADM) method, the Song algorithm, begins with some predetermined minimum step size to follow the input signal. At each sample time the step size is either increased or decreased by one unit to reduce the error. The Space Shuttle algorithm is a similar ADM algorithm; however, if the error switches polarity, the step size drops back to the minimum.

Differential PCM. Differential PCM requires memory to predict the next sample value based on fixed prediction-coefficient values and the present and past input samples, and then transmits the prediction error.

Adaptive differential PCM (ADPCM). This is an even more sophisticated method of encoding the signal, in which the coefficients of the predictor are recomputed every sample time. *See* DATA COMPRESSION.

Quadrature amplitude modulation (QAM). QAM (or quadrature PAM) is a unique format because it simultaneously modulates two carriers of the same frequency, but 90° out of phase with one another. This can be viewed as simultaneous modulation of both amplitude (PAM) and phase (PSK), affording an economical and robust means for transmission of a pair of PCM signals.

In two-dimensional signal space, the allowable amplitudes of the values (numbers) modulating the cosine carrier and the sine carrier may be plotted along the I (in-phase or real) and Q (quadrature or imaginary) axes, respectively. This array of *M* points is known as a constellation. Constellation shapes are usually square $[M = 2 \times 2 = 4, M = 3 \times 3 - 1]$ (a square with the origin removed), $M = 4 \times 4 = 16$, $M = 6 \times 6 - 4 = 32$ (a square with the 4 corners lopped off); or $M = 16 \times 16 = 256$]; sometimes rectangular, such as $M = 2 \times 4 = 8$; or concentric circles with the points on radials at multiples of 45°. However, any configuration may be used.

In the simplest form, a stream of usually Graycoded PCM data words are concatenated (laid endto-end) to form a flowing jumble of zeros and ones. The odd-numbered bits are directed to processor A, the even-numbered bits to processor B. The resulting two data streams are chopped up into two sets of *C*-bit synthetic data words. To the tail end of each synthetic word an extra one is appended as a least significant bit (LSB). The leading, or most significant, bit (MSB) of each word is complemented; that is, each one become a zero and each zero becomes a one. The MSBs will be interpreted as sign bits. Now, for each pair of (C + 1)-bit newly created synthetic words, one from A (the "real" or in-phase component) and one from B (the "imaginary" or quadrature component), we can plot a point in the complex signal space described above. An evenly spaced array of points appears in a $C \times C$ square array in signal space centered on the origin. These new synthetic words are converted to analog values by a pair of digital-to-analog converters. The analog real words are amplitude-modulated onto (multiplied by) a carrier $\cos(2\pi f_c t)$, where f_c is the carrier frequency; similarly, the analog imaginary words are modulated into a second but out-of-phase carrier, sin $(2\pi f_c t)$. The two modulated curriers are in quadrature, that is, 90° out of phase with one another, to minimize interference between the real and imaginary components. Each of the complex modulated carrier segments that contain 2C of the original data bits is called a symbol, and the rate at which they are transmitted is called the symbol rate. See DIGITAL-TO-ANALOG CON-VERTER. Stanley A. White

Bibliography. L. W. Couch II, Digital and Analog Communication Systems, Prentice Hall, Upper Saddle River, NJ, 2001; S. Gibilisco, Handbook of Radio and Wireless Technology, McGraw-Hill, New York, 1999; J. G. Proakis, Digital Communications, 4th ed., McGraw-Hill, New York, 2001; T. S. Rappaport, Wireless Communications, 2d ed., Prentice Hall, Upper Saddle River, NJ, 2002; M. S. Roden, Analog and Digital Communications Systems, 4th ed., Prentice Hall, Englewood Cliffs, NJ, 1995; B. Sklar, Digital Communications: Fundamentals and Applications, 2d ed., Prentice Hall, Upper Saddle River, NJ, 2001; M. E. Van Valkenburg and W. M. Middleton (eds.), Reference Data for Engineers: Radio, Electronics, Computer and Communications, 9th ed., Butterworth-Heinemann, Woburn, MA, 2002.

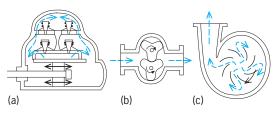
Pumice

A rock froth, formed by the extreme puffing up (vesiculation) of liquid lava by expanding gases liberated from solution in the lava prior to and during solidification. Some varieties will float in water for many weeks before becoming waterlogged. Typical pumice is siliceous (rhyolite or dacite) in composition, but the lightest and most vesicular pumice (known also as reticulite and thread-lace scoria) is of basaltic composition. *See* LAVA; VOLCANIC GLASS.

Gordon A. MacDonald

Pump

A machine that draws a fluid into itself through an entrance port and forces the fluid out through an exhaust port (see **illus.**). A pump may serve to move liquid, as in a cross-country pipeline; to lift liquid, as from a well or to the top of a tall building; or



Pumps. (a) Reciprocating. (b) Rotary. (c) Centrifugal.

to put fluid under pressure, as in a hydraulic brake. These applications depend predominantly upon the discharge characteristic of the pump. A pump may also serve to empty a container, as in a vacuum pump or a sump pump, in which case the application depends primarily on its intake characteristic. *See* CENTRIFUGAL PUMP; COMPRESSOR; DISPLACE-MENT PUMP; FAN; FUEL PUMP; PUMPING MACHINERY; VACUUM PUMP. Elliott F. Wright

Bibliography. B. Glanfield (ed.), *Pump Technology New Challenges, Where Next*, 1989; T. G. Hicks, *Pump Operation and Maintenance*, 1958, reprint 1983: Institute for Power System, *Pumping Manual*, 9th ed., 1995; J. Karassik et al., *Pump Handbook*, 3d ed., 2000; F. Pollak, *Pump User's Handbook*, 3d ed., 1988; A. J. Stepanoff, *Centrifugal and Axial Flow Pumps: Theory, Design and Application*, 2d ed., 1992; H. L. Stewart, *Pumps*, 3d ed., 1978; R. H. Warring, *Pumps Selection, Systems and Applications*, 1982.

Pumped storage

A process, also known as hydroelectric storage, for converting large quantities of electrical energy to potential energy by pumping water to a reservoir at a higher elevation, where it can be stored indefinitely and then released to pass through hydraulic turbines and generate electrical energy. An indirect process is necessary because electrical energy cannot be stored effectively in large quantities. Storage is desirable, as the consumption of electricity varies throughout the day, between weekday and weekend, as well as among seasons. Consequently, much of the generating equipment needed to meet the largest daytime load is unused or lightly loaded at night or on weekends. During those times the excess generating capacity can be used to generate energy for pumping, hence the necessity for storage. Normally, pumping energy can be obtained from economical sources, and its value will be upgraded when used for peak loads.

Operation. In a typical operation, electrical energy generated during the night or weekend is used to pump water from a lower to a higher elevation, where it is stored as potential energy in the upper reservoir. The water can be retained indefinitely without deterioration or significant loss. During the daylight hours when the loads are greatest, the stored water is released to flow from the higher to the lower reservoir through hydraulic turbines that drive

electric generators and is converted to electrical energy. No water is consumed in either the pumping or generating phase. To provide storage or generation merely requires the transfer of water from one reservoir to the other. Pumped storage installations have attained an overall operating efficiency of about 70%. Projected improvements in equipment design promise an efficiency of 75% or more. Postulating one cycle each of generation and pumping per day plus an allowance to change from one mode to the other, the maximum annual generation attainable is 3500 h. *See* HYDRAULIC TURBINE; HYDROELECTRIC GENERATOR; PUMPING MACHINERY.

Description. A typical pumped-storage development consists of two reservoirs of essentially equal volume situated to maximize the difference between their levels. These reservoirs are connected by a system of waterways along which a pumping-generating station is located (Fig. 1). Under favorable geological conditions, the station will be located underground; otherwise it will be situated at the lower reservoir. The principal equipment of the station is the pumping-generating unit. In United States practice, the machinery is reversible and is used for both pumping and generating; it is designed to function as a motor and pump in one direction of rotation and as a turbine and generator in opposite rotation. Transformers, a substation, switchyard, and transmission line are required to transfer the electrical power to and from the station. See ELECTRIC POWER GENERA-TION

The lower reservoir may be formed by impounding a stream or by using an existing body of water. Similarly, an upper reservoir may be created by an embankment across a valley or by a circumferential dike. Frequently, these features produce the most significant environmental impact, which is largely land use and visual. The reservoirs are comparatively small, thus affording some latitude in location to minimize unavoidable effects, such as displacement of developed areas, existing roads, and railways. Problems of emission of particulate matter and gases and of water-temperature rise associated with other generating stations do not exist. On the other hand,

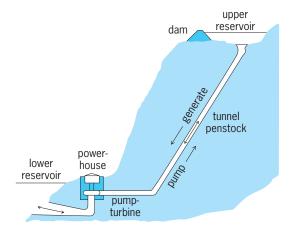


Fig. 1. Diagram of a conventional pumped-storage development.

the reservoirs and surrounding area afford the opportunity to develop recreational facilities such as camp grounds, marinas, boat ramps, picnic areas, and wildlife preserves. In the United States, recreation commands a high priority, and many existing developments have included recreational facilities.

Economics. The economics of pumped storage depend on the system in which it operates. The development must be economically competitive with other types of generation available, namely, nuclear, coal- or oil-fired, gas turbines, hydro, and other storage systems such as compressed air. It has been a generally accepted practice to evaluate the economy of any plant on the basis of its annual cost to the system, but life-cycle costs are now also considered. The annual cost is defined as the sum of fixed costs, operation, maintenance, and fuel. When construction is complete, the fixed costs are established and continue at that level irrespective of the extent of operation. Variations in fuel prices have a major effect. *See* ENERGY SOURCES.

European plants. In European practice, not only system economies receive attention, but also operating requirements and advantages; the latter have materially influenced the design and selection of equipment. European pumped-storage plants have been used not only to provide peak-load power as in the United States, but also to ensure system stability or frequency control. To accomplish this, the units must have very short response times which will enable them to change mode of operation quickly, to follow changes in load, or to provide emergency generation or load. Such operation is the dominant reason for European preference for three-machine sets, that is, separate pump and separate turbine coupled to a motor-generator, as opposed to the two-machine reversible pump-turbine and motor-generator used in the United States. The extent to which European practice incorporates system stability and frequency regulation is shown by the number of pumping and generating starts per day. In one plant, the daily average starts per unit for pumping was 2.5 and for generating 3.5, compared to 1 start for pumping and 1 or possibly 2 per day for generation in typical United States operation.

Deep-underground storage. The deep-underground concept (Fig. 2) is important because of high fuel and construction costs for alternate generation, and the flexibility it affords in selecting a site near load centers. Much research has been directed toward the deep-underground scheme owing to its reduced environmental effect and freedom from topographic restraints. The deep-underground concept is similar to conventional pumped storage, having all of the essential features and utilizing similar equipment. There are two notable variations: the upper reservoir is at ground level, while the lower reservoir is a deep underground cavern. Two stages are developed in series to minimize the size and, hence, the construction cost of the lower reservoir. Inherent advantages of the concept are: material reduction in environmental impact since only the upper

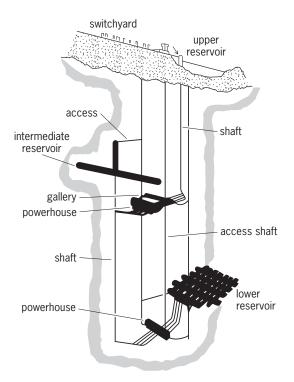


Fig. 2. Diagram of two-stage underground pumpedstorage development.

reservoir is visible, development of heads which utilize the maximum capability of the machines, and elimination of the need for favorable topography. The disadvantages are: the need for large areas to dispose of the excavated material, uncertainties of construction and mining at 3000 ft (900 m) or more below the surface, and substantially increased construction time and costs compared with conventional pumped storage. *See* ENERGY STORAGE; WATER-POWER. Dwight L. Glasscock

Bibliography. American Society of Civil Engineers, *Civil Engineering Guidelines for Planning and Designing Hydroelectric Developments*, vol. 5: *Pumped Storage and Tidal Power*, 1989; American Society of Civil Engineers, *Compendium of Pumped Storage Plants in the United States*, 1993; M. G. Jog, *Hydroelectric and Pumped Storage Plants*, 1989.

Pumping machinery

Devices that convey fluids, chiefly liquids, from a lower to a higher elevation or from a region of lower pressure to one of higher pressure. Pumping machinery may be broadly classified as mechanical or as electromagnetic.

Mechanical pumps. In mechanical pumps the fluid is conveyed by direct contact with a moving part of the pumping machinery. The two basic types are (1) velocity machines, centrifugal or turbine pumps, which impart energy to the fluid primarily by increasing its velocity, then converting part of this energy into pressure or head, and (2) displacement machines with plungers, pistons, cams, or other

confining forms which act directly on the fluid, forcing it to flow against a higher pressure. A pump deep in a well may raise water or oil to the surface. At a ground level location a pump may deliver fluid to a nearby elevated reservoir or, through long pipelines, to a location at similar or different elevation. In a power plant, pumps circulate cooling water or oil at low pressures and transfer water from heaters at moderate pressure to steam generators at pressures of several thousand pounds per square inch. In chemical plants and refineries pumps transfer a great variety of fluids or charge them into reactors at higher pressure. In hydraulic systems, pumps supply energy to a moving stream of oil or water, which is readily controlled, to move a piston or press platen or to rotate a shaft as required by the specific process. See CENTRIFUGAL PUMP; DISPLACEMENT PUMP.

Electromagnetic pumps. Where direct contact between the fluid and the pumping machinery is undesirable, as in atomic energy power plants for circulating liquid metals used as reactor coolants or as solvents for reactor fuels, electromagnetic pumps are used. There are no moving parts in these pumps; no shaft seals are required. The liquid metal passing through the pump becomes, in effect, the rotor circuit of an electric motor. The two basic types are conduction and induction.

The conduction type, used with either direct or alternating current, confines the liquid metal in a narrow passage between the field magnets. Electrodes on each side of this channel apply current through the liquid metal at right angles to the magnetic field and the direction of flow. This current path is like the flow of current in the armature winding of a motor.

The induction or traveling field type operates only on polyphase alternating current. The liquid metal is confined in a thin rectangular or annular passage thermally insulated from the slotted stator. This stator with its windings is similar to the stator of a squirrel-cage motor cut through on one side and rolled out flat. The traveling field induces currents in the liquid metal similar to the currents in a motor armature. *See* ELECTROMAGNETIC PUMP. Elliott F. Wright

Bibliography. N. P. Chopay (ed.), *Fluid Movers*, 1994; T. G. Hicks, *Pump Operation and Maintenance*, 1958, reprint 1983; Institute for Power System, *Pumping Manual*, 9th ed., 1995; I. J. Karassik et al., *Pump Handbook*, 3d ed., 2000; F. Pollak, *Pump User's Handbook*, 3d ed., 1988; A. J. Stepanoff, *Centrifugal and Axial Flow Pumps: Theory, Design and Application*, 2d ed., 1992; H. L. Stewart, *Pumps*, 3d ed., 1978; R. H. Warring, *Pumps Selection, Systems and Applications*, 1982.

Pumpkin

The term commonly applied to the larger, orangecolored fruit of the *Cucurbita* species, used when ripe as a table vegetable, in pies, or for autumn decoration. The flesh is somewhat coarse and strongly flavored, hence generally not served as a baked vegetable. Although some taxonomists would restrict the term pumpkin to the species *Cucurbita pepper* and *C. moschata*, it is also used in referring to *C. mixta*. Popular varieties (cultivars) are Connecticut Field and Small Sugar. Canned pumpkin is usually made from a blend of pumpkins and winter squashes. Cultural practices are similar to those used for squash. Harvesting generally begins when the fruits are mature, usually 4 months after planting. New Jersey, Illinois, and California are important producing states. *See* SQUASH; VIOLALES. H. John Carew

Puna

An alpine biological community in the central portion of the Andes Mountains of South America. Sparsely vegetated, treeless stretches cover high plateau country (altiplano) and slopes of central and southern Peru, Bolivia, northern Chile, and northwestern Argentina. The poor vegetative cover and the puna animals are limited by short seasonal precipitation as well as by the low temperatures of high altitudes. In spite of severe climate and limited use, some portions are inhabited by native Indians.

Altitudinal zonation. Like the paramos of the Northern Andes, punas occur above timberline, and extend upward, in modified form, to perpetual snow. Due to greater heights of the Central Andean peaks, aeolian regions, that is, regions supplied with airborne nutrients above the upper limit of vascular plants, generally the snowline, are more extensive here than above the paramos. Vascular plants reach higher levels in these low-latitude regions, around 30° south of the Equator, than in the paramos, which are equatorial. Thus the snowline, at about 15,700 ft (4800 m) in southern Peru, is above that of the paramos and is still higher farther south. Similar difference in upper limits of plants is found north of the Equator in the Himalaya Mountains. *See* PARAMO.

The altiplano is above intermountain basins, such as the Peruvian Cuzco area at 11,000 ft (3400 m), these basins being well vegetated and intensely cultivated. Some authorities place the beginning of the puna at 13,000 ft (4000 m), approximately the limit of cultivated crops, and call the uppermost slopes with very few plants "cold deserts."

Plant life-forms and species. The low-growing widely spaced plants with considerable bare ground between them, and the lack of much green color most of the year give the vast puna areas a bleak appearance (**Fig. 1**). The appearance of the puna is due not to severe temperatures, but to a short wet season; almost all precipitation, including snow and hail, occurs from January through March and causes these areas to become green. The average temperature of the Bolivian altiplano is 44° F (7°C). Temperatures below zero are rare. A striking contrast with the paramo is the alkaline nature of the puna soil, the paramo being consistently acid. Some shallow

intermittent lakes are used for salt collection by the natives.

Plant adaptations are indicative of the predominantly dry conditions. Bunch grasses, many with stiff needle leaves, are small; xerophilous herbs and resinous shrubs grow as rosettes, cushions, or mats; and succulent cacti occur in dense cushions. A good indicator plant is Pycnopbyllum of the pink family, Caryophyllaceae, which forms pale-green mounds of densely packed stems covered with small, imbricated leaves (Fig. 2). These masses are locally called vareta, although it would seem better to restrict this term to large Azorella cushions of the celery family, Umbelliferae, said to resemble the "vegetable sheep" of New Zealand and to attain larger size in the eastern, somewhat wetter portion of the puna. Many small prostrate shrubs, chiefly of composite family, such as Baccharis caespitosa and Senecio adenophyllus, add to the scanty ground cover and are used, as is the yareta, for altiplano fuels. Puya raimondii



Fig. 1. Vegetation of the puna, with low bunch grass, prostrate shrubs, and cushion growths of cacti appearing white because of dense hairs. This view is from the Lake Titicaca area, Bolivia.



Fig. 2. Mat of *Pycnophyllum*, typical plant of the puna, on rocky soil; Sierra Negra, central Peru.



Fig. 3. *Puya raimondii*, a rare endemic of Peru and Bolivia; Sierra Negra, central Peru.

of the pineapple family, Bromeliaceae (**Fig. 3**), exemplifies the rosette habit so prominent in *Espeletia* of the paramos. This unusual endemic, with heavily cutinized leaves, occurs in a few isolated sites in Peru and Bolivia. On stream banks and occasional seepages, *Distichia muscoides* of the rush family, Juncaceae, carpets the ground. Low shrub thickets called the tola belt occur along the western margin of the puna. This belt borders upon a desert in characteristics but is floristically related to the puna.

Fauna. The puna is interesting as the habitation of the American species of the camel family: llamas (*Lama glama*), beasts of burden since ancient Inca times; alpacas (*L. pacos*); and vicuñas (*Vicugna vicugna*), prized for hides and wool. The vicuña and another species, the guanaco (*L. guanicoe*), are wild. All live chiefly upon grasses, such as *Calamagrostis brevifolia, C. vicunarum*, and *Stipa ichu*, and herbs. Their heavy fleece adapts them to the rapid radiation changes which occur with lack of cloud cover at high altitudes. The dung of these animals is source of fuel for native residents. *See* ALPACA; LLAMA.

Agricultural practice. Unlike the potato *Solanum tuberosum*, so widely grown in cold climates, certain tuber producers are little known, including oca (*Oxalis tuberosa*), ullucu (*Ullucus tuberosus*), and anu (*Tropaeolum tuberosum*), all typically Andean and able to grow at high altitudes. Preservation of tubers by alternate dehydration and freezing, with subsequent mashing and forming into chuñu, is a puna practice dating from humans' early occupation of these high areas. *See* ALTITUDINAL VEGETATION ZONES. Harriet G. Barclay

Bibliography. F. D. Ayres, Giants of the puna, *Nat. Hist.*, 75:54-57, March 1966; H. E. Wright, Jr., and W. H. Osburn (eds.), *Arctic and Alpine Environments*, 1968.

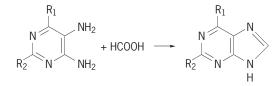
Purine

A heterocyclic organic compound (1) containing

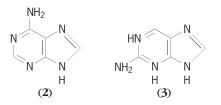


fused pyrimidine and imidazole rings. A number of substituted purine derivatives occur in nature; some, as components of nucleic acids and coenzymes, play vital roles in the genetic and metabolic processes of all living organisms. *See* COENZYME; NUCLEIC ACID.

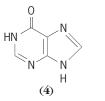
Chemistry. Purines are generally white solids of amphoteric character. Thus, they can form salts with both acids and bases. The presence of conjugated double bonds in purines results in aromatic chemical properties, confers considerable stability, and accounts for their strong ultraviolet absorption spectra. With the exception of the parent compound, most substituted purines have low solubilities in water and organic solvents. While a variety of methods have been used for the synthesis of purine compounds, the Traube method, shown in the reaction below, may be singled out because of its general usefulness.



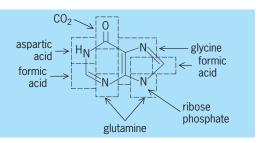
Occurrence. The purine bases, adenine (2) and guanine (3), together with pyrimidines, are fun-



damental components of all nucleic acids. Certain methylated derivatives of adenine and guanine are also present in some nucleic acids in low amounts. In biological systems, hypoxanthine (4), adenine, and



guanine occur mainly as their 9-glycosides, the sugar being either ribose or 2-deoxyribose, although other pentoses are found in rare cases. Such compounds are termed nucleosides generically, and inosine (hypoxanthine nucleoside), adenosine, or guanosine



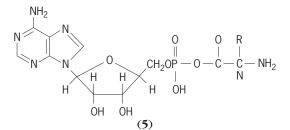
Formation of a purine following enzymatic reactions of its component compounds.

specifically. The name nucleotide refers to phosphorylated derivatives of nucleosides, the principal ones containing 5'-phosphate groups, as in guanosine 5'-phosphate and adenosine 5'-triphosphate (ATP). *See* ADENOSINE TRIPHOSPHATE (ATP).

Nucleic acids are polymers in which nucleotide units are linked by phosphodiester bonds between the 3'-hydroxyl group of one residue and the 5'-hydroxyl group of the next. The methylated xanthines, caffeine (1,3,7-trimethylxanthine), and theobromine (3,7-dimethylxanthine) present in coffee, tea, and cola nuts are responsible for the stimulatory action of these plant products.

Biosynthesis and metabolic functions. Most living organisms are capable of synthesizing purine compounds; in higher animals, this function is largely carried out in the liver and rapidly proliferating tissues such as spleen, bone marrow, and intestinal mucosa. The sequence of enzymatic reactions by which the initial purine product, inosine 5'-phosphate, is formed utilizes glycine, carbon dioxide, formic acid, and amino groups derived from glutamine and aspartic acid (see **illus.**). All of the intermediates in the pathway are derivatives of ribose 5-phosphate. Adenosine 5'-phosphate and guanosine 5'-phosphate are formed from inosine 5'-phosphate by amination reactions requiring aspartic acid or glutamine.

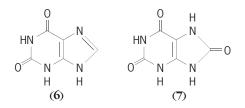
The energy made available from oxidative and nonoxidative metabolism of most foodstuffs in living cells is in part used to condense phosphoric acid and adenosine 5'-phosphate to form ATP. ATP is one of the central compounds in metabolism, acting to transfer, in the form of its energy-rich phosphoanhydride bonds, chemical energy derived from the degradation of foodstuffs to energy-requiring processes of biosynthesis, transport, muscular action, and nerve conduction. As an example, ATP combines with amino acids to form amino acyl adenylates (**5**), the activated intermediates of protein synthesis.



Further, the active forms of several vitamins,

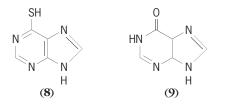
including niacin, pantothenic acid, riboflavin, and cobalamin (B_{12}) , are nucleotides containing adenosine phosphate.

Metabolic degradation of purine derivatives may also occur by hydrolysis of nucleotides and nucleosides to the related free bases. Deamination of adenine and guanine produces hypoxanthine and xanthine (6), both of which may be oxidized to uric acid (7). In most mammals, uric acid is further oxidized to



the excretory product, allantoin. Primates, including humans, are unable to oxidize uric acid and therefore excrete this substance as the end product of purine metabolism. The arthritic condition known as gout results from precipitation of uric acid or its salts in the joints of individuals with abnormally high levels of these materials in serum and tissues. Birds, certain terrestrial reptiles, and arthropods deposit uric acid or guanine in solid form as the major nitrogenous end product of metabolism. Bird guano, composed largely of uric acid, is harvested in enormous quantities for use as a fertilizer. *See* FERTILIZER; URIC ACID.

Medical uses. A vast number of purine-related compounds have been investigated as potential chemotherapeutic agents. In particular, 6-mercaptopurine (8), in the form of its nucleoside phosphate, inhibits several enzymes required for synthesis of adenosine and guanosine nucleotides, and thus proves useful in selectively arresting the growth of tumors. The pyrazolopyrimidine (9) has



been used in gout therapy. As a purine analog, this agent serves to block the biosynthesis of inosine phosphate, as well as the oxidation of hypoxanthine and xanthine to uric acid. As a result of its use, overproduction of uric acid is prevented and the primary cause of gout is removed. *See* CHEMOTHERAPY AND OTHER ANTINEOPLASTIC DRUGS; GOUT; PYRIMIDINE. Standish C. Hartman

Bibliography. L. Stryer, *Biochemistry*, 1995; T. W. Trevor and H. A. Simmonds, *Purines: Basic and Clinical Aspects*, 1991.

Push-pull amplifier

An electronic circuit in which two transistors (or vacuum tubes) are used, one as a source of current and one as a sink, to amplify a signal. One device "pushes"

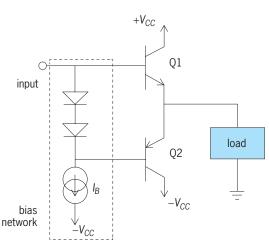


Fig. 1. Complementary-symmetry push-pull output stage. Q1 is an *npn* transistor and Q2 is a *pnp* transistor; $I_{\rm B}$ is a bias current; positive (+ $V_{\rm CC}$) and negative (- $V_{\rm CC}$) power supplies are shown.

current out into the load, while the other "pulls" current from it when necessary. A common example is the complementary-symmetry push-pull output stage widely used to drive loudspeakers (**Fig. 1**), where an *npn* transistor can source (push) current from a positive power supply into the load, or a *pnp* transistor can sink (pull) it into the negative power supply. The circuit functions as an amplifier in that the current levels at the output are larger than those at the input. *See* TRANSISTOR.

Biasing for classes A and B. A so-called bias network in a complementary-symmetry push-pull output stage (Fig. 1) functions to maintain a constant voltage difference between the bases of the two transistors. It can be designed either by setting a bias current, and diode sizes or by replacing it with a different network for class B, class A, or the common compromise, class AB mode of operation. In class B operation, where the bases of the transistors might simply be shorted together, only one transistor is "on" at a time and each is on average "on" for only 50% of the time; when the output current is zero, no current at all flows in the circuit. In class A operation a large voltage is maintained between the bases so that both devices stay "on" at all times, although their currents vary so that the difference flows into the load; and even when the output is zero, a large quiescent current flows from the power supplies. Class B operation is much more efficient than class A, which wastes a large amount of power when the signal is small. However, class B suffers from zero-crossing distortion as the output current passes through zero, because there is generally a delay involved as the input swings far enough to turn one transistor entirely off and then turn the other on. In class AB operation, some intermediate quiescent current is chosen to compromise between power and distortion.

Class AB amplifiers are conventionally used as loudspeaker drivers in audio systems because they are efficient enough to be able to drive the required

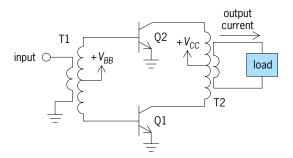


Fig. 2. Transformer-coupled push-pull stage. Both Q1 and Q2 are *npn* transistors, T1 is a phase-splitter transformer, T2 is a center-tapped transformer, and V_{BB} is a bias voltage source. Other symbols are explained for Fig. 1.

maximum output power, often on the order of 100 W, without dissipating excessive heat, but can be biased to have acceptable distortion. Audio signals tend to be near zero most of the time, so good performance near zero output current is critical, and that is where class A amplifiers waste power and class B amplifiers suffer zero-crossing distortion. A class AB push-pull amplifier is also conventionally used as the output stage of a commercial operational amplifier. *See* AUDIO AMPLIFIER; OPERATIONAL AMPLIFIER; POWER AMPLIFIER.

Transformer-coupled amplifier. A second important variant of the push-pull amplifier is the transformercoupled push-pull stage (Fig. 2), where a centertapped transformer is used to avoid the need for complementary device types (the npn and pnp transistors in the complementary-symmetry pushpull output stage). Here, both transistors are type npn, and when one of them turns on, positive current flows in the load (push), whereas when the other transistor turns on negative, load current flows (pull). Both transistors can only sink current, but their effects on the load are opposite because they are connected to opposite ends of the transformer. This configuration is particularly important when vacuum tubes are used instead of transistors, because there is no complementary tube; tubes operate only by electron flow. A phase-splitter transformer is responsible for turning on one transistor at a time, and can be replaced by an active-circuit implementation if appropriate. A bias voltage is also supplied to the two transistors through the phase-splitter transformer.

Push-pull circuits have an odd symmetry in their operation, meaning that their input-output characteristic satisfies f(x) = -f(-x), because of the symmetry of their structure. This property suppresses even-harmonic distortion, but the circuits may still show substantial distortion at odd harmonics of their input frequency, especially if they are biased in class B where there is zero-crossing distortion. *See* AMPLIFIER; DISTORTION (ELECTRONIC CIRCUITS). Martin Snelgrove

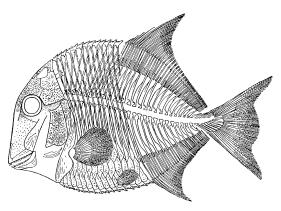
Bibliography. C. A. Holt, *Electronic Circuits, Digital and Analog*, 1978; J. Millman and A. Grabel, *Microelectronics*, 2d ed., 1987; A. S. Sedra and K. C. Smith, *Microelectronic Circuits*, 4th ed., 1997.

Pycnodontiformes

An order of very specialized deep-bodied fishes near the holostean level of organization that are known only from the fossil record. They were a widespread group that first appeared in the Upper Triassic of Europe with the genus *Eomesodon*, flourished during the Jurassic and Cretaceous (*Gyrodus, Coelodus, Gyronchus, Proscinetes, Mesturus*, and others), and persisted to the Upper Eocene (*Pycnodus* and *Palaeobalistum*). Pycnodontiforms are a closely interrelated group most commonly found preserved in marine limestone and associated with coraliferous facies.

They are characterized (see illus.) by a laterally compressed, disk-shaped body; long dorsal and anal fins, with each fin ray supported by its own endoskeletal element; an externally symmetrical tail; an axial skeleton having greatly extended neural and hemal spines that are expanded in some forms, and neural and hemal arches which are well ossifiedsometimes interdigitating around the notochordbut with no vertebral centra. The ganoid scales, if present, are usually articulated in vertical rows by a modified peg and socket mechanism; often the scalation covers only the anterior half of the body. In the head the subopercle is absent as an individual element and there is no interopercle. The mouth is small but with a powerful mandible. There are a few protruding chisel-like teeth at the front of the jaws. In the upper jaw, the maxillary bone is reduced or absent. Seated on an inner (splenial) expansion of each mandible are two or five rows of blunt, hemispherical teeth that oppose a similar battery of vomerine teeth in the roof of the mouth. No replacement teeth are known. The head is greatly modified in support of this dentition, so much so that the visual clues to relationships with other fishes which usually can be found in skull structure are almost completely obscured in this fossil group.

The origin of the pycnodonts remains one of the puzzling problems in paleoichthyology. These fish exhibit a combination of advanced and archaic characters. For example, the skull roof has a supraoccipital, sometimes with a crest developed, while the notochord is unrestricted. The scales are very like



Coelodus costae, a pycnodont from Lower Cretaceous of Italy; length to 4 in. (10 cm).

those of the chondrostean Platysomoidei in gross features but the fins (particularly the caudal) are holostean in nature. The maxilla, when present, is free from the cheek but the preopercle still covers much of the cheek area. The mandible has a high coronoid process, an advanced character; yet the parasphenoid is short and expanded posteriorly as in palaeoniscoids. An attractive hypothesis, of some weight, is the derivation of the pycnodontiforms from a deep-bodied chondrostean stock possibly within the Platysomoidei.

The teeth of pycnodontiforms are adapted for crushing shelled invertebrates, and shell fragments have been found inside the abdominal cavity of some specimens. The body type is an adaptation for life in a confined, quiet water habitat where great maneuverability is necessary, such as the interstices of a coral reef or in an area with much vegetation. A close comparison can be drawn between the modern, foraging type of reef fishes belonging to the teleost level and the pycnodonts. It is probable that the teleost forms as they evolved and spread came into direct contact with some of the pycnodontiforms and competed with them. In like manner, the appearance of the latter may have filled the habitat vacated by the earlier Platysomoidei. See PALAEONISCIFORMES. Ted. M. Cavender

Bibliography. W. N. McFarland et al., *Vertebrate Life*, 1979; A. S. Romer, *Vertebrate Paleontology*, 3d ed., 1966.

Pycnogonida

A class of marine arthropods, consisting of about 600 Recent and 3 Devonian species. The Pycnogonida, or Pantopoda, are commonly called sea spiders.

Morphology. The pycnogonids are characterized by reduction of the body to a series of cylindrical trunk somites supporting the appendages, a large specialized feeding apparatus called the proboscis, gonopores opening on the second joints of the legs, and a reduced abdomen. In many genera, such as *Nymphon* (**Fig. 1**), there are seven pairs of appendages, of which the first four, namely the che-

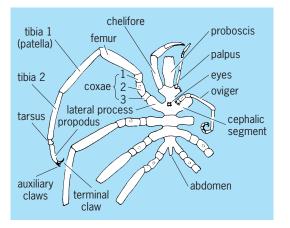


Fig. 1. Nymphon, external features in dorsal view.

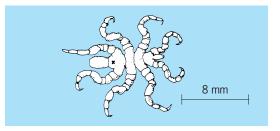


Fig. 2. *Pycnogonum stearnsi*, a common littoral pycnogonid of the Pacific Coast.

lifores, palpi, ovigers, and first walking legs, are on the first or cephalic segment. This segment also bears a dorsal tubercle containing four simple eyes. Each of the remaining three trunk segments bears a single pair of legs. In some genera the chelifores, palpi, or both are absent in the adult, and the ovigers are lacking in the females. Ovigers are always present in males and hold the ball-like egg masses. The extreme of reduction, in which the adults lack anterior appendages, except for ovigers in the male, is represented by Pycnogonum (Fig. 2), the common intertidal genus from which the group is named. In some species there is an additional trunk somite and pair of legs, and one species has 2 extra somites and 12 walking legs. These polymerous forms occur in paucigeneric families with a large number of species to a genus, and are not known in families with a large number of paucispecific genera.

Reduction of the body is accompanied by increased length of the legs, so that a typical pycnogonid appears to be a bundle of long legs. These contain diverticula of the digestive tract and gonads, and the sexual products ripen in the femora. The proboscis is a rigid tripartite structure, often longer than the combined trunk segments, with a three-cornered mouth at the apex, and containing a straining apparatus which enables the animal to feed on the juices of soft-bodied animals. The nervous system follows the usual arthropod pattern of supraesophageal ganglion, or brain, circumesophageal ring, paired ventral ganglia, and ventral nerve cords. The open circulatory system consists of a dorsal tubular heart with two or three pairs of lateral ostia. Respiratory and excretory systems are lacking. The femora of the males contain specialized cement glands which produce an adhesive substance for forming the egg ball.

Affinities. The chelicerate condition of the first pair of appendages, the patellar joint of the legs, and suppressed deuterocentrum are characters suggesting affinity with the Chelicerata, but the multiple genital openings on the legs and the occurrence of polymerous forms are unique among arthropods. The terminal joints of the oviger have intratarsal muscles, another unique feature, while the pretarsus, or propodus, of the legs has both retractor and extensor muscles, in common with both Chelicerata and Crustacea. The protonymphon larva (**Fig. 3**) of the pycnogonida invites comparison with the crustacean nauplius larva, but the three pairs of legs are not biramous, and the resemblances are superficial. The

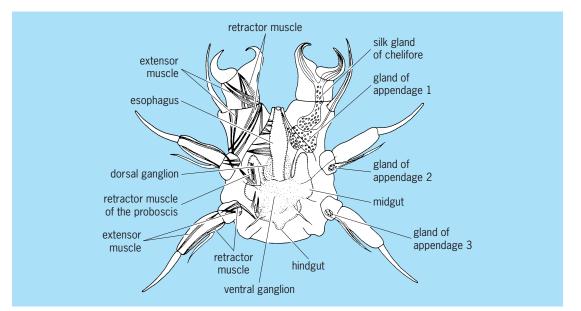


Fig. 3. Protonymphon larva, ventral view.

protonymphon has a number of specialized glands, producing materials for attaching to, or invading, the host, and the structure of the proboscis is well advanced. The common name, sea spider, refers to the appearance of the adults and does not indicate affinity with terrestrial spiders. *See* CHELICERATA.

Mode of life. Pycnogonids are found in all seas except the inner Baltic and Caspian, from intertidal regions to depths of 21,000 ft (6500 m), and one species is bathypelagic at about 3300 ft (1000 m). They are especially common in polar seas. Most of the intertidal species spend their lives in association with coelenterates as encysted, parasitic larval and juvenile stages, or are ectoparasitic as adults, being attached to anemones and hydroids by their claws and proboscides. A few have been found riding in hydromedusae, and some occur in the mantle cavity of bivalves or on nudibranchs and holothurians. Most

of the deep-sea species are known only as adults, and their mode of life is a mystery.

Classification. The Pycnogonida are classified primarily on the presence or absence of various anterior appendages; about 60 genera are recognized, grouped in 8 families, which are listed in the **table**. No ordinal distinctions can be recognized for Recent families, and many genera are artificially defined. In order to include Devonian *Palaeopantopus* in the Pycnogonida, the Recent forms are assigned to Pantopoda, and the fossil to Palaeopantopoda. *See* ARTHROPODA; PALAEOISOPUS. Joel W. Hedgpeth

Bibliography. C. P. Arango, Morphological phylogenetics of the sea spiders (Arthropoda: Pycnogonida), *Organisms Diversity Evolution*, 2:107–125, 2002; F. Arnaud and R. N. Bamber, The biology of Pycnogonida, *Adv. Mar. Biol.*, 24:1–96, 1987; J. A. Dunlop and C. P. Arango, Pycnogonid affinities: A review, *J. Zool.*

Recent families of Pycnogonida					
Family	Chelifores	Palpi	Ovigers	Genera	
Nymphonidae	Present	5-jointed	10-jointed in both sexes	Includes largest genus, Nymphon (over 150 species) and polymerous Pentanymphon and Sexanymphon	
Callipallenidae	Present	Lacking or reduced	10-jointed in both sexes		
Phoxichilidiidae	Present	Lacking	5–9-jointed in male only		
Endeidae	Absent	Absent	7-jointed in male only	Monogeneric, paucispecific, Endeis	
Ammotheidae	Present, usually small and achelate	Well developed, 4–10-jointed	9–10-jointed in both sexes	Many genera	
Austrodecidae	Absent	Present, 5–6-jointed	In both sexes reduced, 4–7-jointed	Monogeneric, Austrodecus	
Colossendeidae	Lacking except in polymerous forms	Long, 9–10-jointed	10-jointed in both sexes	Mostly deep-water forms with large proboscides; genera are Colossendeis, polymerous Pentacolossendeis, Decolopoda, and Dodecolopoda	
Pycnogonidae	Lacking	Lacking	6–9-jointed in male only	Shallow water to shore forms; includes Pycnogonum and polymerous Pentapycnon	

Syst. Evol. Res., 43(1):8–21, 2005; J. W. Hedgpeth, On the phylogeny of the Pycnogonida, *Acta Zool.* (*Stockholm*), 35:193–213, 1954; P. E. King, *Pycnogonids*, 1974; S. M. Manton, *The Arthropoda: Habits, Functional Morphology, and Evolution*, 1978; D. A. Staples, Sea spiders or pycnogonids (phylum Arthropoda), pp. 1040–1072 in *Marine Invertebrates of Southern Australia, Part IIIM*, S. A. Shepherd and M. Davies, eds., South Australian Research and Development Institute, Adelaide, 1997.

Pygasteroida

An order of Eognasthostomata which exhibits various stages in the backward migration of the anus out of the apical system. They have four genital pores, noncrenulate tubercles, and simple ambulacral plates. All members are referred to a single family, the Pygasteridae. They apparently arose from Triassic Pedinidae and occur in the Jurassic and Cretaceous of the Northern Hemisphere. They were formerly classified with other bilaterally symmetrical echinoids in the artificial assemblage Irregularia. *See* ECHINODERMATA; IRREGULARIA. Howard B. Fell

Pyrenulales

An order of the class Ascolichenes, also known as the Pyrenolichenes. As now circumscribed, the Pyrenulales includes only those lichens with perithecia that contain true paraphyses and unitunicate asci. Other pyrenolichens with pseudoparaphyses and bitunicate asci have been transferred to the Pseudosphaeriales. The flask-shaped perithecia are uniformly immersed in the medulla of the thalli with a small ostiole opening at the surface. The asci and paraphyses arise from a blackened hypothecium and line the walls of the perithecium. The spores eventually burst the ascal walls and ooze out through the ostiole in a jelly matrix. Details of ascal development have been described rather fully for Dermatocarpon aquaticum. The Pyrenulales are almost all crustose in growth form, with very simple internal structure. The exception is Dermatocarpaceae, large umbilicate species often confused with the typical rock tripe, Umbilicaria, in the Lecanorales.

There are about 10 families, 50 genera, and more than 1500 species in the Pyrenulales. The major taxonomic criteria for separating genera and species are the septation and color spores, since vegetative characters are so poorly developed. The larger families are as follows.

The Dermatocarpaceae contain four genera with umbilicate or squamulose growth form. Most of the species grow on limestone or calcareous soils.

In the Pyrenulaceae are about 10 genera, all crustose species and most common on tree bark in the tropics. They are closely related to the nonlichenized Sphaeriales, but the perithecia are never produced in true stoma. The Strigulaceae comprise about six genera of crustose species confined chiefly to leaves of evergreen trees in the tropics. These peculiar lichens form extensive crusts on or under the cuticle of leaves without seeming to damage the host plant.

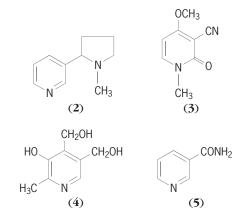
The Verrucariaceae comprise about 8 genera of crustose species typically found on rocks, especially in intertidal or salt-spray zones along rocky coastlines. These are the only truly amphibious lichens. Mason E. Hale

Pyridine

An organic heterocyclic compound containing a triunsaturated six-membered ring of five carbon atoms and one nitrogen atom. Pyridine (1) and



pyridine homologs are obtained by extraction of coal tar or by synthesis. The following are available in commercial quantities: pyridine, 2-, 3-, and 4methylpyridine (also known respectively as α -, β and γ -picoline), 2,4-dimethyl-, 2,6-dimethyl-, and 3,5dimethylpyridine (also known respectively as 2,4-, 2,6-, and 3,5-lutidine), 5-ethyl-2-methylpyridine (also called aldehydecollidine), and 2,4,6-trimethypyridine (also called 2,4,6-collidine). Other pyridine derivatives produced on a large scale include nicotinic acid (pyridine-3-carboxylic acid) for preparation of nicotinamide, nicotine for its insecticidal properties, 2-aminopyridine for synthesis of medicinals, piperidine (hexahydropyridine) as a solvent, and 2-vinylpyridine as a polymerizable monomer. The pyridine system is found in natural products, for example, in nicotine (2) from tobacco, in ricinine (3) from castor bean, in pyridoxine or vitamin B₆ (4), in nicotinamide or niacinamide or vitamin P(5),



and in several groups of alkaloids. *See* HETEROCYCLIC COMPOUNDS.

Properties. Pyridine (1) is a colorless, hygroscopic liquid with a pungent, unpleasant odor. When anhydrous it boils at $115.2-115.3^{\circ}C$ (239.4–239.5°F), its

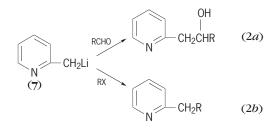
density (20/4) is 0. 98272, and $n_{\rm D}^{20}$ is 1.50920. Pyridine is miscible with organic solvents as well as with water. A constant-boiling mixture, by 92°C (198°F) forms with three molecules of water. Dry pyridine is obtained by treatment with barium oxide followed by calcium hydride or phosphorus pentoxide. Pyridine is a tertiary amine (p K_a 5.17 at 25°C or 78°F) that combines readily with Brönsted and Lewis acids. The pyridine system is aromatic. It is stable to heat, to acid, and to alkali. It undergoes electrophilic substitution with difficulty, with the 3 position favored in sulfonation and nitration. On the other hand, it undergoes nucleophilic substitution with ease at the 2 position, and lightly less readily at the 4 position. Its resonance energy is 35 kcal/mole. Pyridine is used as a solvent for organic and inorganic compounds, as an acid binder, as a basic catalyst, and as a reaction intermediate.

Pyridine is an irritant to skin (eczema) and other tissues (conjunctivitis), and chronic exposure has been known to cause liver and kidney damage. Repeated exposure to atmospheric levels greater than 5 parts per million is considered hazardous.

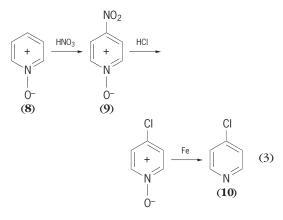
Oxidation of pyridine homologs by nitric acid or by permanganate converts the substituent group in a preparative manner to a carboxylic acid. Reaction at the methyl group of 2- and 4-methypyridine tends to occur more readily than at the methyl group of 3-methylpyridine. Thus, 2- and 4-methylpyridine condense with benzaldehyde to give styryl derivatives [6; reaction (1)], whereas 3-methylpyridine

$$(N \to CH_3 \xrightarrow{C_6H_5CH0} (N \to CH = CHC_6H_5 (1))$$
(6)

does not react. With butyllithium or sodium amide, 2- and 4-methypyridines metalate to give pyridylmethyl metals (7), which react normally with carbonyl compounds [RCHO; reaction (2*a*)] and with alkyl halides [RX; reaction (2*b*)].

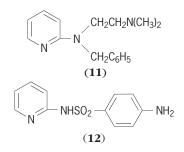


Many halogenated derivatives of pyridine are known. Direct bromination at 300-400°C (570-750°F) places bromine at the 3 position and 5 position. 2-Bromopyridine can be prepared from 2-pyridone and phosphorus oxybromide. 3-Bromopyridine is prepared by application of the Sandmeyer process to 3-aminopyridine or by mercurating and then brominating pyridine. 2-Chloropyridine is formed when *N*-methyl-2pyridone is treated with phosphorus pentachloride. 4-Chloropyridine (10) can be prepared by nitrating pyridine-N-oxide (8), exposing the 4-nitropyridine-N-oxide (9) to the action of concentrated hydrochloric acid, and removing the oxide oxygen by ironacetic acid reduction [reaction (3)], but is usually



made by heating *N*-(4-pyridyl) pyridinium chloride (from pyridine and thionyl chloride). The 2- and 4-halo substituents are more readily replaced by hydroxy, alkoxy, and amino than the 3-halo substituent. Bromopyridines form Grignard and lithium derivatives that react normally.

3-Nitropyridine, from nitration of pyridine, can be converted to 3-aminopyridine. The latter is usually prepared by Hofmann rearrangement of nicotinamide (5). 4-Nitropyridine-*N*-oxide (9) with iron and acetic acid gives 4-aminopyridine. 2-Aminopyridine, prepared on an industrial scale by direct amination of pyridine with sodamide, is utilized in the manufacture of the antihistaminic, pyribenzamine (11), and the bacteriostatic agent, sulfapyridine (12).

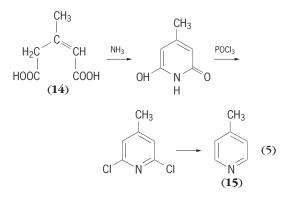


In most of its reactions, 3-hydroxypyridine behaves as a normal phenol, just as 3-aminopyridine behaves as a normal aromatic amine. However, hydroxy or amino groups at the pyridine 2 or 4 position show some reactions that are not characteristic of phenols or aromatic amines. Hydroxyl or amino groups on any pyridine position make electrophilic substitution easier and, as ortho-para directing groups, take control of the orientation. 2-Pyridone is prepared from 2-aminopyridine by a diazotization procedure. 3-Hydroxypyridine is produced by sulfonation of pyridine followed by alkali fusion of the pyridine-3-sulfonic acid.

Preparation. Laboratory synthetic methods can lead to pyridines with no oxygen at positions 2

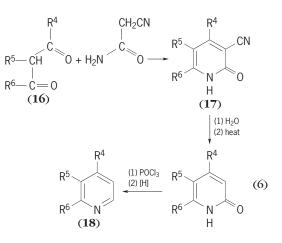
or 6, or to 2-hydroxypyridines (also known as 2pyridone), or to 2,6-dihydroxypyridines (also known as 2,6-pyridinone). These two pyridine derivatives exist almost entirely in their tautomeric forms. For example, 2-hydroxypyridine 13a can tautomerize to 13b as shown in reaction (4). Glutaconic acid cy-

clizes with ammonia to give 6-hydroxy-2-pyridones. The oxygen at the 2 position and 6 position can be removed by standard conversions with phosphorus oxychloride to the 2,6-dichloro derivative, followed by reductive dechlorination. In this way, for example, β -methyl-glutaconic acid (14) can be converted to 4-methylpyridine [15; reaction (5)]. 2-Pyridones (17)

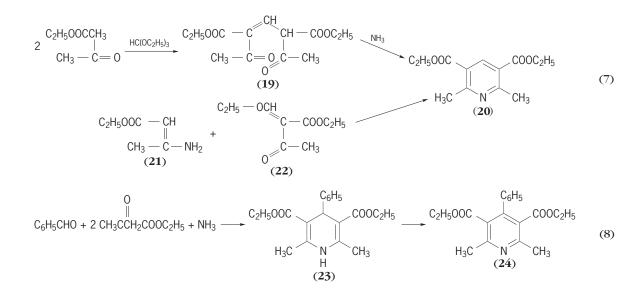


are formed when 1,3-dicarbonyl compounds (16), or their equivalents, react with cyanoacetamide. Subsequent steps remove the cyano groups as well as the

oxygen from (17) to furnish substituted pyridines [18; reaction (6)].



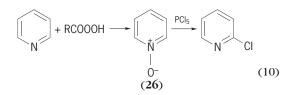
1,5-Dicarbonyl compounds or their equivalents cyclize with ammonia to give pyridines. Thus, the diketone (19) from acetoacetic ester and ethyl orthoformate gives the pyridine (20); in a closely related process, the reaction of ethyl β -aminocrotonate (21) with ethoxymethyleneacetoacetic ester (22) gives the same product [reactions (7)]. The Hantzsch synthesis combines four molecules to form an intermediate dihydropyridine derivative (23), which can be readily oxidized to the corresponding completely aromatic derivative [24; reaction (8)]. The Chichibabin aldehyde-ammonia synthesis involves the condensation of ammonia with aldehydes and ketones. Generally, mixtures of pyridines are obtained, with the course of the reaction depending on such factors as nature and proportion of reactants, reaction time, temperature, and catalyst. In a commercial process, 5-ethyl-2-methylpyridine (25) is obtained in unusually high yields (60-70%) from acetaldehyde



and ammonia [reaction (9)]. Synthetic industrial pyridine and its homologs are prepared from aldehydes and ammonia over a suitable catalyst.

$$NH_3 + 4CH_3CO \longrightarrow CH_3CH_2 \xrightarrow{(9)}_{N \to CH_3} (9)$$

Derivatives. Pyridine compounds containing positively charged nitrogen include simple and quaternary pyridinium salts, acylpyridinium salts, and pyridine-*N*-oxides. The reaction of pyridine with alkylating agents xs quaternary salts, crystalline solids whose aqueous solutions conduct electricity. The quaternary salts are readily oxidized by alkaline ferricyanide to the *N*-substituted-2-pyridones. With acyl halides, pyridine forms *N*-oxide (**26**), prepared by oxidation of pyridine with organic peracids, which, in turn, reacts with phosphorus pentachloride to give 2- and 4-chloropyridine [reaction (10)]

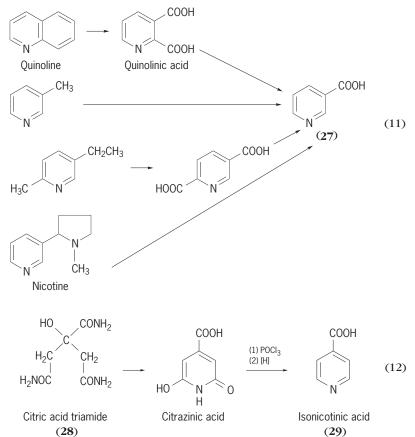


and with acetic anhydride to give 2-acetoxypyridine. Pyridine-*N*-oxides nitrate at the 4 position, and by so doing provide a route to 4-substituted derivates.

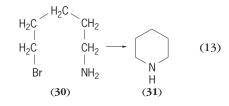
Pyridine aldehydes are prepared by oxidation of groups already on the ring. Acetylpyridines can be synthesized from pyridine carboxylic esters and ethyl acetate by the Claisen condensation. The reactions of pyridine aldehydes and ketones are normal.

Many pyridine carboxylic acids are known. Their reactions as dipolar materials are not exceptional. Thermal decarboxylation is a standard process, with loss of carboxyl from position 2 easier than from 3 or 4, and loss of carboxyl from position 4 easier than from 3. Nicotinic acid (27) is manufactured by oxidation of 3-methylpyridine (β -picoline) or nicotine; it is also obtained by oxidation of either 2-methyl-5-ethylpyridine or quinoline, followed by decarboxylation of the resulting pyridine dicarboxylic acids [reactions (11)]. Pyridine-2-carboxylic acid (a-picolinic acid) can be prepared by oxidation of 2-methylpyridine or by carbonation of 2-pyridyllithium. 4-Pyridine-carboxylic acid (isonicotinic acid) is obtained by oxidation of 4-methylpyridine or by synthesis from citric acid. See (28) to (29), as in reaction (12). The acid hydrazide of isonicotinic acid (Isoniazid) is a tuberculostatic agent.

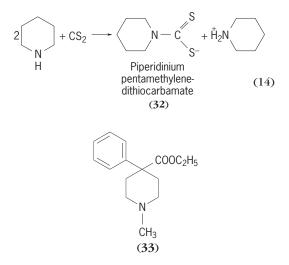
Although dihydro- and tetrahydropyridines are known, the hexahydropyridines, or piperidines, are the most common reduced forms of pyridine. The piperidines may be prepared by reduction of pyridines or by cyclization of bifunctional com-



pounds, for example, the conversion of 5-bromo-1aminopentane (**30**) to piperidine [**30**; reaction (13)].



Piperidine, the parent compound, is a colorless, unpleasant-smelling liquid (bp 105.6°C or 222.1°F),



completely miscible with water. The general properties of piperidine are those of a normal secondary aliphatic amine and, as such, piperidine (p K_a 11.1) is a much stronger base than pyridine (p K_a 5.17). Piperidine carboxylic acids have been investigated in connection with naturally occurring amino acids (for example, pipecolinic acid is piperidine-2-carboxylic acid) as well as with the degradative and synthetic chemistry of quinine. The reaction product (**32**) from piperidine and carbon bisulfide is a rubber accelerator [reaction (14)]. 4,4-Di-substituted piperidines such as Demerol (**33**) are analgesics.

Walter J. Gensler Bibliography. R. A. Abramovitch (ed.), *Pyridine* and Its Derivatives, suppl., vols. 1-4, 1974-1975; D. H. R. Barton et al. (eds.), *Comprehensive Or* ganic Chemistry, 1978; G. R. Newkome (ed.), *Pyridine and Its Derivatives*, vol. 14, pt. 5, 1984; P. Tomasik and Z. Ratajewicz, *Pyridine-Metal Compounds*, vol. 14, pt. 6, 1985.

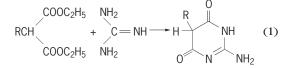
Pyrimidine

A heterocyclic organic compound (1) containing ni-

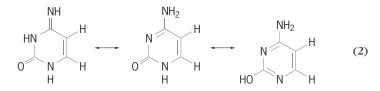


trogen atoms at positions 1 and 3. Naturally occurring derivatives of the parent compound are of considerable biological importance as components of nucleic acids and coenzymes and, in addition, synthetic members of this group have found use as pharmaceuticals. *See* COENZYME.

Chemistry. Pyrimidine (1,3-diazine) and its derivatives are weakly basic substances with characteristic ultraviolet absorption spectra. The diazine structure, like that of benzene, may be regarded as a resonancestabilized system. As a result it is unreactive to all but very vigorous reagents. A classical mode of synthesis of pyrimidine compounds involves condensation of an ethyl malonate with guanidine, urea, or amidine, as in reaction (1). The substituent groups can be

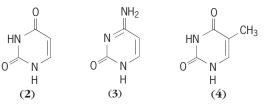


varied either by choice of different starting reagents or by replacement after formation of the ring system. Certain substituted pyrimidines can exist in tautomeric forms; an example is given by reaction (2).

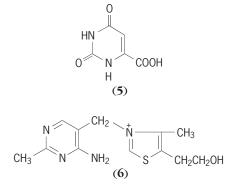


The forms of the substituent groups shown in the middle structure are the most stable.

Occurrence. Pyrimidine compounds which are found universally in living organisms include uracil (2), cytosine (3), and thymine (4). Together with

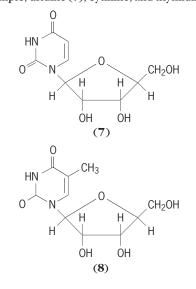


purines these substances make up the "bases" of nucleic acids, uracil and cytosine being found characteristically in ribonucleic acids, with thymine replacing uracil in deoxyribonucleic acids. A number of related pyrimidines, including 5-methyl-cytosine and 5-hydroxymethyl cytosine, also occur in lesser amounts in certain nucleic acids. Other pyrimidines of general natural occurrence are orotic acid (5) and thiamine (vitamin B₁, **6**).



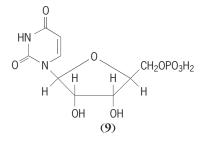
See DEOXYRIBONUCLEIC ACID (DNA); RIBONUCLEIC ACID (RNA).

The pyrimidine bases are normally found in combination with one of two sugars, ribose or 2deoxyribose, in the form of nucleoside derivatives, for example, uridine (7), cytikine, and thymidine (8).



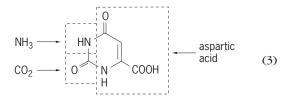
Further, the 5'-hydroxyl group of the sugar residue may exist in ester linkage with phosphoric acid,

forming a nucleoside phosphate, such as uridine 5'-phosphate (uridylic acid, **9**).



The class of compounds composed of bases, sugars, and phosphate is referred to as nucleotides, a term including nucleoside di- and triphosphates and a variety of metabolically important coenzymes. Nucleic acids are polymers of nucleotide units which are bridged by phosphodiester bonds between the 3'-hydroxyl group of one unit and the 5'-hydroxyl group of the next. The specificity of nucleic acids in genetic processes is attributed to the formation of complementary pyrimidine-purine pairs (uracil or thymine with adenine, and cytosine with guanine) which are stabilized by intramolecular hydrogen bonds between the ring nitrogen atoms and a substituent amino and carbonyl groups.

Biosynthesis and metabolic function. The synthesis of pyrimidines takes place in virtually all living organisms, and proceeds by way of the intermediate formation of orotic acid. This material arises from ammonia, carbon dioxide, and aspartic acid in a sequence of four enzyme-catalyzed reactions (3). Uridine 5'-

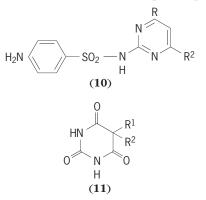


phosphate is formed after addition of the ribose 5phosphate moiety to orotic acid, followed by loss of the carboxyl group. Cytidine nucleotides (such as the 5'-triphosphate) result from amination of uridine 5'-phosphate. The biological phosphorylating agent, adenosine triphosphate, acts to convert the nucleoside 5'-phosphates to the corresponding diand triphosphates. In addition to their direct utilization in nucleic acid synthesis, pyrimidine nucleotides have key metabolic functions in the transfer of activated subunits in the synthesis of many cellular components. As examples, glycogen is built up by incorporation of glucosyl units derived from uridine diphosphoglucose. *See* PURINE.

Cytidine-containing nucleotides play comparable roles in the synthesis of complex lipids found in membranes and nervous tissue. Presumably in such reactions pyrimidine bases are not of primary chemical importance but serve as sites of recognition and specificity in the enzymic processes.

Synthetic pyrimidines of interest. Among the sulfa drugs, the pyrimidine derivatives, sulfadiazine, sulfamerazine, and sulfamethazine, have general for-

mula (10). They inhibit folic acid biosynthesis in microorganisms. The barbiturates (11) possess potent



depressant action on the central nervous system.

This group includes Veronal (R_1 and R_2 = ethyl), phenobarbital ($R_1 = ethyl$, $R_2 = phenyl$), Amytal $(R_1 = \text{ethyl}, R_2 = \text{isoamyl})$, Seconal $(R_1 = \text{allyl},$ $R_2 = 2$ -pentyl), and pentobarbital (Nembutal, $R_1 =$ ethyl, $R_2 = 2$ -pentyl). The effect of many pyrimidinerelated compounds upon nucleotide and nucleic acid metabolism has been examined with the hope of finding useful chemotherapeutic agents. The medical value of such agents has been disappointing even though certain of these have definite biological action. Thus, derivatives of 5-fluorouracil block the methylation step in thymidine phosphate synthesis, and 5-bromouracil and 5-iodouracil are incorporated into deoxyribonucleic acid in place of thymine to produce a polymer of increased susceptibility to mutation. See BARBITURATES; CHEMOTHERAPY AND OTHER ANTINEOPLASTIC DRUGS; MUTATION; NUCLEIC Standish C. Hartman ACID; SULFONAMIDE.

Bibliography. G. M. Blackburn and M. J. Gait, Nucleic Acids in Chemistry and Biology, 2d ed., 1996; M. Florkin and E. H. Stotz (eds.), Comprehensive Biochemistry, vol. 19: Metabolism of Amino Acids, Proteins, Purines, and Pyrimidines, 1981; D. B. Marks, Biochemistry, 3d ed., 1998; K. Mikanagi, Purine and Pyrimidine Metabolism, 1989; E. H. Rodd (ed.), Rodd's Chemistry of Carbon Compounds, vol. 2, pts. A and B, 2d rev. ed., 1968; E. Smith et al., Principles of Biochemistry, 7th ed., 1983.

Pyroclastic rocks

Rocks formed by the accumulation of fragmental volcanic materials (also called pyroclasts) ejected during explosive eruptions. Pyroclasts can vary widely in origin, fragment size and shape, composition, texture, structure (such as their arrangement in the resulting deposit), and other properties (see **table**). Despite such differences, all pyroclastic materials and the great variety of pyroclastic rocks they form share a common, unifying attribute—all are produced by explosive volcanic activity. Pyroclastic deposits are included under a broader category of volcaniclastic deposits, comprising all fragmented volcano-derived material. In some cases, the explosive fragmentation is caused by the interaction of

Average size of pyroclast	Name and shape of pyroclast	Physical state when erupted or deposited	Deposit resulting from pyroclast accumulation	
			If mostly unconsolidated	If mostly consolidate
	Block (angular)	Solid		
	Bomb (rounded or aerodynamically shaped)	Semisolid (still fluid)	Agglomerate, or deposit of blocks or bombs, or pyroclastic breccia	Agglomerate, or pyroclastic breccia
	Cinder, scoria (round to angular)	Semisolid > or liquid		
	Pumice, spatter (round to angular)	Largely liquid		
64 mm*				
	Lapillus [†] (round to angular)	Liquid or solid	Lapilli deposit	Lapillistone
2 mm				
	Coarse ash grain (generally angular)	Liquid or solid	Coarse ash deposit	Coarse ash tuff
1/16 mm				
	Fine ash grain, or volcanic dust (generally angular)	Liquid or solid	Fine ash deposit, or volcanic dust deposit	Fine ash tuff, or dust tuff

water with hot volcanic ejecta, producing hydrovolcanic materials, which are commonly regarded as pyroclastic *sensu lato* and will be discussed as such. *See* CALDERA; VOLCANO.

Pyroclasts are classified into three principal types according to their origin, which can combine in any proportion in any given pyroclastic rock: juvenile, cognate (or accessory), and accidental. Juvenile pyroclasts are produced by the violent shredding and disintegration of the erupting magma (in liquid or plastic state). Typically, juvenile pyroclasts have a glassy texture and contain small cavities (vesicles) that formerly were bubbles of volcanic gas in the magma. The cognate varieties are formed from the fragmentation of solidified volcanic rocks from previous eruptions from the same volcanic system. Accidental pyroclasts are fragments of geologically older rocks unrelated to the volcano and can be of any composition or rock type (igneous, sedimentary, or metamorphic). Cognate and accidental pyroclasts represent fragments of preexisting solid rock that were incorporated into the rising magma before the explosive eruption. Apart from their origin, pyroclasts are subdivided according to their size and shape, texture (glassy or crystalline), mineral or chemical composition, and other characteristics (see table). Of these distinguishing criteria, fragment size (average diameter) is perhaps the most important, resulting in three main subcategories of pyroclasts: bombs or blocks (>64 mm), lapilli (64-2 mm), and ash (<2 mm). See VOLCANIC GLASS.

From studies of pyroclastic rocks formed in the geologic past, combined with direct observations and investigations of pyroclastic deposits formed during historical explosive eruptions, volcanologists have recognized two principal modes of deposition of pyroclastic rocks: fall processes and flowage processes. However, these two processes represent end members of a continuum, and some deposits of pyroclastic rock exhibit features that indicate both kinds of processes were operative to some degree, and sometimes in a complex manner, in their formation. *See* IGNEOUS ROCKS; MAGMA; METAMORPHIC ROCKS; PETROLOGY; SEDIMENTARY ROCKS.

Fall processes and deposits. These processes involve the explosive ejection of pyroclasts high into the air and the fall of these airborne materials, because of gravity, back to the ground, forming pyroclastic fall deposits (sometimes also called air-fall deposits). Tephra refers to any airborne pyroclasts regardless of size, shape, and other physical properties, but it tends mostly to be used for lapilli- and ash-size fragments. Nearly all fall deposits are produced by one or a combination of two processes: (1) fallout of tephra from an eruption column (or eruption plume) jetting upward from the eruptive vent; and (2) fallout of material carried aloft in ash clouds convecting above ground-hugging pyroclastic flows.

Compared with highly destructive pyroclastic flows, the potential hazards to life and property posed by tephra fall processes generally are limited and less severe. The primary concern relates to collapse of roofs because of the weight of the accumulating tephra (especially if wet)—an effect similar to adding the weight of a rapid and intense fall and accumulation of snow.

Fallout from eruption column. An eruption column is the mushroom-shaped plume of volcanic debris and gas that rises above the vent during an explosive eruption. The narrow bottom part of the plume (the mushroom stem) is where the decompressing, expanding volcanic gases are thrusting upward at high velocity. The upper parts of the plume are sustained by thermal energy and buoyancy, before topping out at some height where the bulk density of the plume equals that of the ambient atmosphere. For powerful and sustained explosive eruptions, their eruption columns can rise tens of kilometers high and can

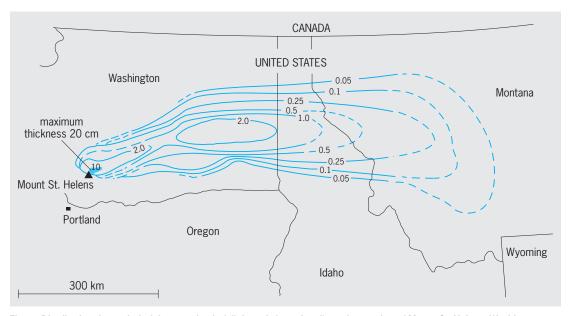


Fig. 1. Distribution downwind of the pyroclastic fall deposit from the climactic eruption of Mount St. Helens, Washington, on May 18, 1980. The contours (isopachs) are drawn through points of measured equal thickness (in centimeters) of this ash layer. (Modified from USGS Prof. Pap. 1250)

remain active for hours to days. The largest and heaviest pyroclasts in the plume fall out closest to the vent, and the size of the tephra and the thickness of the resulting fall deposits typically decrease with increasing distance downwind from the vent. *See* VOL-CANOLOGY.

Tephra fallout from the climactic eruption of Mount St. Helens (Washington) on May 18, 1980, produced by an eruption column more than 20 km high that was active for about 9 h, formed measurable pyroclastic fall deposits hundreds of kilometers downwind, and covered an area of more than 57,000 km² (Fig. 1). While impressive, the fall deposits from this eruption of Mount St. Helens pale by comparison to some of the prehistoric explosive eruptions from this volcano and to those produced by more voluminous historical and prehistoric explosive eruptions at other volcanoes: Mount Pinatubo, Philippines, 1991; Novarupta (Katmai), Alaska, 1912; Mount Mazama (Crater Lake), Oregon, 6800 years ago; Yellowstone caldera, Wyoming, which has produced three huge explosive eruptions during the past 2 million years. Because of their widespread distribution, the pyroclastic fall deposits from such powerfully explosive eruptions provide easily identifiable ash layers in the geologic record, and serve as important time markers in geologic mapping studies to reconstruct the geological histories of regions, even those far distant from the eruptive sources. A diagnostic characteristic of such regional ash marker beds is that they tend to drape the surrounding topography, such that their observed thicknesses-in any given direction and distance from the eruptive vent-are similar, whether measured on topographically high elevations or in valleys and other depressions. Tephra from relatively weak explosive eruptions (such as cinder, scoria, and spatter) fall close to the vent, piling up to form conical structures with slopes near or at the angle of repose. See STRATIGRA-PHY.

Fallout from ash clouds associated with pyroclastic flows. Pyroclastic fall deposits can also be formed from finergrained pyroclasts settling out from ash clouds that typically roil above pyroclastic flows moving along the ground. Much of the material in such clouds is separated out of the pyroclastic flows by volcanic gases flushing upward through the flows and, once airborne, the material can rise high into the air in turbulent, convecting plumes before settling out. Fall deposits formed in this manner are called ashcloud-derived fall deposits or co-ignimbrite ash-fall deposits, as opposed to eruption-plume-derived fall deposits. It is not always possible to distinguish unambiguously between these two types of pyroclastic fall deposits. It also should be emphasized that the fall deposits associated with pyroclastic flows can be as voluminous and widely dispersed as those generated from eruption columns. Many of the volcanic ash layers found in deep-sea sediments are believed to be co-ignimbrite ash fall produced by explosive eruptions on land. For example, a prominent ash layer in deep-sea cores taken from the Indian Ocean was derived from fallout from a gigantic explosive eruption at Lake Toba, Sumatra (Indonesia), about 75,000 years ago. See IGNIMBRITE.

Flowage processes and deposits. Many but not all pyroclastic rocks are formed by the deposition of hot pyroclasts transported in fast-moving surface flows rather than by deposition of fragments, which are largely cooled to ambient temperature by falling through the air. Even though these flowing masses are heavier than air, they are highly mobile because of their high temperature and gas content. They form pyroclastic flows and surges that sweep down steep slopes and valleys at high speeds (from 10 to several hundred meters per second) and can travel great distances (>100 km) from the eruptive centers. Such flows, which vary considerably in bulk density and flow characteristics, are typically composed of two parts: (1) the pyroclastic flow proper—a

ground-hugging basal flow with a high concentration of solids and a density close to that of the resulting deposit; and (2) the associated pyroclastic surge-an overriding less dense, turbulent cloud of ash with a low concentration of solids and density much less than that of the resulting deposit. Pyroclastic surges flow mostly above rather than hug the ground. Pyroclastic flows and surges are also variously termed glowing avalanches or clouds (nuées ardentes in French) or simply ash flows. A less common pyroclastic flow phenomenon is the laterally directed blast (or lateral blast), which appears to share features of both pyroclastic flows and surges. Lateral blasts are generally triggered by the sudden depressurization (uncorking) of a volcanic system by catastrophic collapse at a volcano and associated landsliding.

Because the paths of denser pyroclastic flows are controlled largely by topography, the thicknesses of pyroclastic flow units can vary widely, ranging from less than 0.1 m to more than 100 m where ponded in depressions. If the emplacement temperature and gas content are sufficiently high, the pyroclasts in thick deposits stick together to form welded tuffs or welded ash flows. In steep terrains, pyroclastic flows may drain completely from the upper slopes, so that their deposits are preserved only in the valley floors flanking the volcano. In marked contrast to pyroclastic fall deposits, which invariably become thinner away from the eruptive source, pyroclastic flow

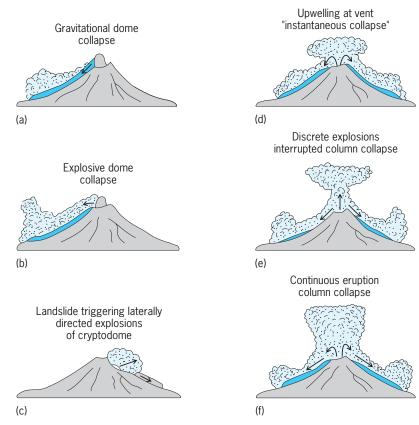


Fig. 2. Observed or postulated mechanisms for the generation of pyroclastic flow and surge deposits (*from W. E. Scott*, 1989). (*a*, *b*) Mechanisms that commonly produce block-and-ash flow deposits. (*b*, *c*) Mechanisms that trigger laterally directed blasts. (*d*–*f*) Mechanisms that represent the most common of the many proposed origins of pyroclastic flow deposits based on collapse of eruption columns.

deposits can become thicker away from the source. Lower-density pyroclastic flows and surges, because they are little influenced by topography, are able to advance great distances over gently sloping ground, and can surmount ridges hundreds of meters high. The more dense pyroclastic flows tend to fill valleys and other topographic depressions, and the deposits of pyroclastic rock they form can attain great thicknesses. In contrast, the less dense pyroclastic surges, even though they can be equally or more greatly traveled, generally leave much thinner deposits, some of which may not be preserved in the geologic record because of postdeposition erosion and reworking. *See* TUFF.

Pyroclastic flows, with their high abundance of hot solid fragments, are characterized by high temperatures (>800°C), whereas pyroclastic surges and laterally directed blasts can be either hot or cold. Hot pyroclastic surges are almost always associated with pyroclastic flows but, because of their lower ratios of hot solids to gas, rarely exceed 300°C. Cold pyroclastic surges (also called base surges) typically are formed by hydrovolcanic explosions caused by the interaction of rising magma with shallow ground water or surface water, or by the sudden heating of water trapped beneath hot lava flows or pyroclastic flows, such as along streams, lakes, oceans, and ice and snow fields. "Hydrovolcanic" is an inclusive term applied to pyroclastic products formed by explosive interaction between hot volcanic material and water. Hydrovolcanic activity, however, has characteristics quite different from pyroclastic activity sensu stricto and commonly produces relatively small volcanic craters (such as maars, tuff rings, and tuff cones) with diameters that are wide relative to the height of their rims.

As with the fall of pyroclastic material into the sea from on-land explosive eruptions, pyroclastic flows and surges from explosive activity of island volcanoes have been observed to enter the sea. Some well-documented examples include the 1883 Krakatau (Java) eruption, the 1902 Mont Pelée eruption, and the 1996-1997 explosive eruptive activity at Soufriére Hills Volcano, Montserrat (Lesser Antilles, Caribbean Sea). The resulting submarine deposits, called subaqueous pyroclastic flows, share many of the features and attributes of their on-land counterparts but differ in details as regards grain-size distribution, texture, and structure. Occurrences of similar-looking pyroclastic rocks, interpreted to be subaqueous in origin, have been recognized throughout the geologic record, dating back more than 2500 million years. Some specialists have proposed that, under certain conditions, pyroclastic flow deposits can also be produced by explosive eruptions at shallow depths underwater. The submarine deposition of pyroclastic rocks can be intermingled with deposition of reworked volcanic debris as well as nonvolcanic sediments, resulting in volcaniclastic aprons that are commonly found offshore of nearocean continental volcanoes, and around volcanic islands and seamounts. See TURBIDITY CURRENT.

Volcanologists have observed or inferred several common flowage processes to generate pyroclastic



Fig. 3. Some variations in pyroclastic rock formations. (a) Layers of pyroclastic fall deposits from the eruption plume of the June 15, 1991, explosive eruption of Mount Pinatubo (Luzon, Philippines). This exposure, located about 9 km (5.5 mi) southeast of the volcano, has a total thickness of about 0.6 m (1.8 ft) (*photograph by W. E. Scott, USGS*). (b) Blockand-ash-flow deposit formed by a collapse of a lava dome at Mount Hood Volcano (Oregon) more than 10,000 years ago; similar deposits were formed as recently as about 200 years ago. Some of the blocks are 1 m (3.3 ft) or greater in diameter; the present summit of the volcano can be seen top center (*photograph by D. R. Crandell, USGS*). (c) Cliff in the Owens River Gorge showing a massive pyroclastic flow deposit (the Bishop Tuff), which was produced about 760,000 years ago by a huge explosion from Long Valley Caldera (east-central California). At this locality, the deposit is about 150 m (500 ft) thick, is strongly welded in places, and shows spectacular columnar jointing formed from posteruption cooling (*photograph by R. I. Tilling, USGS*).

flows, surges, and directed blasts (**Fig. 2**). Pyroclastic flows formed by collapse of lava domes (Fig. 2*a*,*b*), or of thick lava flows on steep slopes, are called block-and-ash-flow deposits, characterized by large hot blocks (sometimes exceeding 5 m in size) in an ash matrix. Depending on the size and the explosive energy release of the eruption, the triggering mechanism, the texture, structure, and other features of the resulting deposits of pyroclastic rocks can exhibit considerable variation (**Fig. 3**).

Because of their mass, high temperature, high gas content, high velocity, and great mobility, the flowage processes that produce pyroclastic rocks are among the most dangerous and deadly of volcano hazards. These processes can have some or all of the following effects on people and property in their path: asphyxiation, burial, incineration, sand-blasting abrasion, and crushing physical impact. Even worse, hot pyroclastic flows and surges can mix with surface water, or water melted from snow and ice, to produce destructive lahars (volcanic mudflows) and floods that can devastate valleys downstream. Volcanic disasters involving pyroclastic flowage processes that have killed 2000 or more people include Mount Vesuvius, Italy, A.D. 79; Mont Pelée, Martinique, 1902; Mount Lamington, Papua New Guinea, 1951; and El Chichón Volcano, Mexico, 1982. Robert I. Tilling

Bibliography. R. A. F. Cas and J. V. Wright, *Volcanic Successions: Modern and Ancient*, Chapman & Hall, London, 1988; R. V. Fisher and H.-U. Schmincke, *Pyroclastic Rocks*, Springer-Verlag, Berlin, 1984;

P. Francis, Volcanoes: A Planetary Perspective, Oxford University Press, 1993; G. Heiken, An Atlas of Volcanic Asb, Smithsonian Contributions to the Earth Sciences, Smithsonian Institution, Washington, DC, 1974; W. E. Scott, Volcanic and related hazards, in R. I. Tilling (ed.), Short Course in Geology, vol. 1: Volcanic Hazards, American Geophysical Union, Washington, DC, 1989; H. Sigurdsson et al. (eds.), Encyclopedia of Volcanoes, Academic Press, San Diego, 2000.

Pyroelectricity

The property of certain crystals to produce a state of electric polarity by a change of temperature. Certain dielectric (electrically nonconducting) crystals develop an electric polarization (dipole moment per unit volume) ΔP when they are subjected to a uniform temperature change ΔT . For a small change ΔT , the components ΔP_i of the polarization vector are given by Eq. (1), where the p_i are constant coeffi-

$$\Delta P_i = p_i \,\Delta T \qquad i = x, y, z \tag{1}$$

cients. This pyroelectric effect occurs only in crystals which lack a center of symmetry and also have polar directions (that is, a polar axis). These conditions are fulfilled for 10 of the 32 crystal classes. Examples of pyroelectric crystals are tournaline, lithium sulfate monohydrate, cane sugar, and ferroelectric barium titanate. *See* CRYSTAL STRUCTURE; DIELECTRIC MATE-RIALS; POLARIZATION OF DIELECTRICS.

Pyroelectric crystals can be regarded as having a built-in or permanent electric polarization. When the crystal is held at constant temperature, this polarization does not manifest itself because it is compensated by free charge carriers that have reached the surface of the crystal by conduction through the crystal and from the surroundings. However, when the temperature of the crystal is raised or lowered, the permanent polarization changes, and this change manifests itself as pyroelectricity. There is no easy way to determine the magnitude of the total permanent polarization, except for those special pyroelectric crystals, called ferroelectrics, in which the polarization can be reversed by an electric field. *See* FERROELECTRICS.

Magnitude. The magnitude of the pyroelectric effect depends upon whether the thermal expansion of the crystal is prevented by clamping or whether the crystal is mechanically unconstrained. In the clamped crystal, the primary pyroelectric effect is observed, whereas in the free crystal, a secondary pyroelectric effect is superposed upon the primary effect. The secondary effect may be regarded as the piezoelectric polarization arising from thermal expansion. The secondary effect is generally much larger than the primary effect. When heated nonuniformly, even piezoelectric crystals produce a polarization due to the temperature gradients and corresponding nonuniform stresses and strains. This phenomenon is called tertiary, or false, pyroelectricity, because it may be found in nonpyroelectric crystals, that is, crystals which have no polar direction. See PIEZOELECTRICITY.

In a typical pyroelectric crystal such as tourmaline, a temperature change of 1.8° F (1°C) produces at room temperature a polarization of about 10^{-5} coulomb/m². The same amount of polarization can be produced in a crystal at constant temperature by applying an external electric field of about 70 kV/m. The pyroelectric effect is much larger in ferroelectrics, in particular at temperatures close to the Curie point. The figure of merit, defined as the pyroelectric coefficient divided by the permittivity, is high for ferroelectrics such as triglycine sulfate. Thermal response times of 2 microseconds and less have been reported. *See* PERMITTIVITY.

Inverse effect. As can be shown in thermodynamics, the pyroelectric effect has an inverse, the linear electrocaloric effect. A temperature change ΔT results when the permanent polarization is altered by an externally applied electric field ΔE , as in Eq. (2).

$$\Delta T = \frac{T}{\rho C_p} \sum_{i} p_i \,\Delta E_i \qquad i = x, y, z \qquad (2)$$

In this expression, ρ is the density, C_P the specific heat at constant pressure, p_i are the pyroelectric coefficients, and T is the absolute temperature. The temperature changes that can be realized in typical pyroelectrics are of the order of magnitude 0.02° F (0.01° C). H. Granicher

Applications. Pyroelectrics have a broad spectrum of potential scientific and technical applications. The most developed is the detection of infrared radiation.

Pyroelectric thermometer. In this instrument a thin pyroelectric plate cut perpendicular to the spontaneous polarization direction is covered with electrodes on both faces. Because of dielectric losses and any bias resistance, there is a resistance *R* in parallel to this special capacitor, and the potential difference ΔV between electrodes is nil at equilibrium. Let $\Delta T(t)$ be the increase of temperature at time *t*; a corresponding increase in spontaneous polarization given by Eq. (3) will appear.

$$\Delta P_s = p \,\Delta T \tag{3}$$

If the electric time constant $\tau' = RC$ (where *C* is the capacitance between the plates) is sufficiently long, the flow of electric charges through *R* may be neglected and voltage given by Eq. (4) will appear,

$$\Delta V = \frac{\Delta P_s}{C_0} \tag{4}$$

proportional to ΔP_s , and hence to ΔT ; here C_0 is the capacitance per unit electrode area. Temperature changes as small as 10^{-5} °F (6 × 10^{-6} °C) can be detected with a vibrating electrometer in a 1-s time interval.

Pyroelectric manometers. A thermal detector in a gas enclosure is sensitive to small variations of pressure, and the enclosure must be evacuated to avoid significant difficulties. This is simply explained by assuming that any rapid compression is adiabatic and thus leads to an increase of temperature. This effect may be used to measure pressure changes with a pyroelectric thermometer. The effect is especially

large in such detectors, where it may be enhanced by their piezoelectric sensitivity to hydrostatic pressure changes.

Pyroelectric radiation detectors. In most pyroelectric radiation detectors, the radiation is absorbed by a thin electrode whose thickness can be adjusted to make the absorption higher than 50% for the whole electromagnetic spectrum. Some dielectric coatings have absorptions of nearly 100% over a wide spectral interval. There is an increase of temperature and thus an electric signal ΔV . These devices generally detect infrared radiation, but have been used from the far ultraviolet to microwaves. Their normalized detectivity D^* at low chopping frequencies (2-15 Hz) is within an order of magnitude of the limit imposed by background radiation at room temperature (80°F or 300 K), given by Eq. (5). The pyroelectric detec-

$$D_{\rm lim}^* (300 \,{\rm K}) \approx 2 \times 10^{10} \,{\rm W}^{-1} \cdot {\rm cm} \cdot {\rm Hz}^{1/2}$$
 (5)

tor is thus comparable to the best thermal detectors (such as the pneumatic detector), and is superior to thermopiles and bolometers. *See* BOLOMETER; THER-MOCOUPLE.

At higher chopping frequencies the detectivity is significantly lower than D_{lim}^* (300 K). The main reason is that the internal resistance of the crystal is not constant but is purely dielectric, that is, inversely proportional to the chopping frequency, ω . Thus, Johnson noise decreases as $\omega^{-1/2}$, whereas it is constant for a bolometer or a thermocouple, and so the detectivity of pyroelectric detector can be expected to be proportional to $\omega^{-1/2}$ as well.

However, the pyroelectric radiation detector is still considerably better than all other thermal detectors, and is, moreover, mass produced and inexpensive. It is used in fast repetitive Fourier-transform spectroscopy, burglar alarms, and so forth. It is also possible to detect fast particles (for example, H⁺ or He⁺) from the temperature increase they produce in the pyroelectric.

In some pyroelectric materials a short pulse of radiation (of the order of 10^{-12} s) can produce an instantaneous electric dipole either in an impurity (such as LiTaO₃: Cu²⁺) or in the crystalline unit cell (because the average value of the dipole moment associated with an anharmonic oscillator can be significant). The result is a quasi-instantaneous pulse of spontaneous polarization followed by an electronic or phonon relaxation, respectively, with a very short time constant (of the order of 10^{-11} s). Special detectors based on this effect can be as fast as quantum detectors. *See* OPTICAL DETECTORS.

Radiometry and pyrometry. Pyroelectric detectors can be used to measure the power generated by a radiation source (in radiometry), or the temperature of a remote hot body (in pyrometry, with corrections due to deviations from the blackbody emission). The temperature change ΔT in Eq. (3) is proportional to the flux of radiant energy impinging on the pyroelectric detector; hence the electric signal ΔV in Eq. (4) is also proportional to this flux. Pyroelectric detectors were first used in the *Nimbus* meteorological satellites [deuterated triglycine sulfate (DTGS) detectors], in the *Pioneer* Venus orbiter, launched on May 20, 1978 (eight DTGS detectors), and in its sounder probe, which descended into the atmosphere of Venus [six lithium tantalate (LiTaO₃) detectors, which gave information on the carbon dioxide emission spectrum of Venus and hence on the temperature at different levels]. *See* PYROMETER; RA-DIOMETRY.

Imaging. An infrared image can be projected on a pyroelectric plate and transformed into a relief of polarization on the surface. The electric charges are read with a beam of electrons as in an ordinary television vidicon. The performances are such that it is actually possible to see in the dark. For instance, temperature differences of 0.9° F (0.5° C) in the landscape can be detected with 15 frames per second. However, a part of the detectivity is lost by the electron beam reading, and still higher performances are expected from linear or two-dimensional arrays of single pyroelectric detectors. *See* INFRARED IMAGING DEVICES.

Pyroelectric probe technique. This technique is one of the best ways to observe ferroelectric domains. A ferroelectric plate is cut perpendicular to the polar axis and coated with electrodes on both sides. A laser spot is swept over one electrode and gives a pyroelectric signal at each inspected point with a plus or minus sign according to the direction of the spontaneous polarization at this point. This technique is capable of inspecting a 100×100 micrometer area in a few seconds with a 2- μ m limit of resolution. With a 200-kHz chopping frequency only a thin superficial layer of the crystal is inspected. (In the case of triglycine sulfate, the thermal diffusion length is 0.64 μ m for that frequency.)

Other applications. Other potential applications of pyroelectricity include solar energy conversion, refrigeration, information storage, geology, and solid-state science.

The pyroelectric radiation detector can be considered as a high-impedance thermoelectric converter, which could be used for solar energy conversion. However, the efficiency achieved so far is low (1%). *See* SOLAR ENERGY.

The use of pyroelectricity for cooling is closely related to adiabatic demagnetization. The temperature has been lowered by $0.5^{\circ}F(0.3 \text{ K})$ at $-442^{\circ}F(10 \text{ K})$ with strontium titanate (SrTiO₃). Larger effects would require new materials. *See* ADIABATIC DEMAGNETIZATION.

When a pyroelectric is also a ferroelectric, the local polarization can be reversed by a laser spot to record information. The reading can also be done with a laser spot. The process is fast $(10^6 \text{ bits s}^{-1})$. The signal is much larger than expected, resulting in a high signal-to-noise ratio. *See* COMPUTER STORAGE TECHNOLOGY.

The interior of the Earth has been suggested to be ferroelectric due to the effect of high temperature and pressure on such compounds as magnesium silicate (MgSiO₃). This could explain the change of acoustic impedance which has been observed at a depth of 400 mi (700 km).

Pyroelectricity may be related to anharmonic

phonons, as discussed above. In other materials it is connected with atomic displacements. The order of magnitude is 10^{-14} m/°C, while the root-meansquare amplitude of vibrations at room temperature is around 10^{-11} m. The Ti⁴⁺ displacement at the barium titanate (BaTiO₃) transition is up to 2 × 10^{-11} m, which explains the large spontaneous polarization. Pyroelectricity is thus a very sensitive method of measuring small atomic displacements and can help specify the structure of a pyroelectric crystal. A. Hadni

Bibliography. A. Hadni, Applications of the pyroelectric effect, *J. Phys.*, E14:1233-1240, 1981; A. Hadni and X. Gerbaux, Infrared and millimeter wave absorber structures for thermal detectors, *Infrared Phys.*, 30:465-478, 1990; S. B. Lang, *Pyroelectricity*, 1981; S. B. Lang, Review of recent work on pyroelectric applications, *Ferroelectrics*, 53:189-195, 1984; M. E. Lines and A. M. Glass, *Principles and Applications of Ferroelectrics and Related Ma terials*, 1977, reprinted 2000; J. F. Nye, *Physical Properties of Crystals: Their Representation by Tensors and Matrices*, 1960, reprinted 1987.

Pyroelectricity (biology)

Electrical polarity in a biological material produced by a change in temperature. Pyroelectricity is probably a basic physical property of all living organisms. First discovered in 1966 in tendon and bone, it has since been shown to exist in most animal and plant tissues and in individual cells. Pyroelectricity appears to play a fundamental part in the growth processes (morphogenesis) and in physiological functions (such as sensory perception) of organisms. *See* PYROELECTRICITY.

The elementary components (for example, molecules) of biological (as well as of nonbiological) pyroelectric structures have a permanent electric dipole moment, and are arranged so that all positive dipole ends point in one direction and all negative dipole ends in the opposite direction. This parallel alignment of elementary dipoles is termed spontaneous polarization (P_s , electric dipole moment per unit volume) because it occurs spontaneously without the action of external fields or forces. In this state of molecular order, the structure concerned has a permanent electric dipole moment on a microscopic and macroscopic level (**Fig. 1***a*). *See* DIPOLE MOMENT.

Spontaneous polarization is temperature-dependent; thus any change in temperature causes a change of the dipole moments, measurable as a change of electric charges at both ends of the polar axis. This is the pyroelectric effect (Fig. 1*b*). All pyroelectric structures are also piezoelectric, but the reverse is not true. *See* FERROELECTRICS; PIEZOELEC-TRICITY.

The existence of an inherent spontaneous polarization (an inherent permanent dipole moment) in biological structures cannot be demonstrated at a constant temperature because the fixed charges on both dipole ends are compensated by free charge carriers within the polar material or from its environment. But the presence of a permanent dipole moment can be established with the aid of the pyroelectric effect by sudden changes of temperature, for example, those produced by the absorption of light flashes (radiant heating, Fig. 1*b*).

Molecular conditions. Prerequisites for the development of spontaneous polarization and pyroelectric activity in biological structures are (1) the presence of a permanent dipole moment in the molecules or molecular aggregates and (2) a molecular shape that favors a parallel alignment as much as possible (or at least does not impede it). Both these conditions are ideally fulfilled in rod-, bar- or board-shaped molecules with a permanent dipole moment along the longitudinal molecular axis. Such molecules readily take on parallel alignment and thus satisfy the physical preconditions for pyroelectric behavior in a defined axial direction.

Several important organic substances have these molecular properties, and therefore behave pyroelectrically in biological structures. The fibrous

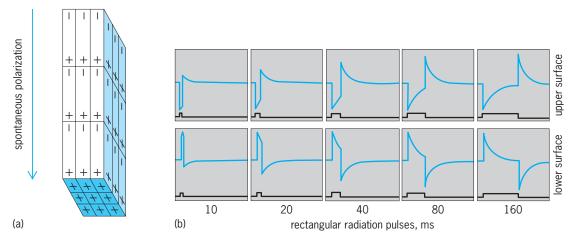


Fig. 1. Schematic diagram of spontaneous polarization and pyroelectric behavior. (a) Spontaneous polarization of a polar (pyroelectric) body. (b) Pyroelectric responses of upper and lower surfaces of a Z-cut tourmaline disk (a polar body) to rectangular radiation pulses. The responses of both surfaces have opposite electric signs (polar behavior). These are traces of a transient recorder; calibration, vertical 50 mV. The voltage-time course of the pyroelectric responses depends on the rate of change of temperature.

protein collagen is one example. Collagenous tissues (such as bone, cartilage, tendon, and ligaments) are thus strongly pyroelectric. *See* COLLAGEN.

Morphological polarity and inherent spontaneous polarization. Spontaneous polarization and pyroelectric behavior occur in the direction of distinct polar axes. A coincidence of these physical axes and axes of morphological polarity in biological structures is demonstrable with pyroelectric methods.

Individual cells. The polar properties of organisms, their organs, and systems originate in the polarity of cells. To establish which processes cause cell polarity and how to account for its stability during the lifetime of the cell is a principal task of biology. The giant cells of some green algae are favorite test objects for the investigation of cell polarity; both longitudinal and radial cell polarity have been demonstrated.

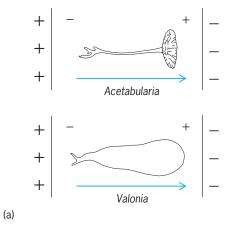
A permanent electric dipole moment coinciding with the (longitudinal) axis of cell polarity has been found in live individual cells of the green algae *Acetabularia mediterranea* and *Valonia utricularis*. The cells show a polar orientation in a high-voltage dc field if the cell temperature is rapidly changed. The apical cell pole then points to the cathode, the basal pole to the anode (**Fig. 2***a*). The polar orientation is maintained if the direction of the dc field is reversed; that means the apical cell pole turns again to the cathode. Under identical experimental conditions a tourmaline crystal rod shows similar polar orientation.

The cell walls of live *Acetabularia* and *Valonia* cells react pyro- or piezoelectrically on their entire surface to thermal or mechanical stimulations (Fig. 2*b*). Pyroelectric and piezoelectric signals of the outer cell wall surface have opposite electric signs to those of the inner surface; the cell wall is thus polar. The direction of the radial polarization is perpendicular to the cell wall surface.

Epidermis of animals and plants. The epidermis, the outermost cellular layer of animals and plants, represents a highly important interface between the living organism and its environment. The epidermis of invertebrates and of plants is usually a monocellular layer. Pyroelectric measurements on numerous living organisms (invertebrates and vertebrates) as well as on a great many living higher plant species confirmed the existence of a spontaneous polarization perpendicular to the epidermal surface and coinciding with the axis of cell polarity in these layers (Fig. 3a). Pyroelectric and piezoelectric responses of the animal and plant epidermis are analogous. Pyroelectric responses of outer and inner epidermal surfaces have opposite electric signs (Fig. 3b), thus proving polar behavior.

The animal and plant epidermis functions on its entire surface as a pyroelectric detector and transducer, converting distinct kinds of physical energy from the environment (for example, rapid changes in temperature, uniaxial and hydrostatic pressure) into electrical energy. Thus the live epidermis has physical properties analogous to nonbiological polar materials which function in pyroelectric detectors and transducers designed for technical applications.

Sensory receptors. Cutaneous receptors are mostly



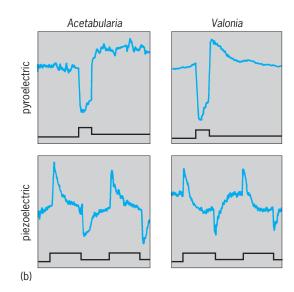


Fig. 2. Longitudinal and radial directions of spontaneous polarization in individual cells. (a) Longitudinal polarization; live cells show polar orientation (apical cell pole toward the cathode) in a dc field if the cells' temperature is rapidly changed. (b) Radial polarization; the cell walls of live *Acetabularia* and *Valonia* cells show pyroelectric and piezoelectric behavior perpendicular to their surfaces. Pyroelectric voltage responses to rectangular radiation pulses of 20 ms; piezoelectric voltage responses to Totage responses to rectangular mechanical pulses of 100 ms. Calibration: pyroelectric, vertical 5 mV; piezoelectric, 20 mV.

transformed epidermal structures. Since all structures of the animal and plant epidermis that have been investigated showed pyroelectric activities, it may be assumed a priori that cutaneous sensory receptors will behave pyroelectrically (and thus also piezoelectrically). Two examples are the auditory membrane of the locust ear and the infrared-sensitive organ of rattlesnakes.

The auditory membrane (tympanum) of the locust ear (**Fig.** 4a-c) is a part of the insect exoskeleton, the epicuticle. The epicuticle of the arthropod integument has pronounced pyroelectric properties; its pyroelectric coefficient can reach or exceed that of tourmaline. The tympanum of the locust ear acts as a piezoelectric mechanoreceptor (Fig. 4d), but it has pyroelectric properties as well. The polar axis of the tympanum is perpendicular to its surface, as it is in the surrounding epicuticle.

The infrared-sensitive organ of rattlesnakes has a

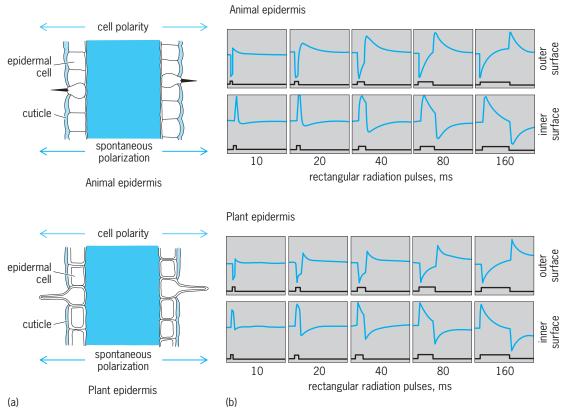


Fig. 3. Spontaneous polarization and pyroelectric behavior of animal and plant epidermis. (a) Uniform direction of spontaneous polarization in the animal and plant epidermis. (b) Opposite electric signs of pyroelectric voltage responses on the outer and the inner epidermal surface (polar behavior); compare Fig. 1b. Animal epidermis: *Periplaneta americana*, cockroach; plant epidermis: *Yucca elephantipes*, yucca. Calibration, vertical 20 mV (*Periplaneta*) and 5 mV (*Yucca*).

similar topography to the locust ear. Its sensitive membrane is suspended between two air-filled cavities and contains numerous nerve endings oriented perpendicular to the membrane's surface. This membrane has strong pyroelectric properties and reacts to minute changes of temperature with electrical signals corresponding to those of technically used pyroelectric infrared detectors, but also has piezoelectric properties.

A physical reaction mechanism common to all sensory receptors is suggested by the following facts. All receptor cells of animals are very similar in their morphological structure and cytochemical organization, in spite of certain differences related to their specific functions. All thermoreceptors, photoreceptors, mechanoreceptors, as well as other receptors, react to the various stimuli (such as variation in temperature, illumination, or mechanical stress) with analogous electric responses (receptor potentials or generator potentials). However, the reactions of one receptor are in most cases not limited to only one external effect. For example, thermoreceptors in the skin react to changes of temperature and mechanical stress with analogous responses; the lateral-line organ of fishes and amphibians can act as a thermal, electrical, chemical, or mechanical receptor; and, in the case of the ampullae of Lorenzini (sense organs found in organisms such as the skate), a function as thermal, electrical, and mechanical receptor has been established. See SENSE ORGAN.

Nervous system. The spinal cord of vertebrates con-

stitutes a permanent electric dipole in a longitudinal direction, the negative pole being oriented cranially and the positive pole caudally (**Fig. 5**). Each segment of the spinal cord cut exactly perpendicular to its longitudinal axis shows pyroelectric voltage responses of opposite signs on its cranial and caudal surfaces. Fresh spinal cord preparations thus show a polar orientation in a high-voltage dc field, if the temperature of the specimens is rapidly changed.

A relationship between the direction of longitudinal spontaneous polarization and the physiological function of the central nervous system is evident. Specific reactions of the Chordata (from *Ampbioxus* up to humans) that are produced by lengthwise application of dc fields (direct currents) in the spinal cord are dependent on the direction of the direct current (electric narcosis in one direction; Fig. 5c and d). This functional polarity of the spinal cord can be explained by the existence of a longitudinal permanent dipole moment.

Experiments on the spinal cord have been continued on the optic nerve (Fig. 5*e*), a brain structure in which the nerve fibers chiefly run in one direction (from the retina of the eye into the brain). In the optic nerve a longitudinal permanent dipole moment of the same direction as in the spinal cord, and pyroelectric and piezoelectric behavior have been established. Thus it is probable that the individual neuron (Fig. 5*f*) has a longitudinal permanent dipole moment and pyroelectric and

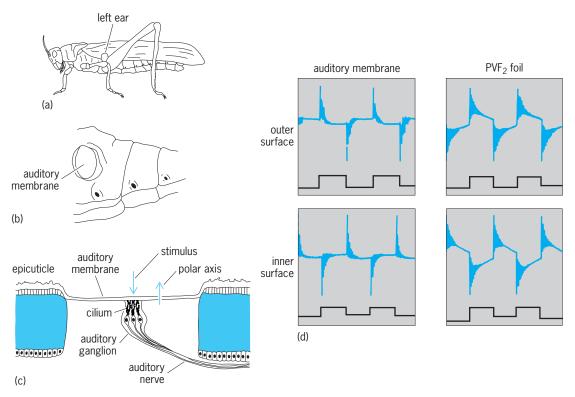


Fig. 4. Piezoelectric behavior of an animal mechanoreceptor (ear of locust). (a) Locusta migratoria, locust, and its left ear (the two ears of locust are situated on both sides of the first abdominal segment). (b) Auditory membrane of the ear. (c) Cross section of the ear. (d) Piezoelectric voltage responses of outer and inner surface of the auditory membrane and (for comparison) of both surfaces of a piezoelectric polyvinylidene (PVF₂) foil to rectangular pressure pulses of 30 ms. Calibration, vertical 10 mV (auditory membrane) and 100 mV (PVF₂).

piezoelectric properties, too.

Skeletal system. In collagenous tissues (such as bone, cartilage, tendon, and ligaments) of the vertebrate skeleton from the Cyclostomata (such as the *Petromyzon* lamprey) up to humans, a spon-

taneous polarization pattern has been established that obeys a natural law and is uniform in homologous structures. This polarization pattern is closely related to the morphological polarity of the tissues concerned and to their embryonic development.

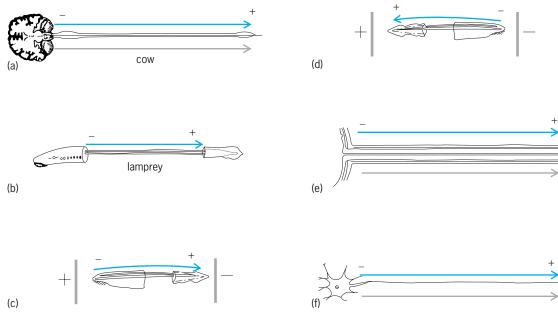


Fig. 5. Longitudinal permanent dipole moment (color arrows) in nerve structures. (*a*, *b*) Spinal cords of vertebrates: cow and lamprey. (*c*) Cranial end of spinal cord oriented in a dc field to anode causes narcosis and (*d*) cranial end of spinal cord in a dc field oriented to cathode causes excitation (shown for one example, *Branchiostoma lanceolatum*). (*e*) Optic nerve; this brain structure is chiefly composed of nerve fibers having one direction. (*f*) Hypothetical direction of longitudinal permanent dipole moment of an individual neurite. Gray arrows indicate direction of growth.

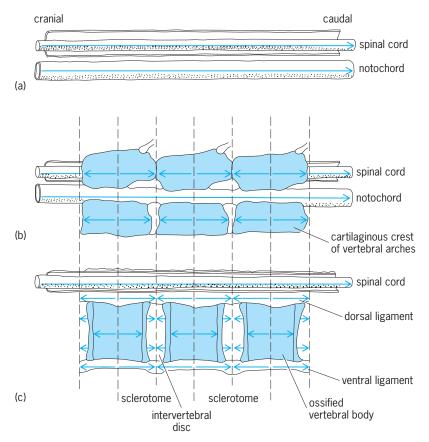


Fig. 6. Directions of spontaneous polarization (arrows) in the axial skeleton of vertebrates. (a) *Petromyzon marinus*, lamprey. (b) *Acipenser sturio*, sturgeon. (c) Human. Boundaries between sclerotomes and the cranial and caudal scleromites are as indicated.

Figure 6 shows experimental results which have been obtained by cutting a continuous series of thin disks at right angles to the longitudinal axis of the tissues under investigation, and by taking individual pyroelectric measurements on each one.

Axial system of Chordata. The chorda dorsalis (notochord), central nervous system, and bilateral row of somites together form the axial system of the chordate germ. The longitudinal polarization of these structures, examined in numerous specimens of juvenile and adult chordates, has corresponding directions throughout the whole range of Chordata, in all stages of development investigated. The direction of the longitudinal dipole moment is the same in notochord and spinal cord; the negative pole lies cranially, the positive pole caudally (Fig. 6a and b). In the somites the direction of the longitudinal dipole moments changes on the boundaries of the sclerotomes (Fig. 6c).

The direction of permanent dipole moments (spontaneous polarization) in all investigated structures of the axial and skeletal systems of vertebrates corresponds with the directions of growth in the structures concerned.

Biological systems as pyroelectric detectors and transducers. Living organisms are able to detect and discriminate between different stimuli in the environment, such as rapid changes of temperature, of illumination, and of hydrostatic and uniaxial pressure. These stimuli represent different forms of energy and

are transduced, or converted, into the nearly uniform type of electrical signals whose voltage-time course frequently depends on dX/dt (X = external stimulus, t = time). Such electrical signals have been recorded on cutaneous sensory receptors, on external nerve endings, on epidermal structures, and even on the cell wall of single-cell organisms.

The mechanisms of detection and transduction in these biological systems, still little understood, may lie in the pyroelectric behavior of the structures. Pyroelectric (and thus piezoelectric) behavior has been proved to exist in most biological systems, which means that these systems should in principle be able to function as pyroelectric detectors and transducers. This function is of exceptional significance in epidermal structures, in sensory receptors, and in the nervous system. Herbert Athenstaedt

Bibliography. H. Athenstaedt, Biological systems as pyroelectric detectors and transducers, *Jap. J. Appl. Phys.*, pp. 177-180, 1985; H. Athenstaedt, Epidermis of human skin: Pyroelectric and piezoelectric sensor layer, *Science*, 216:1018-1020, 1982; H. Athenstaedt, "Functional polarity" of the spinal cord caused by its longitudinal electric dipole moment, *Amer. J. Physiol.*, 247:482-487, 1984; H. Athenstaedt, Pyroelectric and piezoelectric properties of vertebrates, *Ann. N.Y. Acad. Sci.*, 238:68-94, 1974.

Pyrolysis

A chemical process in which a compound is converted to one or more products by heat. By this definition, reactions that occur by heating in the presence of a catalyst, or in the presence of air when oxidation is usually a simultaneous reaction, are excluded. The terms thermolysis or thermal reaction have been used in essentially the same sense as pyrolysis.

A simple example of pyrolysis is the classic experiment in which oxygen was first prepared by heating mercuric oxide [reaction (1)]. Similar reactions

$$HgO \xrightarrow{300^{\circ}C}_{(570^{\circ}F)} Hg + \frac{1}{2}O_2$$
 (1)

occur with numerous other metallic oxides and salts. Thermal decomposition or calcining of limestone (calcium carbonate) is the basic step in the manufacture of lime [reaction (2)].

$$CaCo_3 \xrightarrow[(2200\circ F]{1200\circ C} CaO + CO_2$$
(2)

See LIME (INDUSTRY).

Role in organic chemistry. The term pyrolysis is most commonly associated with thermal reactions of organic compounds. In very early chemical investigations, pyrolysis of material from plant and animal sources provided some of the first clues about constitution, as in the formation of isoprene from the thermal breakdown of rubber. Destructive distillation of wood in the absence of air gave "pyroligneous acid," from which methanol and acetic acid were isolated. Similar treatment of bone yielded pyrrole. A range of substances, including benzene, naphthalene, pyridine, and many other aromatic compounds, was obtained from coal tar, a pyrolysis product of coal. All of these pyrolysis processes lead to formation of volatile products characteristic of the source and also residues of char with high carbon content. *See* DESTRUCTIVE DISTILLATION.

Pyrolysis reactions in contemporary work have been used as preparative methods and as means of generating transient intermediates that can be trapped or observed spectroscopically, or quenched by a further reaction. For preparative purposes, pyrolysis can generally be carried out by a flow process in which the reactant is vaporized with a stream of inert gas through a heated tube, sometimes at reduced pressure. Temperatures range from 400 to 600°C (750 to 1100°F). In flash vacuum pyrolysis (FVP), the apparatus is placed under very low pressure, and the material to be pyrolyzed is vaporized by molecular distillation, usually through a very short zone in a quartz tube at 600-1000°C (1100-1800°F). Under these conditions, contact times are on the order of 0.01 s, and initial unimolecular decompositions can be observed. See CHEMICAL DYNAMICS.

Types of reactions. Thermally activated molecules undergo several types of reactions that can be broadly grouped as dissociation into free radicals, elimination of a small molecule, or carbon-carbon bond fission. In many of these pyrolyses, subsequent reactions occur after the initial event.

Dissociation into radicals. At temperatures of 600-800°C (1100-1500°F), most organic compounds acquire sufficient vibrational energy to cause breaking of bonds with formation of free radicals. Early evidence for the transient existence of free radicals was the formation of a metallic mirror when tetraethyllead [(CH₃)₄Pb] was vaporized through a heated tube, and removal of the mirror when the tube was heated at a point upstream from the metal deposit. These observations indicate the occurrence of overall reaction (3).

$$(CH_3)_4Pb \rightleftharpoons 4CH_3. + Pb$$
 (3)
Methyl
radical

See FREE RADICAL.

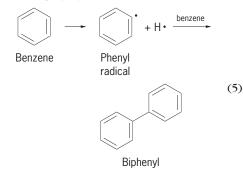
Alkanes undergo rupture of carbon-hydrogen (C-H) and carbon-carbon (C-C) bonds to two radicals that then react to give lower alkanes, alkenes, hydrogen, and also higher-molecular-weight compounds resulting from their recombination [reactions (4)].

$$\begin{array}{c} CH_3CH_2CH_2CH_3 & \xrightarrow{700^{\circ}C} \\ & 2CH_3CH_2 \cdot \rightarrow CH_3CH_3 + H_2C = CH_2 \end{array} (4a) \end{array}$$

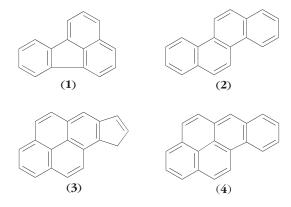
$$\begin{array}{c} \mathsf{CH}_3\mathsf{CH}_2\mathsf{CH}_2\mathsf{CH}_3 \xrightarrow[(1300^\circ\mathsf{F})]{} \\ \mathsf{CH}_3\mathsf{CH}_2 \cdot \mathsf{CHCH}_3 + \mathsf{H} \cdot \to \mathsf{CH}_3\mathsf{CH}_2\mathsf{HC} = \mathsf{CH}_2 + \mathsf{H}_2 \end{array} (4b)$$

These reactions are generally nonselective, since the bond dissociation energies (D°) do not differ greatly; for simple alkanes, $D^{\circ}:C-H = 90-98$ kcal, $D^{\circ}:C-C = 82-90$ kcal. These reactions are the basis of the thermal cracking processes used in petroleum refining. *See* CRACKING.

Pyrolysis of simple aromatic hydrocarbons such as benzene or naphthalene produces aryl radicals, which can attack other hydrocarbon molecules to give bi- and polyaryls, as shown in reaction (5) for

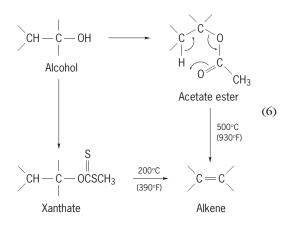


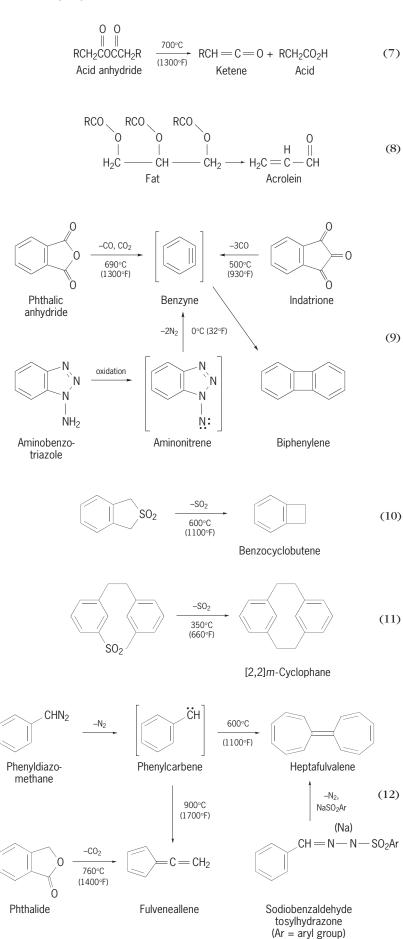
the formation of biphenyl. With alkyl-substituted and polycyclic hydrocarbons, very complex mixtures of pyrolysis products are obtained by a combination of radical substitutions and disproportionation. Among the end products of these reaction sequences are highly condensed ring systems such as fluoroanthene (1), chrysene (2), cyclopentapyrene (3), and benzopyrene (4). These and other polycyclic aromatic



hydrocarbons are present in soot, engine exhausts, and tobacco smoke; they are potent environmental carcinogens. *See* ONCOLOGY.

Elimination. Pyrolytic eliminations can result in formation of a multiple bond by loss of HX from a compound H—C—C—X, where X = any leaving group. A typical example is the pyrolysis of an ester, which is one of the general methods for preparing alkenes. These reactions are commonly carried out with acetate esters at 400-500°C (750-930°F) in a flow reactor and occur by a cyclic transition state [reaction (6)]. Similar elimination of a dithiocarbonate





(xanthate) ester proceeds at 200-300°C (390-570°F). Pyrolytic elimination is particularly useful when acid-catalyzed dehydration of the parent alcohol leads to cationic rearrangement. Another useful application of this process is the production of ketenes from acid anhydrides [reaction (7)]. See ACID ANHYDRIDE; ALKENE; ESTER.

The pungent, lacrimatory odor produced by roasting meat at high temperature is due to the formation of acrolein by pyrolysis of fats, which are triesters of glycerol [reaction (8)].

(8)

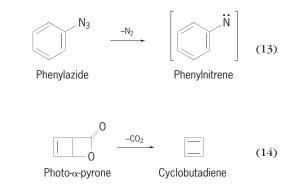
Another type of thermal elimination occurs by loss of a small molecule such as nitrogen (N₂), carbon monoxide (CO), carbon dioxide (CO₂), or sulfur dioxide (SO₂, leading to reactive intermediates such as arynes, diradicals, carbenes, or nitrenes. A few of the pyrolytic routes to benzyne are shown in reactions (9). The conditions vary considerably; thus the nitrene generated from aminobenzotriazole breaks down to benzyne at $0^{\circ}C$ (32°F). Benzyne can be trapped by addition reaction or can dimerize to biphenylene. See REACTIVE INTERME-DIATES

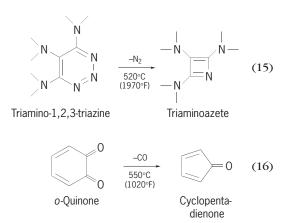
Extrusion of SO₂ by pyrolysis of a cyclic sulfone gives a diradical that can then form a new C-C bond. This reaction is an effective preparative method for the synthesis of four-membered rings [reaction (10)] and also cyclophanes [reaction (11)].

When these thermodynamically stable molecules are lost in high-temperature pyrolysis reactions, the energetic and highly reactive fragments are unencumbered by solvent and encounter few intermolecular collisions. Therefore, they can give rise to unusual and otherwise inaccessible products, which may be thought of as a shallow minimum on a potential energy surface.

An example is phenylcarbene, which can be generated by pyrolysis of phenyldiazomethane or other compounds. The carbene can rearrange by either ring contraction or expansion, depending on the source and conditions, to give fulvene or fulvalene products [reactions (12)]. Similarly, nitrenes are formed by pyrolysis of azides [reaction (13)].

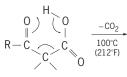
Unstable and elusive compounds from flash vacuum pyrolysis can be collected at 20 to 77 K (-253 to -196°C or -423 to -321°F) in an inert matrix for direct spectroscopic observation. By these techniques, cyclobutadiene, triaminoazete, and cyclopentadienone have been characterized [reactions (14)-(16)].



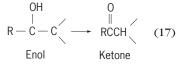


See MATRIX ISOLATION.

Fission of carbon-carbon bonds. A number of pyrolytic reactions involve cleavage of specific C-C bonds in a carbon chain or ring. Fragmentation accompanied by transfer of hydrogen is a general reaction that occurs by a cyclic process. An example is decarboxylation of acids that contain a carbonyl group in the β position, which lose CO2 on relatively mild heating [reaction (17)]. Acids with a double or triple carbon-to-carbon



β-Ketoacid



bond undergo decarboxylation at 300-400°C (570°- 750° F), as in reaction (18). This type of reaction also

$$RC = C - CO_{2}H \xrightarrow{400^{\circ}C} RCHC = C + CO_{2} \quad (18)$$

occurs at higher temperatures with unsaturated alcohols [reaction (19)] and, by transfer of hydrogen

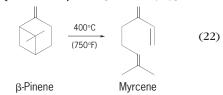
$$RC = C - CO_{2}H \xrightarrow{450^{\circ}C}_{(840^{\circ}F)}$$
$$RCH = C = C + H_{2}C = 0 \quad (19)$$

from a C-H bond, with unsaturated ethers [reaction (20)].

...

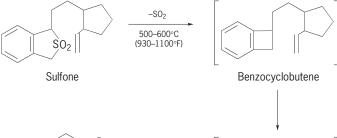
Cleavage of a ring frequently occurs on pyrolysis. With alicyclic or heterocyclic four-membered rings, cleavage into two fragments is the reverse of 2 + 2 cycloaddition, as illustrated by the cracking of diketene [reaction (21)]. Pyrolysis is an important reaction in

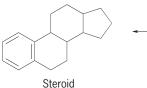
the chemistry of the pine terpenes, as in the conversion of β -pinene to myrcene [reaction (22)].

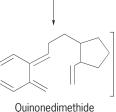


See PINE TERPENE.

Benzocyclobutenes undergo ring opening to o-quinone dimethides. By combining this reaction in tandem with formation of the benzocyclobutene and a final Diels-Alder reaction, a versatile one-step synthetic method for the steroid ring system has been developed [reaction (23)].

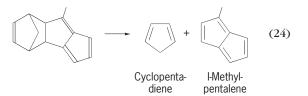






See DIELS-ALDER REACTION.

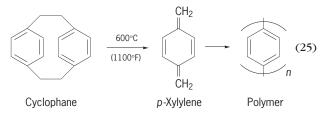
Cleavage of a cyclohexene by reverse 4 + 2 cycloaddition is a very general pyrolytic process. Cyclopentadiene dimerizes at room temperature, and this versatile diene is prepared as needed by pyrolysis of the dimer. This ring scission has been employed in the synthesis of the extremely unstable pentalene by flash vacuum pyrolysis of a tetracyclic precursor, in which a cyclopentadiene unit acts as a protective group until the last step [reaction (24).



Another application of pyrolytic-ring cracking is the reaction of [2,2]-*p*-cyclophane at 600° C (1100°F). The resulting *p*-xylylene is passed in the vapor phase



over surfaces on which a film of the resulting polymer is deposited, forming a tough protective coating [reaction (25)].



Many thermal reactions involve isomerization without elimination or fragmentation. These processes can occur by way of intermediates such as diradicals, as in the pyrolysis of pinene [reaction (22)], or they may be concerted pericyclic reactions. An example of the latter is the Claisen-Cope rearrangement of phenyl or vinyl ethers [reaction (26)] and other



1,5-diene systems. These reactions can be carried out by relatively mild heating, and they are very useful in synthesis. *See* ORGANIC SYNTHESIS; PERICYCLIC REACTION.

Analytical applications. Thermal breakdown of complex structures leads to very complex mixtures of products arising from concurrent dissociation, elimination, and bond fission. Separation of these mixtures provides a characteristic pyrogram that is valuable as an analytical method, particularly for polymeric materials of both biological and synthetic origin. In this application, a small sample is heated on a hot filament or by laser. The pyrolysis products are then analyzed by gas chromatography (Py-Gc), mass spectrometry (Py-Ms), or a combination of both techniques. *See* GAS CHROMATOGRAPHY; MASS SPECTROMETRY; THERMAL ANALYSIS.

Instrumentation with appropriate interfaces and data-handling systems has been developed to permit rapid and sensitive detection of pyrolysis products for a number of applications. One example is the optimization of conditions in petroleum cracking to produce a desired product from varied crude oils. The profile of pyrolysis fragments from a polymer can also be used to detect impurities.

Analytical pyrolysis has also been used in the identification of microorganisms. In this application, a sample of whole cells is pyrolyzed under rigidly controlled conditions, and the entire mixture of volatile products arising from proteins, polynucleotides, cellwall carbohydrates, and other constituents is analyzed. The Py-Gc or Py-Ms pyrograms obtained from gas chromatography or mass spectrometry contain over 100 identifiable degradation products, ranging from simple alkanes and aromatic hydrocarbons to furans and pyrroles. Many of these compounds are common to all organisms, but computer analysis of the pyrogram can provide a distinctive fingerprint. James A. Moore Bibliography. R. F. C. Brown, *Pyrolytic Methods in Organic Chemistry* 1980; W. J. Irwin, *Analytical Pyrolysis*, 1982; V. E. Wiersum, Preparative flash vacuum pyrolysis, *Aldrich Chimica Acta*, 17:31-40, 1984.

Pyrometallurgy

The branch of extractive metallurgy in which processes employing chemical reactions at elevated temperatures are used to extract metals from raw materials, such as ores and concentrates, and to treat recycled scrap metal. The use of heat for pottery making dates as far back as 3500-4000 B.C., and is believed to have gradually led to the first high-temperature furnaces used for smelting copper. Small bowllike furnaces were used around 3500 B.C. for reducing oxidic copper ores mixed with charcoal at high temperatures (1100-1200°C; 2000-2200°F) to metallic copper. The historical metal ages show the development of metallurgy and the beginning of the practice and science of pyrometallurgy. The Chalcolithic Age (roughly 5000-2200 B.C.) led to the Bronze Age (approximately 3000-1300 B.C.), which gave way to the Iron Age (about 1000 B.C. to A.D. 400). The techniques of pyrometallurgy were gradually perfected, especially during the late nineteenth and the twentieth centuries as knowledge, mechanical equipment, and materials of construction improved. Pyrometallurgy continues to be the principal means of metal production.

The simplest reaction in pyrometallurgy for the production of metal M from the compound MX (such as an oxide or sulfide in a concentrate) using a reducing or oxidizing agent R can be written as $MX + R \rightarrow$ M + RX. In order for this reaction to proceed, the free-energy term (ΔG°) must be negative. For a practical process, the reaction should proceed rapidly, and M should easily separate from the compound MX. The use of high temperatures prompts higher reaction rates and allows for a sufficiently negative value of ΔG° . During smelting, a liquid slag phase of a lower-density collecting RX forms, which easily separates from the denser liquid metal M. In the case of treating sulfides, R is normally oxygen, so RX would be SO₂, and any accompanying unwanted metals, such as iron, zinc, or nickel, in the feed copper concentrate, for example form oxides such as FeO. In high-temperature metallurgical processing, chemical reactions are rapid, allowing for the processing of large tonnages of metal in a given furnace. Reaction equilibria change as temperature increases so that processes impossible at low temperature become almost spontaneous at higher temperature. In addition, production of the metal as a liquid or a gas facilitates the physical separation of the metal from other materials, which are collected in a liquid slag for disposal. The temperatures used in pyrometallurgy range from 150-200°C (300-400°F) to over 1600°C (2900°F), depending on the specific operation. See FREE ENERGY.

For metal production, the pyrometallurgical

operation commences with either a raw material obtained by mining and subsequent mineral and ore processing steps to produce a concentrate, or a recycled material such as separated materials from scrapped automobiles, machinery, or computers. Recycled materials are increasingly important.

Pyrometallurgical preparation processes convert raw materials to forms suitable for future processing, such as roasting of zinc sulfide concentrates to convert the sulfides to oxides for further processing, sintering to oxidize and agglomerate sulfide concentrates to coarser oxides for the blast furnace smelting of lead, or chlorination reactions for certain metals such as titanium. Reduction processes reduce metallic oxides and compounds to metal, as in the iron blast furnace which reduces iron oxide to pig iron, or the lead blast furnace which reduces lead oxide sinter to lead metal. Oxidizing processes oxidize the feed material to an intermediate or a semifinished metal product, such as smelting processes which produce a high-grade copper matte (a copper sulfide rich-iron sulfide material), or converting copper(I) sulfide (Cu₂S) to blister copper (an impure form of copper). Refining processes remove the last of the impurities from a crude metal, as in fractional distillation to remove lead or cadmium from impure zinc, the gaseous reduction (removal of oxygen) of anode copper prior to casting, or copper fire-refining (the refining of molten copper using fluxes such as silica or soda-ash).

The complete production scheme, from raw material (concentrate or recycled scrap) to refined metal, may use pyrometallurgical processes for all steps (steel, lead, tin, zinc), or only the primary extraction processes may be pyrometallurgical with other methods, such as electrorefining, used for final refining (copper or nickel). Alternatively, the primary step may be nonpyrometallurgical (in many cases a leaching step), followed by pyrometallurgical processing (aluminum, magnesium, titanium), or vice versa, such as the production of zinc by first roasting sulfide concentrates, followed by leaching and electrowinning (electrochemical metal extraction from solution). In some cases, certain pyrometallurgical processes are used in predominantly nonpyrometallurgical treatment schemes. See ELECTROMETAL-LURGY; IRON METALLURGY; METALLURGY; PYROMET-Phillip J. Mackey ALLURGY, NONFERROUS.

Pyrometer

A temperature-measuring device, originally an instrument that measures temperatures beyond the range of thermometers, but now in addition a device that measures thermal radiation in any temperature range. This article discusses radiation pyrometers. For other temperature-measuring devices *see* BOLOMETER; THERMISTOR; THERMOCOUPLE.

Technique. Figure 1 shows a very simple type of radiation pyrometer. Part of the thermal radiation that is emitted by a hot object is intercepted by a lens and focused onto a thermopile. The resultant

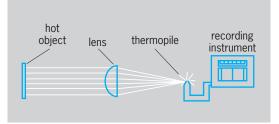


Fig. 1. Elementary radiation pyrometer. (After D. M. Considine, ed., Process Instruments and Controls Handbook, McGraw-Hill, 1957)

heating of the thermopile causes it to generate an electrical signal (proportional to the thermal radiation) which can be displayed on a recorder.

Unfortunately, the thermal radiation emitted by the object depends not only on its temperature but also on its surface characteristics. The radiation existing inside hot, opaque objects is so-called blackbody radiation, which is a unique function of temperature and wavelength and is the same for all opaque materials. However, such radiation, when it attempts to escape from the object, is partly reflected at the surface. The fraction reflected back into the interior (and hence also the fraction emitted) depends on the type of material, surface roughness, surface films, and microstructure.

The pyrometer readings, then, are ambiguous, since a measured rate of thermal radiation may be due to a temperature of $752^{\circ}F$ (400°C) and an internal reflectance of 0.80 (emittance of 0.20) on the one hand, or to a temperature of $932^{\circ}F$ (500°C) and an internal reflectance of 0.90 (emittance of 0.10) on the other hand. In order to use the output of the pyrometer as a measure of target temperature, the effect of the surface characteristics must be eliminated from the measurements. *See* BLACKBODY.

This elimination is accomplished as follows. A cavity is formed in an opaque material and the pyrometer is sighted on a small opening extending from the cavity to the surface. The opening has no surface reflection, since the surface has been eliminated. Such a source is called a blackbody source, and is said to have an emittance of 1.00. Any other object, when its surface temperature is measured, has some actual surface reflectance and therefore has an emittance less than 1.00.

By attaching thermocouples to the blackbody source, a curve of pyrometer output voltage versus blackbody temperature can be constructed. Similarly, for a given material (without the cavity) thermocouples can be attached to the surface and the radiation output to the pyrometer output voltage determined at known material temperatures. The emittance factor is taken to be this voltage divided by the voltage output observed for the blackbody at the same temperature. Then, in making measurements on this material remotely and without contact (as on a moving process line), the voltage readings are divided by the proper emittance factor to obtain the true temperature from the blackbody curve. *See* THERMOCOUPLE. The effects of target size and target distance on the pyrometer readings must also be eliminated. One solution is to design the field of view to increase as the square of the target distance. Thus, as long as the field of view is filled by the target, the increasing target size just compensates for the decreased signal per unit target area as the target distance increases.

Types. Pyrometers can be classified generally into types requiring that the field of view be filled, such as narrow-band and total-radiation pyrometers; and types not requiring that the field of view be filled, such as optical and ratio pyrometers. The latter depend upon making some sort of comparison between two or more signals.

Narrow-band (spectral) pyrometers. These pyrometers (**Fig. 2**) respond to a limited band of wavelengths. (For example, a typical lead sulfide pyrometer will have a filter that limits its spectral response to the 1.8–2.7-micrometer region.) The output may vary as the 10th to 15th power of the temperature. Time constants are of the order of milliseconds (as in lead sulfide) or microseconds (as in silicon cells). Fields of view are of the order of 1-cm diameter at 1 m. The detectors are usually of the photoconductive or photovoltaic type. *See* RADIOMETRY.

Total-radiation pyrometers. The lens, or window and mirror, because of limited spectral transmission, allows only broad-band response, rather than "total" radiation response. Hence the output usually varies as about the 4.5 or 5th power. Detectors are usually of the thermal type (that is, the detector generates an output because of a rise in its temperature), such as a thermopile. Time constants are usually of the order of 1 s or longer. Fields of view of about 2 in. (5 cm) diameter at 40 in. (1 m) are typical (**Fig. 3**).

Optical pyrometer. This instrument should more

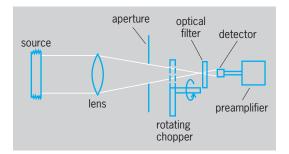


Fig. 2. Narrow-band pyrometer.

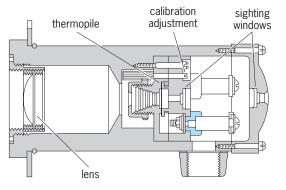


Fig. 3. Radiation pyrometer. (Honeywell Inc.)

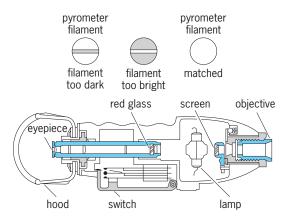


Fig. 4. Optical pyrometer. (Leeds and Northrup Co.)

strictly be called the disappearing-filament pyrometer. In operation, an image of the target is focused in the plane of a wire that can be heated electrically. A rheostat is used to adjust the current through the wire until the wire blends into the image of the target (equal brightness condition), and the temperature is then read from a calibrated dial on the rheostat (**Fig. 4**).

Ratio, or "two-color," pyrometer. This pyrometer makes measurements in two wavelength regions and electronically takes the ratio of these measurements. If the emittance is the same for both wavelengths, the emittance cancels out of the result, and the true temperature of the target is obtained. This so-called graybody assumption is sufficiently valid in some cases so that the "color temperature" measured by a ratio pyrometer is close to the true temperature.

A given target may be said to have a certain "brightness temperature," "radiation temperature," "color temperature" and, of course, true temperature, depending on how the temperature is measured. These may all differ before the readings are corrected for emittance. The terms luminance temperature and distribution temperature have been adopted to replace brightness temperature and color temperature, respectively.

Radiation pyrometers are particularly useful for (1) temperatures above the practical operating ranges of thermocouples; (2) environments that contaminate and limit the useful life of thermocouples; (3) moving targets; (4) visible targets not easily accessible; (5) targets that would be damaged by contact or insertion; (6) average temperatures of large areas; and (7) rapidly changing temperatures (using fast response detectors).

Errors in temperatures measured by radiation pyrometers can be caused by (1) uncertainty in the emittance value of the target; (2) dust, steam, or smoke in the radiation path; (3) dirt on lens or window; (4) water or dirt on the target surface; and (5) reflection of spurious radiation from the target. *See* TEMPERATURE MEASUREMENT; THERMOME-TER. Thomas P. Murray

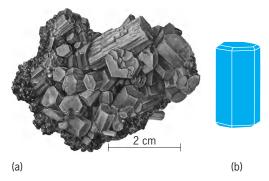
Bibliography. American Institute of Physics, Temperature: Its Measurement and Control in Science and Industry, vol. 5, 1982; D. P. Dewitt, Theory and Practice of Radiation Thermometry, 1988; T. D. McGee, *Principles and Methods of Temperature Measuremen*, 1988; E. C. Magison, *Temperature Measurement in Industry*, 1990; L. Michalski, K. Eckersdorf, and J. McGhee, *Temperature Measurement*, 2d ed., 2001.

Pyromorphite

A mineral series in the apatite group, or in the larger grouping of phosphate, arsenate, and vanadate-type minerals. In this series lead (Pb) substitutes for calcium (Ca) of the apatite formula $Ca_5(PO_4)_3(F,OH,Cl)$, and little fluorine (F) or hydroxide (OH) is present. Substitution between phosphorus (P) and arsenic (As) gives rise to a complete mineral series with pyromorphite, $Pb_5(PO_4)_3Cl$, and mimetite, $Pb_5(ASO_4)Cl$, as pure end members. Substitution between As and vanadium (V) produces another series between mimetite and vanadinite, $Pb_5(VO_4)_3Cl$.

Pyromorphite generally is a chloride-phosphatearsenate of lead, $Pb_5(PO_4, AsO_4)_3Cl$. The half of the series in which P > As is called pyromorphite; the remainder of the series with As > P is termed mimetite. Vanadinite is a chloride-vanadate of lead. P and As may replace V in amounts up to P:V = 1:47 and As:V = 1:1.

The pyromorphite series crystallizes in the hexagonal system (see **illus.**). Crystals are prismatic;



Pyromorphite. (a) Hexagonal crystals, Nassau (specimen from Department of Geology, Bryn Mawr College). (b) Crystal habit (after C. Klein and C. S. Hurlbut, Jr., Manual of Mineralogy, 21st ed., John Wiley and Sons, 1993)

pyramidal faces may be present. Other forms are granular, globular, and botryoidal. Pyromorphite colors range through green, yellow, and brown; vanadinite occurs in shades of yellow, brown, and red.

Pyromorphites are widely distributed as secondary minerals in oxidized lead deposits. Pyromorphite is a minor ore of lead; vanadinite is a source of vanadium and minor ore of lead. *See* APATITE; LEAD; VANADIUM. Wayne R. Lowell

Pyrophyllite

A hydrated aluminum silicate with composition $Al_2Si_4O_{10}(OH)_2$. The mineral is commonly white, grayish, greenish, or brownish, with a pearly to waxy appearance and greasy feel. It occurs as compact masses, as radiating aggregates (see **illus**.),



Specimen of pyrophyllite. (Pennsylvania State University)

and as foliated masses. Pyrophyllite belongs to the layer silicate (phyllosilicate) group of minerals. The atomic (crystal) structure is similar to that of talc, Mg₃Si₄O₁₀(OH)₂, with two of the Mg positions occupied by Al and the third position vacant. The layers are electrically neutral and are held together by weak secondary bonds and long-range ionic (electrostatic) forces. The mineral is soft (hardness $1-1^{1}/_{2}$ on Mohs scale) and has easy cleavage parallel to the structural layers. The mineral is highly stable to acids. Thermal decomposition begins around 750–800°C (1400–1500°F), and recrystallization to mullite and cristobalite occurs at 1200° C (2200° F) and higher temperatures.

Pyrophyllite is used principally for refractory materials, that is, those for use at high temperatures, and in other ceramic applications. In the United States it has been used mainly in making refractories, ceramic materials, and as a carrier for insecticides. The main sources for pyrophyllite in the United States are in North Carolina. An unusual form from the Transvaal, called African wonderstone, is fine-grained and dark gray to black and has the property that it can be shaped mechanically and drilled into complex shapes; after firing it becomes a useful refractory material for special applications. *See* CERAMICS; RE-FRACTORY; SILICATE MINERALS. George W. Brindley

Bibliography. Bureau of Mines, U.S. Department of the Interior, *Mineral Commodity Summaries*, annually; Bureau of Mines, U.S. Department of the Interior, *Minerals Yearbook*; W. A. Deer, R. A. Howie, and J. Zussman, *Rock-Forming Minerals*, vol. 3, 1962.

Pyrotechnics

Mixtures of substances that produce noise, light, heat, smoke, or motion when ignited. They are used in matches, incendiaries, and other igniters; in fireworks and flares; in fuses and other initiators for primary explosives; in delay trains; for powering mechanical devices; and for dispersing materials such as insecticides.

Greek fire, a pyrotechnic whose composition was a military secret and was lost, was used by the Greeks in naval warfare in A.D. 673. But black powder, an intimate mixture of potassium nitrate, charcoal, and sulfur, was perhaps the earliest pyrotechnic and remains the most important one. It was discovered prior to A.D. 1000 in China during the Sung dynasty, where it was used in rockets and fireworks. Its introduction to Europe prior to 1242, when Roger Bacon recorded its formula, eventually revolutionized warfare. Its first use in mines was reported in Hungary in 1677. Its use in blasting rock was also revolutionary, enabling rapid fragmentation of ore and other rock with a great reduction in manual labor.

The making of fireworks developed over the centuries as an art whose secrets were passed down from generation to generation. Pyrotechnics still contain much art even in such highly technical applications as the *Apollo* spacecraft, which contained over 200 pyrotechnic devices, including lightweight, remotely operated electrical switches and mechanical releases required to function only once.

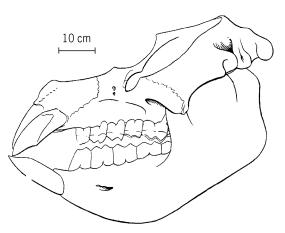
A pyrotechnic mixture contains a fuel and an oxidizer, usually another ingredient to give a special effect, and often a binder. The oxidizer is usually a nitrate, perchlorate, chlorate, or peroxide of potassium, barium, or strontium. Fuels may be sulfur, charcoal, boron, magnesium, aluminum, titanium, or antimony sulfide. Examples of binders that are also fuels are dextrin and natural polysaccharides such as red gum. Salts of strontium, calcium, barium, copper, and sodium and powdered magnesium metal when combined with a fuel and oxidizer can give the special effects of scarlet, brick red, green, blue-green, yellow and white flames, respectively; iron and aluminum granules can give gold and white sparks, respectively; organic dyes and a coolant such as sodium bicarbonate can give colored smokes; potassium benzoate can give unstable burning which results in a whistling sound, and powdered titanium metal can give very rapid burning which results in a booming sound. Mixtures such as boron with red lead have no gaseous reaction products and are useful for confined delay trains where gasesous products would be disruptive.

Many pyrotechnic mixtures are easily ignited by impact, friction, flame, sparks, or static electricity. Even those that burn quietly in small quantities can explode violently when ignited under confinement or in larger quantities. Their manufacture is done with elaborate safety precautions owing to the hazards involved. D. Linn Coursen

Bibliography. R. T. Barbour, *Pyrotechnics in Industry*, 1981; J. A. Conkling, Chemistry of fireworks, *Chem. Eng. News*, 59(26):24-32, 1981; R. Lancaster, *Fireworks Principles and Practice*, 3d ed., 1998; C. F. Ruggieri, *Principles of Pyrotechnics*, 1994.

Pyrotheria

An extinct order of primitive, mastodonlike, herbivorous, hoofed mammals restricted to the Eocene and Oligocene deposits of South America. There is only one family (Pyrotheriidae) in the order and four gen-



Skull and jaw of *Pyrotherium sorondi*, an early Oligocene pyrothere from South America. 10 cm = 25.4 in.

era (*Pyrotherium*, *Propyrotherium*, *Carlozittelia*, and *Griphodon*) in the family.

The characters of this group superficially resembling those in early proboscideans are nasal openings over orbits indicating the presence of a trunk, strong neck musculature, and six upper and four lower bilophodont cheek teeth. They also had four upper and two lower chisellike incisor tusks (see **illus**.). They are distantly related to members of the superorder Paenungulata, including Proboscidea, Xenungulata, and others. *See* PROBOSCIDEA; XENUNGULATA. Gideon T. James

Bibliography. M. J. Benton, Vertebrate Paleontology, 1991; R. L. Carroll, Vertebrate Paleontology and Evolution, 1988.

Pyroxene

A large, geologically significant group of dark, rockforming silicate minerals. Pyroxene is found in abundance in a wide variety of igneous and metamorphic rocks. Because of their structural complexity and their diversity of chemical composition and geologic occurrence, these minerals have been intensively studied by using a wide variety of modern analytical techniques. Knowledge of pyroxene compositions, crystal structures, phase relations, and detailed microstructures provides important information about the origin and thermal history of rocks in which they occur.

Crystal structures. All silicate minerals in the Earth's crust contain silicon cations (Si^{4+}) bonded to four nearest-neighbor oxygen anions (O^{2-}) arranged at the corners of a tetrahedron centered on the Si^{4+} cation. Silicate tetrahedra may polymerize by means of one or more oxygen anions bonding to two Si^{4+} cations. If each tetrahedron shares two of its four anions with adjacent tetrahedra, single silicate chains are formed. The crystallographic repeats of these single chains may range from two to nine tetrahedra; the pyroxene group includes only those with a repeat of two. The two unshared oxygen anions of each silicate tetrahedron are bonded to divalent (for example, Mg^{2+} , Fe^{2+}) or trivalent (for example, Al^{3+} , Fe^{3+})

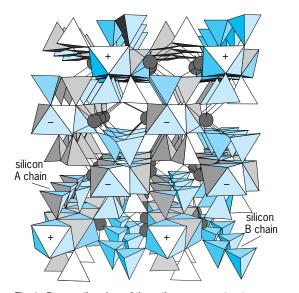


Fig. 1. Perspective view of the orthopyroxene structure. Crystallographic orientation: *a* axis vertical, *b* axis horizontal, *c* axis parallel to the direction of tetrahedral silicate chains. Pyroxene I beam consists of a strip of MI octahedra sandwiched between a nearly straight silicon A chain and a kinked silicon B chain. Relative I-beam orientation is indicated with +1 or -1 on M1 octahedral faces. Shaded circles are M2 sites, showing bonds to nearest-neighbor oxygens.

cations occupying M1 sites, forming octahedral coordination polyhedra sharing edges in chains parallel to the tetrahedral chains (M1 octahedra; **Fig. 1**). The oxygen anions also bond to divalent cations (Ca²⁺, Mg²⁺, Fe²⁺) or sodium (Na⁺) cations that occupy M2 sites. The composite building block of all pyroxene structures, called an I beam, consists of a chain of edge-sharing M1 octahedra and two attached silicate tetrahedral chains, one above and one below the octahedral chain (Fig. 1). The M2 cations located along the edges of the I beams cross-link them to form the three-dimensional structure. The I beams can be stacked in several ways, leading to pyroxene structures with different space groups and unit-cell sizes. *See* COORDINATION CHEMISTRY.

Crystal chemistry. The general chemical formula for pyroxenes is $M2M1T_2O_6$, where T represents the tetrahedrally coordinated sites, occupied primarily by silicon cations (Si⁴⁺). Names of specific end-member pyroxenes are assigned based on composition and structure type. Extensive solid solution exists between many end-member compositions (see **table** and **Fig. 2**). *See* SOLID SOLUTION.

Those pyroxenes containing primarily calcium (Ca^{2+}) or sodium (Na^+) cations in the M2 site are monoclinic. Some omphacites (Fig. 2*b*) exhibit ordering of Ca^{2+} and Na^+ on M2 sites, along with concomitant ordering of magnesium (Mg^{2+}) and aluminum (Al^{3+}) cations on M1 sites. Pyroxenes containing primarily Mg^{2+} or iron(II) (Fe²⁺) cations in the M2 site are orthorhombic at low temperatures, but they may transform to monoclinic at higher temperature. These clinopyroxene phases exhibit the usual monoclinic I-beam stacking sequence but have two crystallographically distinct silicate chains (A and B; Fig. 1) that differ in their degree of straight-

Important end-member pyroxene minerals		
Name	Space group	Formula
Clinoentstatite High clinoenstatite Enstatite Protoenstatite	P2 ₁ /c C2/c Pbca Pbcn	Mg ₂ Si ₂ O ₆
Clinoferrosilite Ferrosilite	P2₁/c] Pbca ∫	$\mathrm{Fe_2}^{2+}\mathrm{Si_2O_6}$
Diopside	C2/c	CaMgSi ₂ O ₆
Headenbergite	C2/c	CaFe ²⁺ Si ₂ O ₆
Jadeite	C2/c	NaAlSi ₂ O ₆
Aegirine	C2/c	NaFe ³⁺ Si ₂ O ₆
Ca-Tschermak's pyroxene	C2/c	CaAl(SiAl)O ₆
Johannsenite	C2/c	CaMnSi ₂ O ₆
Spodumene	C2/c	LiAlSi ₂ O ₆

ness, similar to the orthopyroxenes. They will transform rapidly and reversibly at higher temperatures to a symmetry where all silicate chains are equivalent. The extent of polymorphism a particular pyroxene may undergo thus depends on the content of the M2 sites: those containing significant amounts of Ca^{2+} or Na⁺ cations are always monoclinic, whereas those containing little or no Ca^{2+} or Na⁺ may be orthorhombic or monoclinic; and the particular monoclinic symmetry depends on temperature. *See* CRYS-TAL STRUCTURE; CRYSTALLOGRAPHY.

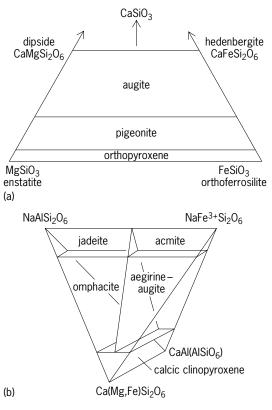


Fig. 2. Nomenclature for common pyroxene solid solutions. (a) Rock-forming pyroxene quadrilateral with end members enstatite-ferrosilite-diopside-hedenbergite. (b) A clinopyroxene composition tetrahedron, with end members diopside + hedenbergite, jadeite, aegirine, and Ca-Tschermak's pyroxene. (After W. D. Nesse, Introduction to Optical Mineralogy, 2d ed., Oxford University Press, 1991)

Within the rock-forming pyroxene quadrilateral (Fig. 2*a*), the Mg²⁺ \leftrightarrow Fe²⁺ solid solution is complete, except that pure ferrosilite (Fe₂Si₂O₆) does not occur naturally but can be synthesized at high pressure. However, there is a strong tendency for the Ca-poor clino- and orthopyroxenes and the Ca-rich clinopyroxenes to be immiscible in each other. This immiscibility decreases with increasing temperature, and is clearly related to the differing M2 coordination requirements of Ca²⁺ versus Mg²⁺ and Fe²⁺. As a consequence, most igneous pyroxenes crystallizing at high temperatures exhibit layers precipitated from the solid crystalline phase (exsolution lamellae) of augite in a pigeonite or orthopyroxene matrix, or vice versa. Pigeonite may transform to orthopyroxene during cooling, further complicating analysis of textural relationships. The orientation of exsolution lamellae of orthopyroxene in clinopyroxene (or the reverse) is parallel to their common (100) planes, whereas exsolution between two clinopyroxenes (pigeonite-augite) is oriented along (001). Detailed study of these exsolution phenomena can provide substantial insight into rock crystallization history. See PIGEONITE; SOLID SOLUTION.

The distribution of Mg^{2+} and Fe^{2+} cations between M1 and M2 in the enstatite-ferrosilite series varies with temperature. At low temperatures Fe^{2+} preferentially occupies M2 sites, but as the temperature increases the preference is weakened and disorder prevails. Examination of the distribution of Fe^{2+} over M1 and M2 in these phases using x-ray diffraction techniques or Mössbauer spectroscopy can provide further information about cooling history. *See* MÖSSBAUER EFFECT; SPECTROSCOPY; X-RAY DIFFRACTION.

In addition to the more important pyroxene compositions (see table and Fig. 2), pyroxenes with a significant amount of manganese cations (Mn^{2+}) in M1 or M2 or both are known. Zinc cations (Zn^{2+}) and some titanium cations (either Ti³⁺ or Ti⁴⁺) can be accommodated in M1, as can chromium (Cr³⁺) or scandium (Sc³⁺) substituting for iron(III) cations (Fe³⁺) in sodic pyroxenes.

Physical properties. Common pyroxenes have specific gravity ranging from about 3.2 (enstatite, diopside) to 4.0 (ferrosilite). Hardnesses on the Mohs scale range from 5 to 6. Iron-free pyroxenes may be colorless (enstatite, diopside, jadeite); as iron content increases, colors range from light green or yellow through dark green or greenish brown, to brown, greenish black, or black (orthopyroxene, pigeonite, augite, hedenbergite, aegirine). Spodumene may be colorless, yellowish emerald green (hiddenite), or lilac pink (kunzite). *See* SPODUMENE.

Occurrence. Pyroxenes in the rock-forming quadrilateral (Fig. 2*a*) are essential constituents of ferromagnesian igneous rocks such as gabbros and their extrusive equivalents, basalts, as well as most peridotites. These rocks generally contain either augite or pigeonite or orthopyroxene, or mixtures of augite with pigeonite or orthopyroxene, or occasionally all three. Complex exsolution relationships may cause all three phases to be present in the same mineral grain. Pyroxenes may also be present as the dark constituents of more silicic diorites and andesites. Aegirine or aegirine-augite is frequently found in alkaline igneous rocks such as nepheline syenites and some alkali granites. *See* ANDESITE; BASALT; DIORITE; GABBRO; GRANITE; IGNEOUS ROCKS; PERIDOTITE.

Pyroxenes, especially those of the diopsidehedenbergite series, are found in medium- to highgrade metamorphic rocks of the amphibolite and granulite facies. In calc-silicate rocks formed by regional metamorphism of silty carbonate rocks and in skarns formed by contact metamorphism of carbonate rocks, these pyroxenes can form through reaction (1).

$$\begin{array}{lll} \mbox{CaMg(CO_3)}_2 + 2\mbox{SiO}_2 \end{tabular} & \mbox{CaMgSi}_2\mbox{O}_6 + 2\mbox{CO}_2 & (1) \\ \mbox{Dolomite} & \mbox{Quartz} & \mbox{Diopside} & \mbox{Carbon} \\ \mbox{dioxide} & \end{tabular} \end{array}$$

See METAMORPHISM.

Jadeite is found in glaucophane schists, where under moderate pressure it forms from the breakdown of albite feldspar according to reaction (2).

$$NaAlSi_3O_8 \rightarrow NaAlSi_2O_6 + SiO_2$$
 (2)
Albite Jadeite Quartz

Gem-quality jadeite is one of two materials used as jade in jewelry and artifacts (the other is nephrite, an amphibole). The closely related omphacite is an essential constituent of eclogites (high-grade metamorphosed basalts and gabbros) where it occurs with pyrope garnet. Manganese (Mn)pyroxenes are restricted to skarn-type metamorphosed manganese deposits. Spodumene is found exclusively in lithium-bearing granite pegmatites, where it can occur in very large crystals, often mined as a source of lithium. Clinoenstatite is rare in terrestrial environments but is a common constituent of stony meteorites. See AMPHIBOLITE; ECLOGITE; FACIES (GEOLOGY); FELDSPAR; GLAUCOPHANE; GRAN-ULITE; JADE; METAMORPHIC ROCKS; OMPHACITE; PEG-MATITE; SCHIST; SKARN.

The peridotites found in the Earth's upper mantle contain Mg-rich, Ca-poor pyroxenes, in addition to olivine and other minor minerals. At successively greater depths in the mantle, these Mg-rich pyroxenes will transform sequentially to spinel (Mg₂SiO₄) plus stishovite (SiO₂), an ilmenite structure, or a garnet structure, depending on temperature; and finally at depths of around 360–420 mi (600–700 km) to an MgSiO₃ perovskite structure thought to be the major stable constituent of the lower mantle to depths of around 1700 mi (2900 km) at the core-mantle boundary. *See* ASTHENOSPHERE; EARTH INTERIOR; GARNET; ILMENITE; OLIVINE; PEROVSKITE; SILICATE MINERALS; SPINEL; STISHOVITE. Charles W. Burnham

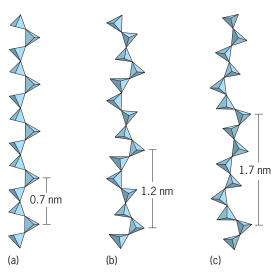
Bibliography. W. A. Deer, R. A. Howie, and J. Zussman, *An Introduction to the Rock-Forming Minerals*, 2d ed., 1996; N. Morimoto et al., Nomenclature of pyroxenes, *Amer. Mineralog.*, 73:1123-1133, 1988; C. T. Prewitt (ed.), Pyroxenes, *Reviews in Mineralogy*, vol. 7, 1980.

Pyroxenite

A heavy, dark-colored, phaneritic (visibly crystalline) igneous rock composed largely of pyroxene with smaller amounts of olivine or hornblende. Pyroxenite composed largely of orthopyroxene occurs with anorthosite and peridotite in large, banded gabbro bodies. These rocks are formed by crystallization of gabbroic magma (rock melt). Some of these pyroxenite masses are rich sources of chromium. Certain pyroxenites composed largely of clinopyroxene are also of magmatic origin, but many probably represent products of reaction between magma and limestone. Other pyroxene-rich rocks have formed through the processes of metamorphism and metasomatism. See GABBRO; IGNEOUS ROCKS; PERIDOTITE; PYROXENE. Carleton A. Chapman

Pyroxenoid

A group of silicate minerals whose physical properties resemble those of pyroxenes. In contrast with the two-tetrahedra periodicity of pyroxene single silicate chains, the pyroxenoid crystal structures contain single chains of $(SiO_4)^4$ silicate tetrahedra having repeat periodicities ranging from three to nine (see illus.). The tetrahedron is a widely used geometric representation for the basic building block of most silicate minerals, in which all silicon cations (Si⁴⁺) are bonded to four oxygen anions arranged as if they were at the corners of a tetrahedron. In pyroxenoids, as in other single-chain silicates, two of the four oxygen anions in each tetrahedron are shared between two Si⁴⁺ cations to form the single chains, and the other two oxygen anions of each tetrahedron are bonded to divalent cations, such as calcium (Ca^{2+}) , iron (Fe^{2+}) , or manganese (Mn^{2+}) . These divalent cations bond to six (or sometimes seven or eight) oxygen anions, forming octahedral (or irregu-



Tetrahedral silicate chains in three pyroxenoid structures: (a) wollastonite with three-tetrahedra periodicity, (b) rhodonite with five-tetrahedra periodicity, and (c) pyroxmangite-pyroxferroite with seven-tetrahedra periodicity.

Pyroxenoid minerals	
Name	Formula
Anhydrous	
Wollastonite	CaSiO ₃
Bustamite-	
ferrobustamite	~ Ca(Mn,Fe ²⁺)Si ₂ O ₆
Rhodonite	$\sim CaMn_4Si_5O_{15}$
Pyroxmangite-	
pyroxferroite	~ Ca(Mn,Fe ²⁺) ₆ Si ₇ O ₂₁
Ferrosilite III	FeSiO ₃
Hydrous	
Pectolite-serandite	(Ca,Mn) ₂ NaHSi ₃ O ₉
Babingtonite	Ca ₂ (Fe,Mn,Mg)Fe ³⁺ HSi ₅ O ₁₅
Nambulite-marsturite	(Li,Na)(Mn,Ca)₅HSi₅O ₁₅
Santaclaraite	CaMn ₄ HSi ₅ O ₁₅ (OH)·H ₂ O

lar seven- or eight-cornered) coordination polyhedra. *See* PYROXENE; SILICATE MINERALS.

The pyroxenoid structures all have composite structural units consisting of strips of octahedra (or larger polyhedra) two or more units wide formed by sharing of polyhedral edges, to which the silicate tetrahedral chains are attached on both top and bottom. The repeat periodicity of the octahedral strips is the same as that of the silicate chains to which they are attached. These composite units are cross-linked to form the three-dimensional crystal structures. Although various crystallographic settings have been employed in the past, there has evolved general agreement that the most useful orientation selects the unit-cell c axis to be parallel to the silicate chain and octahedral strip direction, and sets the (100) plane parallel to the plane of the octahedral strips, which also is a plane of approximately close-packed oxygen atoms; this setting is analogous to that used for pyroxenes. The pyroxenoid minerals are all triclinic, with either $C\overline{1}$ or $I\overline{1}$ space-group symmetry depending on the stacking of the composite units, and c-axis lengths ranging from about 0.71 nanometer for three-repeat silicate tetrahedral chains to about 2.3 nm for nine-repeat chains. See CRYSTAL STRUCTURE.

Types. There are two series of pyroxenoid minerals, one anhydrous and one hydrous (see table). The anhydrous pyroxenoids are significantly more abundant. A general formula for anhydrous pyroxenoids is (Ca, Mn, Fe²⁺)SiO₃. Silicate-chain repeat length is inversely proportional to mean divalent cation size, for example, wollastonite with three-repeat chains, and ferrosilite III with nine-repeat chains. Wollastonite (CaSiO₃) has two other polymorphs: parawollastonite is a rare pyroxenoid having a different stacking sequence of the composite tetrahedraloctahedral units; and pseudowollastonite is a ring silicate and not, therefore, a pyroxenoid. Both bustamite (CaMnSi₂O₆) and ferrobustamite (CaFeSi₂O₆) are polymorphous with pyroxenes: johannsenite and hedenbergite respectively. The MnSiO₃ composition can exist as rhodonite, pyroxmangite, pyroxene, or garnet, depending on temperature and pressure. Ferrosilite III has been synthesized in the laboratory as a pyroxenoid polymorph of the pyroxenes clinoferrosilite and orthoferrosilite; it has not been found occurring naturally. The hydrogen in hydrous pyroxenoids is hydrogen-bonded between two oxygen atoms; additional hydrogen in santaclaraite is bound as hydroxyl (OH) and as a water molecule (H_2O) .

Occurrence. Wollastonite is the most abundant pyroxenoid, occurring primarily as a consequence of high-grade metamorphism of siliceous limestone $[CaCO_3$ with silicon dioxide $(SiO_2)]$. Under high-temperature conditions, it can form from the decarbonation reaction below. Wollastonite is mined

$$CaCO_3 + SiO_2 \longrightarrow CaSiO_3 + CO_2$$

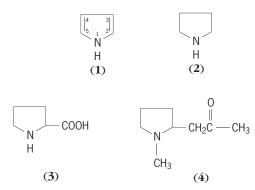
Carbon
dioxide

in New York and California and, to a lesser extent, in Mexico, Finland, Kenya, South-West Africa, and central Asia. It is used commercially in the ceramic industry, as a filler in high-quality lightcolored exterior paints, as a strengthening additive in some plastics, and in ceramic-bonded abrasives. Rhodonite and other manganese-rich pyroxenoids are found chiefly in association with some metamorphosed manganese deposits; particularly wellformed rhodonite crystals are sold as semiprecious gems, with value related primarily to their pink color. Pyroxferroite was first found as a minor phase in lunar basalts; it has also been found in pegmatites in Japan. *See* RHODONITE; WOLLASTONITE.

Charles W. Burnham Bibliography. W. A. Deer et al., *Rock-Forming Minerals*, vol. 2A, 2d ed., 1978; W. A. Deer, R. A. Howie, and J. Zussman, *An Introduction to Rock-Forming Minerals*, 2d ed., 1992.

Pyrrole

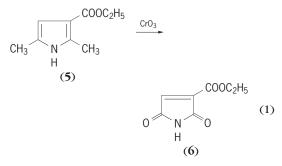
One of a group of organic compounds containing a doubly unsaturated five-membered ring in which nitrogen occupies one of the ring positions. Pyrrole (1) is a representative compound. The pyrrole system is found in the green leaf pigment, chlorophyll, in the red blood pigment, hemoglobin, and in the blue dye, indigo. Interest in these colored bodies has been largely responsible for the intensive study of pyrroles. Tetrahydropyrrole, or pyrrolidine (2), is part of the structures of two protein amino acids, proline (3) and hydroxyproline, and of hygrine (4),



which is an alkaloid from Peruvian coca. *See* HETE-ROCYCLIC COMPOUNDS; INDOLE; PORPHYRIN.

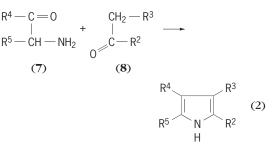
Properties. Pyrrole (1) is a liquid, bp 130°C $(266^{\circ}F)$, n^{20}_{D} 1.5085 (1.5098), and density (20/4) 0.948 (0.969), that darkens and resinifies on standing in air, and that polymerizes quickly when treated with mineral acid. Polyalkyl pyrroles are not so sensitive and negatively substituted pyrroles are even less so. Pyrrole is a planar, aromatic compound, with an experimental resonance energy of 22-27 kcal/mole. Familiar substitution processes, such as halogenation, nitration, sulfonation, and acylation, can be realized. Substitution generally occurs more readily than in the corresponding benzene analog. The entering group favors the 2 or 5 position. Pyrrole, by virtue of its heterocyclic nitrogen, is very weakly basic and is comparable in this respect to urea or to semicarbazide. The hydrogen at the 1 position is removable as a proton, and accordingly, pyrrole is also an acid, although a weak one.

Pyrroles are not as resistant to oxidation as the analogous benzene compounds. Controlled chromic acid oxidation converts pyrroles with or without groups in the α positions to the corresponding maleimides. In this way for example, ethyl-2,5-dimethylpyrrole-3-carboxylate (5) is oxidized to a maleimide derivative [6; reaction (1)]. Conditions are

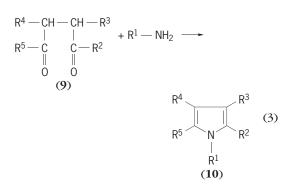


available for useful reductions of groups attached to the nucleus. Zinc and hydrochloric acid can reduce the pyrrole ring to dihydropyrrole (pyrroline). Raney nickel under rigorous conditions, or platinum or palladium under milder conditions, serve as catalysts in the hydrogenation of pyrroles to pyrrolidines.

Preparation. The Knorr synthesis [reaction (2)],



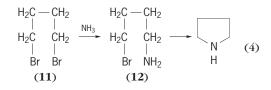
probably the most versatile pyrrole synthesis, combines an α -aminoketone (7) with an α -methylene carbonyl compound (8). The condensation is carried out either in glacial acetic acid or in aqueous alkali. The R groups may be varied widely; however, best results are obtained when R³ of the carbonyl compounds (8) is an activating group, for example, carbethoxy. Another useful general synthesis, the Paal-Knorr synthesis [reaction (3)], converts a



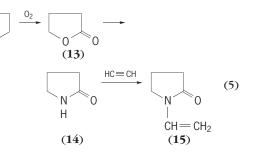
1,4-dicarbonyl compound (9) by cyclization with ammonia ($\mathbb{R}^1 = \mathbb{H}$) or with a primary amine to a pyrrole (10). Pyrrole itself is obtained by pyrolysis of ammonium mucate (saccharate). Pyrolysis of primary amine salts of mucic acid gives 1-substituted pyrroles.

Derivatives. Chloro-, bromo-, and iodopyrroles have been prepared, with bromopyrroles comprising the largest group. Halopyrroles carrying an electronegative group, such as carbethoxy, are more stable than the same compounds without the group. 2,3,4,5-Tetraiodopyrrole, formed by iodination of pyrrole or of tetrachloromercuripyrrole, is useful as an antiseptic in the same areas as iodoform. Pyrroles with ethylmagnesium bromide react as active hydrogen compounds to give Grignard derivatives. Ethyl formate reacts with such pyrryl Grignard reagents to form pyrrole aldehydes; chloroformic ester reacts to give pyrrolecarboxylic esters. Pyrrole aldehydes can also be formed by formylation of a vacant pyrrole position with hydrocyanic acid and hydrogen chloride. Acylpyrroles, that is, pyrryl ketones, are prepared satisfactorily and under mild conditions by Friedel-Crafts acylation of the pyrrole nucleus.

Pyrrolidine (2), bp $87-88^{\circ}$ C (189-190°F), can be prepared by catalytic hydrogenation of pyrrole or by ring-closure reactions [reaction (4)]. Either



1,4-dibromobutane (11) or 4-bromobutylamine (12) can serve as starting material. 2-Ketopyrrolidine, or pyrrolidone (14), is of considerable interest in connection with the preparation of polyvinylpyrrolidone. Pyrrolidone, which can be formed from tetrahydrofuran by autoxidation in the presence of a cobalt catalyst and treatment of the resulting γ -butyrolactone (13) with ammonia, is combined with acetylene to form vinylpyrrolidone (15) [reaction (5)]. Polymerization of this material furnishes polyvinylpyrrolidone, a material of relatively high



molecular weight, which is suitable for maintaining osmotic pressure in blood and so acting as an extender for plasma or whole blood. *See* PYRIDINE. Walter J. Gensler

Bibliography. D. H. R. Barton et al. (eds.), *Comprehensive Organic Chemistry*, 1978; F. A. Carey and R. J. Sundberg, *Advanced Organic Chemistry*, pt. A: *Structure and Mechanisms*, 4th ed., 2000.

Pythagorean theorem

This theorem states that in any right triangle the square on the hypotenuse is equal to the sum of the squares on the other two sides: $r^2 = x^2 + y^2$. More than 100 different proofs have been given for this extremely important theorem of euclidean plane geometry, which is named for the Greek philosopher and mathematician Pythagoras (sixth century B.C.). One of these is based on a comparison of areas in the squares in **Fig. 1**.

The three-dimensional Pythagorean theorem (**Fig. 2**) may be phrased "the square of the diagonal of a rectangular box is equal to the sum of the squares

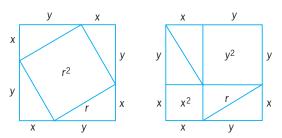


Fig. 1. Proof of Pythagorean theorem based on comparison of areas of squares.

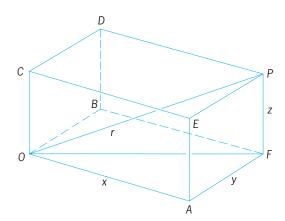


Fig. 2. Three-dimensional Pythagorean theorem.

of three adjacent edges that meet at a vertex: $r^2 = x^2 + y^2 + z^2$."

Squares that are equal to the sum of two or three squares were known to Pythagoras. For example, $3^2 + 4^2 = 5^2$ and $1^2 + 4^2 + 8^2 = 9^2$. The name pythagorean triple is given to any triple (x, y, r) of whole numbers such that $x^2 + y^2 = r^2$, and the name Pythagorean quadruple to such integral quadruples (x, y, z, r) for which $x^2 + y^2 + z^2 = r^2$. All pythagorean

triples in which *x*, *y*, *r* are without common factor and *x* is odd are obtained by replacing the letters *a* and *b* in the triple $(a^2 - b^2, 2ab, a^2 + b^2)$ by whole numbers that have odd sum and no common factor; and similarly all such pythagorean quadruples are found by replacing *a*, *b*, *c*, *d* in the quadruple $(a^2 - b^2 + c^2 - d^2, 2ab + 2cd, 2bc - 2ad, a^2 + b^2 + c^2 + d^2)$ by whole numbers that have odd sum and no common factor. J. Sutherland Frame



Q (electricity)

Often called the quality factor of a circuit, Q is defined in various ways, depending upon the particular application. In the simple *RL* and *RC* series circuits, Q is the ratio of reactance to resistance, as in Eqs. (1),

$$Q = \frac{X_L}{R} \quad Q = \frac{X_C}{R} \quad \text{(a numerical value)} \tag{1}$$

where X_L is the inductive reactance, X_C is the capacitive reactance, and R is the resistance. An important application lies in the dissipation factor or loss angle when the constants of a coil or capacitor are measured by means of the alternating-current bridge.

Q has greater practical significance with respect to the resonant circuit, and a basic definition is given by Eq. (2), where Q_0 means evaluation at resonance.

$$Q_0 = 2\pi \frac{\text{max. stored energy per cycle}}{\text{energy lost per cycle}}$$
(2)

For certain circuits, such as cavity resonators, this is the only meaning Q can have.

For the *RLC* series resonant circuit with resonant frequency f_0 , Eq. (3) holds, where *R* is the total

$$Q_0 = \frac{2\pi f_0 L}{R} = \frac{1}{2\pi f_0 CR}$$
(3)

circuit resistance, L is the inductance, and C is the capacitance. Q_0 is the Q of the coil if it contains practically the total resistance R. The greater the value of Q_0 , the sharper will be the resonance peak.

The practical case of a coil of high Q_0 in parallel with a capacitor also leads to $Q_0 = 2\pi f_0 L/R$. *R* is the total series resistance of the loop, although the capacitor branch usually has negligible resistance.

In terms of the resonance curve, Eq. (4) holds,

$$Q_0 = \frac{f_0}{f_2 - f_1} \tag{4}$$

where f_0 is the frequency at resonance, and f_1 and f_2 are the frequencies at the half-power points. *See* RESONANCE (ALTERNATING-CURRENT CIRCUITS). Burtis L. Robertson

Q meter

A direct-reading instrument formerly widely used to measure the *Q* of an electrical component or passive network of components at radio frequencies. The instrument measures the voltage applied to a component or circuit together with the voltage across a key component and calculates the *Q* of the circuit from their ratio. An impedance-measuring instrument (bridge or analyzer) is more likely to be employed now. *See* ELECTRICAL MEASUREMENTS; Q (ELECTRICITY); RADIO-FREQUENCY IMPEDANCE MEA-SUREMENTS. Bryan P. Kibble

Bibliography. D. A. Bell, *Electronic Instrumentation and Measurements*, 2d ed., 1994; A. J. Diefenderfer and B. E. Holton, *Principles of Electronic Instrumentation*, 3d ed., 1994; L. Schnell (ed.), *Technology of Electrical Measurements*, 1993.

Quadrature

The condition in which the phase angle between two alternating quantities is 90° , corresponding to one-quarter of an electrical cycle. The electric and magnetic fields of electromagnetic radiation are in space quadrature, which means that they are at right angles in space. *See* ELECTROMAGNETIC RADIATION.

The current and voltage of an ideal coil are in quadrature because the current through the coil lags behind the voltage across the coil by exactly 90° . The current and voltage of an ideal capacitor are also in quadrature, but here the current through the capacitor leads the voltage across the capacitor by 90° .

In these last two cases the current and voltage are in time quadrature. *See* ALTERNATING-CURRENT CIR-CUIT THEORY. John Markus

Quadric surface

A surface defined analytically by an equation of the second degree in three variables. If these variables are x, y, z, such an equation has the form of Eq. (1).

$$ax^{2} + by^{2} + cz^{2} + 2exy + 2fxz + 2gyz + 2px + 2qy + 2rz + d = 0$$
(1)

Every plane section of such a surface is a conic. To simplify the notation, the corresponding capital letters are used to denote the cofactors of the coefficients *a*, *b*, *c*, *d*, *e*, *f*, *g*, *p*, *q*, *r* in the determinant shown in Eq. (2). Thus $D = abc + 2efg - ag^2 - bf^2$ –

$$\Delta = \begin{vmatrix} a, & e & f & p \\ e & b & g & q \\ f & g & c & r \\ p & q & r & d \end{vmatrix}$$
(2)

 ce^2 . By λ_1 , λ_2 , λ_3 are denoted the three roots of the polynomial whose product is *D*, shown in Eq. (3).

$$\phi(\lambda) = \begin{vmatrix} a - \lambda & e & f \\ e & b - \lambda & g \\ f & g & c - \lambda \end{vmatrix}$$
(3)

Quadrics are classified as central if $D \neq 0$, and noncentral if D = 0; nondegenerate if $\Delta \neq 0$, and degenerate if $\Delta = 0$. For further subclassifications it can be assumed that the coefficients in Eq. (1) are real numbers. The roots of $\phi(\lambda)$ will then be real also.

Central quadrics. The central quadric has a center of symmetry at the point x = P/D, y = Q/D, z = R/D. Its equation is freed of linear terms by the translation of axes x' = x - P/D, y' = y - Q/D, z' = z - R/D and becomes Eq. (4).

$$x'^{2} + by'^{2} + cz'^{2} + 2ex'y' + 2fx'z' + 2gy'z' = -\frac{\Delta}{D} \quad (4)$$

C

By an appropriate rotation of axes this equation can be reduced further to the form of Eq. (5). If $\Delta = 0$,

$$\lambda_1 X^2 + \lambda_2 Y^2 + \lambda_3 Z^2 = -\frac{\Delta}{D}$$
(5)
$$\phi(\lambda_i) = 0$$

the degenerate central quadric of Eq. (1) is a quadric cone. Whenever the real roots λ_i of $\phi(\lambda) = 0$ are not all of the same sign, the quadric cone is real and is a surface formed of lines connecting the points of a real conic to a point outside its plane, called the vertex. If the roots λ_i are all of the same sign, the quadric cone is imaginary except for a single real point, the vertex.

If $\Delta \neq 0$, the nondegenerate central quadric of (1) is a real ellipsoid, a hyperboloid of one sheet, a hyperboloid of two sheets, or an imaginary ellipsoid according to whether three, two, one, or none of

the λ_i are equal, the central quadric is a spheroid or a hyperboloid of revolution. If the three roots λ_i are equal, it is a sphere.

Noncentral quadrics. If D = 0, the quadric of (1) is noncentral, and $\phi(\lambda) = 0$ has a root $\lambda_3 = 0$. A suitable rotation of axes reduces Eq. (1) to Eq. (6), where $\Delta =$

$$\lambda_1 X^2 + \lambda_2 Y^2 + 2\mu_1 X + 2\mu_2 Y + 2\mu_3 Z + d = 0$$
(6)

 $-\lambda_1\lambda_2\mu_3^2$. A translation further reduces μ_1 to 0 if $\lambda_1 \neq 0$, μ_2 to 0 if $\lambda_2 \neq 0$, and *d* to 0 if $\mu_3 \neq 0$. A nondegenerate noncentral quadric ($D = 0, \Delta \neq 0$) is an elliptic paraboloid if $\lambda_1\lambda_2 > 0$, or a hyperbolic paraboloid if $\lambda_1\lambda_2 < 0$. A degenerate noncentral quadric ($D = \Delta = 0$) is either a quadric cylinder (if at least one of the cofactors in Δ is not 0) or a pair of planes, a single plane or a line, or it has no real locus. If $\lambda_1\lambda_2 > 0$ and $\mu_3 = 0$, the surface may be a real elliptic cylinder, depending on the value of *d*.

If $\lambda_1\lambda_2 < 0$ and $\mu_3 = 0$, the surface is a hyperbolic cylinder or a pair of planes. If either $\lambda_1\mu_3$ or $\lambda_2\mu_3$ is 0, but not both, the surface is a parabolic cylinder. *See* CYLINDER. J. Sutherland Frame

Qualitative chemical analysis

The branch of chemistry concerned with identifying the elements and compounds present in a sample of matter. Inorganic qualitative analysis traditionally used classical "wet" methods to detect elements or groups of chemically similar elements, but instrumental methods have largely superseded the test-tube methods. Methods for the detection of organic compounds or classes of compounds have become increasingly available and important in organic, forensic, and clinical chemistry. Once it is known which elements and compounds are present, the role of quantitative analysis is to determine the composition of the sample. *See* ANALYTICAL CHEM-ISTRY; QUANTITATIVE CHEMICAL ANALYSIS.

A considerable part is played by the human senses as analytical tools. There are surprisingly sensitive standard procedures which employ the human sense of smell for the qualitative detection and semiquantitative estimation of extremely small concentrations of objectionable odor-causing substances in water supplies. Perfumers develop a skill based on the sense of smell that is remarkable in identifying and making use of odoriferous substances, both to blend perfumes and to disguise objectionable odors. Tea tasters and wine experts depend on a developed skill that is a combination of taste and smell. The trained eye can serve as a relatively good spectrophotometer, both qualitatively in regard to absorption maxima and to some extent quantitatively to measure color intensity. Prospectors and geologists greatly rely on visual skills for the recognition of ores and rock types.

Inorganic Analysis

The operating principles of all systematic inorganic qualitative analysis schemes for the elements are

similar: separation into groups by reagents producing a phase change; isolation of individual elements within a group by selective reactions; and confirmation of the presence of individual elements by specific tests. In the classical cation scheme, the wide range of solubilities of sulfide salts of the metals was exploited through control of sulfide ion concentration by pH control. Outlines for short qualitative schemes for several metals and three common anions of interest in the environment are presented in Figs. 1 and 2. The cation scheme has been adapted from a more extensive non-hydrogen sulfide scheme. The outlines show the reagents and steps used, and incomplete chemical equations are presented to illustrate the reactions and conditions employed. Although they avoid the use of hydrogen sulfide or thioacetamide (important reagents in earlier methods), the schemes include most of the basic qualitative analysis techniques of selective and pHcontrolled precipitation, oxidation and reduction, complex ion formation, gas evolution, and a flame test.

Sample size. Through usage and tradition, descriptive terms for sample sizes have been:

macro	0.1 gram or more
semimicro	0.01 to 0.1 gram
micro	1 milligram (10^{-3} g)
ultramicro	1 microgram (10 ⁻⁶ g)
submicrogram	less than 1 microgram

For defining the smallest amount of a substance that

Cations included: Hg_2^{2+} , Cr^{3+} , Sb^{3+} , Pb^{2+} , Sr^{2+} , Cd^{2+} , Cu^{2+} Step 1: Add HCI	* , Hg²+ , As³+
Precipitate: $Hg_2CI_2 \downarrow$ (white) $Hg_2CI_2 + NH_4OH \rightarrow HgO \downarrow$ (black mixture with $HgNH_2CI$ confirms $Hg_2^{2^+}$)	Solution: Cr ³⁺ , Sb ³⁺ , Pb ²⁺ , Sr ²⁺ , Cd ²⁺ , Cu ²⁺ , Hg ²⁺ , As ³⁺
Step 2: Adjust pH to 3.5–4.0, add NH ₄ benzoate and Na ^{$+$} be	nzoate
$\begin{array}{l} \label{eq:precipitate: Cr(OH) (benzoate)_2 \downarrow, SbOCI \downarrow \\ Cr(OH) (benzoate)_2 + OH^- + H_2O_2 \rightarrow CrO_4^{2-}$ \\ CrO_4^{2^-} + Pb^{2^+} \rightarrow PbCrO_4 \downarrow (yellow, confirms Cr^{3^+})$ \\ SbOCI + Na^+ tartrate \rightarrow Sb-tartrate complex, acidify with HNO_3$ \\ Sb-tartrate complex + H_2S \rightarrow Sb_2S_3 \downarrow (orange, confirms Sb^{3^+})$ \\ \end{array}$	Solution: Pb ²⁺ , Sr ²⁺ , Cd ²⁺ , Cu ²⁺ , Hg ²⁺ , As ³⁺
Step 3: Add NaF, adjust pH to 3.4	V
$\begin{array}{l} \label{eq:precipitate: PbF_2 \downarrow, $SrF_2 \downarrow}\\ PbF_2 + 2NH_4^* acetate \rightarrow Pb(acetate)_2$\\ Pb(acetate)_2 + S^{2^-} \rightarrow PbS \downarrow (black, confirms Pb^{2*})$\\ SrF_2 + 2HCI \rightarrow SrCI_2$\\ SrCI_2, flame test \rightarrow crimson flame confirms Sr^{2*} \end{array}$	Solution: Cd ²⁺ , Cu ²⁺ , Hg ²⁺ , As ³⁺
Step 4: Make alkaline with NaOH, boil free of NH ₃	
$\begin{array}{l} \mbox{Precipitate: } Cd(OH)_2 \downarrow, Cu(OH)_2 \downarrow, HgO \downarrow \\ Cd(OH)_2 + 2HCI \rightarrow Cd^{2+}, treat with NH_4OH, acidify with \\ HCI \\ Cd^{2+} + S^{2-} \rightarrow CdS \downarrow (yellow, confirms Cd^{2+}) \\ Cu(OH)_2 + 2HCI \rightarrow Cu^{2+}, treat with NH_4OH, acidify with \\ HCI \\ 2Cu^{2+} + 4SCN^- \rightarrow (SCN)_2 + 2CuSCN \downarrow (white, indicative of Cu^{2+}) \\ 2CuSCN + HNO_3, oxidize \rightarrow 2Cu^{2+} \\ Cu^{2+} + 4NH_3 \rightarrow [Cu(NH_3)_4]^{2+} (deep blue, confirms \\ Cu^{2+}) \\ HgO + 2HCI \rightarrow HgCI_2 \\ HgCI_2 + NH_4OH \rightarrow HgNH_2CI \downarrow \\ HgNH_2CI + HNO_3 \rightarrow HGCI_2; two tests follow: \\ a. 2HgCI_2 + Sn^{2+} \rightarrow 2HgO (white turning gray \\ confirms Hg^{2+}) \\ b. HgCI_2 + 2\Gamma \rightarrow HgI_2 \downarrow (red-orange, confirms Hg^{2+}) \\ \end{array}$	Solution: As ³⁺
Step 5: Acidify with 12 <i>M</i> HCl, boil to reduce volume, add 15.	MNH₄OH, 6MNaOH, and Al granules
	$\begin{array}{l} \text{Solution: } \text{As}^{3^+} + 4\text{OH}^- \rightarrow \text{AsO}_2^- + 2\text{H}_2\text{O} \\ 2\text{AsO}_2^- + 2\text{Al} + 3\text{H}_2\text{O} \rightarrow 2\text{AsH}_3 \uparrow \text{ (gas)} \\ 2\text{AsH}_3 + 6\text{Ag}^+ \rightarrow 6\text{AgO} \downarrow \text{ (black precipitate} \\ \text{on } \text{AgNO}_3\text{-treated paper strip confirms } \text{As}^{3^+} \text{)} \end{array}$

Fig. 1. Outline of a scheme of analysis for cations of interest in the environment. (After J. L. Lambert and C. E. Meloan, A simple qualitative analysis scheme for several environmentally important elements, J. Chem. Educ., 54:249, 1977)

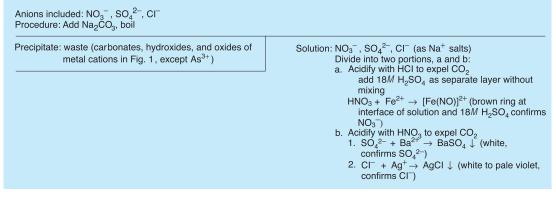


Fig. 2. Outline of a scheme of analysis for three common anions. (After J. L. Lambert and C. E. Meloan, A simple qualitative analysis scheme for several environmentally important elements, Chem. Educ., 54:249, 1977)

can be detected by a given method, the term "limit of identification" is used. Under favorable conditions, an extremely sensitive method can detect as little as 10^{-15} (the ability to detect a single cesium atom in gaseous methane by a laser-induced photoionization procedure has been reported, although it hardly qualifies as an analytical method). If the smallest amount is present in extremely dilute solution, "limiting concentration" defines the lowest concentration of the substance in solution that will give the test. If another substance is present that gives similar results or interferes with the test, "limiting proportion" defines the ratio of the mass of the substance sought to the largest mass of interfering substance that can be present without producing spurious results.

Visual tests. Spot tests are selective or specific single qualitative chemical tests carried out on a spot plate (a glass or porcelain plate with small depressions in which drop-size reactions can be carried out), on paper, or on a microscope slide. On paper or specially prepared adsorbent surfaces, spot tests become one- or two-dimensional through the use of solvent migration and differential adsorption (thinlayer chromatography). Containing selected indicator dyes, pH indicator paper strips are widely used for pH estimation. Solid reagent monitoring devices and indicator tubes are used for the detection and estimation of pollutant gases in air. *See* CHROMATOG-RAPHY; PH.

By use of the microscope, crystal size and habit can be used for qualitative identification. With the addition of polarized light, chemical microscopy becomes a versatile method widely employed in forensic analysis, geology, ceramics, and for the identification of particulate matter in air. *See* CHEMICAL MI-CROSCOPY; FORENSIC CHEMISTRY.

Instrumental methods. Generally speaking, instrumental methods of analysis have been developed primarily for quantitative analysis, but any quantitative method can be adapted to qualitative analysis. Some, such as electrochemical methods, are not often used in qualitative analysis (an exception would be the ubiquitous pH meter). Others, such as column chromatography (gas and liquid) and mass spectrometry, are costly and complicated but capable of providing unique results. *Spectroscopy.* The first instrumental method emission spectroscopy—began as a qualitative tool. Emission spectroscopy still is important in establishing the presence or absence of a suspected element in forensic analysis. The simple qualitative flame test developed into flame photometry and subsequently into atomic absorption spectrometry, which is one of the most widely used analytical methods today. Absorption spectrometry (particularly in the infrared region), mass spectrometry, and nuclear magnetic resonance are extensively used for identification and structural analysis or organic compounds. *See* ATOMIC SPECTROMETRY; EMISSION SPECTROCHEMICAL ANALYSIS.

X-rays, electrons, and ions. Bombardment of surfaces by x-rays, electrons, and positive ions has given rise to a number of methods useful in analytical chemistry. X-ray diffraction is used to determine crystal structure and to identify crystalline substances by means of their diffraction patterns. X-ray fluorescence analysis employs x-rays to excite emission of characteristic x-rays by elements. The electron microprobe uses electron bombardment in a similar manner to excite x-ray emission. Bombarding electrons can also cause emission of electrons having energies characteristic of the target atoms. Bombardment by H⁺ or He⁺ ions permits identification and determination of elements by measurement of backscattering of the bombarding ions, emission of secondary ions for analysis by a mass spectrometer, or emission of characteristic photons for spectrophotometric analysis. See SECONDARY ION MASS SPECTROMETRY (SIMS); X-RAY DIFFRACTION; X-RAY FLUORESCENCE ANALYSIS.

Activation analysis. Neutron activation analysis, wherein neutron bombardment induces radioactivity in isotopes of elements not naturally radioactive, involves measurement of the characteristic gamma radiation and modes of decay produced in the target atoms. This bombardment method is not restricted to surface analysis, and is independent of the chemical combinations of the target elements. Some elements can be detected in solids in concentrations as low as 1 in 10^{12} by this method. Mössbauer spectroscopy, by contrast, involves selective absorption of gamma radiation from an excited isotope by

the stable isotope of the same atomic and mass number, and is exceedingly sensitive to the chemical environment of the target atoms. *See* ACTIVATION ANALYSIS. Jack L. Lambert

Organic Analysis

Qualitative analysis of an organic compound is the process by which the characterization of its class and structure is determined. Due to the numerous classes of organic compounds (alcohols, aldehydes, ketones, and so forth) and the complexity of their molecular structures, a systematic analytical procedure is often required. A typical procedure entails an initial assignment of compound classification, followed by a complete identification of the molecular structure.

Physical methods. The initial step is an examination of the physical characteristics. Color, odor, and physical state can be valuable clues. For example, nitro compounds are often highly colored. The color of a sample is often misleading. Air oxidation and sample decomposition can result in color formation. Some compound classes have characteristic odors. While many esters have a sweet odor, thiols yield a very disagreeable odor. Caution should be exercised when investigating the odor of an unknown compound. Direct inhalation should be avoided.

The physical constants of an unknown compound provide pertinent data for the analyst. Constants such as melting point, boiling point, specific gravity, and refractive index are commonly measured. The melting point is usually determined by packing a small quantity of the solid in a capillary tube. The tube is then heated in either an oil bath (Thiel tube method) or an aluminum block (electrically heated) to the melting-point temperature. For a more precise measurement, a microscope equipped with a hot stage can be employed. The temperature range over which a solid melts is a measure of compound purity. Samples with wide ranges should be recrystallized. Boiling-point measurements are conducted by utilizing a small-scale distillation apparatus. A "center cut" from the distillation is taken for the boiling point. Wide boiling-point ranges also indicate sample impurity. The measurements of specific gravity and refractive index are less common due to their limited utility. Specific gravity is determined with a gravitometer or a pycnometer. Refractive index can be measured with a refractometer. For solid samples, chemical microscopy (Becke test) may be employed.

Chemical methods. The preliminary chemical tests which are applied are elemental analysis procedures. The elements generally associated with carbon, hydrogen, and oxygen are sulfur, nitrogen, and the halogens. The analyst is usually interested in the latter group. These elements are converted to watersoluble ionic compounds via sodium fusion. The resulting products are then detected with wet chemical tests.

A sample is fused with sodium metal by adding it to the sodium and heating to the melting temperature. After a short heating period, the solution is cooled, the excess sodium is dissolved with ethanol, and the fusion products are dissolved in water. Sulfur can be detected by the formation of a lead sulfide precipitate upon the addition of lead acetate. A variety of procedures for the detection of nitrogen are known. The resulting products are usually precipitates or colored solutions. The results of nitrogen tests are sometimes inconclusive due to the difficulty in detecting nitrogen in nitro compounds. The halogens can be detected by the evidence of silver halide precipitates upon the addition of silver nitrate. The identity of the specific halogen can be determined by more selective reactions or by x-ray diffraction.

The solubility of a compound in various liquids provides information concerning the molecular weight and functional groups present in the compound. A systematic solubility study is outlined in **Fig. 3**. On the basis of solubility, assumptions can often be made that reduce the number of tests necessary to classify a compound. Amines are basic compounds, whereas phenols are weakly acidic.

In order to indicate the presence or absence of a functional group, specific classification reactions are tested. The products of these tests are usually precipitates or colored solutions. The reactions are simple, are rapid, and require a small quantity of sample. No single test is conclusive evidence. A judicious choice of reactions can confirm or negate the presence of a functional group. A list of the popular classification tests is presented in **Table 1**.

Once thee identity of a functional group has been established, the structure must be determined. Derivatization of the unknown is one procedure employed for complete identification. Derivatives prepared from the unknown are compared with those of known compounds. The agreement of melting points is the basis of an identification. The majority of derivatives are solids and are produced in high yield. This allows for a facile isolation and purification. Extensive tabulations of derivative melting points have been compiled. A list of derivatization reagents are presented in **Table 2**.

Chemical structure. Instrumental methods are commonly applied for functional group determination and structure identification. Revolutionary developments in instrumentation and data interpretation have greatly simplified the analytical process. Many of the time-consuming procedures have been replaced by instrumental techniques.

Absorption spectroscopy methods are among the most important techniques. Whenever a molecule is exposed to electromagnetic radiation, certain wavelengths cause vibrational, rotational, or electronic effects within the molecule. The radiation required to cause these effects is absorbed. The nature and configuration of the atoms determine which specific wavelengths are absorbed. Radiation in the ultraviolet and visible regions is associated with electronic effects, and infrared radiation is associated with rotational and vibrational effects.

Infrared absorption spectroscopy is perhaps the most valuable instrument for functional group

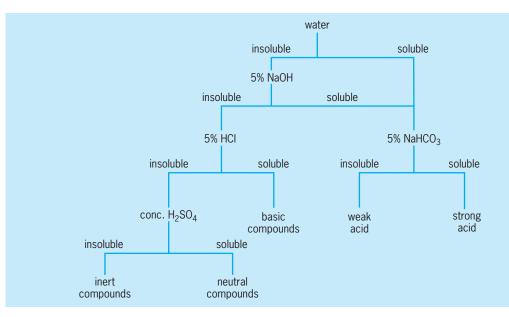


Fig. 3. Diagram of a typical solubility scheme.

determination. A beam of infrared radiation is passed through the sample. The transparent radiation is dispersed by means of a prism or diffraction grating. With the aid of a slit, the intensity of a narrow wavelength region is measured. The entire infrared region is scanned, and an absorbance-versus-wavelength scan is obtained. Based on the wavelengths at which radiation is absorbed, a functional group can be identified. *See* INFRARED SPECTROSCOPY.

The principle of ultraviolet absorption spectroscopy is similar to that of infrared spectroscopy, except that ultraviolet radiation is monitored. Unlike infrared scans, where numerous sharp absorption bands are obtained, ultraviolet scans normally exhibit one or two broad bands of limited utility for qualitative purposes. Visible absorption spectroscopy suffers the same limitations.

Raman spectroscopy is slightly different from the other spectroscopic techniques. A sample is irradi-

ated by a monochromatic source. Depending upon the vibrational and rotational energies of the functional groups in the sample, light is scattered from the sample in a way which is characteristic of the functional groups. Raman spectroscopy is a fine complement to infrared spectroscopy because the two are sensitive for different groups. This technique has not enjoyed widespread application as a qualitative tool due to the high cost of the instrumentation and the difficulty in obtaining good spectra. *See* RAMAN EFFECT.

Instrumental techniques are also valuable for structure identification. The most popular are infrared spectroscopy, nuclear magnetic resonance, and mass spectrometry. Other, less common techniques are electron spin resonance, x-ray diffraction, and nuclear quadrupole spectroscopy.

In addition to the value of infrared for functional group analysis, an infrared spectrum can identify

TABLE 1. Common classification reactions		
Compound	Reagent	Comments
Alcohols	Hydrochloric acid/zinc chloride	Forms alkyl halide
	Ceric nitrate	Color change yellow to red
	Chromic anhydride	Distinguishes 1° and 2° from 3° alcohols
Aldehydes	Benedict's solution	Colored precipitate
-	Tollens' reagent	Forms silver mirror
	Fuchsin-aldehyde	Color change
Aldehydes and	2,4-Dinitrophenyl hydrazine	Hydrazone precipitate
ketones	Hydroxylamine hydrochloride	HCI evolved
Amines	Benzenesulfonyl chloride	Hinsberg method: distinguishes 1°, 2°, and 3° amines
	Acid chloride	Amide formation
Esters	Hydroxamic acid test	Magenta color formation
	Sodium hydroxide	Saponification
Ethers	Hydriodic acid	Zeisel method: light orange color produced
Nitro compounds	Zinc and ammonium chloride/Tollens' reagent	Silver mirror
	Ferrous hydroxide	Red-brown precipitate
Unsaturation	Potassium permanganate	Loss of color
	Bromine in carbon tetrachloride	Loss of color

Compound	Reagent	Derivative
Alcohols	Phenyl isocyanate	Phenyl urethane
	3,5-Dinitrobenzyl chloride	3,5-Dinitrobenzoate
Aldehydes and ketones	2,4-Dinitrophenyl hydrazine	2,4-Dinitrophenyl hydrazon
	Semicarbazide hydrochloride	Semicarbazone
Alkyl and aryl halides	Magnesium/mercuric chloride	Alkyl mercuric halide
Amines 1° and 2°	Benzenesulfonyl chloride	Benzenesulfonamide
	Phenyl isothiocyanate	Phenyl urea
	Acetic anhydride	Acetamide
	Benzoyl chloride	Benzamide
Amines 3°	Methyl iodide	Quaternary ammonium salt
	Benzyl chloride	,
Aromatic hydrocarbons	H_2SO_4/HNO_3	Nitro compound
Esters	Benzylamine/ammonium chloride	N-Benzylamide
	Hydrazine	Acid hydrazide
Ethers, aromatic	Bromine	Bromo compound
,	Picric acid	Picrate
Ethers, aliphatic	3,5-Dinitrobenzyol chloride/zinc chloride	3,5-Dinitrobenzoate
Phenols	Bromine	Bromo compounds
	α -Naphthyl isocyanate	α -Naphthyl urethane
	Acetic anhydride	Acetate

a compound. The spectral region $6.0-14.0 \ \mu m$ is known as the fingerprint region. Comparison to known spectra can identify an unknown compound. Extensive files of infrared spectra are available for this purpose.

Mass spectroscopy is a powerful tool for structure identification. A spectrum is obtained by bombarding a sample with an electron beam. The resulting fragments are accelerated through a magnetic field and are deflected from their original path. The deflection is a function of both mass and charge. The separated fragments are then detected. Based on the fragmentation pattern and the mass-to-charge ratio of the fragments, structural information is obtained. A fingerprint identification method can also be employed. *See* MASS SPECTROMETRY.

Nuclear magnetic resonance is another powerful structural tool. As in other spectroscopic methods, a sample is irradiated and the absorbed radiation is monitored. Similar to electrons, certain nuclei (1H, ¹³C, and so on) can be considered to spin. The rotation of the charged particle creates a magnetic moment about the axis of spin. When a sample is placed in a magnetic field, the nuclei are aligned in a given direction. In order to reverse the direction, energy is required. The changing of spin states results in absorption of the applied radio-frequency radiation. Either the radio frequency or the magnetic field can be held constant while the other is changed and the radiation absorption is monitored. In practice it is more convenient to hold the radio frequency constant and vary the magnetic-field strength. The energy required to change spin states is also dependent upon neighboring groups in the molecule. Based on the number and intensity of absorptions and the field strength required, structural information can be obtained. See NUCLEAR MAGNETIC RESONANCE (NMR).

Physical structure. In addition to chemical structure, the physical structure of a sample can be important. Certain materials behave differently depending on characteristics such as crystallinity, crosslinking, and their association as dimers, trimers, and so on. Thermal analysis is a useful procedure for examining these structural characteristics. Among the most popular methods are thermogravimetry, differential thermal analysis, differential scanning calorimetry, and thermal mechanical analysis. *See* CALORIMETRY.

Although the implementation of instrumental methods have greatly simplified qualitative analytical procedures, no individual instrument is capable of complete identification for all samples. A complement of instrumental and wet chemical techniques is generally required for adequate proof of identification. *See* SPECTROSCOPY. Sidney Siggia; Kenneth Longmoore

Bibliography. F. Feigl, Spot Tests in Inorganic Analysis, 6th ed., 1972; W. Kemp, Qualitative Organic Analysis: Spectrochemical Techniques, 2d ed., 1988; J. Lagowski and H. Sorum, Introduction to Semimicro Qualitative Analysis, 8th ed., 2004; R. L. Shriner et al., The Systematic Identification of Organic Compounds, 8th ed., 2003; S. Siggia and J. G. Hanna, Quantitative Organic Analysis via Functional Groups, 4th ed., 1979, reprint 1988; G. I. Svehla, Vogel's Qualitative Inorganic Analysis, 7th ed., 1996.

Quality control

The operational techniques and the activities that sustain the quality of a product or service in order to satisfy given requirements. Quality control is a major component of total quality management and is applicable to all phases of the product life cycle: design, development, manufacturing, delivery and installation, and operation and maintenance.

The quality-control cycle consists of four steps: quality planning, data collection, data analysis, and implementation. Quality planning consists of defining measurable quality objectives. Quality objectives are specific to the product or service and to the phase in their life cycle, and they should reflect the customer's requirements.

The collection of data about product characteristics that are relevant to the quality objectives is a key element of quality control. These data include quantitative measurements (measurement by variables), as well as determination of compliance with given standards, specifications, and required product features (measurement by attributes). Measurements may be objective, that is, of physical characteristics, which are often used in the control of the quality of services. Since quality control was originally developed for mass manufacturing, which relied on division of labor, measurements were often done by a separate department. However, in the culture of Total Quality Management, inspection is often done by the same individual or team producing the item.

The data are analyzed in order to identify situations that may have an adverse effect on quality and may require corrective or preventive action.

The implementation of those actions as indicated by the analysis of the data is undertaken, including modifications of the product design or the production process, to achieve continuous and sustainable improvement in the product and in customer satisfaction.

The methods and techniques for data analysis in quality control are generic and can be applied to a variety of situations. The techniques are divided into three main categories: diagnostic techniques; process control, which includes process capability assessment and control charts; and acceptance sampling.

Diagnostic techniques. Diagnostic techniques serve to identify and pinpoint problems or potential problems that affect the quality of processes and products, and include the use of flowcharts, cause-and-effect diagrams, histograms, Pareto diagrams, location diagrams, scatter plots, and boxplots.

Flowcharts. A flowchart is a graphical description of a process in terms of its constituent steps and the sequence among them (Fig. 1). Various symbols are used to represent the steps and the types of relationships among them. The most common types of steps are processing, waiting, transportation, inspection, and storage. Flowcharts are useful in evaluating the way that a process is actually executed relative to its original design. Preparing a flowchart for the process is often a good first step toward improving it. Often, when the process is clearly defined and documented, bottlenecks and redundant steps become apparent, and opportunities for improvement are obvious. It is also helpful to classify the various steps of the process into three groups: (1) Value-adding steps create or modify the product features that are required to meet customer expectations, and typically include manufacturing, assembly, and delivery activities. These steps are the essence of the process and should be carried out as efficiently and effectively as possible. (2) Process-sustaining steps are required to manage the process, such as record-

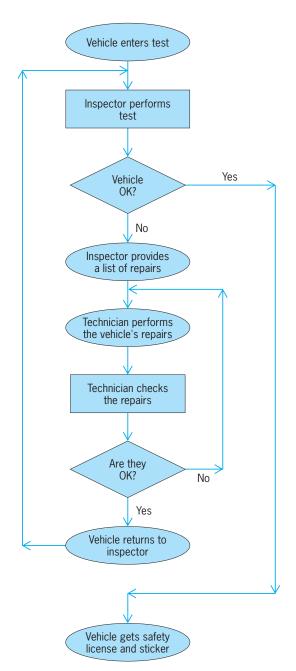


Fig. 1. Flowchart used to monitor a vehicle safety test.

keeping, production control, and invoicing, and they should be minimized to the extent possible. (3) Nonvalue-adding steps, such as waiting, in-process storage, rework, and certain types of inspection, should be eliminated from the process.

Cause-and-effect diagrams. These are also called fishbone diagrams because of their shape, or Ishikawa diagrams in honor of the Japanese engineer who made their use popular. The cause-and-effect diagram is used to describe the relationship among the factors that may contribute to a quality problem. Construction of the cause-and-effect diagram begins by identifying the variable or phenomenon of interest. This might be the yield of a process, the response rate from a customer survey, the performance of a particular assembly, the variation of a process, or any other

variable that bears on the quality of the product or service. An arrow is drawn from left to right, with the tip pointing to the variable. Then, main contributory factors are drawn as arrows leading into the main arrow, and secondary and lower-order contributory factors are shown as feeding into their respective arrows (**Fig. 2**).

Histogram. A histogram is a frequency distribution plot. The horizontal axis represents different values or categories, and the vertical axis represents the number of occurrences observed for each value or category (**Fig. 3**). Histograms are helpful for organizing the data collected from measurements taken on a product or process. The histogram can then be analyzed for any unusual pattern, such as skewed distribution, multimodal distribution, and long tail. These patterns are analyzed for indications of problems and are interpreted based on the specific quality objectives of the product.

Pareto diagram. This is a histogram of defects or quality problems classified by type and sorted in descending order of frequency (**Fig. 4**). The Pareto diagram, named after the Italian economist Vilfredo Pareto, serves to focus attention on the main sources of problems. A variation of this type of technique is called the Pareto pyramid, where two Pareto diagrams, representing the situation before and after an intervention, are shown side by side with the axes transposed. *See* PARETO'S LAW.

Location diagram. This type of chart shows the distribution of defects by their location on the process or the product. For example, it can show the layout of a printed circuit board, with the number of soldering defects shown next to each soldering point (**Fig. 5**).

Scatter plot. A scatter plot is a two-dimensional plot showing the relationship between two variables, each plotted along one axis (**Fig.** 6). The pattern of points suggests the type of correlation between the two variables (positive, negative, nonlinear) and its strength. Correlation between two variables can be exploited to improve the performance of the dependent variable. This is achieved by manipulating the independent variable in a direction such that it will cause the dependent variable to change in the desired direction.

Box plot. This plot summarizes the results of successive sample observations from the process taken over time. The horizontal axis represents the sample number (or time). On the vertical axis, each box summarizes a sample with five numbers: the upper and lower quartiles, which form the ends of the box; the median, which is the line dividing the box; and the minimum and maximum values, which appear at the end of the whiskers. Box plots analyze the process behavior over time and the relationship between its mean and its variability.

Process control. Process-control methods are applicable to systems that produce a stream of product units, either goods or services. They serve to control the processes that affect those product characteristics that are relevant to quality as defined in the quality objectives. For example, in a system that produces

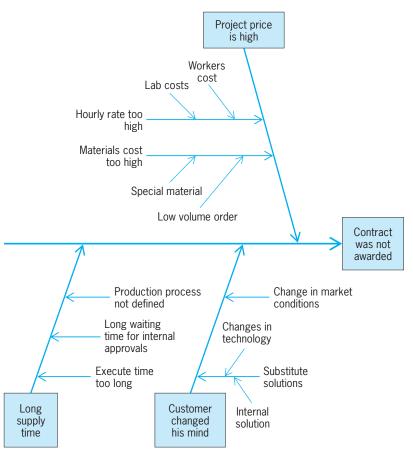


Fig. 2. Cause-and-effect diagram.

metal parts, some of the processes that might need to be controlled are cutting, machining, deburring, bending, and coating. The relevant product characteristics are typically spelled out in the specifications in terms of physical dimensions, position of features,

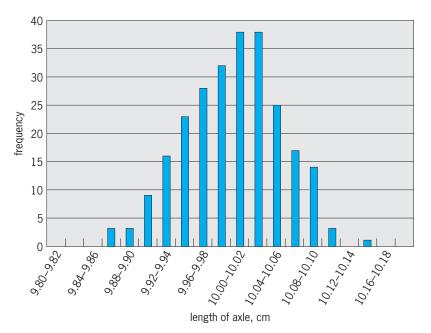


Fig. 3. Histogram showing the length distribution of a manufactured axle.

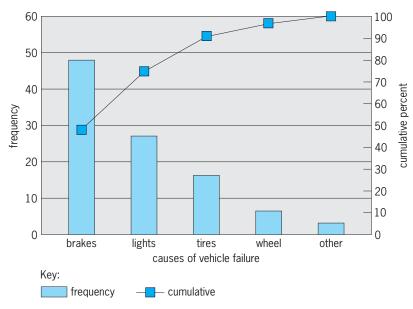


Fig. 4. Pareto diagram showing causes and failure rates in a vehicle test.

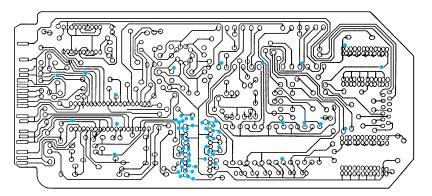


Fig. 5. Location diagram showing the places that are not functioning as expected (white circles) in a jet engine.

surface smoothness, material hardness, paint thickness, and so on. In a system that produces a service, such as a telephone help line, the relevant processes could be answering the call, identifying the problem, and solving the problem. The characteristics that are relevant to quality as perceived by the customer might include response time, number of referrals, frequency of repeat calls for the same problem, and elapsed time to closure.

Process control focuses on keeping the process operating at a level that can meet quality objectives, while accounting for random variations over which there is no control. There are two main aspects to process control: control charts and capability analysis. Control charts are designed to ascertain the statistical stability of the process and to detect changes in its level or variability that are due to assignable causes and can be corrected. Capability analysis considers the ability of the process to meet quality objectives as implied by the product specifications.

Process-control techniques were originally developed for manufactured goods, but they can be applied to a variety of situations as long as the statistical distribution of the characteristics of interest can be approximated by the normal distribution. In other cases, the principles still apply, but the formula may need to be modified to reflect the specific mathematical expression of the probability distribution functions. *See* PROCESS CONTROL.

Control charts. Every process has some inherent variability due to random factors over which there is no control and which cannot economically be eliminated from the process. This inherent variability is the aggregate result of many individual causes, each having a small impact. Control charts are designed to detect excessive variability due to specific assignable causes that can be corrected; they serve to determine whether a process is or is not in a state of statistical control. Being in statistical control means that the extent of variation of the output of the process does not exceed that which is expected based on the natural statistical variability of the process.

Samples are drawn from the process output at regular intervals, and a statistic is calculated from the observations in the sample and is charted over time. A nonrandom pattern of points, such as points outside the control limits, indicates lack of statistical control. If a nonrandom pattern is detected, the process should be stopped and an investigation should be done to determine the causes for the change. Once the causes of the change have been ascertained and any required corrective action has been taken, the process can be resumed. *See* CONTROL CHART.

Process capability analysis. The capability of a process is defined as the range of values that the process generates while operating under statistical control. Typically, the process capability is \pm three standard deviations from the process mean, for a total of six standard deviations. The standard deviation of the process should be estimated from historical data collected while the process was operating under normal, stable circumstances. Control charts can be used to establish the statistical stability of the process and the absence of assignable causes.

The process capability is compared to the tolerance spread, that is, the range of values that meet the specifications for the product and are deemed acceptable by the customer. This range is determined

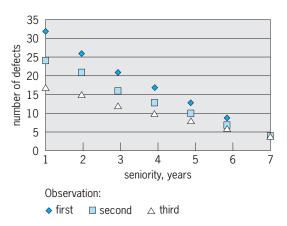


Fig. 6. Scatter plot showing a negative correlation between number of defects and worker seniority.

by the upper specification limit (USL) and the lower specification limit (LSL), which together define the range of values that will meet the requirements from the product. The process capability index (C_p) is defined as the ratio between the tolerance spread (USL-LSL) and the process capability [six standard deviations (6σ)] as in the equation below. If the capability

$$C_p = \frac{\text{USL-LSL}}{6\sigma}$$

index is smaller than 1, a certain fraction of the process output will fall outside the tolerance spread and will not be acceptable, entailing lower quality and higher costs through rework and scrap. When the capability index is equal to 1, if the process is perfectly centered in the tolerance interval, and assuming the normal distribution, virtually all the process output will fall within specifications and the fraction of units that fail to conform to the specifications will be very small, about 0.27%. However, if the process mean shifts to either side, the fraction nonconforming will increase rapidly. For example, a shift of $1/_2$ standard deviation results in 0.62% nonconforming, and a shift of 1 standard deviation results in 2.28% nonconforming. In general, the capability index should be greater than 1, with some recommended values being at least 1.33 (tolerance spread of $\pm 4\sigma$) and possibly 2.0 (tolerance spread of $\pm 6\sigma$).

There are two avenues for improving the capability index: reducing the natural variability of the process, and increasing the tolerance spread. Reduction of the process variability can be achieved by operating the process under more consistent conditions (such as uniform inputs, better-trained operators, or more frequent adjustments). A focus on continuous process improvement can help in that respect by eliminating sources of variation and promoting more consistent execution. Some of the diagnostic techniques, such as the Pareto diagram and the causeand-effect diagram, can help identify the aspects in the process that can yield effective variability reductions.

The second avenue, widening the tolerance spread, requires the involvement of the customers to validate the specifications. Often it is possible to modify the design of the product and to loosen its specifications without adversely affecting its users.

Acceptance sampling. Acceptance sampling refers to the procedures used to decide whether or not to accept product lots or batches based on the results of the inspection of samples drawn from the lots. Acceptance sampling techniques were originally developed for use by customers of manufactured products while inspecting lots delivered by their suppliers. These techniques are particularly well suited to situations where a decision on the quality level of product lots and their subsequent disposition needs to be made, but it is not economic or feasible to inspect the entire production output.

Sampling by attributes. The most commonly used acceptance sampling plans are based on inspection by attributes. In inspection by attributes, each product unit is classified as either possessing the required attributes and therefore conforming to specifications, or lacking at least one of the required attributes and therefore nonconforming or defective. The recipient of the product (that is, the consumer) specifies what constitutes the acceptable quality level (AQL) in terms of the fraction defective in a lot. Lots that do not exceed the AQL are accepted in their entirety. Those that do exceed it are considered not acceptable and are either returned to the producer or screened to remove all defective units and only then accepted.

The sampling plan is defined in terms of the sample size (n) and the acceptance number (c). Randomly selected samples of n units are drawn from each lot and inspected. Only if the number of defective units found in the sample does not exceed c is the lot accepted. Because of the random variability inherent in the sampling process, two types of inference errors are possible: rejection of a lot of acceptable quality, and acceptance of a lot of unacceptable quality. The second type of error requires the customer to define what fraction defective constitutes unacceptable quality. This fraction is called the limiting quality level (LQL); and if expressed as a percentage, it may also be referred to as the lot tolerance percent defective (LTPD). The rejection of a lot with fraction defective not exceeding the AQL is called a type 1 error, and its probability of occurrence is called the producer risk (α). The acceptance of a lot with fraction defective greater than the LQL is called a type 2 error, and its associated probability is called the consumer risk (β).

The entire range of performance of a given sampling plan for various incoming quality levels is summarized by the operating characteristic curve (**Fig.** 7). The operating characteristic curve shows on the vertical axis the probability of accepting a lot as a function of its fraction defective. The probabilities are calculated with the appropriate probability distribution function.

Rectifying inspection. Rectifying inspection is the procedure of screening lots that were not accepted by the sampling inspection plan. Screening means 100% inspection of the lot and removal of all the nonconforming units, followed by acceptance of the remaining units in the lot. If it is necessary to maintain a

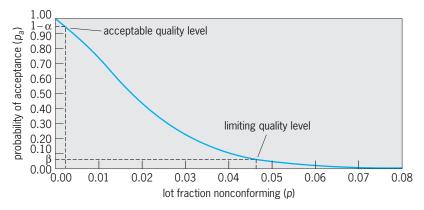


Fig. 7. Sampling by attributes. Operating characteristic curve for the sampling plan: sample size (n) = 100 and acceptance number (c) = 1.

constant lot size, the nonconforming units are replaced by conforming ones.

Under rectifying inspection, the average quality of the accepted lots varies, depending on the incoming fraction nonconforming. If the incoming fraction nonconforming is less than the AQL, most lots will be accepted according to the sampling plan, and the average fraction nonconforming will be low. If the incoming fraction nonconforming is significantly higher than the AQL, most lots will be rejected and screened, which, by removing all nonconforming units, lowers the average outgoing fraction nonconforming. Between these two extremes, the average outgoing fraction nonconforming will be higher. This fraction, which is called the average outgoing quality (AOQ), is a function of the incoming fraction nonconforming and of the given sampling plan.

Variables sampling plans. Variables sampling plans use the actual measurements taken on the units in the sample rather than the number of units that conform to specifications. Variables sampling plans are more complex to design and to administer, and the measurements are typically more costly than inspection by attributes. They also require the assumption that the observations are drawn from a normal distribution. On the positive side, variables sampling plans require a much smaller sample size than attributes sampling plans that provide the same protection in terms of AQL, producer risk, LQL, and consumer risk; and the actual measurements provide more information regarding how close the lot is to meeting specifications. However, if the lots come from a process that runs continuously, control charts are a preferable method for controlling quality.

Double and multiple sampling plans. Double sampling plans consist of two samples and two sets of acceptance and rejection numbers. The sample sizes need not be equal, but in many instances either they are equal or the second sample is twice as large as the first one. A double sampling plan is carried out as follows: (1) A sample of n_1 units is inspected. (2) If the number of nonconforming units does not exceed the first acceptance number (c_1) , the lot is accepted. (3) If the number of nonconforming units exceeds the first rejection number (r_1) , the lot is rejected. (4) Otherwise, a second sample of n_2 units is inspected. (5) If the total number of nonconforming units does not exceed the second acceptance number (c_2) , the lot is accepted. (6) Otherwise, the lot is rejected. For the same protection in terms of AQL, producer risk, LQL, and consumer risk, double sampling plans result in a smaller average sampling number than single sampling plans. The reduction in inspection effort is even greater if the plan incorporates curtailing. Curtailing is the practice of stopping the inspection as soon as the number of nonconforming units found reaches the rejection number for the second sample (r_2) , which is equal to $c_2 + 1$.

Multiple sampling plans are an extension of the same concept, involving a series of samples of equal size and their associated acceptance and rejection numbers. To prevent multiple plans from continuing indefinitely, they are designed such that a decision is made by the time the total number of units inspected is twice the sample size of a single sampling plan that offers the same protection to consumer and producer. Multiple sampling plans result in an even lower average sampling number, but they require more decisions and are more complex to design and administer.

Sequential sampling plans. In sequential sampling, units are inspected one at a time, and a decision to accept or reject the lot is made based on the cumulative number of nonconforming units found. The decision numbers are calculated by using the sequential probability ratio (SPR). If the sequential probability ratio is greater than or equal to $(1 - \beta)/\alpha$, the lot is rejected. If sequential probability ratio is smaller than or equal to $\beta/(1 - \alpha)$, the lot is accepted. If neither is true, an additional unit is inspected, and the sequential probability ratio is recalculated. Here again, provision should be made to truncate the test when the total number inspected reaches a certain value, typically a multiple of the sample size of an equivalent single sampling plan or of the average sampling number of an equivalent double or multiple sampling plan.

Chain sampling plans. Chain sampling is a procedure in which the decision to accept or reject a lot depends in part on the results of inspecting the preceding lots. When these procedures were originally developed, it was suggested that two acceptance numbers, 0 and 1, be used. If a number of previous lots (i) had no nonconforming units, the current lot will be accepted if the sample contains no more than one nonconforming unit. Otherwise, the lot will be accepted only if the sample has no nonconforming units. The design parameters for this type of plan are the sample size n and the number of immediately preceding lots. This type of plan increases the probability of acceptance when the fraction nonconforming is in general low, and reduces inspection costs by taking into account the immediate history of the product.

Continuous sampling plans. Continuous sampling plans are designed for processes that generate a continuous stream of product units that does not lend itself to grouping into lots or even meaningful samples. Continuous sampling plans entail periods of screening and periods of sampling. The procedure begins with 100% inspection and, after *i* consecutive conforming units are found, switches to inspection of a randomly selected fraction (*f*) of the units. When a nonconforming unit is found, it reverts to 100% inspection.

Standards. Following the recommendation from the American Society for Quality Control, the American National Standards Institute adopted a standard called Sampling Procedures and Tables for Inspection by Attributes. This system of tables is based on a standard used by military procurement, and it covers single, double, and multiple sampling for defectives (nonconforming units) and for defects (nonconformities).

The tables are indexed by AQL, lot size, and inspection level. The inspection level determines the relationship between the sample size and the lot size, which affects the probability of rejecting lots with quality worse than the AQL. There are three general inspection levels: normal, tightened, and reduced. Once the AQL has been agreed upon by the consumer and the supplier, and the sampling plans have been selected based on the number of units in the lots submitted for acceptance, inspection begins at the normal level. If two or more lots out of any five consecutive lots are rejected, inspection switches to the tightened level, which uses a larger sample size and has a higher probability of rejection. If five consecutive lots are accepted while on tightened inspection, the system allows return to normal inspection. If, while under normal inspection, ten consecutive lots are accepted, inspection can switch to the reduced inspection level, which uses a smaller sample size. If a lot is rejected while on reduced inspection, the system prescribes return to the normal inspection level.

This system of tables is intended for use to protect the consumer in situations where there is a continuing series of lots. A similar system of tables is available for acceptance sampling by variables. *See* CON-TROL CHART; INSPECTION AND TESTING; STATISTICS. Tzvi Baz

Bibliography. L. S. Aft, Fundamentals of Industrial Quality Control, CRC Press, Boca Raton, FL, 1998; J. Bentley, Introduction to Reliability and Quality Engineering, Addison-Wesley, Harlow, 1999; F. W. Breyfogle, Statistical Methods for Testing, Development and Manufacturing, 1992; J. W. Cortada and J. A. Woods, The McGraw-Hill Encyclopedia of Quality Terms and Concepts, McGraw-Hill, New York, 1995; J. R. Evans and W. M. Lindsay, The Management and Control of Quality, South-Western College Publishers, Cincinnati, 1999; V. E. Sower, M. J. Savoie, and S. Renick, An Introduction to Quality Management and Engineering, Prentice Hall, Upper Saddle River, NJ, 1999; D. Straker, A Toolbook for Quality Improvement and Problem Solving, Prentice Hall, London, 1995.

Quantitative chemical analysis

The branch of analytical chemistry that deals with the determination of the relative or absolute amounts of the various constituents of materials.

Methods

Methods may be broadly grouped as chemical or physical, classical or instrumental, general or selective, and simple or compound. Any given method is classifiable within each of these groups. All techniques require calibration with respect to some standard of known composition. Caution is necessary to prevent substances other than those being determined from falsely giving signals attributable to the sought-for substance, which is called the analyte. *See* CALIBRATION.

Scale of determination. Selection of a technique is based in part on the size of sample available,

the quantity of analyte expected to be in the sample, the precision and accuracy of the technique, and the speed of analysis required. The size of the sample dictates whether a determination is macroscopic (>0.1 g), semimicroscopic (0.01-0.1 g), microscopic (0.001-0.01 g), or ultramicroscopic (<0.001 g). Expected analyte levels may be major constituent (>10%), minor constituent (0.1-10%), or trace constituent (<0.1%).

Chemical and physical methods. Chemical methods employ reactions between the analyte and some other species, followed by observation of some quantitative result. Physical methods employ the interaction of the species of interest with electromagnetic radiation, sound, heat, or some other entity with which chemical reaction does not occur. A chemical method for determining the acidity of a solution would be the use of a series of indicator compounds which change color at specific pH or acidity values. A physical method for the same determination would involve the use of a glass indicator electrode to measure the adsorption of hydrogen ion to the glass by sensing the electrical potential of the glass. See ACID-BASE INDICATOR; ION-SELECTIVE MEM-BRANES AND ELECTRODES; PH.

Classical and instrumental methods. In a classical method, human senses are used to detect a signal related to the quantity of the analyte. In an instrumental method, mechanical (usually electronic) means are used to detect the signal. Iron may be determined classically by titration of Fe(II) with Ce(IV) by using Ferroin as a visual indicator of the equivalence point. Iron may be determined instrumentally by producing a cloud of iron atoms in an electrical plasma and observing the light emitted by the iron at specific wavelengths. The light intensity at these wavelengths is proportional to the amount of iron introduced into the plasma.

General and selective methods. General methods respond to a wide range of elements or compounds, whereas selective methods respond only to a limited range of species. Measurement of the thermal conductivity of a gas stream can indicate the presence of any analyte species in the stream, so long as the thermal conductivity of the analyte is different from that of the pure stream. For carrier streams of inert gases, this means that most molecular analytes will give rise to signals. Measurement of the ability of the gas to attach electrons (an electron capture detector) will indicate only the presence of highly electronegative species such as the halogens.

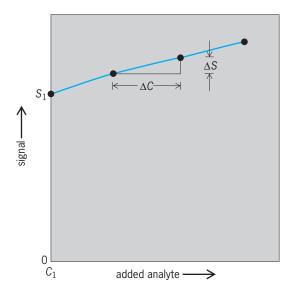
Simple and compound techniques. A simple technique is one in which a single measurement operation is performed. A compound technique requires several operations to complete the analysis. Measurement of pH is a simple technique. Chromatographic separation of a mixture followed by thermal conductivity detection of the presence of a species, followed in turn by mass spectrometric detection of the molecular weight and fragmentation of the species, is an example of a compound technique.

Quantitation. Direct measurement of a signal related to concentration or activity of the chemical

species of interest is the most intuitive approach. Generally, a linear relationship between signal (or its logarithm) and concentration (or its logarithm) is sought. The relationship between signal and concentration is called a working curve. The slope of the line describing the relationship is known as sensitivity. The smallest quantity which is measurably different from the absence of analyte is the detection limit. Although a linear working curve is the simplest form to use, nonlinear curves may still be employed (either graphically or with the use of computers).

Titration. This is the process by which an unknown quantity of analyte (generally in solution) is determined by adding to it a standard reagent with which it reacts in a definite and known proportion. A classical or instrumental means is provided to indicate when the standard reagent has consumed exactly the amount of analyte initially present. Each determination is performed with a reagent whose concentration is directly traceable to a primary standard, so that accuracy is frequently superior to that of other methods. *See* STOICHIOMETRY; TITRATION.

Standard additions. Sensitivity of a method may be related not only to the method and analyte of interest, but also to the presence of other species in the sample. The materials other than the analyte constitute a sample matrix. For example, seawater is a matrix quite different from distilled water because of the large amount of dissolved electrolyte present. If a signal can be derived from the analyte which is proportional to the amount of the analyte, but the sensitivity of the signal to concentration varies as a function of the sample matrix, the method of standard additions may be useful (see illus.). The signal from the analyte is measured, after which small, known quantities of analyte are successively added to the sample, and the signal remeasured. The sensitivity of the method can thus be obtained, and the initial, preaddition signal interpreted to give the amount of analyte (C_1) initially present by using the relationship shown in the



Working curve for the method of standard additions.

equation below, where S_1 is the initial signal; C_1 the

$$C_1 = \frac{\Delta C}{\Delta S} \cdot S_1$$

initial quantity of analyte; ΔC the standard addition of analyte; and ΔS the change in the signal caused by the addition. The sensitivity is $\Delta C/\Delta S$.

This approach is applicable only in the absence of reagent blanks, that is, signals caused by the presence of analyte in the reagent used for the determination.

Internal standard. In many methods, aliquots of samples are introduced into the measurement instrument. The signal from the analyte may vary with sample uptake rate or volume. To compensate for such effects, an internal standard, or species other than the analyte, may be added to the analyte in a known concentration prior to determination. The signal due to the internal standard is measured simultaneously with the analyte signal. Variation of the signal of the internal standard is interpreted to indicate the variation in sample uptake, which should be the same for both analyte and internal standard. The ratio of analyte signal to internal standard signal is independent of sample uptake. Thus the ratio of analyte to internal standard signal is used to establish the working curve, rather than using analyte signal alone.

Null comparison. If the relationship between signal and analyte concentration is nonlinear, quantitation may require the use of null comparison. A signal is observed from the analyte and from a standard whose concentration can be adjusted in a known way. When the signal from the adjustable standard equals the signal from the analyte, the two have identical concentrations. This condition is called a null, as there is no detectable difference between the sample and reference signals.

Interferences. It is good laboratory practice to check every quantitative measurement for the influence of species other than the one being sought. For example, a glass electrode designed to sense hydrogen ion will also respond to high concentrations of sodium ion. The degree to which a given sensor responds to one species in preference to another is called the selectivity coefficient. If a general detector is desired, this coefficient ideally should be 1.0. If a species specific detector is desired, the coefficient should be infinite. A signal of unknown or general origin which appears to underlie the analyte signal is known as background. This quantity, together with that for the reagent blank, may be subtracted from the raw signal and thus be compensated. However, variations in their level may prevent reliable compensation, particularly when small quantities of analyte are to be determined. Alexander Scheeline

Analysis of Organic Materials

For determination of elements, functional groups, or molecules in organic materials, the type of analysis used is determined by the information required. If the total nitrogen content is desired, elemental analysis is used; if only amino nitrogen is desired, then only the amino group is determined. Organic analyses are made on a wide range of materials, from pure compounds to mixtures such as blood and fertilizer. The identification of structures in conjunction with purity criteria, analyses of isotopic mixtures of organic molecules, and other examples of structure-related problems are increasingly important analytically. Proton nuclear magnetic resonance and other classes of spectrometric measurement are prominent for such purposes.

Determination of metals. This may be done in two ways. In the first method the sample is moistened with sulfuric acid and heated in oxygen to obtain metal oxides or sulfates, which are weighed. Some metals such as gold and silver are weighed as metals. The second method is based on destruction of the organic portion of the sample by heating with nitric and sulfuric acids, followed by determination of the metals by regular procedures. This method is required for metals which are easily volatilized, such as arsenic.

Carbon and hydrogen. These elements are usually determined simultaneously by burning a sample in a stream of oxygen to form carbon dioxide and water. This is done at $600^{\circ}C (1100^{\circ}F)$ in the presence of platinum as a catalyst. If nitrogen is present in the sample, the gas stream is passed through lead peroxide to remove the oxides of nitrogen. Silver wool adsorbs halogens and oxides of sulfur. The water formed is absorbed on a drying agent, such as magnesium perchlorate and from the increase in weight the amount of hydrogen in the sample can be calculated. Carbon dioxide is absorbed by an alkaline solid such as sodium hydroxide on asbestos fibers, and from the increase in weight the amount of carbon can be calculated.

Carbon alone is determined by wet combustion. The sample is heated in a mixture of sulfuric acid and potassium or silver dichromate. The carbon dioxide formed is determined from the amount of sodium hydroxide with which it combined, or by measuring directly the volume of the carbon dioxide. Reaction chromatography provides a newer basis for rapid carbon and hydrogen analyses on small samples of most classes or organic compounds. The weighed sample is reacted or combusted, and the reaction products are passed through suitable chromatographic columns. The products are measured from the areas beneath their chromatographic peaks.

Oxygen. Direct determination of oxygen in organic materials is accomplished by heating the sample in a stream of nitrogen to form water, oxides of carbon, and hydrocarbons. On passage of this mixture through graphite at $1150^{\circ}C$ ($2100^{\circ}F$), all the oxygen is converted to carbon monoxide. The carbon monoxide is oxidized with iodine pentoxide to form carbon dioxide and iodine. Either the iodine or the carbon dioxide may then be measured. This procedure is difficult to perform, so that oxygen in pure compounds is usually obtained by difference rather than by direct measurement.

Nitrogen. Two methods are used to determine nitrogen. In the Dumas method the sample is mixed with copper oxide and heated in a stream of pure carbon dioxide. The elemental nitrogen formed is collected over 50% potassium hydroxide solution, and its volume is measured. This procedure determines total nitrogen in most samples. In the Kjeldahl method the sample is heated in concentrated sulfuric acid containing a catalyst; this procedure converts the nitrogen to ammonia. On the addition of sodium hydroxide followed by boiling, ammonia is distilled into dilute boric acid solution, which is then titrated with acid. The choice of the catalyst is critical because different nitrogen compounds require a variety of catalysts. Esters of nitric and nitrous acids cannot be analyzed by this procedure.

Other elements. Sulfur, halogens, and phosphorus are determined by conversion to sulfuric acid, halogen acids, and phosphoric acids, which are measured by standard inorganic procedures. In the Carius method the sample is heated with fuming nitric acid in a sealed tube. In the combustion method the sample is burned in oxygen, oxides of sulfur are absorbed in neutral hydrogen peroxide, and halogens are absorbed in sodium bisulfite solution. In the bomb method for sulfur the sample is ignited with sodium peroxide and sugar.

Functional groups. Groups such as carboxyl, hydroxyl, nitro, and amide are determined by chemical reactions which are characteristic for each group. These reactions include neutralization, oxidationreduction, precipitation, condensation, and gas evolution. These methods are based on consumption of a reagent measured by direct titration, or in some cases by determination of excess reagent. For others, a product formed in the reaction is measured. Since most organic compounds are insoluble in water, organic solvents are used. These solvents enhance the acidity or basicity of some groups and also permit some reactions which are not possible in water. Many common organic functionalities can be used quantitatively to undergo reactions in which water is stoichiometrically consumed or evolved. The titrimetric measurement of such water with Karl Fischer reagent provides a useful basis for determining such functional groups.

Other methods. Many organic compounds absorb energy in the ultraviolet or infrared regions. Since this absorption is characteristic for each type of molecule, a direct analysis based on energy absorption is frequently possible. The variation in mass among hydrocarbons is the basis for the massspectrometric analysis of petroleum samples. These and other instrumental methods are based on the properties of molecules.

Physical methods of separation are commonly required to obtain molecules free from interferences. Normal distillation, vapor distillation, azeotropic distillation, ion exchange, chromatography, extraction of solids and liquids, and diffusion have a place in quantitative analysis.

In an analytical perspective, the concurrent separation, characterization, and quantitative measurement of the component molecules from very small samples results from such techniques as mass spectrometry or vapor chromatography. With preconcentration or larger samples, similar techniques detect and measure trace components at ppm (parts per million) to ppb (parts per billion) levels. For complex samples, the preseparated species from a chromatographic column may be introduced at the inlet of a mass spectrometer.

Many enzyme-substrate combinations can be combined with selective-ion electrode responses for potentiometric measurement. For example, the ureasecatalyzed hydrolysis of urea produces ammonium ion, which is sensed by an ammonium ion-selective electrode. If an enzyme-gel layer surrounds the electrode, its potential responds to external concentrations of urea. With other techniques, often spectrophotometric, the progress of enzymatic or other organic reactions may be followed with time, and provide a kinetic-based method of analysis. *See* ANA-LYTICAL CHEMISTRY.

For further information on common quantitative techniques *see* ACTIVATION ANALYSIS; CALORIME-TRY; CHEMICAL MICROSCOPY; CHROMATOGRA-PHY; COMBUSTION; ELECTRON SPECTROSCOPY; ELECTROPHORESIS; GAS CHROMATOGRAPHY; GEL PERMEATION CHROMATOGRAPHY; IMMUNOASSAY; IMMUNOELECTROPHORESIS; ION EXCHANGE; ISOTOPE DILUTION TECHNIQUES; KINETIC METHODS OF ANAL-YSIS; MASS SPECTROMETRY; NUCLEAR MAGNETIC RESONANCE (NMR); POLARIMETRIC ANALYSIS; SPEC-TROSCOPY; X-RAY SPECTROMETRY. Charles L. Rulfs

Bibliography. American Chemical Society, Analytical Chemistry, monthly; T. R. Crompton, Comprebensive Organometallic Analysis, 1987; J. S. Fritz and G. H. Schenk, Jr., Quantitative Analytical Chemistry, 5th ed., 1987; D. C. Harris, Quantitative Chemical Analysis, 7th ed., 2006; R. W. Ramette, Chemical Equilibrium and Analysis, 1981; S. Siggia and J. B. Hanna, Quantitative Organic Analysis via Functional Groups, 4th ed., 1979, reprint 1988; R. M. Silverstein et al., Spectrometric Identification of Organic Compounds, 7th ed., 2005; H. H. Willard et al., Instrumental Methods of Analysis, 7th ed., 1988.

Quantized electronic structure (QUEST)

A material that confines electrons in such a small space that their wavelike behavior becomes important and their properties are strongly modified by quantum-mechanical effects. Such structures occur in nature, as in the case of atoms, but can be synthesized artificially with great flexibility of design and applications. They have been fabricated most frequently with layered semiconductor materials. Generally, the confinement regions for electrons in these structures are 1-100 nanometers in size. The allowable energy levels, motion, and optical properties of the electrons are strongly affected by the quantummechanical effects. The structures are referred to as quantum wells, wires, and dots, depending on whether electrons are confined with respect to motion in one, two, or three dimensions (Fig. 1). Multiple closely spaced wells between which electrons can move by quantum-mechanical tunneling through

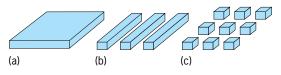


Fig. 1. Geometric configurations of (a) quantum well, (b) quantum wires, and (c) quantum dots.

intervening thin barrier-material layers are referred to as superlattices. *See* QUANTUM MECHANICS.

The most frequently used fabrication technique for quantized electronic structures is epitaxial growth of thin single-crystal semiconductor layers by molecular-beam epitaxy or by chemical vapor growth techniques. These artificially synthesized quantum structures find major application in highperformance transistors such as the microwave high-electron-mobility transistor (HEMT), and in high-performance solid-state lasers such as the semiconductor quantum-well laser. They also have important scientific applications for the study of fundamental two-dimensional, one-dimensional, and zero-dimensional physics problems in which particles are confined so that they have free motion in only two, one, or zero directions. Chemically formed nanocrystals, carbon nanotubes, zeolite cage compounds, and carbon buckyball C60 molecules are also important quantized electronic structures.

Properties. Electrons that are confined in quantum structures can have only certain energies instead of the continuous distributions of energies that are available to particles in free space. For particles strongly confined to a single layer of thickness L and not moving parallel to the layer, the allowed energy levels relative to the bottom of the well are $E_n = b^2/(8m^*L^2)$, $4b^2/(8m^*L^2)$, $9b^2/(8m^*L^2)$, ..., $n^{2}h^{2}/(8m^{*}L^{2})$, where n = 1, 2, 3, ... are integers, *b* is Planck's constant, and m^* is the effective mass with which electrons move in the layer. These energies correspond to electron states in which the electrons form standing waves of electron density with $n = 1, 2, 3, \dots$ peaks in electron density (Fig. 2). These are the allowed states of the confined electrons. See NONRELATIVISTIC OUANTUM THEORY.

For electron effective mass $m^* = 0.1$ free electron masses (which is typical in semiconductors,

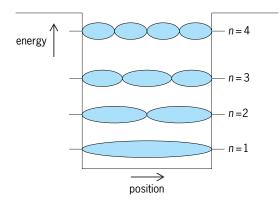


Fig. 2. Electron energy levels and electron standing waves for n = 1, 2, 3, and 4 quantum states in a quantum well.

where the electron inside the crystal behaves as if it were lighter than a free electron) and L = 10 nm, the allowed electron energies in electronvolts are $E_n = 0.0376n^2$ eV. Such quantized energy levels have clear signatures in the discrete lines of the absorption spectrum of light in high-quality gallium arsenide and other semiconductor quantum wells.

The potential energy for an electron in a quantum well and the potential energy for electrons in the barrier (Fig. 2) are basically determined by the electron affinities of the materials from which the well and barrier layers are composed. The electron affinity is the difference between the energy of an electron at rest in vacuum and the energy of an electron at the bottom of the conduction band of a material. The quantum wells also provide confining potentials for electrons in the valence bands of quantized electronic structures. When electrons are removed from the valence band of a semiconductor quantum well, the resulting hole in the valence-band quantum well leads to quantized hole phenomena that are directly analogous to all of the effects discussed above. See BAND THEORY OF SOLIDS; ELECTRON AFFINITY; HOLE STATES IN SOLIDS.

Potential well-shape profiles other than square or rectangular can be produced by growth of mixed alloys of the constituent semiconductors. A particularly basic example is the parabolic quantum well in which the confining potential for electrons varies as the square of the distance from the center of the well and thus closely mimics a simple harmonic oscillator potential for electron confinement. Well shape can also be changed by the application of an electric field perpendicular to the quantum well. Potentials in quantized electronic structures can also be shaped with the electric fields of charged donor and acceptor ions, and form the basis for so-called nipi doping superlattices. *See* ACCEPTOR ATOM; ANHARMONIC OS-CILLATOR; DONOR ATOM.

Quantum wires and quantum dots provide additional confinement of electrons in quantized electronic structures and further sharpening of the spectrum of allowed energy levels. Wires and dots can be formed in a thin quantum well layer by lithographic etching of the layer or by placement of metallic gate electrodes on top of the layer. Electrons passing through short quantum wires known as quantum point contacts have shown sharp quantization of the electrical conductance in units of e^2/b , where e is the electron charge. Closely spaced pairs of constrictions produce quantum-dot-like structures where electrons have shown unique Coulomb blockade behavior, in which the electrostatic energy needed to place an additional electron on a dot is clearly observable.

Fabrication. Quantized electronic structures are most frequently formed by epitaxial (that is, crystalline film) growth of layers of the silicon-germanium (group IV) system of semiconductors; in the phosphide, arsenide, and antimonide families of group III-V semiconductors; and in the telluride families of II-VI semiconductors. *See* SEMICONDUC-TOR.

In the molecular-beam-epitaxy growth of quantized electronic structures, beams of the constituent atoms and molecules of the semiconductor structures are formed by thermal evaporation of elemental sources of the atoms and molecules in ultrahigh vacuum. In the chemical-vapor-based growth techniques, by contrast, gases of compounds containing the desired elements react on hot substrate surfaces to produce the quantum structures. In each case, films can be grown at rates of the order of one monolayer of deposition per second. Film surfaces can be produced that are smooth at the scale of plus or minus one monolayer. Smoothness is important in quantized electronic structures because smooth films have sharper energy levels and stronger quantum behavior. See CRYSTAL GROWTH; SEMICONDUC-TOR HETEROSTRUCTURES.

Film thickness is controllable with monolayer precision by molecular-beam shutter operation in molecular-beam epitaxy and by valve openings and closures in the chemical-vapor techniques. Thickness control is also of importance because it determines the magnitude of quantized energies and the operating frequencies or wavelengths for many devices. Fortunately, it is possible to grow strained layers with different lattice constants on top of each other, and strain is frequently usable to fine-tune the properties of the quantum structures.

Quantum dot and wire formation pose a special challenge because lateral dimensional control during epitaxy is less precise than thickness control. Ingenious ways that have been demonstrated to achieve small lateral quantum confinement structures are overgrowth of cleaved edges of quantum wells, growth of repeated alternate submonolayer coverages on regularly stepped surfaces, and nucleation of small islands during growth of thin epitaxial layers of different lattice spacing than the substrate material. Chemical means of producing quantumsize particles have also rapidly developed, particularly the growth of sulfide and selenide nanocrystals and the growth of cage compounds that form templates for particle or wire formation. The buckyball family of closed cage molecules of symmetrically arranged carbon atoms has been extended to carbon nanotubes. These crystalline nanotubes can be millions of times longer than their diameters. They can be formed by vaporization of graphite with lasers or arcs and also by chemical vapor deposition. See FULLERENE; ZEOLITE.

Applications. Both the optical and the electronic applications of quantized electronic structures are important.

Optical applications. The optical applications are based on the interactions between light and electrons in the quantum structures. The absorption of a photon by an electron in a quantum well raises the electron from occupied quantum states to unoccupied quantum states. Electrons and holes in quantum wells may also recombine, with the resultant emission of photons from the quantized electronic structure as the electron drops from a higher state to a lower state. *See* ELECTRON-HOLE RECOMBINATION.

The photon emission is the basis for quantumwell semiconductor lasers, which have widespread applications in optical fiber communications and compact disk and laser disk optical recording. Quantum-well lasers operate by electrically injecting or pumping electrons into the lowest-conductionband (n = 1) quantum-well state, where they recombine with holes in the highest-valence-band (n = 1) quantum-well state (that is, the electrons drop to an empty n = 1 valence-band state; **Fig. 3***a*), producing the emission of photons. These photons stimulate further photon emission and produce high-efficiency lasing. *See* COMPACT DISK; LASER; OPTICAL COMMUNI-CATIONS; OPTICAL RECORDING.

The photon absorption is the basis for quantumwell photodetectors and light modulators. In the quantum-well infrared photodetector an electron is promoted from lower (say, n = 1) to higher (say, n =2) conduction band quantum-well states (Fig. 3b) by absorption of an infrared photon. An electron in the higher state can travel more freely across the barriers, enabling it to escape from the well and be collected in a detector circuit. Changes in quantum-well shapes produced by externally applied electric fields can change the absorption wavelengths for light in a quantized electronic structure. The shift in optical absorption wavelength with electric field is known as the quantum-confined Stark effect. It forms the basis for semiconductor light modulators and semiconductor optical logic devices. See OPTICAL DETEC-TORS; OPTICAL MODULATORS; STARK EFFECT.

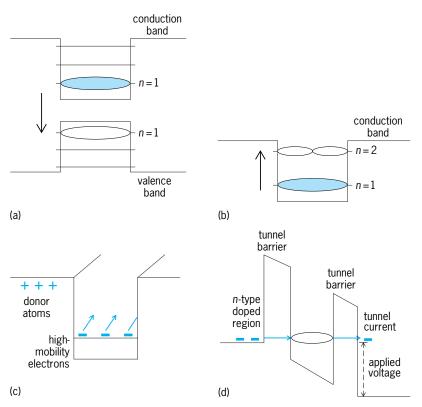


Fig. 3. Principles of operation of quantum-well devices. (a) Quantum-well laser. (b) Quantum-well infrared detector. (c) High-electron-mobility transistor (HEMT or MODFET). Electrons in the quantum well that came from donor atoms in the barrier are free to move with high mobility in the direction perpendicular to the page. (d) Resonant tunneling device.

Electrical applications. Modulation doping is a special way of introducing electrons into quantum wells for electrical applications. The electrons come from donor atoms lying in adjacent barrier layers (Fig. 3c). Modulation doping is distinguished from conventional uniform doping in that it produces carriers in the quantum well without introducing impurity dopant atoms into the well. Since there are no impurity atoms to collide with in the well, electrons there are free to move with high mobility along the quantum-well layer. Resistance to electric current flow is thus much reduced relative to electrical resistance in conventional semiconductors. This enhances the low-noise and high-speed applications of quantum wells and is the basis of the high-electronmobility transistor (HEMT), which is also known as the modulation-doped field-effect transistor (MOD-FET). HEMTs are widely used in microwave receivers for direct reception of satellite television broadcasts. See TRANSISTOR.

Modulation-doped structures have also shown a remarkable quantization of the Hall resistance, known as the quantum Hall effect. In this quantum effect, the lateral voltage induced in a flowing electron gas by a magnetic field, divided by the longitudinal current along the two-dimensional electron gas, is precisely quantized with values $R_H = b/(ne^2) =$ 25,812.81/*n* ohms, where $n = 1, 2, 3, \dots$ This ratio is so precisely determined and measurable that it has been adopted as the basis for the standard of resistance and the ohm. It was discovered in the two-dimensional electron gas formed at the metaloxide interface of a metal-oxide-semiconductor fieldeffect transistor (MOSFET). At very low temperatures (less than 1K) and high magnetic fields (greater than 1 tesla), the electrons in the two-dimensional electron gas in high-mobility materials can form ordered states that give rise to fractional quantization of the Hall resistance. The electrons then behave as if they had fractional charges. See ELECTRICAL UNITS AND STANDARDS; HALL EFFECT; RESISTANCE MEASURE-MENT.

Electrical conductivity in carbon nanotubes occurs without doping and results from the absence of any energy gap in the electronic energy band structure of the nanotubes and the presence of allowed states at the Fermi energy. Individual nanotubes can be electrically contacted. Simple quantum wire transistors displaying quantized electron motion have been formed from single nanotubes.

Quantum-mechanical tunneling is another important property of quantized electronic structures. Tunneling of electrons through thin barrier layers between quantum wells is a purely quantummechanical effect without any real analog in classical physics or classical mechanics. It results from the fact that electrons have wavelike properties and that the particle waves can penetrate into the barrier layers. This produces a substantial probability that the particle wave can penetrate entirely through a barrier layer and emerge as a propagating particle on the opposite side of the barrier. The penetration probability has an exponential drop-off with barrier thickness. The tunneling is greatest for low barriers and thin barriers.

This effect finds application in resonant tunnel devices, which can show strong negative resistance in their electrical properties. In such a device (Fig. 3d), electrons from an n-type doped region penetrate the barrier layers of a quantum well by tunneling. The tunneling current is greatest when the tunneling electrons are at the same energy as the quantum-well energy. The tunneling current actually drops at higher applied voltages, where the incident electrons are no longer at the same energy as the quantum-state energy, thus producing the negative resistance characteristic of the resonant tunneling diode. See ARTIFICIALLY LAYERED STRUC-TURES; NANOSTRUCTURE; NEGATIVE-RESISTANCE CIR-CUITS; RESONANCE (QUANTUM MECHANICS); TUNNEL-ING IN SOLIDS. Arthur C. Gossard

Bibliography. J. H. Davies, *The Physics of Low-Dimensional Semiconductors: An Introduction*, Cambridge University Press, 1998; A. C. Gossard (ed.), *Epitaxial Microstructures*, Academic Press, 1994; M. J. Kelly, *Low-Dimensional Semicon-ductors: Materials, Physics, Technology, Devices*, Clarendon Press, 1995; A. Shik, *Quantum Wells: Physics and Electronics of Two-Dimensional Systems*, World Scientific, 1997.

Quantized vortices

A type of flow pattern exhibited by superfluids, such as liquid ⁴He below 2.17 K. The term vortex designates the familiar whirlpool pattern where the fluid moves circularly around a central line and the velocity diminishes inversely proportionally to the distance from the center. The strength of a vortex is determined by the circulation k, which is the line integral of the velocity around any path enclosing the central line. For an ordinary vortex, k can possess any value; for a superfluid vortex, k is restricted to a quantized multiple of Planck's constant b divided by m, the mass of the helium atom. Hence the expression quantized vortex line. Although b and m are microscopic (that is, atomic) quantities, their ratio b/m is rather large, being equal to 10^{-3} cm² · s⁻¹. See VORTEX.

Theory. The possible existence of quantized vortices was suggested on theoretical grounds as early as 1949. A superfluid is believed to be characterized by a macroscopic (that is, large-scale) quantummechanical wave function ψ . This wave function locks the superfluid into a coherent state. Considerations on how a uniform flow might affect ψ lead to realization that the superfluid velocity v_s is intimately connected to the manner in which ψ might vary in space. From elementary considerations, one then deduces that quantized vortices should exist.

Since the velocity around the vortex increases without limit as the center is approached, the superfluid density and thus ψ must vanish at the center in order to avoid an infinite energy. Thus the central core of the vortex marks the zeros, or nodal lines, in the macroscopic wave function. *See* NONRELATIVIS-TIC QUANTUM THEORY; QUANTUM MECHANICS.

Production. Quantized vortex lines are usually produced by rotating a vessel containing superfluid helium. At very low rotation speeds, no vortices exist: the superfluid remains at rest while the vessel rotates. At a certain speed the first vortex appears and corresponds to the first excited rotational state of the system. If the container continues to accelerate, additional quantized vortices will appear. At any given speed the vortices form a regular array which rotates with the vessel.

Experimental investigation. Quantized vortex lines were first detected in the mid-1950s by their influences on superfluid thermal waves traveling across the lines. Shortly thereafter the quantum of circulation was directly measured by studying the precessional motion of a fine vibrating wire immersed in rotating superfluid helium. If no circulation is present, the plane of vibration of a circular-cross-section wire would remain stationary. In the presence of circulation, the vibration plane precesses at a rate determined by the circulation. It has been found that the stable values of circulation are close to b/m.

Mechanical properties. The mechanical properties of vortices have been studied by detecting various types of bending waves in the vortices. In one early experiment, well-defined normal modes were observed, which were consistent with predictions. Since the effective tension of a single vortex is very small, these mechanical measurements have usually been concerned with systems containing many lines. However, some progress has been made in measuring the very small angular momentum of a single line.

Use of ions to probe vortices. In the late 1950s it was discovered that electrons in liquid helium form tiny charged bubbles which can become trapped on the vortex core but can move quite freely along the line. These electron bubbles (often referred to as ions) have been one of the most useful probes of quantized vortices. The most accurate quantitative proof of the quantized nature of the vortices emerged from the discovery that at low temperatures (below 0.5 K) a rapidly moving ion can create a vortex ring. These rings were found to move precisely like classical smoke rings except that the circulation κ was equal to h/m to within 1%.

Detection of single vortex lines. Researchers have been able to use ions to detect single quantized vortex lines. In one experiment the vortices in a rotating vessel were charged with ions, and subsequently the amount of trapped charge was measured. As the vessel slowly accelerates from rest, the trapped charge is seen to increase as a stepwise function of rotation speed, thus proving that the superfluid comes into rotation in a sequence of quantum steps.

In an extension of the ion technique, the trapped ions are pulled out at the top of the vortex lines, accelerated, and focused onto a phosphor screen. The pattern of light thus produced on the phosphor

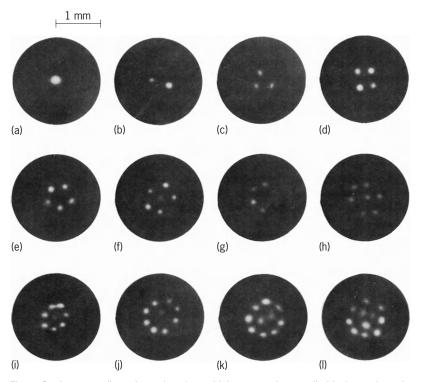


Fig. 1. Stationary configurations of vortices which appear when a cylindrical container of superfluid ⁴He is rotated about its axis. As the rotation speed increases from a to *I*, more vortices appear and the patterns become more complex. (From E. J. Yarmchuk, M. J. V. Gordon, and R. E. Packard, Observation of stationary vortex arrays in rotating superfluid helium, Phys. Rev. Lett., 43:214–217, 1979)

is a map of the position of the vortices where they contact the liquid meniscus (**Fig. 1**). The arrays of vortices agree well with theoretical predictions.

Vortices in helium-3. A second isotope of helium, ³He, also exists in superfluid states but only at temperatures below 2×10^{-3} K, 150,000 times colder than room temperature. Superfluid ³He is associated with a pairing of two atoms, a mechanism similar to that responsible for the phenomenon of superconductivity. There have been several varieties of

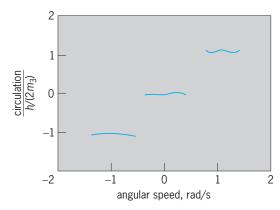


Fig. 2. Measured fluid circulation around a thin wire immersed in superfluid ³He. The horizontal axis of the graph is the angular speed of the rotating refrigerator. There are three stables values of fluid circulation: $h/(2m_3)$, 0, and $-h/(2m_3)$. (After J. C. Davis et al., Observation of quantized circulation in superfluid ³He-B, Phys. Rev. Lett., 66:329–332, 1991)

quantized vortices predicted for the ³He superfluids. The simplest type of vortex is similar to that in ⁴He except that the circulation quantum is expected to equal $b/(2m_3)$, where m_3 is the mass of the ³He atom and the factor 2 arises from the pairing of the atoms. In a direct measurement of the superfluid flow around a fine wire, the circulation was found to be quantized, as expected, in units of $b/(2m_3)$ [**Fig. 2**]. *See* SUPERCONDUCTIVITY.

Experiments using magnetic resonance techniques have confirmed the prediction that several varieties of ³He vortices exist. Which particular vortex type exists depends on the temperature and pressure within the superfluid. Using these nuclear magnetism techniques, the appearance of individual ³He vortices has been detected in a small vessel whose rotational speed is steadily changing. *See* MAGNETIC RESONANCE.

Vortices in neutron stars. Numerous sources of very regular, pulsed electromagnetic radiation have been detected. These so-called pulsars are believed to be rotating neutron stars. In addition to the two helium isotopes, it is believed that the interiors of these neutron stars are also superfluid. Since these pulsars are rotating objects, the neutron star's interior is expected to be filled with an array of quantized vortices, and some features of the neutron star's motion have been explained in terms of the related motion in the vortices. A laboratory-sized rotating superfluid ⁴He analog of a neutron star has been developed, and exhibits features similar to those seen in pulsar signals, supporting the idea that pulsars are indeed rotating superfluid neutron stars. See NEUTRON STAR; PULSAR. Richard E. Packard

Bibliography. M. Krusius, Vortices in rotating ³He, the vortices of superfluid ³He, *J. Low Temp. Phys.*, 91:233-273, 1993; D. R. Tilley and J. Tilley, *Superfluidity and Superconductivity*, 3d ed., 1990; E. J. Yarmchuk and R. E. Packard, Photographic studies of quantized vortex lines, *J. Low Temp. Phys.*, 46:479-515, 1982.

Quantum (physics)

A term characterizing an excitation in a wave or field, connoting fundamental particlelike properties such as energy or mass, momentum, and angular momentum for this excitation. In general, any field or wave equation that is quantized, including systems already treated in quantum mechanics that are second-quantized, leads to a particle interpretation for the excitations which are called quanta of the field. This term historically was first applied to indivisible amounts of electromagnetic, or light, energy usually referred to as photons.

The photon, or quantum of the electromagnetic field, is a massless particle, best interpreted as such by quantizing Maxwell's equations. Analogously, the electron can be said to be the quantum of the Dirac field through second quantization of the Dirac equation, which also leads to the prediction of the existence of the positron as another quantum of this field with the same mass but with a charge opposite to that of the electron. In similar fashion, quantization of the gravitational field equations suggests the existence of the graviton, a particle which is the quantum of gravitational field. The pi meson or pion was theoretically predicted as the quantum of the nuclear force field by H. Yukawa; other particles associated with nuclear forces have since been found. Another quantum is the quantized lattice vibration, or phonon, which can be interpreted as a quantized sound wave since it travels through a quantum solid or fluid, or through nuclear matter, in the same manner as sound through air. The energy of a phonon is hv, b being Planck's constant (called the quantum of action in "old" quantum mechanics) and v the frequency of oscillation of the matter through which the wave propagates. The energy of a photon is also bv, with *b* being Planck's constant and ν the frequency of the light wave.

The use of the word quantum as an adjective (quantum mechanics, quantum electrodynamics) implies that the particular subject is to be treated according to the modern rules that have evolved for quantized systems. *See* ELEMENTARY PARTICLE; GRAV-ITATION; GRAVITON; MAXWELL'S EQUATIONS; MESON; NONRELATIVISTIC QUANTUM THEORY; PHONON; PHO-TON; QUANTUM ELECTRODYNAMICS; QUANTUM FIELD THEORY; QUANTUM MECHANICS. Kenneth E. Lassila

Quantum acoustics

The investigation of the effects of the laws of quantum mechanics on the propagation and absorption of sound. At the present stage of development of physical science, quantum mechanics is the most fundamental theory of physical phenomena. However, for many applications in the everyday world, a sufficiently accurate description of nature is provided by classical mechanics. Quantum acoustics refers to acoustic experiments that are carried out under conditions such that the results can be understood only in terms of quantum theory. As a general tendency, quantum effects become more important in acoustic experiments that are performed with higher-frequency sound waves or that are carried out at lower temperatures. See NONRELATIVISTIC QUANTUM THEORY; QUANTUM MECHANICS.

Sound and phonons. To understand the quantum nature of sound, it is valuable to consider the origins of the quantum theory of light. According to the classical theory of electromagnetism, the energy of a light wave can have any value. However, experiments in the latter part of the nineteenth century showed that the classical theory of electromagnetism combined with the laws of statistical mechanics could not explain the spectrum of light emitted by a heated surface. To reconcile theory and experiment, Max Planck in 1900 proposed that the energy of a light wave is quantized. This means that only certain values of the energy of the wave are al-

lowed. These allowed values are given by Eq. (1),

E

$$= n\hbar\omega$$
 (1)

where \hbar is Planck's constant divided by 2π , ω is the angular frequency of the light wave, and *n* is an integer. The modern interpretation of this formula describes this energy in terms of elementary quanta called photons. Each photon has an energy $\hbar\omega$, and the wave is made up of *n* photons. It was later realized that this same quantization of energy should apply to sound waves. These fundamental units of sound energy were given the name phonons. *See* HEAT RADIATION; LIGHT; PHONON; PHOTON.

Complementarity principle. According to quantum theory, a sound wave should be considered to be a beam of phonon quanta. However, for many practical applications in acoustics, such as the design of musical instruments or the construction of concert halls, the classical theory of sound holds extremely well and it is unnecessary to describe the sound wave in terms of phonons. This is an example of the complementarity principle, which states that under suitable conditions the predictions of classical and quantum theory agree, even though the nature of the two theories may first seem to be quite different.

Size of quanta. For a sound wave of given frequency, the energy $\hbar \omega$ of a single phonon can be calculated. The numerical value of Planck's constant is 6.6 \times 10^{-34} J \cdot s. Thus, for a sound wave of frequency 1000 Hz, the energy of a sound quantum is only 6.6×10^{-31} J. Hence, a wave of this frequency and of total energy as small as 10⁻¹² J still contains more than 1018 quanta. Consequently, for practical purposes the restriction that the energy of the wave be quantized in units of the phonon energy is not important. However, because the amount of energy in each quantum increases with increasing frequency, the effects of the quantization are much more likely to be significant when the acoustic wave has a very high frequency. With modern techniques it is possible to generate and detect sound waves whose frequency extends far above the audible range, and it is in experiments conducted with these methods that the quantum nature of sound has been convincingly demonstrated. See ULTRASONICS.

Phonoatomic effect. One of the earliest experiments that confirmed the idea that light consists of photons was the photoelectric effect. In this effect the energy of light quanta is used to eject electrons from the surface of a metal. Only if the frequency of the light is sufficiently high do the photons have enough energy to knock the electrons out of the metal. The detection of this threshold frequency is thus strong evidence for quantization of energy. An analogous experiment has been performed with sound (Fig. 1). In the experiment, sound is generated by a source S in a vessel of liquid helium maintained at very low temperature, less than 0.1 K above absolute zero. At this temperature, even very high-frequency sound waves will travel a long distance through the liquid without appreciable loss of energy. In liquid helium, each atom is bound to the other atoms in

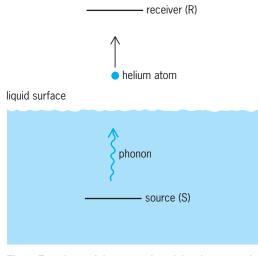


Fig. 1. Experimental demonstration of the phono-atomic effect in liquid helium.

the liquid by an energy that is unusually small compared to that of other liquids. This binding energy E_{bind} is such that $E_{\text{bind}}/k_B = 7.1$ K (where k_B is Boltzmann's constant). For sound of sufficiently high frequency, the sound quanta when they arrive at the surface of the liquid have sufficient energy to knock helium atoms out of the liquid. These atoms can be detected by a suitable receiver *R* placed above the surface of the liquid. The energy E_{kinetic} of an ejected atom is equal to the energy of a phonon minus the energy E_{bind} , and so the velocity of the atom is given by Eq. (2), where *m* is the mass of a helium atom.

$$v_{\rm atom} = \sqrt{\frac{2(\hbar\omega - E_{\rm bind})}{m}} \tag{2}$$

The time of arrival of the ejected atoms at the receiver R is determined by this velocity and hence is dependent on the frequency of the sound wave. Experiments have confirmed that the helium atoms arrive at the time expected based on this theory. *See* PHOTOEMISSION.

Phonon-phonon interactions. Sound waves in solids are attenuated during their propagation by a wide variety of physical processes. In many materials the most important mechanisms are related to impurities or defects in the solid, such as cracks, grain boundaries, or dislocations. Even when sound travels through a perfect crystal containing no defects, it is found that a measurable attenuation still occurs. In insulating crystals, where there are no free electrons, this attenuation is due to an interaction between the sound wave and the random thermal vibrations of the atoms in the solid. These random vibrations, which constitute the heat energy of the solid, are also quantized and are called thermal phonons. The attenuation of the sound wave can be attributed to collisions with the thermal phonons in which some of the sound quanta are scattered out of the sound beam. This mechanism is referred to as phononphonon scattering. See CRYSTAL DEFECTS; LATTICE VI-BRATIONS; SOUND ABSORPTION.

In a linear elastic solid the elastic stress is exactly proportional to the strain, a phenomenon that is called Hooke's law. If Hooke's law holds, the presence of one wave does not affect the propagation of another, and so there are no interactions between phonons. Phonon-phonon scattering occurs because real solids always exhibit some deviations from linear elastic behavior. This nonlinearity is called anharmonicity. *See* HOOKE'S LAW; NONLINEAR ACOUSTICS.

The most important type of collision (**Fig. 2**) between phonons is the three-phonon process. In this interaction a sound phonon collides with a thermal phonon, and one new phonon is produced. The collision has to conserve energy, and so Eq. (3) must

$$E_{\text{sound}} + E_1 = E_2 \tag{3}$$

be satisfied, where E_{sound} , E_1 , and E_2 are the energies of the sound phonon and the two other phonons. Since it is also necessary to conserve momentum, Eq. (4) must be satisfied, where \vec{p}_{sound} , \vec{p}_1 , and \vec{p}_2 are

$$\vec{p}_{\text{sound}} + \vec{p}_1 = \vec{p}_2 \tag{4}$$

the momentum vectors of the three phonons. The momentum of a phonon is related to its wave vector \vec{k} by Eq. (5).

$$\vec{p} = \hbar \vec{k} \tag{5}$$

See CONSERVATION OF ENERGY; CONSERVATION OF MOMENTUM.

The attenuation of a sound wave that occurs as a result of these collisions increases with increasing frequency of the sound wave. It also increases rapidly with temperature because the higher the temperature the more thermal phonons are present in a solid.

Even when the temperature is very low and there are very few thermal phonons, anharmonicity can still give rise to an attenuation of a sound wave. This is because the sound phonons can spontaneously decay into phonons of lower frequencies (Fig. 2*b*). In

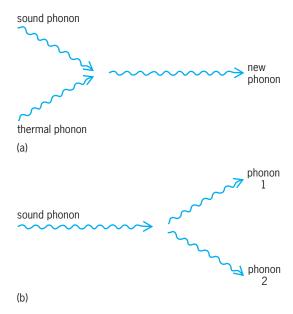


Fig. 2. Phonon-phonon processes. (a) A sound phonon collides with a thermal phonon to produce another phonon. (b) A sound phonon decays into phonons 1 and 2.

this process the conditions of conservation of energy and momentum are given by Eqs. (6) and (7), with

$$E_{\text{sound}} = E_1 + E_2 \tag{6}$$

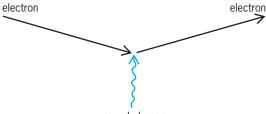
$$\vec{p}_{\text{sound}} = \vec{p}_1 + \vec{p}_2 \tag{7}$$

 E_1 and E_2 the energies, and \vec{p}_1 and \vec{p}_2 the momenta of the phonons produced in the decay. The rate at which this decay occurs is proportional to the frequency of the sound wave to the fifth power, and so the attenuation is important only for sound waves of very high frequency. If the number of phonons is sufficiently large, it is possible under some circumstances for there to be a large buildup in the population of some of the decay phonons. This process is called parametric amplification. *See* PARAMETRIC AMPLIFIER.

Sound attenuation by electrons. In a metal a certain number of electrons from each atom are able to move freely through the crystal. The sound quanta can be absorbed by these electrons (**Fig. 3**). This process also leads to an attenuation of the sound wave. Just as for the phonon-phonon interaction, energy and momentum must be conserved in the process. It can be shown that because of these conditions the sound phonons are absorbed primarily by electrons that are moving at right angles to the direction of propagation of the sound wave.

Resonant absorption of phonons. Many solids contain defects or impurities that can absorb energy from a sound wave. The simplest example is a paramagnetic impurity in a crystal. When an atom with an unpaired atomic electron is placed in a magnetic field, the atom can exist in one of two different energy states; these states correspond to different orientations of the spin of the electron with respect to the direction of the magnetic field. If the energy of the phonons in a sound wave is the same as the difference in energy between these two states, the phonons can be absorbed by the impurity atom. In this process the atom makes a transition from the lower energy state to the upper. This process is called resonance absorption because there must be a precise matching of the sound energy and the spacing between the two states.

This same general type of phenomenon can occur in several other situations. Some crystals are composed of atoms that have nuclei that possess a magnetic moment. In a magnetic field these nuclei have several equally spaced energy states. The sound phonons can be absorbed by these nuclei if the phonon energy is the same as the difference in en-



sound phonon

Fig. 3. Absorption of a sound phonon by an electron in a metal.

ergy of the different nuclear states. For magnetic fields that can be readily produced in the laboratory, this type of absorption occurs for phonons of frequency in the range of roughly 5 to 100 MHz. The attenuation is quite small and can be detected only in specially designed experiments. In semiconducting crystals that have been doped with donor or acceptor impurities, the electrons associated with these impurities often have several different energy states even in the absence of a magnetic field. Sound phonons can also be resonantly absorbed by these electrons.

In the above examples of resonance absorption, the attenuation of the sound wave occurs only if the sound wave has one particular frequency. An especially interesting example of resonance absorption occurs in glasses and other amorphous materials. In these solids there is no regular crystal structure, and so every atom is in a slightly different environment. It is believed that in all glasses there are a small number of atoms that interact with their neighbors in such a way that each one of the atoms can exist in two states that have nearly the same energy. Such an atom can absorb energy from a sound wave if the energy of the phonon quanta matches the difference in the energies of the two states of the atom. However, because of the random nature of the glass this energy difference is different for different atoms. As a result, regardless of the frequency of the sound wave, there will always be some atoms in the glass that are able to absorb the sound phonons. Thus, resonance absorption in glasses occurs over a broad band of frequencies. See AMORPHOUS SOLID; GLASS.

Phonon maser. Resonance absorption occurs when a paramagnetic impurity makes a transition from a low energy state to a higher state and a sound phonon is absorbed. If by some means it is possible to prepare a crystal in which the majority of the impurities are in the higher of their two states, the sound phonons will induce transitions from the higher state to the lower state. New phonons will be emitted, and the sound wave will be amplified. This process is analogous to the amplification of microwave electromagnetic radiation that occurs in a maser. *See* MASER. Humphrey J. Maris

Bibliography. R. T. Beyer and S. V. Letcher, *Physical Ultrasonics*, 1969; M. J. Crocker (ed.), *Handbook of Acoustics*, Wiley, 1998; S. Hunklinger et al. (eds.), *Phonons '89*, 1990; H. J. Maris, Phonon interactions in liquid helium, *Rev. Mod. Phys.*, 49:341–359, 1977; M. Meissner et al. (eds.), *Phonon Scattering in Condensed Matter VII*, 1993; R. Peierls, *Quantum Theory of Solids*, 1955, paperback 2001; W. A. Phillips (ed.), *Amorphous Solids: Low-Temperature Properties*, Springer-Verlag, Berlin, 1981; M. N. Wybourne et al. (eds.), Phonons 2001, *Physica B*, vols. 316–317, 2002.

Quantum anomalies

Phenomena that arise when a quantity that vanishes according to the dynamical rules of classical physics acquires a finite value when quantum rules are used. For example, the classical Poisson bracket for some entities may vanish; yet the corresponding quantum commutator may be nonzero-this is a commutator anomaly. Alternatively, the flow of some material current may satisfy a continuity equation by virtue of the classical equations of motion, indicating conservative flow; but upon quantization the continuity equation may fail and the flow may no longer be conservative in the quantum theory-this is an anomalous divergence (of the current in question). Since the forms of Poisson brackets and quantum commutators as well as the occurrence of continuity equations for currents are related to symmetries and conservation laws of the theory, quantum anomalies serve to break some symmetries and destroy some conservation laws of classical models. This violation of symmetry is not driven by explicit symmetry-breaking terms in the dynamical equations-rather the quantization procedure itself violates the classical symmetry. The mathematical reason for this phenomenon is that classical dynamics, involving a finite number of degrees of freedom, usually leads to a quantum theory on an infinite-dimensional vector space (Hilbert space), and this "infinity" gives rise to novel effects. See CANONICAL COORDINATES AND TRANS-FORMATIONS; CONSERVATION LAWS (PHYSICS); EQUA-TION OF CONTINUITY; HILBERT SPACE; NONRELATIVIS-TIC QUANTUM THEORY; SYMMETRY LAWS (PHYSICS).

Example from quantum mechanics. While most instances of anomalies arise in quantum field theory, the simplest example is provided by the quantum mechanics of a point particle moving on a plane, described by a two-vector $\mathbf{r}(t)$, and governed by a potential proportional to a two-dimensional delta function, $\delta^2(\mathbf{r})$. Within classical mechanics, the two-dimensional delta-function potential does not deflect the particle's motion. But solving the quantum-mechanical Schrödinger equation yields a bound state and an isotropic scattering cross section $\sigma(\theta)$, independent of the scattering angle θ , that is given by Eq. (1). Here, -B is the energy of the bound state,

$$\sigma(\theta) = \frac{\lambda}{[\pi^2 + \ln^2(B/E)]} \tag{1}$$

E is the energy of the scattered particle, and λ is its quantum-mechanical wavelength. This example also shows that a symmetry of the classical system is broken: When the spatial variable **r** is rescaled to s**r**, and the time variable t is rescaled as s^2t , where s is a number that can have any value, classical equations of motion are unchanged. This invariance of dynamics, called scale symmetry, holds because the model with a delta-function potential has no intrinsic definition of spatial and temporal scales. Consequences of scale symmetry are that σ/λ must have no energy variation, and that bound states cannot exist. Yet the quantum theory for the planar delta-function potential leads to a bound state and to a cross section where σ/λ varies with energy. The model exhibits what is known as a scale anomaly, and its scale symmetry is broken quantum-mechanically. The mathematical reason for the anomaly is the singular nature of the

hamiltonian quantum operator. Owing to its $\mathbf{r} = 0$ short-distance singularity in the delta-function potential, the hamiltonian operator is not well defined on the quantum Hilbert space. Moreover, the mathematical procedure for converting this ill-defined operator into a well-defined one (called self-adjoint extension) necessarily breaks scale invariance. *See* OPERATOR THEORY; SCHRÖDINGER'S WAVE EQUATION.

Role in quantum field theory. While this example exhibits all the features of quantum-mechanical or anomalous symmetry breaking, the physically interesting setting for these phenomena is in quantum field theory, especially as applied to elementary particle physics, where the mechanism serves as an important source for symmetry breaking. Physicists are mostly satisfied that the ultimate laws of nature enjoy a high degree of symmetry. This ancient esthetic prejudice is built into field theoretic models of fundamental particle physics by endowing these models with a high degree of symmetry, whose presence also allows the control and renormalization of the ultraviolet infinities that afflict quantum field theory. The symmetries of the "standard" particle physics model include a scale symmetry (analogous to that described previously), a non-abelian gauge symmetry (matrix-valued generalization of the electromagnetic gauge invariance), and a chiral symmetry associated with massless spin-1/2 fermions-ingredients of the standard model. This symmetry ensures that a chiral current exists, seemingly obeying a continuity equation, since a classical calculation finds the current to be divergence-free. See ELEMENTARY PARTI-CLE; GAUGE THEORY; QUANTUM FIELD THEORY; STAN-DARD MODEL

In spite of physicists' preference that descriptions of nature be enhanced by a large amount of symmetry, actual physical phenomena rarely exhibit overwhelming symmetry. Therefore, at the very same time that a theory is constructed with intrinsic symmetry, a way must be found to break the symmetry in the physical consequences of the model. One mechanism for achieving this is spontaneous symmetry breaking. Another is quantum-mechanical symmetry breaking through anomalies. Two instances of anomalous symmetry breaking in the standard model of particle physics are described below.

Scale anomalies. While the standard model appears to possess scale symmetry, nature clearly exhibits various scales, so it is not scale-invariant. The potential conflict between theory and experiment is avoided because scale invariance is anomalously broken. The true quantum behavior (as opposed to the classical behavior) of quantum-mechanically broken scale symmetry is described by the renormalization group. One aspect is that strengths of interactions depend on the energy at which they are measured, and the dependence is inverse-logarithmic, as in the quantum-mechanical example. This variation has been observed experimentally. *See* RENORMALIZA-TION.

Chiral anomalies. The previously mentioned chiral symmetry is absent from some channels in the fermion sector. The relevant current j_{thiral}^{μ} , which

would satisfy a continuity equation if the symmetry were present, in fact possesses a nonvanishing, anomalous divergence, which obeys Eq. (2). Here **E**

$$\frac{\partial}{\partial x^{\mu}} j^{\mu}_{\text{chiral}}(x) \propto \mathbf{E} \cdot \mathbf{B} \neq 0$$
 (2)

and B are the electric and magnetic (abelian and nonabelian) gauge fields present in the model. The proportionality constant is determined by the charges of the fermions (quarks and leptons) in the relevant channel. This result is gratifying, for if the chiral current were conserved, the amplitude for neutral pion decay (to two photons) would be much suppressed relative to the observed value. Instead, with the parameters of the standard model determining the proportionality constant for this anomalous divergence of j^{μ}_{chiral} in the channel containing the neutral pion, the decay amplitude can be calculated and excellent agreement with experiment is obtained. In fact, the chiral current in all fermionic channels potentially can acquire an anomalous divergence. This would spell disaster for the theory, because consistency and renormalizablility require that the current be divergence-free in those channels to which the gauge fields of the theory couple. But each fermion contributes independently the value of its charge q_i to the proportionality constant in the anomalous divergence relevant to gauge-field-coupled channels. So there is a possibility of cancellation if $\sum_{i} q_i = 0$. With the known charges for each family of quarks and leptons, the sum indeed vanishes. This is striking evidence that anomaly phenomena exist in nature, and are not mere constructions of mathematical physicists.

The symmetry structure of a theory is encoded in the symmetry group's Lie algebra, which is realized by Poisson brackets in a classical theory and by commutators in the quantum theory. The Lie algebra of the apparent chiral symmetry in the standard model is indeed realized by the Poisson brackets, which are calculated according to the rules of classical physics. On the other hand, calculation of quantum commutators exposes a discrepancy: there are commutator anomalies, reflecting the anomalous breaking of chiral symmetry. Thus anomalous current divergences and anomalous commutators are two sides of the same coin. *See* LIE GROUP.

Relation to infinities. The technical reason for the emergence of scale and chiral anomalies in quantum field theory is the presence of ultraviolet, short-distance infinities in the formalism. These infinities must be controlled and renormalized, but that procedure cannot be carried out in a manner consistent with the symmetries. Thus anomalous symmetry breaking is a consequence of the infinities in quantum field theory. Yet the anomalies themselves are finite quantities. Indeed the chiral anomalies, both in current divergences and in commutators, have a mathematical significance related to topological properties of the gauge fields that occur in the anomaly equations; for example, the $\mathbf{E} \cdot \mathbf{B}$ term in the anomalous divergence of the chiral current is the Chern-Pontryagin density, whose space-time integral is an integer-valued measure of the amount of "twist" in **E** and **B**.

Role in other branches of physics. Quantum anomalies play a role in various other branches of physics, in which quantum field theory finds application, including condensed matter, supersymmetry, string theory, and motion in curved space-time. For example, the requirement that chiral anomalies cancel in string theory or that scale anomalies cancel in supersymmetric theories (analogous to the cancellation in gauge field channels of the standard model) severely constrains the structure and determines the model almost uniquely. This fact suggests that anomalies reflect a subtle and deep feature in the mathematical description of nature, which has been only partially uncovered and understood. See SPACE-TIME; SUPER-STRING THEORY; SUPERSYMMETRY. Roman Jackiw

Bibliography. R. Bertlmann, Anomalies in Quantum Field Theory, Oxford University Press, 2000; B. R. Holstein, Anomalies for pedestrians, Amer. J. Phys., 61:142-147, 1993; S. Treiman et al., Current Algebra and Anomalies, Princeton University Press/World Scientific, 1985; S. Weinberg, The Quantum Theory of Fields, vol. 1: Foundations, 1995, vol. 2: Modern Applications, 1996, Cambridge University Press.

Quantum chemistry

A branch of chemistry concerned with the application of quantum mechanics to chemical problems. More specifically, it is concerned with the electronic structure of molecules. Since 1960 the ease with which the quantum chemist may obtain reliable approximate solutions to the nonrelativistic Schrödinger equation has improved by at least eight orders of magnitude. The following is a brief review of developments in ab initio molecular electronic structure theory since 1960. The term "ab initio" implies that no approximations have been made in the one- and two-electron integrals, shown in Eqs. (1) and (2), arising from the ordinary nonrelativistic hamiltonian operator, Eq. (3). In contrast,

$$I(i \mid j) = \int \phi_i^*(1) \left\{ \frac{-\nabla_1^2}{2} - \sum_A \frac{Z_A}{r_{1A}} \right\} \cdot \phi_j(1) \, dv(1) \quad (1)$$

$$(ij \mid kl) = \int \phi_i^*(1)\phi_j^*(2) \frac{1}{r_{12}} \cdot \phi_k(1)\phi_l(2) \, dv(1) \, dv(2) \quad (2)$$

$$H = \sum_{i} \left\{ \frac{-\nabla_{i}^{2}}{2} - \sum_{A} \frac{Z_{A}}{r_{iA}} \right\} + \sum_{i} \sum_{j>i} \frac{1}{r_{ij}} \quad (3)$$

semiempirical methods resort to various approximate schemes, especially in evaluating the twoelectron integrals [(ij | kl); also known as electron repulsion integrals]. The present discussion is primarily restricted to the method that dominates the field of quantum chemistry, namely, the Hartree-Fock or self-consistent-field approximation and the family of methods grouped under the category of density functional theory. *See* HAMILTON'S EQUATIONS OF MOTION; NONRELATIVISTIC QUANTUM THEORY; QUANTUM MECHANICS.

Elementary theory. For closed-shell molecules, the form of the Hartree-Fock wave function is given by Eq. (4), in which A(n), the antisymmetrizer for n

$$\psi_{\rm HF} = A(n)\phi_1(1)\phi_2(2)\dots\phi_n(n)$$
 (4)

electrons, has the effect of making a Slater determinant out of the orbital product on which it operates. The ϕ 's are spin orbitals, products of a spatial orbital χ and a one-electron spin function α or β . For any given molecular system there are an infinite number of wave functions of form (4), but the Hartree-Fock wave function is the one for which the orbitals ϕ have been varied to yield the lowest possible energy [Eq. (5)].

$$E = \int \psi_{\rm HF}^* H \psi_{\rm HF} \, d\tau \tag{5}$$

The resulting Hartree-Fock equations are relatively tractable due to the simple form of the energy E for single determinant wave functions [Eq. (6)].

$$E_{\rm HF} = \sum_{i} I(i \mid i) + \sum_{i} \sum_{j > i} [(ij \mid ij) - (ij \mid ji)] \quad (6)$$

To make this discussion more concrete, it should be noted that for singlet methylene (the CH_2 molecule), the Hartree-Fock wave function is of the form given in Eq. (7).

$$\psi_{\rm HF} = A(8) \ 1a_1\alpha(1) \ 1a_1\beta(2) \ 2a_1\alpha(3) \ 2a_1\beta(4)$$
$$1b_2\alpha(5) \ 1b_2\beta(6) \ 3a_1\alpha(7) \ 3a_1\beta(8) \quad (7)$$

The same energy expression, Eq. (6), is also applicable to any open-shell system for which the openshell electrons all have parallel spins. This follows from the fact that such Hartree-Fock wave functions can always be expressed as a single Slater determinant. A simple example is triplet methylene, shown in Eq. (8), for which the outer two $3a_1$ and $1b_1$ or-

$$\psi_{\rm HF} = A(8) \ 1a_1\alpha(1) \ 1a_1\beta(2) \ 2a_1\alpha(3) \ 2a_1\beta(4)$$
$$1b_2\alpha(5) \ 1b_2\beta(6) \ 3a_1\alpha(7) \ 1b_1\alpha(8) \tag{8}$$

bitals have parallel spins. For clarity it is often helpful to abbreviate Eq. (8) as Eq. (9). Although solution of

$$\psi = 1a_1^2 \, 2a_1^2 \, 1b_2^2 \, 3a_1\alpha \, 1b_1\alpha \tag{9}$$

the Hartree-Fock equations for an open-shell system such as triplet methylene is more difficult than for the analogous closed-shell system, Eq. (7), the procedures are well established.

In fact, methods are readily available for the solution of the Hartree-Fock equations for any system for which the energy expression involves only Coulomb and exchange integrals, Eqs. (10) and (11).

$$J_{ij} = (ij \mid ij) \tag{10}$$

$$K_{ij} = (ij \mid ji) \tag{11}$$

Open-shell singlets are a class of systems that can be treated in this way, and one such example is the first excited singlet state (of ${}^{1}B_{1}$ symmetry) of methylene, Eq. (12). In addition, these same generalized

$$\psi_{\rm HF} = \frac{1}{\sqrt{2}} \, 1a_1^2 \, 2a_1^2 \, 1b_2^2 \, 3a_1\alpha \, 1b_1\beta \\ - \frac{1}{\sqrt{2}} \, 1a_1^2 \, 2a_1^2 \, 1b_2^2 \, 3a_1\beta \, 1b_1\alpha \quad (12)$$

Hartree-Fock procedures can be used for certain classes of multiconfiguration Hartree-Fock wave functions.

In recent years, density functional methods have come to play an important role in molecular electronic structure theory. Density functional theory can be traced to a critical 1951 paper by John Slater. In order to avoid the evaluation of the exchange integrals in Eq. (11), Slater proposed an approximation to Hartree-Fock theory in which exchange was treated in a statistical manner. For two decades, this statistical exchange approximation was viewed as a theoretically inferior but computationally advantageous method, primarily used in applications to the band structure of solids, for which ab initio methods were not feasible.

In 1965, W. Kohn and L. J. Sham wedded Slater's statistical exchange to a potentially more powerful theoretical methodology, again with many similarities to ordinary Hartree-Fock theory. However, in principle, the Kohn-Sham method could go beyond Hartree-Fock theory to include electron correlation. Today, the most popular methods for molecular electronic structure theory are hybrid methods involving a mixture of Hartree-Fock theory and density functional theory. For example, the popular B3LYP method combines A. D. Becke's three-parameter treatment of exchange with C. Lee, W. Yang, and R. G. Parr's description of electron correlation. The B3LYP method is semiempirical in the sense that the three parameters in the treatment of exchange were chosen to reproduce certain experimental thermochemical data. Becke's treatment of exchange is a hybrid in that it includes 20% exact exchange; that is, 20% of the exchange functional is the Hartree-Fock exchange. Although the relationship of methods such as B3LYP to an exact quantum-mechanical theory is unclear, it is undeniable that B3LYP and related methods are very efficient and very useful in solving chemistry problems.

Basis sets. To solve the Hartree-Fock or Kohn-Sham equations exactly, one must either expand the spatial orbitals χ in a complete set of analytic basis functions or obtain strictly numerical (that is, tabulated) orbitals. The former approach is impossible from a practical point of view for systems with more than two electrons, and the latter has been accomplished only for atoms and, more recently, for some

TABLE 1. Minimum basis set of Slater functions for the carbon atom					
Label	Analytic form	Exponent ζ^*			
1s	$(\zeta^3/\pi)^{1/2} \exp(-\zeta r)$	5.673			
2s	$(\zeta^5/3\pi)^{1/2} r \exp(-\zeta r)$	1.608			
$2p_x$	$(\zeta^5/\pi)^{1/2} x \exp(-\zeta r)$	1.568			
$2p_v$	$(\zeta^5/\pi)^{1/2} y \exp(-\zeta r)$	1.568			
2pz	$(\zeta^5/\pi)^{1/2} z \exp(-\zeta r)$	1.568			
* The orbital exponents ς are optimum for the $^{3}\textsc{P}$ ground state of the carbon atom.					

diatomic molecules. Therefore the exact solution of the Hartree-Fock (or Kohn-Sham) equations is abandoned for polyatomic molecules. Instead, one adopts an incomplete (but reasonable) set of analytic basis functions and solves for the best variational [that is, lowest-energy; Eq. (5)] wave function of form (4). C. C. J. Roothaan and coworkers at the University of Chicago referred to such a wave function as being of self-consistent-field (SCF) quality. At the limit of very large basis sets, then, it is reasonable to refer to the resulting SCF wave function as near Hartree-Fock.

For large chemical systems, only minimum basis sets (MBS) are typically used in ab initio theoretical studies. The term "large" includes molecular systems with 1000 or more electrons. A large molecule treated by MBS-SCF methods as early as 1991 is the six-base-pair DNA fragment $C_{116}H_{138}N_{46}O_{68}P_{10}$. Significantly larger systems can be treated today with state-of-the-art quantum-chemical methods. A minimum basis set includes one function for each orbital occupied in the ground state of each atom included in the molecule. For the first row atoms B, C, N, O, and F, this means that a minimum basis set includes 1s, 2s, $2p_x$, $2p_y$, and $2p_z$ functions.

Historically, minimum basis sets have been composed of Slater functions, such as those seen in **Table 1** for the carbon atom. However, experience has shown that the evaluation of the molecular integrals, Eq. (2), arising when Slater functions are employed is extremely time-consuming. Therefore each Slater function in a minimum basis set is typically replaced by a linear combination of three or four gaussian functions. The resulting chemical predictions obtained with such STO-3G (Slater-type orbital-three gaussian functions) or STO-4G basis sets are usually indistinguishable from the corresponding Slater function results.

Minimum basis sets are inadequate for certain types of chemical predictions. Therefore a basis twice as large, and appropriately designated doublezeta (DZ), is often used in theoretical studies. Here, however, it is not as fruitful to expand each Slater function as a linear combination of gaussians. Instead gaussian functions $x^p y^q z^r e^{-\alpha r^2}$ are used directly in atomic self-consistent-field calculations and then contracted according to the atomic results. Among the most widely used contracted gaussian double-zeta basis sets are those of T. H. Dunning. His double-zeta basis has 9s and 5p original (or primitive) gaussian functions, and is contracted to 4*s* and 2*p*. Thus the basis may be designated C(9s5p/4s2p). A substantial molecular system studied as early as 1995 with the double-zeta basis set is the giant fullerene C_{240} , a molecule with 1440 electrons. By 1999 G. E. Scuseria and coworkers were able to report a DZ basis set study of a 1026-atom (6767 basis functions) RNA fragment.

Just as the double-zeta basis logically follows the minimum set, the logical extension of the doublezeta set involves the addition of polarization functions. Polarization functions are of higher orbital angular momentum than the functions occupied in the atomic self-consistent-field wave function. That is, for carbon, d, f, g, \ldots functions will be polarization functions. Fortunately, d functions are far more important than f, f functions are far more important than g, and so on. For most chemical applications a double-zeta plus polarization (DZ + P) basis including a single set of five d functions ($d_{x^2-y^2}, d_{z^2}, d_{xy}, d_{xz}, d_{yz}$) will be quite adequate for first-row atoms.

Structural predictions. Ab initio theoretical methods have had the greatest impact on chemistry in the area of structural predictions. A good illustration of this is the methylene radical for which S. F. Boys reported the first ab initio study in 1960. Boys predicted the structure of triplet methylene to be $r_e(C-H) = 0.112$ nanometer, $\theta_e(HCH) = 129^\circ$. Unfortunately, however, the work of Boys was largely denigrated due to the 1961 spectroscopic conclusion of G. Herzberg that the lowest triplet state of CH₂ was linear. Herzberg's conclusion was greatly strengthened by a very influential 1962 semiempirical study of H. C. Longuet-Higgens, who concurred that the ground state of CH₂ was linear.

It was not until 1970 that a definitive theoretical prediction of the nonlinearity of ${}^{3}B_{1}$ CH₂ appeared. The prediction was swiftly verified by independent electron spin resonance experiments. For many chemists, the structure of triplet methylene was the first genuine example of the usefulness of ab initio theoretical chemistry.

Turning from the specific to the more general, the most encouraging aspect of ab initio geometry predictions is their, perhaps, surprising reliability. Essentially all molecular structures appear to be reliably predicted at the Hartree-Fock level of theory. Even more encouraging, many structures are accurately reproduced by using only minimumbasis-set self-consistent-field methods. This is especially true for hydrocarbons. A fairly typical example is methylenecyclopropane (**Fig. 1**), with

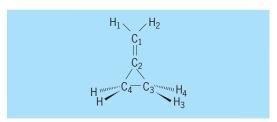


Fig. 1. Methylenecyclopropane structure.

Parameter*	STO-3G	Experimen
$r(C_1 = C_2)$	0.1298 nm	0.1332 nn
$r(C_2 - C_3)$	0.1474 nm	0.1457 nn
$r(C_3 - C_4)$	0.1522 nm	0.1542 nn
$r(C_1 - H_1)$	0.1083 nm	0.1088 nn
$r(C_3 - H_3)$	0.1083 nm	0.109 nm
$\theta(H_1C_1H_2)$	116.0°	114.3°
$\theta(H_3C_3H_4)$	113.6°	113.5°
$\theta(H_{34}C_3C_4)$	149.4°	150.8°

its minimum-basis-set self-consistent-field structure compared with experiment in **Table 2**. Carboncarbon bond distances differ typically by 0.002 nm from experiment, and angles are rarely in error by more than a few degrees. Even for severely strained molecules such as bicyclo[1.1.0] butane, very reasonable agreement with the experimental structure is obtained. It is noteworthy that experimental geometries are available for less than half of the C₄ hydrocarbons studied theoretically to date. Thus, for many purposes, theory may be considered complementary to experiment in the area of structure prediction.

For molecules including atoms in addition to C and H, minimum-basis-set self-consistent-field results are sometimes less reliable. For example, the F_2N_2 molecule has minimum-basis-set self-consistent-field bond distances $r_e(N-F) = 0.1384$ nm, $r_e(N=N) = 0.1214$ nm, which are respectively 0.0107 nm longer and 0.0169 nm shorter than the experimental values. Fortunately, vastly improved agreement with experiment results is obtained when a larger basis set is adopted for F_2N_2 .

In general, double-zeta self-consistent-field structure predictions are considerably more reliable than those based on minimum basis sets. A noteworthy exception is the water molecule, for which minimum-basis-set self-consistent-field yields a bond angle of 100.0° and double-zeta self-consistent-field predicts 112.6°, compared to the well-known experimental value of 104.5° . More typical are the HF and F₂ molecules, for which the minimum-basis-set, doublezeta, and experimental bond distances are 0.0956, 0.0922, and 0.0917 nm (HF); and 0.1315, 0.1400, and 0.1417 nm (F₂). In fact it can be argued that if the theory will not be pursued beyond the Hartree-Fock (single-configuration) approximation, doublezeta self-consistent-field is often a reasonable stopping point.

Transition states are typically more sensitive to basis set than equilibrium geometries. This is true because potential energy surfaces are often rather flat in the vicinity of a saddle point (transition state). An example is the carboxime-cyanic acid rearrangement given in reaction (13).

H

$$HONC \to HOCN \tag{13}$$

Minimum-basis-set self-consistent-field and doublezeta (technically 431G) self-consistent-field transition state geometries are compared in **Fig. 2**. There it is seen that the minimum-basis-set and doublezeta predictions are very different. Neither transitionstate structure is reliable, as is seen in Fig. 2 by comparison with results from more complete basis sets. It should be noted that for many other transition states, better agreement is found between the minimum and double-zeta basis sets. More typical structural variations are ~0.005 nm in internuclear separations and 5° in angles.

As larger basis sets within the Hartree-Fock formalism are considered, better agreement with experiment is frequently obtained. As implied above, the water molecule bond angle is much improved at the double-zeta plus polarization level, to 106.1° . However, it is often the case that adding polarization functions has only a marginal effect on predicted geometries. A reasonably typical comparison is given by the NH₂F and PH₂F molecules. In this example, the only polarization functions added were sets of d functions on the central N or P atom. The only pronounced improvement with respect to experiment is for the P-F separation in PH₂F, and this is improved by 0.0034 nm when d functions are added to phosphorus. The good agreement with experiment for NH₂F and PH₂F suggests a high degree of reliability for the comparable NHF2 and PHF2 predictions, where no experimental structures have been determined.

An interesting comparison of the three frequently used basis sets is given in **Table 3** for the linear HCNO molecule. The most sensitive geometrical parameter is the N—O bond distance, for which the

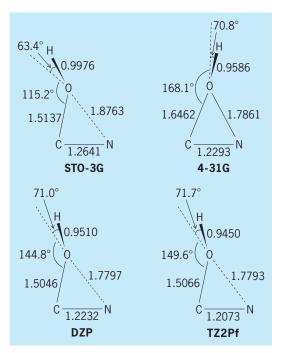


Fig. 2. Comparison of self-consistent-field transition states for the HONC \rightarrow HOCN reaction using four different basis sets.

TABLE 3. Equilibrium geometry of formonitrile oxide, HCNO, from self-consistent-field theory and experiment*						
	MBS	4-31G	DZP	Experiment		
r _e (H—C) r _e (C—N) r _e (N—O)	0.1065 0.1155 0.1294	0.1049 0.1133 0.1255	0.1059 0.1129 0.1201	0.1027 0.1168 0.1199		
*Values are in nanometers.						

minimum basis set is 0.0095 nm too long, the doublezeta still 0.0056 nm too long, but the double-zeta plus polarization result is in nearly perfect agreement with experiment. For the CN distance, the minimum-basis-set treatment actually gives the best agreement with experiment. The experimental microwave spectrum is difficult to unravel for a quasilinear molecule such as HCNO, and it has been suggested that the double-zeta plus polarization prediction for the C—H distance may be more reliable than experiment.

An example of a sizable molecule for which complete geometry optimization became routine in 1998, even using a substantial triple-zeta plus polarization basis set, is the digallium compound $Na_2{Ga_2[C_6H_3(C_6H_5)_2]_2}$ [Fig. 3]. The complete structural optimization (50 independent degrees of freedom) was carried out using the B3LYP method in conjunction with a basis set of 836 contracted gaussian functions.

Energetic predictions. Among the most chemically important energetic quantities are conformational energy changes, exothermicities or heats of reaction, dissociation energies, and activation energies or barrier heights. In general, only the first of these, and sometimes the second, is reliably predicted at the Hartree-Fock level of theory. In other words, energetic quantities are often sensitive to the effects of electron correlation.

Conformational energy changes are, almost without exception, properly reproduced within the Hartree-Fock formalism. In fact, certain types of barriers, typified by the ethane rotational barrier, are quite satisfactorily predicted at the minimum-basisset self-consistent-field level of theory. More sensitive problems, such as the ammonia inversion barrier and the rotational barrier of hydrogen peroxide, demand the inclusion of polarization basis functions.

Although Hartree-Fock exothermicities are often unreliable, there is at least one fairly large class of reactions for which consistently good agreement with experiment has been found. Generally speaking, heats of reaction for systems having closed-shell reactants and products are often predicted successfully. More specifically, even better agreement with experiment is found for isodesmic reactions, where the number of bonds of each type is conserved. In fact, reasonable predictions are often made at the minimum-basis-set self-consistent-field level for isodesmic reactions. Further, such information can sometimes be used indirectly (or in conjunction with other thermochemical information) to predict quantities that might be very difficult to evaluate by more straightforward ab initio methods.

The dissociation energies of covalent molecules are generally predicted poorly by singleconfiguration self-consistent-field methods. Certainly the best-known example is the F₂ molecule, for which the molecular Hartree-Fock energy lies about 1 eV above the Hartree-Fock energy of two fluorine atoms. This problem is often mistakenly attributed to the "perverse" nature of the fluorine atom. In fact, the near-Hartree-Fock dissociation energy of N₂ is 5.27 eV, only about half of the experimental value, 9.91 eV. In addition, the near-Hartree-Fock dissociation energy of O2 is 1.43 eV, only one-third of the experimental value, 5.21 eV. Thus the Hartree-Fock dissociation energies of covalent molecules are consistently much less than experiment.

Another frequent failing of the Hartree-Fock method is in the prediction of the barrier heights or activation energies of chemical reactions. However, it must be noted that there are many classes of reactions for which Hartree-Fock theory does yield meaningful barrier heights. Two well-studied examples are the isomerizations shown in reactions (14) and (15). For the HNC rearrangement, comparison

$$HNC \rightarrow HCN$$
 (14)

$$CH_3NC \rightarrow CH_3CN$$
 (15)

between self-consistent-field (40 kilocalories or 167 kilojoules) and configuration interaction (CI; 36 kcal or 151 kJ) barriers reveals good qualitative agreement. However, the inclusion of *d* functions in

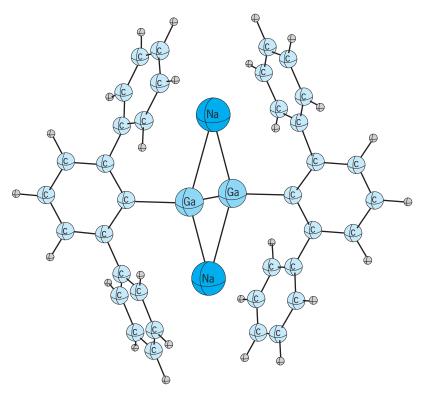


Fig. 3. Example of a sizable molecule for which complete DFT geometrical optimization with a substantial basis set became routine in 1998.

the C and N basis sets appears to be very important. For example, for the methyl isocyanide rearrangement, the self-consistent-field barrier decreases from 60 to 45 kcal (251 to 188 kJ) when polarization functions are added to a double-zeta basis. The remaining discrepancy with the experimental activation energy of 38 kcal (159 kJ) is about equally due to correlation effects and the fact that the zero-point vibrational energy at the transition state is ~3 kcal (13 kJ) less than for CH₃NC. Hydrogen isocyanide is one of the first three simple interstellar molecules (HNC, HCO⁺, and HN₂⁺) to be identified by ab initio theory prior to their laboratory detection.

Generally speaking, unimolecular reactions seem to be treated more reliably by Hartree-Fock methods than bimolecular systems. A second example is the geometrical isomerization of cyclopropane. This system has been studied in considerable detail by L. Salem and coworkers, who were the first to totally resolve the structure of the transition state within the full 21-dimensional hypersurface. Although their work involved only a minimum basis set, it goes slightly beyond the self-consistent-field model in that 3×3 configuration interaction was included. The predicted barrier height is 53 kcal (222 kJ), in reasonable agreement with the experimental value, 64 kcal (268 kJ). Subsequent studies with much larger basis sets show quantitative agreement with experiment.

For many attractive potential energy surfaces, that is, those having no barrier or activation energy at all, Hartree-Fock methods are frequently reliable. An example that has been carefully documented is reaction (16). One of the most important features of

$$H + Li_2 \rightarrow LiH + Li$$
 (16)

the H + Li₂ surface is the fact that the $C_{2\nu}$ HLi₂ structure is a chemically bound entity. Self-consistentfield theory suggests that the dissociation energy relative to LiH + Li is 20 kcal (84 kJ), in excellent agreement with the large-scale configuration interaction result of 22 kcal (93 kJ).

Hybrid Hartree-Fock/density functional methods such as B3LYP usually do a far better job than Hartree-Fock theory for thermochemistry. Often, diatomic molecule dissociation energies have errors an order of magnitude smaller with B3LYP than with Hartree-Fock. Moreover, the density functional methods are typically less sensitive to basis set than is Hartree-Fock theory. Thus it seems likely that the hybrid Hartree-Fock/density function methods will dominate quantum chemistry for the foreseeable future.

Beyond Hartree-Fock and density functional theory. The most important post-Hartree-Fock methods for quantum chemistry are perturbation theory and the configuration interaction (CI) and coupled cluster (CC) methods. These three rigorous approaches may be labeled "convergent" quantum-mechanical methods, as each is ultimately capable of yielding exact solutions to Schrödinger's equation. The present discussion focuses on coupled cluster theory. The coupled cluster method treats excitations based on the number of electrons by which they differ from the Hartree-Fock reference function. Thus the CCSD method incorporates amplitudes differing by single (S) and double (D) excitations from Hartree-Fock. The CCSDT method adds all triple excitations to the CCSD treatment. As one goes to higher and higher excitations (for example, CCSDTQ includes all configurations differing by one, two, three, or four electrons from the Hartree-Fock reference configuration), one approaches the exact quantum-mechanical result.

In fact, coupled cluster theory beyond CCSD becomes impractical for large molecular systems. Thus, although the coupled cluster path to exact results is clear, it becomes a difficult road to follow. Triple excitations are sufficiently important that effective coupled cluster methods have been developed in which the effects of triples are approximated. The best of these methods, CCSD(T), is the closest thing to a panacea that exists today in quantum chemistry for very difficult problems involving smaller molecules. However, the range of applicability of the theoretically superior CCSD(T) method is much narrower than that of the popular hybrid Hartree-Fock/density functional methods. For example, the DZP basis-set coupled cluster treatments of ferrocene, Fe(C₅H₅)₂, in 1995 and the small fullerene C20 in 1996 were considered state-of-the-art and required the use of supercomputers. In contrast, these systems were readily (though not as reliably) treated in the year 2000 with personal computers costing less than \$500 when methods such as B3LYP were employed. Thus, for most chemists the Hartree-Fock and density functional methods are likely to play central roles in molecular electronic structure theory for many years to come. See CHEMICAL BONDING; COMPUTA-TIONAL CHEMISTRY; MOLECULAR ORBITAL THEORY; MOLECULAR STRUCTURE AND SPECTRA; RESONANCE (MOLECULAR STRUCTURE); STRUCTURAL CHEMISTRY. Henry F. Schaefer III

Bibliography. R. F. Bader, Atoms in Molecules: A Quantum Theory, 1994; C. J. Barden and H. F. Schaefer, Quantum chemistry in the twenty-first century, Pure Appl. Chem., 72:1405-1423, 2000; I. N. Levine, Quantum Chemistry, 5th ed., 2000; J. P. Lowe, Quantum Chemistry, 3d ed., 2005; H. F. Schaefer, The reachable dream, J. Mol. Struc., 398:199-209, 1997; T. Veszpremi and M. Feher, Quantum Chemistry: Fundamentals to Applications, 1999; Y. Yamaguchi et al., A New Dimension to Quantum Chemistry, 1994.

Quantum chromodynamics

A theory of the strong (nuclear) interactions among quarks, which are regarded as fundamental constituents of matter, structureless and indivisible at the current resolution of about 10^{-18} m. Quantum chromodynamics (QCD) explains why quarks combine in certain configurations to form the observed patterns of subnuclear particles, such as the proton and pi meson. According to this picture, the strong interactions among quarks are mediated by a set of force

particles known as gluons. Quantum chromodynamics successfully accounts for many features of highenergy hard scattering of the strongly interacting particles. Interactions among gluons may lead to new structures that correspond to as-yet-undiscovered particles. The long-studied nuclear force that binds protons and neutrons together in atomic nuclei is regarded as a collective effect of the elementary interactions among constituents of the composite protons and neutrons. *See* NUCLEAR STRUCTURE.

By construction, quantum chromodynamics embodies many features abstracted from empirical observations, but the strength of the theory lies in the predictions it makes and the new understanding it brings. The property of asymptotic freedom implies a realm in which reliable calculations can be extracted using traditional Feynman-diagram techniques. These have been subjected to numerous quantitative tests with impressive success. The development of lattice gauge theory has made possible a growing range of computations of particle properties that emerge from the confinement of quarks. Part of the esthetic appeal of the theory is due to the fact that quantum chromodynamics is nearly identical in mathematical structure to quantum electrodynamics (QED) and to the unified theory of weak and electromagnetic interactions. This resemblance encourages the hope that a unified description of the strong, weak, and electromagnetic interactions may be at hand. See ELECTROWEAK INTERACTION; FEYN-MAN DIAGRAM; QUANTUM ELECTRODYNAMICS; WEAK NUCLEAR INTERACTIONS.

Gauge theories. At the heart of current theories of the fundamental interactions is the idea of gauge invariance. Gauge theories constructed to embody various symmetry principles have emerged as the correct quantum-mechanical descriptions of the strong, weak, and electromagnetic interactions, at energies up to at least 1 TeV (10^{12} eV). *See* SYMMETRY LAWS (PHYSICS).

Electromagnetism. The simplest example of a gauge theory is electromagnetism. It may be derived from a symmetry principle as follows. Quantum-mechanical observables do not depend upon the phase of the complex wave function that describes the state of a system. Therefore, the phase of a wave function can be rotated by an amount that is the same at all times and all places without affecting the physical consequences of the theory. The choice of phase is thus conventional, as opposed to observable. This is known as a global symmetry principle. It is natural to ask whether it should not be possible to choose this arbitrary convention independently at each point of space-time, again without affecting the physical consequences of the theory. It turns out to be possible to construct a quantum theory that is invariant under local (that is, position- and time-dependent) phase rotations that are proportional to the electric charge of the particles, but only if the theory contains an electromagnetic field with precisely the observed properties as summarized by Maxwell's equations. In the quantum theory, a massless spin-1 particle identified as the photon mediates the electromagnetic interaction. The interactions of matter with electromagnetism thus are essentially prescribed by the requirement of local phase invariance.

Local gauge invariance. Local phase rotations of the kind described above are the simplest examples of local gauge transformations. For a continuous symmetry, global gauge invariance implies the existence of a set of conserved currents. In the case of electromagnetism, it is the electromagnetic current that is conserved. A local gauge invariance requires, in addition, the existence of a massless gauge field corresponding to the set of phase transformations that forms the one-parameter unitary group U(1). The question arises as to whether the notion of local gauge invariance can be used to deduce the theory of nuclear forces.

Yang-Mills theory. This question was addressed in 1954 by C. N. Yang and R. L. Mills and, independently, by R. Shaw. Early in the study of nuclear forces it was established that the nuclear interaction is chargeindependent; it acts with the same strength between proton and proton, or proton and neutron, or neutron and neutron. This may be understood by thinking of the proton and neutron as representing two states of the same particle, called the nucleon. Just as an electron can be in a state with spin up or spin down, a nucleon can be in a state with the internal quantum number isospin up (defined as the proton) or isospin down (defined as the neutron). Charge independence then would reflect the invariance of the strong interactions under isospin rotations, characterized by the group SU(2). If isospin is regarded as a gauge group, local gauge invariance requires the existence of three massless spin-1 gauge particles, corresponding to the three generators of SU(2). The interactions of the gauge particles with nucleons are prescribed by the gauge principle. All of this is entirely parallel to the theory of electromagnetism. What distinguishes this SU(2) gauge theory from its U(1) counterpart is that the SU(2) gauge fields carry isospin and thus couple among themselves, whereas the photon is electrically neutral and does not interact with itself. Interacting gauge fields are an attribute of any theory based upon a nonabelian gauge group. See I-SPIN.

Spontaneous symmetry breaking. Its mathematical properties notwithstanding, the Yang-Mills theory was unacceptable as a description of nuclear forces because they are mediated by massive particles, whereas the gauge particles are required to be massless. Beginning in the late 1950s, a succession of gauge theories of the weak interactions appeared. At first these too foundered on the prediction of massless gauge bosons, but it was ultimately learned that spontaneous breakdown of the gauge symmetry would endow the gauge bosons with masses. All the elements were successfully combined in 1967 in the theory of weak and electromagnetic interactions proposed separately by S. Weinberg and A. Salam. When in 1971 it was demonstrated that spontaneously broken gauge theories were renormalizable, and hence calculable in the same sense as quantum electrodynamics, experimental interest in the predictions of the Weinberg-Salam model was stimulated, and theoretical enthusiasm for gauge theories in general was renewed. Spontaneous symmetry breaking was not a cure for the shortcomings of gauge theories of the strong interactions. Instead, thanks to parallel developments described below, a new candidate emerged for the strong gauge group, and with it arose the idea of quantum chromodynamics. *See* GAUGE THEORY; RENORMALIZATION; SYMMETRY BREAKING.

Color. It was shown in 1963 that the observed pattern of strongly interacting particles, or hadrons, could be explained if the hadrons were composed of fundamental constituents called quarks. According to the quark model, a baryon such as the proton, which has half-integral spin in units of \hbar (Planck's constant *b* divided by 2π), is made up of three quarks, each with spin 1/2. An integral-spin meson, such as the pi meson, is made up of one quark and one antiquark. Three varieties (or flavors) of quarks, denoted up, down, and strange, could be combined to make all of the known hadrons, in precisely the families identified according to the eightfold way.

Although the idea that the strongly interacting particles are built up of quarks brought new order to hadron spectroscopy and suggested new relations among mesons and baryons, the constituent description brought with it a number of puzzles. An obvious question concerns the rules by which the hadrons are built up out of quarks. Mesons are composed of a quark and an antiquark, while baryons are made of three quarks. What prevents two-quark or fourquark combinations? Within this innocent question lurks a serious problem of principle. The Pauli exclusion principle of quantum mechanics is the basis for understanding the periodic table of the elements. It restricts the configurations of electrons within atoms and of protons and neutrons within nuclei. It should be a reliable guide to the spectrum of hadrons as well. According to the Pauli principle, identical spin- $\frac{1}{2}$ particles cannot occupy the same quantum state. As a consequence, the observed baryons such as Δ^{++} (*uuu*) and Ω^{-} (sss), which would be composed of three identical quarks in the same state, would seem to be forbidden configurations.

To comply with the Pauli principle, it is necessary to make the three otherwise identical quarks distinguishable by supposing that every flavor of quark exists in three varieties, fancifully labeled by the colors red, green, and blue. Then each baryon can be constructed as a "colorless" (or "white") state of a red quark, a green quark, and a blue quark. Similarly, a meson will be a colorless quark-antiquark combination. The rule for constructing hadrons may then be rephrased as the statement that only colorless states can be isolated. *See* COLOR (QUANTUM ME-CHANICS).

A second issue is raised by the fact that free quarks have not been observed. This suggested that the interaction between quarks must be extraordinarily strong, and perhaps permanently confining. That free quarks are not seen is of course consistent with the idea that colored states cannot exist in isolation. On the other hand, the quark-parton model description of violent collisions rests on the assumption that quarks within hadrons may be regarded as essentially free, as explained below.

This paradoxical state of affairs may be visualized as follows. A hadron may be thought of as a bubble within which the constituent quarks are imprisoned. The quarks move freely within the bubble but cannot escape from it. This picture yields an operational understanding of many aspects of hadron structure and interactions, but it falls far short of a dynamical explanation for the puzzling behavior of quarks.

The quarks are the constituents of strongly interacting particles. The leptons, of which the electron and neutrino are the most common examples, are the fundamental particles that do not interact strongly. Each lepton flavor appears in only a single species. In other words, the leptons are colorless. In other respects, leptons resemble quarks: they are spin-1/2 particles that have no internal structure at the current limits of resolution. Color may therefore be regarded as the strong-interaction analog of electric charge. Color cannot be created or destroyed by any of the known interactions. Like electric charge, it is said to be conserved. *See* LEPTON.

In light of evidence that color could be regarded as the conserved charge of the strong interactions, it was natural to seek a gauge symmetry that would have color conservation as its consequence. An obvious candidate for the gauge symmetry group is the unitary group SU(3), now to be applied to color rather than flavor. The theory of strong interactions among quarks that is prescribed by local color gauge symmetry is known as quantum chromodynamics. The mediators of the strong interactions are eight massless spin-1 bosons, one for each generator of the symmetry group. These strong-force particles are named gluons because they make up the "glue" that binds quarks together into hadrons. Gluons also carry color (in eight color-anticolor combinations) and hence have strong interactions among themselves

Asymptotic freedom. The theoretical description of the strong interactions has historically been inhibited by the very strength of the interaction, which renders low-order perturbative calculations untrustworthy. However, in 1973 H. David Politzer, David J. Gross, and Frank Wilczek found that in many circumstances the effective strength of the interaction in Yang-Mills theories becomes increasingly feeble at short distances. For quantum chromodynamics, this remarkable observation implies that the interaction between quarks becomes weak at small separations. This discovery raises the hope that some aspects of the strong interactions might be treated by using familiar computational techniques that are predicated upon the smallness of the interaction strength.

The physical basis for the change in the strength of the strong interaction as a function of the distance may be understood by examining the corresponding question in electrodynamics. If an electric charge is placed in surroundings in which other

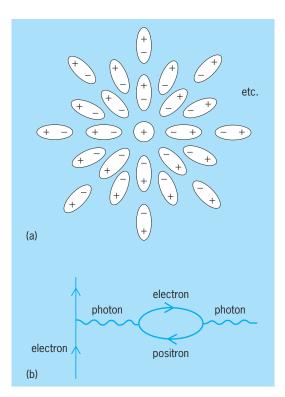


Fig. 1. Screening in electrodynamics. (a) Polarization of a dielectric medium by a test charge. (b) Feynman diagram contributing to vacuum polarization.

charges are free to move about, the effects of the charge may be modified. An example is a medium composed of many molecules, each of which has a positively charged end and a negatively charged end. Introducing a test charge polarizes the medium (**Fig. 1***a*). Charges of opposite sign are oriented toward the test charge, while those with like charge are repelled. This screening effect means that the influence of the charge is diminished by the surrounding medium. Viewed from afar (but within the medium), the charge appears smaller in magnitude than its true or unscreened value.

A closely related phenomenon in quantum electrodynamics is known as vacuum polarization (Fig. 1b). The vacuum, or empty space, is normally thought of as the essence of nothingness. However, in quantum theory the vacuum is a complicated and seething medium in which "virtual" pairs of charged particles, predominantly electrons and positrons, have a fleeting existence. These ephemeral vacuum fluctuations are polarizable in the same way as the molecules of the above example, and they serve to screen the charge at large distances. Consequently, in quantum electrodynamics it is also expected that the effective electric charge should increase at short distances. The behavior of the effective charge in quantum electrodynamics is opposite to that required to understand the strong interactions, where the interaction between quarks must diminish in strength at short distance.

The situation in quantum chromodynamics is complicated by the fact that gluons carry the strong (color) charge, whereas photons are electrically neutral. This means that, in addition to a color polarization phenomenon like the related charge screening of quantum electrodynamics (Fig. 2a and b), it is possible for the color charge of a quark to be shared with the gluon cloud (Fig. 2c and d). Because the color charge is spread out rather than localized, the effective color charge will tend to appear larger at long distances and smaller at short distances. The outcome of the competition between these opposing tendencies depends on the number of gluon species that can share the color charge and on the number of quark types that can screen the color charge. If the color gauge group is SU(3), the net effect is one of antiscreening, that is, of a smaller effective charge at short distances, provided the number of quark flavors is less than 17. Six quark flavors are known. Extremely

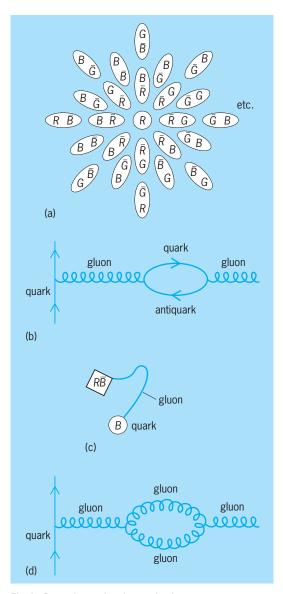


Fig. 2. Screening and antiscreening in quantum chromodynamics. (a) Screening of a colored quark by vacuum polarization. Color charges are denoted by $R = \text{red}, \overline{R} = \text{anti-red}, B = \text{blue}, \overline{B} = \text{anti-blue}, G = \text{green}, \overline{G} = \text{anti-green}.$ (b) Feynman diagram contributing to color polarization of the vacuum. (c) Delocalization of the color charge by gluon radiation. (d) Feynman diagram contributing to color antiscreening.

close to a quark, the effective color charge becomes vanishingly small, so that nearby quarks behave as if they were noninteracting free particles. This is the origin of the term asymptotic freedom.

Experimental consequences. Asymptotic freedom has the immediate consequence, within the context of quantum chromodynamics, of providing a partial justification of the parton model description of violent scattering processes. The parton model was invented to explain features of electron-proton collisions in which many particles are created. Experiments first carried out in 1968 indicated that in these deeply inelastic collisions the electron "sees" the proton, not as an amorphous whole, but as a collection of structureless entities that have been identified as quarks. According to this picture, an electron scatters from a single parton, free from the influence of any other parton. This is reminiscent of electron scattering from nuclei, in which an electron may scatter from an individual proton or neutron as if it were a free particle. There is an important difference. Protons and neutrons are lightly bound in nuclei and may be liberated in violent collisions. In contrast, free quarks have not been observed, so they must be regarded as very deeply bound within hadrons. Asymptotic freedom offers a resolution to the paradox of quasifree quarks that are permanently confined. At the short distances probed in deep inelastic scattering, the effective color charge is weak, so the strong interactions between quarks can largely be neglected. As quarks are separated, the effective color charge grows, so infinite energy would be required to extract a quark from an isolated hadron. This characteristic offers a mechanism for quark confinement, but an analytic proof has not yet been constructed. A critical insight of quantum chromodynamics is that most of a proton's mass arises from the energy stored up in confining its component quarks in a small volume, not from the masses of the quarks.

In the regime of short distances probed in violent high-energy collisions, the strong interactions are sufficiently feeble that reaction rates may be calculated by using the diagrammatic methods developed for quantum electrodynamics. These calculations reproduce the simple quark-parton model results as first approximations. For example, in electron-positron annihilations into hadrons, the quark-antiquark production rate correctly anticipates both the structure of the dominant two-jet events and the approximate rate of hadron production. The strong-interaction corrections to this process include the process in which a gluon is radiated by one of the outgoing quarks. Like the quarks, the gluon materializes as a jet of hadrons. The resulting three-jet events are commonplace in electronpositron annihilations.

In high-energy proton-antiproton collisions, the hard scatterings of the partons (**Fig. 3***a*) lead to striking jets of hadrons at large angles to the direction defined by the incident proton and antiproton beams (Fig. 3*b*). Quantum chromodynamics, supplemented by measurements of the parton distributions within protons, accounts for the features of jet production in great detail.

Although quantum chromodynamics provides support for the spirit of the parton model, it also exposes the incompleteness of the parton model description. Because quantum chromodynamics is an interacting field theory of quarks and gluons, probes of different wavelengths, analogous to microscopes of different resolving power, may map out different structures within the proton. Under increasing magnification (**Fig. 4**), quarks undergo virtual dissociation into quarks and gluons. The structure of the proton resolved in high-energy scattering experiments changes in the way predicted by quantum chromodynamics.

Quarkonium. The light hadrons are sufficiently large that the forces between the quarks within them

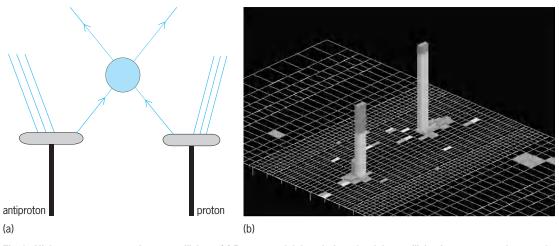


Fig. 3. High-energy proton-antiproton collisions. (a) Parton-model description of a violent collision between constituents of a proton and an antiproton. The active particles may be quarks, antiquarks, or gluons. (b) The most energetic two-jet event yet observed, as recorded in the Collider Detector at Fermilab. A proton and antiproton, each with energy 980 GeV, collide head on. In this event, a quark from the proton scatters at large angle off an antiquark from the antiproton. Two jets of hadrons emerge, carrying 1364 GeV at right angles to the initial beam directions. The event display depicts the energy deposited in individual cells of the cylindrical detector, unrolled. (*Courtesy CDF Collaboration*, http://www-cdf.fnal.gov)

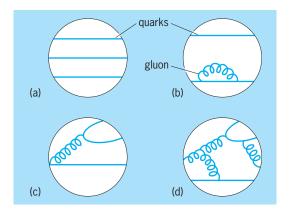


Fig. 4. Structure of the proton as observed with probes of increasing resolution. (a) Quarks appear as simple and noninteracting objects under "low magnification." (b-d) Quarks reveal increasing complexity because of their interactions with gluons as resolving power is improved.

are quite formidable. A fundamental description of the spectroscopy of light hadrons therefore relies on large-scale computer simulations. It is, however, possible to imagine special situations in which hadron spectroscopy is completely tractable by using available theoretical techniques. It was suggested in 1974 that the bound system of an extremely massive quark with its antiquark would be so small that the strong force would be extremely feeble. In this case, the binding between quark and antiquark is mediated by the exchange of a single massless gluon, and the spectrum of bound states resembles that of an exotic atom composed of an electron and an antielectron (positron) bound electromagnetically in a Coulomb potential generated by the exchange of a massless photon (**Fig. 5***a*). Since the electron-positron atom is known as positronium, the heavy quark-antiquark atom has been called quarkonium. Two families of heavy quark-antiquark bound states, the ψ/J system composed of charmed quarks (Fig. 5*b*) and the Υ system made up of *b* quarks, have been discovered. Both have level schemes characteristic of atomic spectra, which have been analyzed by using tools of nonrelativistic quantum mechanics developed for ordinary atoms. The atomic analogy has proved extremely fruitful for studying the strong interaction. *See* CHARM; J/PSI PARTICLE; POSITRONIUM.

Lattice models. To deal with the existence and properties of the hadrons themselves, it is necessary to devise a new computational approach that does not break down when the interaction becomes strong. The most promising method has been the crystal lattice formulation of the theory. By localizing quarks on individual lattice sites, it is possible to use many of the Monte Carlo simulation techniques developed in statistical physics for the study of spin systems such as magnetic substances. Modern calculations include the influence of dynamical quarkantiquark pairs. Lattice quantum chromodynamics analyses support the color-confinement paradigm. *See* ISING MODEL; STATISTICAL MECHANICS.

Quark-gluon plasma. It seems likely that when hadronic matter is compressed to very great densities and heated to extremely high temperatures hadrons will lose their individual identities. When the hadronic bubbles overlap and merge, quarks and gluons may be free to migrate over great distances.

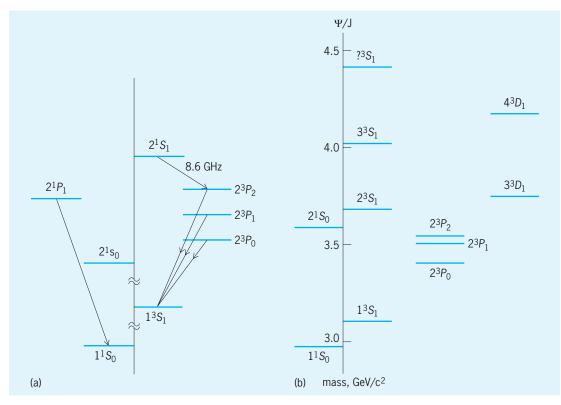


Fig. 5. Spectra of bound particle-antiparticle systems. (a) Positronium. (b) ψ/J family, known as charmonium.

The resulting new states of matter may exist in the cores of collapsing supernovae and neutron stars. The possibility of creating new quantum chromodynamic states of matter in the laboratory is under active experimental study in collisions of relativistic heavy ions. *See* NEUTRON STAR; QUARK-GLUON PLASMA; SUPERNOVA.

Grand unification. Because quantum chromodynamics and the unified theory of weak and electromagnetic interactions have the same gauge theory structure, it is appealing to speculate that all three interactions have a common origin in a single gauge symmetry. The construction of such grand unification theories is a very active area of theoretical research. *See* ELEMENTARY PARTICLE; FUNDAMENTAL INTERACTIONS; GLUONS; GRAND UNIFICATION THEORIES; QUANTUM FIELD THEORY; QUARKS; STANDARD MODEL; STRONG NUCLEAR INTERACTIONS. C. Quigg

Bibliography. I. J. R. Aitchison and A. J. G. Hey, *Gauge Theories in Particle Physics*, vol. 2, 3d ed., Taylor & Francis, London, 2002; R. K. Ellis, W. J. Stirling, and B. R. Webber, *QCD and Collider Physics*, Cambridge University Press, 1996, paper 2003; I. Montvay and G. Münster, *Quantum Fields on a Lattice*, Cambridge University Press, 1994, paper 1997; T. Muta, *Foundations of Quantum Chromodynamics*, 2d ed., World Scientific, Singapore, 1998; Y. Nambu, The confinement of quarks, *Sci. Amer.*, 235(5):48-60, November 1976; C. Quigg, *Gauge Theories of the Strong, Weak, and Electromagnetic Interactions*, Perseus/Westview, New York and Boulder, 1983, reprint 1997.

Quantum computation

Whereas classical computers perform operations on classical bits, which can be in one of two discrete states, 0 or 1, quantum computers perform operations on quantum bits, or qubits, which can be put into any superposition of two quantum states, $|0\rangle$ and $|1\rangle$. Quantum computers hold the promise to perform classically intractable tasks-computations that require astronomical time or hardware resources on the fastest classical computer-in minuscule time with minuscule resources. The building blocks of a quantum computer are atomic particles which, following the laws of quantum physics, behave like waves and exhibit interference phenomena as if they were in several locations at the same time. By equating different locations-for example, an electron in the lowest orbit or in an excited orbit of an atomto binary digits 0 or 1, one may interpret the timeevolving state of the particles as executing several computations at the same time. One set of locations at a given time describes the result of one computation. Thus one atom can do two computations at once; two atoms can do four; three atoms can do eight. The challenge is to coerce the atoms to follow trajectories that amount to meaningful computations and to read out a definite result from the multitude of computations occurring in parallel. The control of trajectories is the hardware part of the challenge; the design of useful trajectories-algorithms that are

superior to classical algorithms—is the software part of the challenge. *See* DIGITAL COMPUTER; SUPERCOMPUTER.

Principles. A classical computer manipulates strings of N classical bits, (n_1, \ldots, n_N) with $n_i = 0$ or 1 (j = 1, ..., N), in such a way that intermediate states of the computation are also strings of classical bits. A quantum computer manipulates states of N two-level atoms, nuclear spins, or other entities, $|n_1,\ldots,n_N\rangle$, with $n_j = 0$ if the *j*th atom is in the ground state and $n_i = 1$ if it is in the excited state, in such a way that intermediate states are superpositions of the states $|n_1, \ldots, n_N\rangle$. The 2^N states $|n_1, \ldots, n_N\rangle$ (the computational basis) are product states in which each atom is in either the ground state or excited state, and n_i is called the value of the *j*th qubit; they represent the strings of classical bits. The superpositions include states in which an atom no longer has a sharp value of n_i (it has indefinite bit value), and states in which an atom no longer exists in a state separate from the other atoms (entangled states); both have no classical counterpart. A quantum computation starts with a product state $|n_1, \ldots, n_N\rangle$; allows the state to evolve according to the Schrödinger equation (1),

$$i\hbar \frac{d}{dt} |\psi(t)\rangle = H(t) |\psi(t)\rangle \tag{1}$$

with initial condition $|\psi(0)\rangle = |n_1, \ldots, n_N\rangle$ and time-dependent Hamiltonian *H*(*t*) [energy operator] driving the coupled atoms; and ends with the measurement of the values of the qubits of the state $|\psi(t)\rangle$. The Hamiltonian generates the unitary time evolution operator U(t) which takes the initial state into the final state according to $|\psi(t)\rangle =$ $U(t) |\psi(0)\rangle$. The measurement transforms $|\psi(t)\rangle$ into the output state $|n'_1, \ldots, n'_N\rangle$ with probability $|\langle n'_1, \ldots, n'_N | \psi(t) \rangle|^2$. The output is probabilistic because quantum measurements are so. The computation $|n_1,\ldots,n_N\rangle \rightarrow |\psi(t)\rangle$ is a unitary transformation, hence reversible; the readout $|\psi(t)\rangle \rightarrow |n'_1, \ldots, n'_N\rangle$ is a projection (called the collapse of the wave function), hence irreversible. Thus, to perform a specific computation, one must drive the atoms with a specific Hamiltonian; to read out the result, one must send the atoms through a series of state detectors. See BIT; SCHRÖDINGER'S WAVE EQUATION.

Power of quantum computation. A quantum computer is more powerful than a classical computer for two reasons:

1. The quantum state space is much larger than the classical state space: *N* qubits can be in an infinite number of different states (any point on the unit sphere of the complex Hilbert space spanned by the 2^N basis vectors $|n_1, ..., n_N\rangle$); *N* classical bits can be in only 2^N different states (the points where the 2^N coordinate axes intersect the unit sphere; for N = 1: **Fig. 1**). If states on the unit sphere can be resolved to within accuracy ε ($\varepsilon \ll 1$), the sphere hosts a number of the order of $\varepsilon^{-(2N-1)}$ [written $O(\varepsilon^{-(2N-1)})$] distinct states. Thus a quantum computer can store and access an exponentially large number of states compared to a classical computer. 2. The quantum computer operates in a massively parallel way: If the initial state is the uniform superposition of all basis states, Eq. (2), the time evolution

$$|\psi(0)\rangle = 2^{-N/2} \sum_{n_1,\dots,n_N=0,1} |n_1,\dots,n_N\rangle$$
 (2)

computes simultaneously $U(t)|n_1, \ldots, n_N\rangle$ for all 2^N possible inputs $|n_1, \ldots, n_N\rangle$ by linearity of U(t). The matrix element $\langle n'_1, \ldots, n'_N | U(t) | n_1, \ldots, n_N \rangle$ is the probability amplitude that the computation converts the input $|n_1, \ldots, n_N\rangle$ into the output $|n'_1, \ldots, n'_N\rangle$, along all possible classical computational paths in parallel (as given by Feynman's path integral). A classical computation can follow only a single path. The aim is to choose U(t) so that computational paths of no interest cancel each other by destructive interference, and paths of interest add constructively, that is, to select relevant computations by quantum interference (**Fig. 2**). *See* FEYNMAN INTEGRAL; INTER-FERENCE OF WAVES; SUPERPOSITION PRINCIPLE.

Logic gates. Any computation U(t), also called a quantum circuit, can be approximated by sequential application of a finite set of unitary transformations that operate on only one or two qubits. Such transformations are called logic gates. A logic gate assigns to any input, consisting of a one-qubit or two-qubit state, a unique output, in terms of a one-qubit or two-qubit state. An example for such a universal set of quantum logic gates are the unitary transformations given in Eqs. (3) [Hadamard gate], (4) [*T* gate], and (5) [controlled-not gate], where $|n_i\rangle \langle n'_i|$ acts only on

$$H_{j} = (|0_{j}\rangle\langle 0_{j}| + |0_{j}\rangle\langle 1_{j}| + |1_{j}\rangle\langle 0_{j}| - |1_{j}\rangle\langle 1_{j}|)/\sqrt{2}$$
(3)

$$T_j = |0_j\rangle\langle 0_j| + e^{i\pi/4}|1_j\rangle\langle 1_j| \tag{4}$$

 $C_{jk} = |0_j 0_k \rangle \langle 0_j 0_k | + |0_j 1_k \rangle \langle 0_j 1_k |$ + |1_j 0_k \rangle \langle 1_j 1_k | + |1_j 1_k \rangle \langle 1_j 0_k | \qquad (5)

qubit j, and $|n_i n_k\rangle \langle n'_i n'_k|$ only on qubits j and k. They correspond to the logic gates in a classical computer, but are reversible, unlike the classical "and" and "exclusive or" gates, which are irreversible. A logic gate is called reversible if one can reconstruct the input when one knows the output. The Hadamard gate transforms the states $|0_i\rangle$ and $|1_i\rangle$ into the superpositions $(|0_i\rangle \pm |1_i\rangle)/\sqrt{2}$; the T gate shifts the phase of the excited state relative to the ground state by n/4; and the controlled-not gate flips the target qubit k if and only if the control qubit *j* is in the excited state. The three gates are the analog of an optical beam splitter, which splits an incoming wave into a transmitted wave and a diffracted outgoing wave; a phase shifter (refractive medium); and a conditional halfwave plate, respectively. Only the Hadamard gate creates multiple computational paths; the other two transform a single basis state into a single basis state. To approximate a general U(t) to within accuracy ε requires $O(4^N N^2 [\ln (4^N N^2 / \varepsilon)]^s)$ gates, where $s \approx$ 2 (this is the Solovay-Kitaev theorem); the leading factor 4^N is set by the number of matrix elements of U(t). For special computations, however, often a

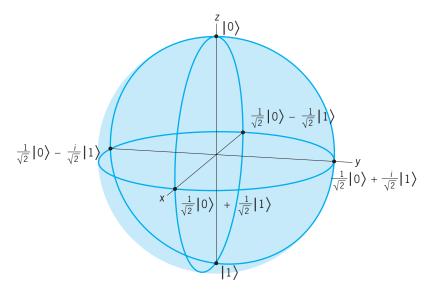


Fig. 1. A classical bit can be in only state $|0\rangle$ or $|1\rangle$ (the "north" and "south" poles of the sphere). A quantum bit can be in any superposition $a|0\rangle + b|1\rangle$, where $|a|^2 + |b|^2 = 1$ (an arbitrary point on the sphere). The sphere is called the Bloch sphere, and the polar and azimuthal angles, θ and φ , of a point on the sphere are given by $a = \cos(\theta/2)$ and $b = \exp(i\varphi) \sin(\theta/2)$.

much smaller number of gates, independent of ε , suffices. *See* LOGIC CIRCUITS.

Fast quantum algorithms. A widely used class of encryption systems for public transmission of sensitive data derive their security from the difficulty of factoring a large, publicly transmitted integer. The fastest known classical algorithm factors an N-bit number in time $O[\exp(\text{const} \times N^{1/3}(\ln/\ln N)^{2/3})]$. Peter Shor's quantum algorithm for the same task requires only time $O[N^2 (\ln N)(\ln N)]$. The key task in Shor's algorithm is to find the period of a function related to the number to be factored, and the algorithm does this using the discrete Fourier transform: The values at which the transform does not vanish yield the period-similarly to the way in which Bragg peaks in x-ray diffraction, at which the x-ray intensity does not vanish, yield the crystal periodicity. The quantum fast Fourier transform, with no approximation, requires $O(N^2)$ operations, which is much less than the $O(2^N N)$ operations in the classical fast Fourier transform and much less than the Solovay-Kitaev bound. The period is obtained from $O(\ln N)$ repeated measurements of the qubits of the state representing the

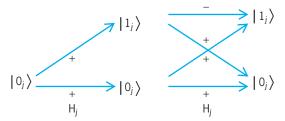


Fig. 2. Application of the Hadamard gate, Eq. (3), to state $|n_j\rangle$ creates two computational paths, one leading to state $|0_j\rangle$ with probability amplitude $1/\sqrt{2}$, and the other to $|1_j\rangle$ with amplitude $1/\sqrt{2}$ if $n_j = 0$, and amplitude $-1/\sqrt{2}$ if $n_j = 1$. Application of the Hadamard gate twice to $|0_j\rangle$ creates four paths, leading to $|0_j\rangle$ with amplitude (+1)(+1)/2 + (+1)(-1)/2 = 1, and to $|1_j\rangle$ with amplitude (+1)(+1)/2 + (+1)(-1)/2 = 0 (cancellation of trajectories leading to $|1_j\rangle$).

Fourier transform. *See* CRYPTOGRAPHY; FOURIER SERIES AND TRANSFORMS; X-RAY DIFFRACTION.

Other quantum algorithms that outperform classical algorithms by orders of magnitude are Lov K. Grover's algorithm for "finding a needle in a haystack" (search of an item in a database of 2^N items) in $O(2^{N/2})$ instead of $O(2^N)$ steps; estimation of the median and mean of 2^N items to precision ε in $O(N/\varepsilon)$ instead of $O(N/\varepsilon^2)$ steps; search of the minimum of a function sampled at 2^N points in $O(2^{N/2})$ instead of $O(2^N)$ steps; search of two distinct pre-images giving the same image of a two-to-one function sampled at 2^N points, in $O(2^{N/3})$ instead of $O(2^{N/2})$ steps; the Deutsch-Jozsa algorithm to determine if either [1] 2^N numbers are either all 0 or all 1 (a constant function), or [2] half of these numbers are 0 and half are 1 (a balanced function), in one instead of up to $2^{N-1} + 1$ steps; and various allocation tasks and game-theoretic strategies. For N = 1, the Deutsch-Jozsa algorithm can be visualized as putting two slabs of material, which may be thin ("0") or thick ("1"), into the apertures of a double-slit apparatus and measuring the interference pattern of a light wave passing through the two slits. If the pattern is unchanged compared to the empty apparatus, the slabs have the same thickness (constant function); or else they have different thickness (balanced function).

Quantum error correction. Noise from imperfect computer operation poses no fundamental barrier to large-scale computations. A quantum errorcorrection code encodes the N logical qubits into N' carrier qubits (N' > N); runs the carrier qubits through a group of accordingly encoded logic gates; transforms the noisy state by appropriate projection operators (which perform syndrome measurements to detect errors) and unitary operators (which recover the original carrier qubits) into an errorcorrected state; and feeds the state into the next group of gates. Such periodic error correction prevents accumulation of errors in the state. At the end of the computation, the carrier qubits are decoded. The encoding spreads the state of the *N*logical qubits over all N' carrier qubits so that, when the syndromes (which provide information about errors) are measured, no information about the state of the logical qubits is revealed: The projections preserve superpositions of the logical qubits, and the state of the carrier qubits is highly entangled even if the logicalqubit state is not. Remarkably, a discrete set of corrections can correct a continuum of errors. For a code to correct any error on any M carrier qubits, a necessary condition is $N' \ge 4M + N$ (the Knill-Laflamme bound), and a sufficient condition for large N is given by Eq. (6) [the Gilbert-Varshamov bound]. For exam-

$$N/N' < 1 - 2[-x\log_2 x - (1-x)\log_2(1-x)]_{x=2M/N'}$$
(6)

ple, a code exists that encodes 1 logical qubit (N = 1) into 5 carrier qubits (N' = 5) and corrects any error on any 1 carrier qubit (M = 1) with 16 pairs of syndrome measurements and recovery operations (thus achieving equality in the Knill-Laflamme bound).

Different codes require different encodings of gates, and of interest are encodings for which an error in an input carrier qubit or in the gate operation propagates only to a small number of output carrier qubits. Specifically, an encoded gate is called fault-tolerant if a failure with probability p of any single component (for example, one of the 10 "wires" feeding two logical qubits, each encoded by 5 carrier qubits, into a controlled-not gate) introduces an error in 2 or more carrier qubits in any logical output qubit with probability cp^2 at most, with *c* a constant and p small. Such a gate, when followed by error correction with M = 1, yields an error-free output with probability $1 - cp^2$; that is, it reduces the error probability from p to cp^2 if p < 1/c. Fault-tolerant Hadamard, T, and controlled-not gates exist with $c \approx$ 10^4 and $d \approx 10^2$, where d is the number of operations on carrier qubits needed to encode and error-correct the gate. By hierarchical fault-tolerant encoding of all gates, a computation involving L gates can be carried out to within accuracy ε using only $O(L[\ln(L/\varepsilon)]^{\log_2 d})$ operations on carrier qubits if p < 1/c. This is the threshold theorem for quantum computation. Thus, if the noise in individual carrier qubits is low enough, so that $p < 1/c \approx 10^{-4}$, arbitrarily large computations can be performed because the overhead for error correction grows only polynomial-logarithmically with the size of the computation, L. See FAULT-TOLERANT SYSTEMS; INFORMATION THEORY.

Experimental state of the art. A remarkable array of experimental realizations of quantum computing devices and implementations of algorithms have been achieved. As of 2006, records with respect to the number of qubits that can be controlled and prepared in well-defined states are a 12-qubit nuclear magnetic resonance device (NMR, four ¹H, six ¹³C, and two ¹⁴N nuclei in *l*-histidine, each in its spinup or spin-down state, driven by radio-frequency pulses), an 8-qubit ion-trap device (IT, eight ⁴⁰Ca⁺ ions in a linear electromagnetic trap, each in the electronic ground state or an excited state, driven by optical pumping), a 3-qubit device based on cavity quantum electrodynamics [CQED, three rubidium (Rb) atoms, each in one of two Rydberg states and coupled to a cavity mode with zero or one photon, driven by microwave pulses], various 2-qubit devices based on superconducting circuits (SC), and a 1qubit quantum-dot device [QD, two gallium arsenide (GaAs) quantum dots in a spin singlet or triplet state, driven by nanosecond voltage pulses]. IT, SC, and QD devices have the potential for scalable or solidstate architectures (less-than-exponential growth of the physical hardware necessary to manipulate large numbers of qubits). Successful computations carried out include Shor's algorithm on a 7-qubit NMR device, factoring the number 15 with a sequence of 300 radio-frequency pulses; the Deutsch-Jozsa algorithm on a 5-qubit NMR device; finding the order of a permutation on a 5-qubit NMR device; correction of any one-qubit error on a 5-qubit NMR device; spin-flip error correction on a 3-qubit IT device; and Grover's algorithm on a 2-qubit IT device. See CAVITY RESONATOR; NUCLEAR MAGNETIC RESONANCE (NMR); PARTICLE TRAP; QUANTIZED ELECTRONIC STRUCTURE

(QUEST); QUANTUM ELECTRODYNAMICS; RYDBERG ATOM; SUPERCONDUCTIVITY.

General hardware requirements. Any experimental realization faces three challenges:

1. The device must be able to control the state of each qubit separately, while allowing neighboring qubits to interact with each other during the operation of 2-qubit gates. Control of individual qubits is achieved by different chemical shifts in NMR, by laser beams driving ions spatially separated by tens of micrometers in IT, and by pulses addressing successive atoms traveling at spatial separation of several centimeters in CQED. Qubits interact via spin-spin coupling in NMR, via a shared phonon state of the ions in IT, and via electric dipole coupling between each atom and the cavity mode in CQED.

2. The system must be switchable, so that interactions executing a prescribed sequence of gate operations [the Hamiltonian H(t)] can be turned on and off by external control. Switching is done by magnetic field pulses in NMR, and by laser pulses in IT and CQED.

3. The computer must be well isolated from the environment so that the decoherence time t_d , the time at which the computer and the environment depart significantly from a product state (entanglement of the computer with the environment), is long compared to t_g , the time it takes to operate a single gate. At time t_d , a generic qubit state $a|0_i\rangle + b|1_i\rangle$ will have degraded into the mixture with the density matrix operator $|a|^2 |0_i\rangle \langle 0_i| + |b|^2 |1_i\rangle \langle 1_i|$, which no longer contains the interference terms necessary for quantum computation. Good isolation is provided by long spin-spin and spin-lattice relaxation times in NMR ($t_d = 10^{-2}$ to 10^8 s, $t_g = 10^{-6}$ to 10^{-3} s), longlived states and stable trap and laser operation in IT $(t_d = 10^{-1} \text{ to } 10^0 \text{ s}, t_g = 10^{-7} \text{ to } 10^{-5} \text{ s})$, low spontaneous emission rate of Rydberg states and low photon escape rate from the cavity in CQED ($t_d = 10^{-3}$ to 10° s, $t_g = 10^{-5}$ to 10^{-4} s), and spin-echo pulses in QD $(t_d = 10^{-6} \text{ to } 10^{-5} \text{ s}, t_g = 10^{-8} \text{ to } 10^{-7} \text{ s}).$ See DENSITY MATRIX; QUANTUM THEORY OF MEASUREMENT.

But the condition p < 1/c in the threshold theorem is more demanding than $t_g < t_d$. Let $|\psi_i(t)\rangle$ be the state of the computer and environment interacting with each other, starting from a product state at t = 0; let $|\psi_n(t)\rangle$ be the same state, but noninteracting (product state at all times); and assume the overlap $F(t) = |\langle \psi_i(t) | \psi_n(t) \rangle|$ (called the fidelity) decays exponentially with time. The decay constant is the decoherence time, $F(t) = \exp(-t/t_d)$, and the threshold condition requires $1 - \exp(-2t_g/t_d) < 1/c$ [because $p = 1 - F^2(t_g)$], giving $t_g < t_d/(2c)$ for large c. Thus the computer should remain isolated from the environment for about 10^4 gate operations for sustained computations. Present devices can execute about 300 operations and thus are far from this goal.

Outlook. Numerous other approaches have been proposed and are under investigation, for quantum computers with much longer decoherence times and control over many more qubits (a long-term goal is 200 qubits). These include solid-state NMR; solid-state CQED; electron spin resonance of donor atoms in semiconductors; electronic excitation in

quantum dots; ion traps with hot ions; electrons trapped on superfluid helium films; neutral atoms, including Bose-Einstein condensates, trapped in an optical lattice; linear optics (beam splitters, phase shifters, and single-photon sources); and topological quantum computing, using braided world lines of quasi-particles in the fractional quantum Hall effect as decoherence-free qubits. With so many lines of developments, quantum computing holds a rich future. *See* BOSE-EINSTEIN CONDENSATION; ELECTRON PARA-MAGNETIC RESONANCE (EPR) SPECTROSCOPY; HALL EFFECT; LIQUID HELIUM; NONRELATIVISTIC QUANTUM THEORY; QUANTUM MECHANICS. Peter Pfeifer

Bibliography. G. P. Berman et al., Introduction to Quantum Computers, World Scientific, Singapore, 1998; H. O. Everitt (ed.), Experimental Aspects of Quantum Computing, Springer, New York, 2005; G. Greenstein and A. G. Zajonc, The Quantum Challenge, 2d ed., Jones and Bartlett, Boston, 2006; A. Yu. Kitaev, A. H. Shen, and M. N. Vyalyi, Classical and Quantum Computation, American Mathematical Society, Providence, 2002; M. Nakahara et al. (eds.), Physical Realizations of Quantum Computing: Are the DiVicenzo Criteria Fulfilled in 2004?, World Scientific, Hackensack, NJ, 2006; M. A. Nielsen and I. L. Chuang, Quantum Computation and Quantum Information, Cambridge University Press, 2000; W.-H. Steeb and Y. Hardy, Problems and Solutions in Quantum Computing and Quantum Information, 2d ed., World Scientific, Hackensack, NJ, 2006.

Quantum electrodynamics

The field of physics that studies the interaction of electromagnetic radiation with electrically charged matter within the framework of relativity and quantum mechanics. It is the fundamental theory underlying all disciplines of science concerned with electromagnetism, such as atomic physics, chemistry, biology, the theory of bulk matter, and electromagnetic radiation.

Quantum mechanics and relativity. Efforts to formulate quantum electrodynamics (QED) were initiated by P. A. M. Dirac, W. Heisenberg, and W. Pauli soon after quantum mechanics was established. The first step was to remedy the obvious shortcoming of quantum mechanics: that it applies only to the case where particle speeds are small compared with that of light, c. This led to Dirac's discovery of a relativistic wave equation, in which the wave function has four components and is multiplied by certain 4×4 matrices. His equation incorporates in a natural manner the observed electron-spin angular momentum, which implies that the electron is a tiny magnet. The strength of this magnet (magnetic moment) was predicted by Dirac and agreed with observations then available. A detailed prediction of the hydrogen spectrum was also in good agreement with experiment. See ATOMIC STRUCTURE AND SPECTRA; ELECTRON SPIN; MATRIX THEORY.

In order to go beyond this initial success and calculate higher-order effects, however, the interaction of charge and electromagnetic field had to be treated dynamically. To begin with, a good theoretical framework had to be found for describing the waveparticle duality of light, that is, the experimentally well-established fact that light behaves like a particle (photon) in some cases but like a wave in others. Similarly, the electron manifests wave-particle duality, another observed fact. Once this problem was settled, the next question was how to deal with the interaction of charge and electromagnetic field. It is here that the theory ran into severe difficulties. Its predictions often diverged when attempts were made to calculate beyond lowest-order approximations. This inhibited the further development of the theory for nearly 20 years. Stimulated by spectroscopic experiments vastly refined by microwave technology developed during World War II, however, S. Tomonaga, R. P. Feynman, and J. Schwinger discovered that the difficulties disappear if all observable quantities are expressed in terms of the experimentally measured charge and mass of the electron. With the discovery of this procedure, called renormalization, quantum electrodynamics became a theory in which all higher-order corrections are finite and well defined. See NONRELATIVISTIC QUANTUM THE-ORY; PHOTON; QUANTUM MECHANICS; RELATIVISTIC QUANTUM THEORY; RELATIVITY; WAVE MECHANICS.

Scope of QED. Quantum electrodynamics turned out to be much more than a marriage of quantum mechanics and relativity. The wave-particle duality is now fully incorporated. Both electrons and photons are treated as quantized fields. The theory makes it possible to handle not only the dynamics of electrons but also the creation and destruction of electronpositron pairs, a phenomenon unheard of in classical physics. What makes quantum electrodynamics particularly intriguing is that it is the first physical theory ever developed that has no obvious intrinsic limitation and describes physical quantities from first principle. The predictive power of quantum electrodynamics may be appreciated by considering what is needed to describe the hydrogen atom by ordinary quantum mechanics to the same precision as that of quantum electrodynamics. For this purpose, a large number of new terms must be added to the potentialenergy operator of quantum mechanics. Unfortunately, quantum mechanics itself is powerless in selecting these terms. Meanwhile, quantum electrodynamics accounts for everything with just one term. See ELECTRON-POSITRON PAIR PRODUCTION.

However, nature accommodates forces other than the electromagnetic force, such as those responsible for radioactive disintegration of heavy nuclei (called the weak force) and the force that binds the nucleus together (called the strong force). A theory called the standard model, has been developed which unifies the three forces and accounts for all experimental data from very low to extremely high energies. This does not mean, however, that quantum electrodynamics fails at high energies. It simply means that the real world has forces other than electromagnetism. Predictions of quantum electrodynamics itself have thus far been confirmed by experiments without exception. Breakdown of quantum electrodynamics in any form would imply that there is something in nature besides relativity and quantum mechanics, the two pillars of modern physics. Since the standard model is an adaptation of quantum electrodynamics to include other forces, it would also fail if quantum electrodynamics fails. This is why it is important to keep examining the validity of quantum electrodynamics. *See* FUNDAMENTAL INTERACTIONS; STANDARD MODEL; STRONG NUCLEAR INTERACTIONS; WEAK NUCLEAR INTERACTIONS.

That it has been possible to examine quantum electrodynamics so closely is due to the fortuitous circumstance that the interaction of electromagnetic field and charge is relatively weak and its effect can be evaluated systematically as a small deviation from the noninteracting case. Therefore, the properties of noninteracting (or free) fields will be examined before interactions are considered. *See* PERTURBATION (QUANTUM MECHANICS).

Free electromagnetic field. The simplest electrodynamic system is that of radiation in an empty space, free of any charge, described classically by the Maxwell equations (which are consistent with relativity). In classical physics, the radiation is regarded as confined in a box and expressed as a superposition of normal modes, analogous to the dynamics of a string on a violin, each mode having a well-defined wavelength and oscillating with a well-defined frequency. The complete field can be constructed by specifying the amplitude of each normal mode and adding up the contributions of all modes. In other words, the dynamics of the electromagnetic field is equivalent to that of normal-mode amplitudes. See CAVITY RESONATOR; ELECTROMAGNETIC RADIATION; MAXWELL'S EQUATIONS; VIBRATION.

In quantum theory, any normal mode of the radiation field can be found only in a discrete set of quantum states with energy given by Eq. (1), where

$$E_n = b\nu \left(n + \frac{1}{2} \right) \tag{1}$$

b is Planck's constant, ν is the oscillator frequency, and n is restricted to integers 0, 1, 2, This state is said to consist of n photons each of energy bv. The total energy in the field is the sum of energies of all photons. The state with no photons present, called the vacuum, is characterized by assigning all oscillators to the lowest energy state n = 0. The total energy of such a state is the sum over all the normal modes of the quantity 1/2bv. This sum is not zero; in fact, it is infinite. This infinity is interpreted as a consequence of quantum fluctuations (the ephemeral presence of photons allowed by the uncertainty principle). It causes no problem since what is observable is the difference of energies between states, in which the infinities cancel out. See HARMONIC OSCILLATOR; UNCERTAINTY PRINCIPLE.

Free electron field. As mentioned above, Dirac's equation successfully explains many aspects of electrons and atoms. On closer scrutiny, however, it has some strange features. As noted above, his wave function has four components. While the first

681

two components correspond to the usual positive energy states, the last two are associated with the states of negative energy, which are unknown in classical physics and seem to have pathological properties. Dirac found, however, that they can be reinterpreted as positive energy states of the electron's antiparticle, the positron, which behaves exactly like an electron except that its charge has opposite sign. This bold prediction was strikingly successful with the subsequent discovery of positron. Dirac's successful reinterpretation required, however, that his "wave function" be treated not as a probability amplitude analogous to the Schrödinger wave function but as a quantized field with infinitely many degrees of freedom. Just like the electromagnetic field, the electron field can be written as a superposition of quantized normal modes, except that each mode can accommodate at most one electron or one positron in accord with the Pauli exclusion principle familiar in chemistry. The state with the lowest energy is one in which neither electron nor positron is present, analogous to the state of the electromagnetic field with no photons present, discussed above. Naturally, such a state is also called a vacuum. Like the photon vacuum, the electron vacuum is filled with fluctuating electrons and positrons. This means that quantum electrodynamics must always deal with (infinitely) many particles, even if the states of just a few particles are under consideration. This is where quantum electrodynamics differs strongly from the ordinary quantum mechanics. See ANTIMATTER; EX-CLUSION PRINCIPLE.

Interaction of light and electron. The classical Maxwell equations contain in a well-defined way the current and charge of particles as the source of radiation. The quantum-electrodynamics version of the charge-current density is constructed from the electron field and its hermitian conjugate, and is proportional to a quantity, e_0 , referred to as the bare charge of an electron. The photon field is represented by a four-vector potential. The interaction energy can be written as the integral over space of a product of the components of these two fields.

The equation of motion that follows from this expression for the interaction energy is nonlinear, and has thus resisted exact solutions. Fortunately, the dynamics of the coupled electron-photon system can often be described to a good approximation by the first few terms of a power series expansion in a small dimensionless parameter. The calculation of expansion terms is facilitated greatly by a pictorial representation (Fig. 1) of the interaction energy, where two straight lines (representing electrons or positrons) and one wavy line (representing a photon) meet at a space-time point, called a vertex. One such diagram (Fig. 1a) represents a situation in which an incident electron (entering from the bottom of the picture) absorbs an incident photon (also entering from the bottom) to produce a final electron (leaving from the top of the picture). Here a convention is adopted in which time proceeds upward from the bottom toward the top of page. Also, electrons and positrons are distinguished by upward

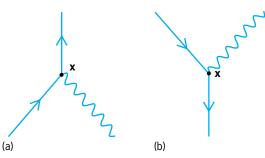


Fig. 1. Diagrams of photon absorption and emission. (a) Absorption of a photon (represented by a wavy line) by an electron (represented by a solid line with an upward arrow). (b) Emission of a photon by a positron (represented by a solid line with a downward arrow). The photon meets the electron or positron at a space-time point, *x*, called a vertex.

and downward arrows, respectively, on both initial and final solid lines. Another such diagram (Fig. 1*b*) corresponds to emission of a photon by a positron, and six more diagrams can be drawn corresponding to other situations such as creation of an electronpositron pair by a photon.

Feynman diagrams. Actually, the processes discussed above (Fig. 1) cannot occur by themselves since they violate the law of energy-momentum conservation. Nevertheless, they do serve as building blocks out of which allowed processes are constructed. All these processes can be represented by a set of diagrams obtained by connecting the vertices to each other according to a certain rule. In terms of these pictures, referred to as Feynman diagrams, a given process can be readily visualized. Feynman introduced a set of rules (called Feynman rules) which make it possible to write expressions called transition amplitudes. The absolute square of the sum of transition amplitudes gives the probability with which this process will take place. In many problems, it suffices to use only a few of these diagrams to obtain results with sufficient precision.

In general, Feynman diagrams contain several solid lines or wavy lines which originate in the distant past (the bottom of the figure) and represent incident particles. They will subsequently interact (at several vertices). After that, other solid or wavy lines emerge and move toward the distant future (the top of the figure). For example, (**Fig. 2**) an electron and a photon can enter the diagram, interact at two space-time points, **x** and **y**, and then leave. The line originating in the past or emerging to the future is referred

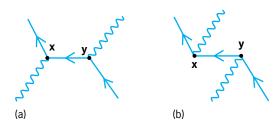


Fig. 2. Diagrams representing scattering of a photon on an electron caused by an exchange of a virtual electron. The electron and photon interact at space-time points x and y.

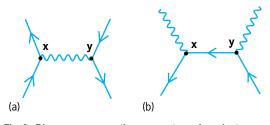


Fig. 3. Diagrams representing encounters of an electron and a positron. (a) Scattering of a positron and an electron caused by an exchange of a virtual photon. (b) Mutual annihilation of an incident positron and electron into two photons.

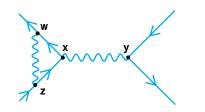


Fig. 4. Example of a diagram of order e_0^4 , involving two virtual photons and two virtual electrons. The diagram has a closed path of internal lines connecting the vertices *x*, *z*, and *w*.

to as an external line. The line connecting **x** and **y** is referred to as internal or virtual electron line. To each internal line corresponds a function of x-y, called an electron propagator. It provides a measure of the likelihood that an electron at point \mathbf{x} will traverse freely to point y. Internal wavy lines can also occur, which correspond to virtual photons or photon propagators. An electron and a positron may undergo scattering (Fig. 3a), or they may interact and subsequently disappear (Fig. 3b), producing two photons in the final state. Since the corresponding pictures have two space-time points at which interaction occurs, the amplitude has a strength proportional to e_0^2 (where e_0 is the bare charge that appears in the charge-current density). There are also diagrams proportional to e_0^4 , including some that give the so-called radiative corrections to electronpositron scattering (Fig. 4). Although such radiative corrections may be less important than diagrams proportional to e_0^2 (Fig. 3*a*) because of the extra two powers of e_0 , their contribution must be included if a comparison with precise experimental results is desired. Another feature of Feynman diagrams is that they may have a closed path of internal lines connecting vertices (Fig. 4). Whenever a closed path occurs, the amplitude contains a four-dimensional integral. The situation becomes very complicated when there are multiple closed paths. See FEYNMAN DIAGRAM; PROPAGATOR (FIELD THEORY).

Renormalization. It turns out that some of the integrals associated with closed paths diverge; that is, they are infinite. The simplest examples of divergent integrals are given by the self-energy and vertex correction diagrams (**Fig. 5**). Feynman diagrams containing any self-energy or vertex correction diagrams as subdiagrams are divergent. These are the divergences that initially hampered the formulation of quantum electrodynamics until the concept of renormalization was introduced. The basic idea of renormalization is as follows. For any theory, say quantum electrodynamics, there are certain parameters such as electron mass m_0 (called bare mass) and electron charge e_0 (called bare charge) which are needed to describe the theory. In the absence of any interaction, only m_0 is present. When an interaction is included a new parameter, e_0 , appears. Now m_0 and e_0 are presumed to be the values that these parameters would have if the interaction was reduced to zero. They are not the values of the electron mass or charge which an experimentalist would determine in some measurement. When Feynman diagrams are evaluated, they are expressed in terms of the parameter m_0 and e_0 . However, this is clearly unsatisfactory since they are not measured or "physical" values of the parameters. Thus *m* and *e*, the measured values, must be introduced instead. Now, clearly m and e must include the effect of interaction since the experimentalist cannot turn them off. In order to express answers in terms of the measured quantities, it is necessary to replace m_0 and e_0 by writing them as m and e plus correction terms. This leads to some new terms to be computed. These new terms, presumed to be divergent in perturbation theory, now bring about a cancellation of the infinite results when they are expressed as functions of physical mass mand physical charge e. The process of renormalization of mass and charge, although not entirely satisfactory from a mathematical point of view, leads to predictions that agree very well with experiment. See RENORMALIZATION.

High-precision tests. High-precision tests have provided excellent confirmation for the validity of the renormalization theory of quantum electrodynamics. In the high-energy regime, the most extensive tests of quantum electrodynamics have been carried out by using electron-positron colliding-beam facilities at various high-energy physics laboratories. These measurements have confirmed the predictions of quantum electrodynamics at center-of-mass energies up to 1.8×10^{11} electronvolts (180 GeV). The uncertainty principle implies that this is equivalent to saying that quantum electrodynamics is valid down to about 10⁻¹⁷ meter, a distance 100 times shorter than the radius of the proton. This is remarkable in view of the fact that quantum electrodynamics was tested to only 10^{-10} m when it was first formulated. At energies around 90 GeV, the contribution

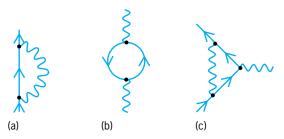


Fig. 5. Examples of divergent diagrams. (a) Electron self-energy diagram. (b) Photon self-energy diagram. (c) Vertex correction diagram.

of the weak boson Z^0 dominates those of quantum electrodynamics. As was mentioned above, however, this does not imply the failure of quantum electrodynamics at short distances. *See* INTERMEDIATE VECTOR BOSON; PARTICLE ACCELERATOR.

High-precision tests of quantum electrodynamics have also been carried out at low energies by using various simple atomic systems. The most accurate is that of the measurement of the magnetic moment of the electron, or the gyromagnetic ratio g, the ratio of spin and rotation frequencies. According to the Dirac equation, g must be exactly equal to 2. However, it was found in a 1947 experiment that g is larger than 2 by about 0.1%. Schwinger showed that this is due to the radiative correction of the electron in a magnetic field. The agreement of his calculation with experiment was one of the crucial successes of the then newly developed theory of quantum electrodynamics. By 1987, measurements of $a_e \equiv (g - g)$ 2)/2, utilizing radio-frequency resonance transitions between quantized levels of an electron in a magnetic field, reached a precision of 3.7 parts in 109, as given in Eqs. (2). These measurements show that the

$$a_{e^{-}}^{\exp} = 0.001\ 159\ 652\ 188\ 4(43)$$

$$a_{e^{+}}^{\exp} = 0.001\ 159\ 652\ 187\ 9(43)$$
(2)

electron (e^-) and its antiparticle, the positron (e^+), have identical magnetic moments, providing an example of the validity of a general theorem of quantum field theory, including quantum electrodynamics, called the CPT theorem. (The numeral 43 in the parentheses represents measurement uncertainty of $\pm 43 \times 10^{-13}$.) See CPT THEOREM.

In 2006, the precision of the measurement of a_e was improved significantly by a substantial refinement of the experiment, yielding the value given in Eq. (3), which is about six times more accurate than

$$a_{a}^{\exp} = 0.001\,159\,652\,180\,85(76)$$
 (3)

the 1987 measurement quoted in Eqs. (2).

Quantum electrodynamics calculations involving many Feynman diagrams provide the coefficients for the expansion of a_e in powers of the ratio α/π , where α is the fine-structure constant, as given in Eq. (4).

$$a_e = A^{(2)}(\alpha/\pi) + A^{(4)}(\alpha/\pi)^2 + A^{(6)}(\alpha/\pi)^3 + \cdots$$
(4)

The coefficients $A^{(2)}$, $A^{(4)}$, and $A^{(6)}$ are known analytically; their numerical values are given in Eqs. (5).

$$A^{(2)} = 0.5$$

$$A^{(4)} = -0.328\,478\,965\,579\dots$$

$$A^{(6)} = 1.181\,241\,456\,587\dots$$
(5)

Thus far the value of $A^{(8)}$ has been obtained only by numerical integration, yielding Eq. (6), where

$$A^{(8)} = -1.7283(35) \tag{6}$$

35 is the uncertainty in the last two digits. Including contributions of small quantum electrodynamics terms and non-quantum electrodynamics effects, and using the most precise value of the fine-structure constant available at present (2006), given in Eq. (7),

$$\alpha^{-1} = 137.035\ 998\ 78(91) \tag{7}$$

one obtains the theoretical prediction of a_e given in Eq. (8), in good agreement with experiment, where

$$a_e^{\text{theory}} = 0.001\ 159\ 652\ 188\ 70\ (40)\ (771)$$
 (8)

40 is the remaining theoretical uncertainty and 771 comes from the uncertainty in the measurement of the fine-structure constant α quoted in Eq. (7). The gyromagnetic ratio *g* is thus correctly predicted by quantum electrodynamics to 12 significant digits. The comparison of a_e^{\exp} with a_e^{theory} presented here constitutes the most precise confrontation of any experiment with a theoretical prediction in the history of science. Furthermore, it provides a new value of the fine-structure constant, given by Eq. (9),

$$\alpha^{-1}(g-2) = 137.035\ 999\ 710\ (96) \tag{9}$$

with a relative uncertainty of 0.7 part per billion (10^9) , which is an order of magnitude more precise than any other measurement, including Eq. (7). Further improvement in the non-quantum electrodynamics measurement of α is crucial in pushing the test of the validity of quantum electrodynamics further. At present, other low-energy tests of quantum electrodynamics are less accurate, but their precision is improving with the development of the frequency comb technique and atom interferometry in laser spectroscopy. *See* ELECTRON SPIN; FREQUENCY COMB; LASER SPECTROSCOPY; QUANTUM FIELD THEORY.

Bibliography. J. Bjorken and S. Drell, *Relativistic Quantum Fields*, 1965; J. Bjorken and S. Drell, *Relativistic Quantum Mechanics*, 1965; R. P. Feynman, *QED: The Strange Theory of Light and Matter*, rev. ed., 2006; C. Itzykson and J.-B. Zuber, *Quantum Field Theory*, 1980, paperback, 2006; T. Kinoshita, *Quantum Electrodynamics*, 1990; J. Schwinger (ed.), *Selected Papers on Quantum Electrodynamics*, 1958; G. Sterman, *An Introduction to Quantum Field Theory*, 1993; S. Weinberg, *The Quantum Theory of Fields*, vol. 1: *Foundations*, 1995.

Quantum electronics

Quantum electronics is concerned with the interaction of electromagnetic radiation and matter and with the use of radiation-matter interactions for applications in devices such as lasers and masers. The term "quantum electronics" for this field was coined in 1958 or 1959 by Charles H. Townes, who invented the maser, as this field in its early days brought together ideas from electrical engineering and quantum physics in the development of the maser and the laser. The field has grown immensely, and today quantum electronics is concerned not just with masers and lasers but with a wide range of topics related to linear as well as nonlinear radiation-matter interactions, quantum-mechanical properties of radiation, noise and photon statistics, physical properties of solid-state materials relevant to radiation interaction, and devices and systems that utilize these basic principles for practical applications. *See* LASER; MASER.

Many electronic devices, such as semiconductor resonant tunneling diodes or the superconductor Josephson junctions, may be considered quantum electronic devices, since their operation is dictated by the principles of quantum mechanics. However, quantum electronics is understood to be concerned with only those devices whose operation involves radiation-matter interactions, such as lasers, parametric oscillators, and optoelectronic devices, such as light modulators, photodetectors, and optical switches. *See* ELECTROOPTICS; OPTICAL BISTABIL-ITY; OPTICAL DETECTORS; OPTICAL MODULATORS.

Radiation absorption and emission. Some of the most fundamental ideas in quantum electronics are related to the absorption and emission of radiation by matter. The early quantum theory of radiation developed by Max Planck in 1901 postulated that radiation can be absorbed or emitted by matter in only discrete bundles or packets that were later called photons, with each photon having energy *bf*, where *b* is Planck's constant and *f* is the frequency of the radiation. Soon after in 1917, Albert Einstein discovered that photon emission by matter can be a spontaneous process or a process stimulated by other photons. Stimulated absorption of photons and spontaneous and stimulated emission of photons by atoms are illustrated in Fig. 1. The atoms are assumed to have two energy levels, and an electron in an atom can be in either the lower or the upper energy level. In each process, energy conservation requires that the energy difference between the upper and the lower energy level is equal to the photon energy $(E_2 - E_1 =$

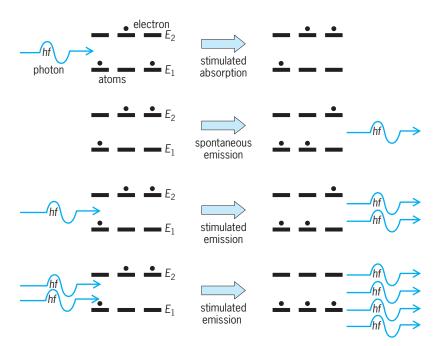


Fig. 1. Stimulated and spontaneous emission of photons by atoms.

bf). Figure 1 shows that the number of photons can be increased substantially through stimulated emission, as each generated photon can in turn stimulate the emission of additional photons. This photon multiplication through stimulated emission is the basic principle behind the operation of lasers and masers. The acronym "laser" (or "maser") stands for light (or microwave) amplification through stimulated emission of radiation. *See* PHOTOEMISSION; PHOTON.

Laser operation. In a laser, photons are confined inside a cavity along with atoms that can emit photons. These atoms can be in the form of molecules (as in dye lasers), solid-state crystals (as in semiconductor lasers), separate atoms (as in gas lasers), or impurities in a host material (such as chromium atoms in aluminum oxide in ruby lasers). Photons can be generated inside the cavity via spontaneous and stimulated emission by the atoms. Photons can be lost from the cavity via stimulated absorption by the atoms. In addition, photons can escape from the cavity. When the rate at which photons are generated inside the cavity is larger than the rate at which photons are lost from the cavity, the number of photons inside the cavity increases dramatically and this is called lasing. Lasing requires that the rate of stimulated emission of photons by the atoms exceed the rate of stimulated absorption, and this is achieved by ensuring that electrons in a majority of the atoms are in the upper energy level. This condition is called population inversion. Energy must be expended from outside in order to maintain population inversion during lasing

Coherent radiation. A feature of lasers that makes radiation from a laser source different from radiation from a blackbody source is the coherence of the laser radiation. Coherent radiation does not have any significant phase jumps in time (or phase noise). In semiclassical terms, the coherence of laser radiation can be related to the fact that radiation emitted via a stimulated process has the same phase as that of the radiation that caused the stimulation. Therefore no phase noise is introduced during radiation amplification through stimulated emission. The small amount of phase noise in laser radiation is due to the spontaneous emission of radiation. Coherent electromagnetic radiation at low frequencies (less than 10¹¹ Hz) can also be generated with oscillating electrical currents, as in antennas. Lasers and masers can provide high-power coherent radiation all the way from microwave frequencies (109 Hz) to x-ray frequencies (10^{17} Hz) .

Nonlinear optical effects. Soon after the invention of the laser, it was discovered that a wide assortment of interesting and useful nonlinear optical phenomena could be produced with appropriate atoms or materials using light from lasers because of the unprecedented high power and coherence of laser radiation. Nonlinear optical effects can be understood in terms of the atomic or material polarization caused by the strong electric fields of laser radiation. It is well known from the classical theory of radiation that oscillating charge dipoles radiate electromagnetic energy just like oscillating currents. In fact, stimulated emission by atoms in a laser can also be understood in terms of radiation emitted by oscillating charge dipoles induced in the atoms in the presence of the stimulating radiation. A volume density of charge dipoles induced in a material medium by an electric field E(t) is described by the time-dependent material polarization P(t) which can be expanded in a Taylor series in powers of the electric field strength E(t) as

$$p(t) = \varepsilon_0 [\chi^1 E(t) + \chi^2 E^2(t) + \chi^3 E^3(t) + \cdots]$$

where ε_0 is the permittivity of free-space, and χ^1 , χ^2 , and χ^3 are the first-, second-, and third-order optical susceptibilities of the material, respectively. The first-order susceptibility describes the linear optical response of the material, and the secondand third-order susceptibilities describe the nonlinear optical response of the material. Stimulated emission of radiation in lasers can be described in terms of the first-order linear optical susceptibility. Atoms and materials with large values of the second- and third-order susceptibilities have been used to produce a wide variety of nonlinear optical phenomena such as harmonic generation, sum and difference frequency generation, parametric amplification and oscillation, four-wave mixing, self-phase modulation, and Raman and Brillouin scattering (Fig. 2). Nonlinear optics has become a major part of the field of quantum electronics. See DIFFERENTIATION; DIPOLE; NONLINEAR OPTICS; POLARIZATION OF DIELECTRICS; RAMAN EFFECT; SCATTERING OF ELECTROMAGNETIC RADIATION; SERIES.

Ultrafast optics. Ultrafast optics has been another notable development in the field of quantum electronics since the 1980s. High-intensity single-cycle optical pulses have been produced by mode-locking lasers. Most lasers have many different discrete frequencies all lasing at the same time. These frequencies are almost equally spaced and correspond to the Fabry-Perot modes of the laser cavity. The phases of the different lasing modes inside a laser cavity are usually unrelated. Using a variety of linear and nonlinear optical techniques, the phases of the different lasing modes can be locked with respect to each other. When this happens, light output from the laser is not in the form of a continuous beam but in the form of short periodic pulses. Ultrashort pulses from mode-locked lasers have enabled the study of dynamical processes in physics, chemistry, and biology with femtosecond time resolution. In addition, the high temporal stability of the periodic pulse train from mode-locked lasers has enabled their use in optical frequency metrology applications with performances comparable to atomic clocks. See FRE-QUENCY COMB; INTERFEROMETRY; OPTICAL PULSES; ULTRAFAST MOLECULAR PROCESSES.

Optics on microchips. With the advance in microand nanofabrication techniques since the 1980s, quantum electronics has developed a promising new frontier. It has become possible to generate, amplify, and control coherent radiation over a wide range of frequencies on microchips. Micrometer-scale lasers

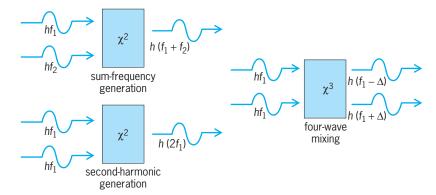


Fig. 2. Some nonlinear optical processes.

and optoelectronic devices are now possible. The development of inexpensive semiconductor lasers and integrated optoelectronic devices has been partly responsible for the wide deployment of fiber-optic communication links. The development of photonic very large scale integrated (VLSI) circuits is an active area of research. Such photonic circuits could provide ultra-high-speed information processing capabilities on microchips. *See* INTEGRATED CIRCUITS; INTEGRATED OP-TICS; MICROLITHOGRAPHY; NANOTECHNOLOGY; OP-TICAL COMMUNICATIONS.

Bibliography. R. Boyd, *Nonlinear Optics*, 2d ed., Academic Press, New York, 2003; L. A. Coldren and S. W. Corzine, *Diode Lasers and Photonic Integrated Circuits*, Wiley, New York, 1995; *IEEE Journal of Quantum Electronics*, IEEE Press, New York, published monthly; R. Loudon, *The Quantum Theory of Light*, 3d ed., Oxford University Press, New York, 2000; A. E. Siegman, *Lasers*, University Science Books, Sausalito, CA, 1986; A. Yariv, *Quantum Electronics*, 3d ed., Wiley, NY, 1989.

Quantum field theory

The quantum-mechanical theory of physical systems whose dynamical variables are local functions of space and time. As distinguished from the quantum mechanics of atoms, quantum field theories describe systems with an infinite number of degrees of freedom. Such theories provide the natural language for describing the interactions and properties of elementary particles, and have proved to be successful in providing the basis for the fundamental theories of the interactions of matter. The present understanding of the basic forces of nature is based on quantum field theories of the strong, weak, electromagnetic, and gravitational interactions. Quantum field theory is also useful in the study of many-body systems, especially in situations where the characteristic length of a system is large compared to its microscopic scale. See NONRELATIVISTIC QUANTUM THEORY; QUANTUM MECHANICS.

Quantum electrodynamics. Quantum field theory originated in the attempt, in the late 1920s, to unify P. A. M. Dirac's relativistic electron theory and J. C. Maxwell's classical electrodynamics in a quantum

theory of interacting photon and electron fields. This effort was completed in the 1950s and was extremely successful. At present the quantitative predictions of the theory are largely limited to perturbative expansions (believed to be asymptotic) in powers of the fine-structure constant. However, because of the extremely small value of this parameter, $\alpha = e^2/\hbar c \approx 1/137$ (where *e* is the electron charge, \hbar is Planck's constant divided by 2π , and c is the speed of light), such an expansion is quite adequate for most purposes. The remarkable agreement of the predictions of quantum electrodynamics with high-precision experiments (sometimes to an accuracy of 1 part in 1012) provides strong evidence for the validity of the basic tenets of relativistic quantum field theory. See CLASSICAL FIELD THEORY; ELECTROMAGNETIC RADIATION; MAXWELL'S EQUATIONS; PERTURBATION (QUANTUM MECHANICS); PHOTON; QUANTUM ELECTRODYNAMICS; RELATIVIS-TIC ELECTRODYNAMICS; RELATIVISTIC QUANTUM THE-ORY.

Theories of other interactions. Quantum field theory also provides the natural framework for the treatment of the weak, strong, and gravitational interactions.

Weak interactions. The first of such applications was Fermi's theory of the weak interactions, responsible for radioactivity, in which a hamiltonian was constructed to describe beta decay as a product of four fermion fields, one for each lepton or nucleon. This theory has been superseded by the modern electroweak theory that unifies the weak and the electromagnetic interactions into a common framework. This theory is a generalization of Maxwell's electrodynamics which was the first example of a gauge theory, based on a continuous local symmetry. In the case of electromagnetism the local gauge symmetry is the space-time-dependent change of the phase of a charged field. The existence of massless spin-1 particles, photons, is one of the consequences of the gauge symmetry. The electroweak theory is based on generalizing this symmetry to space-time-dependent transformations of the labels of the fields, based on the group SU(2) \times U(1). However, unlike electromagnetism, part of this extended symmetry is not shared by the ground state of the system. This phenomenon of spontaneous symmetry breaking produces masses for all the elementary fermions and for the gauge bosons that are the carriers of the weak interactions, the W^{\pm} and Z bosons. (This is known as the Higgs mechanism.) The electroweak theory has been confirmed by many precision tests, and almost all of its essential ingredients have been verified. See ELECTROWEAK INTERACTION; GAUGE THEORY; IN-TERMEDIATE VECTOR BOSON; SYMMETRY BREAKING; SYMMETRY LAWS (PHYSICS); WEAK NUCLEAR INTERAC-TIONS.

Strong interactions. The application of quantum field theory to the strong or nuclear interactions dates from H. Yukawa's hypothesis that the short-range nuclear forces arise from the exchange of massive particles that are the quanta of local fields coupled to the nucleons, much as the electromagnetic inter-

actions arise from the exchange of massless photons that are the quanta of the electromagnetic field. The modern theory of the strong interactions, quantum chromodynamics, completed in the early 1970s, is also based on a local gauge theory. This is a theory of spin- $\frac{1}{2}$ quarks invariant under an internal local SU(3) (color) gauge group. The observed hadrons (such as the proton and neutron) are SU(3) colorneutral bound states of the quarks whose interactions are dictated by the gauge fields (gluons). This theory exhibits almost-free-field behavior of quarks and gluons over distances and times short compared to the size of a hadron (asymptotic freedom), and a strong binding of quarks at large separations that results in the absence of colored states (confinement). See ELEMENTARY PARTICLE; GLUONS; MESON; QUAN-TUM CHROMODYNAMICS; QUARKS.

Gravitation. Quantum field theory has been tested down to distances of 10^{-20} m. There appears to be no reason why it should not continue to work down to Planck's length, $(G\hbar/c^3)^{1/2} \approx 10^{-35}$ m (where G is the gravitational constant), where the quantum effects of gravity become important. In the case of gravity, A. Einstein's theory of general relativity already provides a very successful classical field theory. However, the union of quantum mechanics and general relativity raises conceptual problems that seem to call for a radical reexamination of the foundations of quantum field theory. Among these problems are the lack of renormalizability of standard perturbative quantization schemes, and the quantum-mechanical implications of the ubiquitous singularities that occur in classical general relativity. See FUNDAMENTAL INTERACTIONS; GRAVITATION; QUANTUM GRAVITATION; RELATIVITY.

General properties. A field theory can be specified by either lagrangian or hamiltonian methods, the former having the advantage of making manifest the symmetries of the theory and the latter being the basis of canonical quantization, the generalization of canonical quantization for systems with a finite number of degrees of freedom. An interacting relativistic field theory necessarily involves creation and annihilation processes and is therefore a many-body theory.

The problem of constructing exact solutions of the coupled nonlinear equations for interacting fourdimensional field theories has so far proved too formidable for solution. However, there do exist both soluble field-theoretic models and existence proofs for others in three or two space-time dimensions. In four dimensions, however, certain assumptions must be made (the axioms of quantum field theory) in order to obtain physical predictions. These include assuming that a nondegenerate, Lorentz-invariant vacuum state exists; that the spectra of the energies and momenta lie in the forward light cone; and that a discrete spectrum of single-particle states exists. In addition, it is assumed that, for large positive (or negative) times, the matrix elements of the fields under consideration (appropriately smeared out) between normalizable physical states approach constant multiples of the matrix elements of free fields. By using this condition, which expresses the intuitive idea that any system with only short-range interactions will consist of freely moving particles at asymptotic times, all matrix elements of physical interest can be expressed in terms of matrix elements of products of field operators between vacuum states. In particular, the S matrix, which yields the probability amplitude for the transition from an initial state of free particles to a final state of free particles, can be related to these matrix elements. *See* SCATTERING MATRIX.

Many properties of the S matrix can be established without recourse to the specific dynamics of the fields under consideration, as long as they are local fields satisfying the microcausality condition. This condition expresses the simultaneous observability of field operators at spacelike separations, and is a necessary condition in a Lorentz-invariant theory in which signals do not propagate with a velocity faster than the velocity of light. It can be shown that the axioms of quantum field theory lead to many general properties that are independent of specific dynamical models. One of the most general results is the CPT theorem. This theorem states that the laws of physics must be invariant under a simultaneous reversal of time, charge conjugation (which reverses the sign of all conserved charges and interchanges particles with antiparticles), and parity, even though none of these individual transformations are exact symmetries of nature. A consequence of this theorem is that the masses of particles and antiparticles must be identical; this has been tested in one case to an accuracy of 1 part in 10⁻⁷. Another general result is the correlation between spin and statistics, which states that integer-spin and half-integer-spin particles must obey Bose-Einstein and Fermi-Dirac statistics respectively; that is, collections of such particles must be described by totally symmetric and totally antisymmetric wave functions respectively. See BOSE-EINSTEIN STATISTICS; CPT THEOREM; FERMI-DIRAC STATISTICS; QUANTUM STATISTICS.

Another class of general results involves the analytic properties of scattering amplitudes and Green's functions as functions of complex energy and momentum, which together with derived or assumed bounds on the asymptotic behavior of the scattering amplitude provide the basis for the derivation of dispersion relations. These dispersion relations are essentially consequences of microcausality, and their verification is strong evidence for the basic structure of quantum field theory. *See* DISPERSION RELATIONS; GREEN'S FUNCTION.

Other general features follow from symmetry properties. In quantum-mechanical systems with a finite number of degrees of freedom, a continuous symmetry implies the existence of a conserved charge that commutes with the hamiltonian. The consequences of such a symmetry are that the particles of the theory fall into degenerate multiplets with the same mass and their scattering amplitudes are linearly related. In quantum field theory, another realization of the symmetry is possible and often occurs, one in which the vacuum or ground state does not respect the symmetry. An example of symmetry breaking is provided by the phenomenon of spontaneous magnetization, wherein a magnetic material exhibits at low temperature a magnetic field pointing in a specific direction, even though the microscopic laws are rotationally invariant. In such a case the physical manifestations of the symmetry are quite different. Spontaneous symmetry breaking implies the existence of massless particles (called Goldstone bosons) with the quantum numbers of the conserved charge. The particles of the theory need not form degenerate multiplets, and the symmetry gives rise to low-energy theorems for the scattering of the Goldstone bosons. If, however, the symmetry is a local gauge symmetry, the Goldstone bosons are absent but the corresponding gauge bosons acquire a mass. This is the Higgs mechanism, and it underlies the unified electroweak theory, in which the exact SU(2) \times U(1) symmetry is spontaneously broken down to the U(1) gauge symmetry of electrodynamics.

Feynman diagrams. In a given lagrangian quantum field theory, Green's functions and the S matrix can be calculated perturbatively in the nonlinear coupling constants by constructing a perturbation series expansion for the unitary transformation that relates the fields under consideration to free fields. Individual terms in the expansion of Green's functions and scattering amplitudes can be represented by Feynman diagrams. These diagrams provide an instructive physical picture of the process under consideration and, with the aid of Feynman rules that associate a specific integral with a given diagram, are of enormous aid in performing calculations. *See* FEYNMAN DIAGRAM.

Path-integral formalism. An alternate formulation of quantum field theory is based on Feynman's representation of transition amplitudes as sums over path histories weighted by the classical action. Although it is difficult to give a rigorous definition of the functional integrals that appear in this formulation, it has proved extremely useful for deriving consistent perturbative expansions for gauge field theories and as a starting point for semiclassical approximations. The path integral is also the starting point for numerical approaches to quantum field theory. Continuous space and time is replaced by a finite-volume lattice, in which case the path integral reduces to a finite product of ordinary integrals. Various numerical techniques (such as Monte Carlo integration) are then used to evaluate physical observables, by extrapolating the lattice calculations to the limit of an infinite volume and vanishing lattice spacing. Such methods have been used in quantum chromodynamics to calculate the hadronic mass spectrum, with some success. See FEYNMAN INTEGRAL; MONTE CARLO METHOD.

Renormalization. Quantum field theory is plagued by many divergences such as those arising from the existence of an infinite number of degrees of freedom (which can be dealt with by careful control of the infinite-volume limit), infrared divergences arising from the existence of massless particles (which can be dealt with by careful treatment of asymptotic states), and ultraviolet divergences arising from the pointlike structure of local field theory. The last are dealt with by the process of renormalization, which eliminates the original (or bare) parameters in the lagrangian (coupling constants and masses) in favor of those with physical significance. The meaning of renormalization is that the parameters in lagrangian refer to the properties of the theory at arbitrarily small distances, whereas the physical parameters measured at normal distances are modified by the interactions. Thus, in quantum electrodynamics charge renormalization is due to the screening of the electric bare charge by virtual electron-positron pairs (vacuum polarization), resulting in a diminished physical electric charge. In a renormalizable field theory, all ultraviolet divergences can be systematically eliminated by perturbatively adjusting the bare parameters to yield finite physical parameters. Nonrenormalized field theories, however require the specification of an infinite number of parameters in perturbation theory or complete nonperturbative control over their short-distance behavior. See RENORMALIZATION.

Renormalization group. The renormalization group describes how the dynamics of a system evolves as the scale of the phenomena being observed is changed. A scale-dependent hamiltonian is constructed to describe the system at varying distances. A change of the scale of length induces a transformation of this effective hamiltonian. These transformations often have the property of approaching a fixed point after many iterations. In this case, the large- or small-distance structure of the theory will be relatively independent of its structure at intermediate distances and, if the fixed-point hamiltonian is simple enough, can be explicitly calculated. These techniques have been enormously successful in applications of quantum field theory to the study of critical behavior in statistical mechanics. In the case of the strong interactions the asymptotic freedom of quantum chromodynamics means that the effective dynamics of quarks and gluons at very high energies is that of an almost-free theory, and thus can be calculated. Many of these high-energy predictions have been verified by experiment. See CRITICAL PHE-NOMENA.

The renormalization group also provides a perspective on models of low-energy physics as effective field theories. It is assumed that the final theory of physics (a unified theory of all interactions) can be described at some very high energy scale, Λ , by a local quantum field theory. In the absence of further dynamical information, this theory will be described by the most general possible local quantum field theory, constrained only by the assumed symmetries of the theory. It will also be assumed that all coupling constants that parametrize the strengths of the interactions at the scale Λ are numbers of order one, times appropriate powers of Λ , \hbar , and *c*. Starting with this infinite-parameter theory at high energies, it is possible to integrate out the high-momentum modes of all the fields, thus deriving an effective theory that describes low-energy physics at energies *E* much less than Λ . The effective theory will depend on the scale of energy E. The renormalization group describes the change in the effective theory as E is changed. The characteristic behavior of the solutions of the renormalization group equations, which describe the flow in the space of theories, is that the solutions approach a finite-dimensional submanifold in this infinite-dimensional space, which is parametrized by the renormalizable couplings. A nonrenormalizable interaction is one whose coupling constant has dimensions of powers of the inverse energy, and thus has a coupling of order g/Λ^d , where *d* is the dimension of the coupling, Λ is the cutoff, and g is a pure number of order one. A renormalizable coupling, however, is characterized by a dimensionless parameter, g', of order one. To compare the two at low energies of order E, dimensionless ratio must be formed, given by the equation below.

$$\frac{g(E/\Lambda)^d}{g'} \sim (E/\Lambda)^d$$

The effect of the nonrenormalizable interactions on low-energy processes will therefore be of order $(E/\Lambda)^d$. Thus the nonrenormalizable interactions (of which there might be an infinite number) can be totally neglected at low energies, and the low-energy theory, up to small corrections, will then be described by the most general renormalizable quantum field theory consistent with the symmetries, with a finite number of parameters. David Gross

Bibliography. J. Bjorken and S. Drell, *Relativistic Quantum Fields*, 1965; C. Itzkykson and J. Zuber, *Quantum Field Theory*, 1980; M. Kaku, *Quantum Field Theory*, 1993; F. Mandl and G. Shaw, *Quantum Field Theory*, rev. ed., 1993; P. Ramond, *Field Theory: A Modern Primer*, 2d rev. ed., 1990; J. Zinn-Justin, *Quantum Field Theory and Critical Phenomena*, 4th ed., 2002.

Quantum gravitation

The quantum theory of the gravitational field; also, the study of quantum fields in a curved space-time. In classical general relativity, the gravitational field is represented by the metric tensor $g_{\mu\nu}$ of space-time. This tensor satisfies Einstein's field equation, with the energy-momentum tensor of matter and radiation as a source. However, the equations of motion for the matter and radiation fields also depend on the metric.

Classical field theories such as Maxwell's electromagnetism or the classical description of particle dynamics are approximations valid only at the level of large-scale macroscopic observations. At a fundamental level, elementary interactions of particles and fields must be described by relativistic quantum mechanics, in terms of quantum fields. Because the geometry of space-time in general relativity is inextricably connected to the dynamics of matter and radiation, a consistent theory of the metric in interaction with quantum fields is possible only if the metric itself is quantized. *See* MAXWELL'S EQUA-TIONS; QUANTUM FIELD THEORY; RELATIVISTIC QUAN-TUM THEORY; RELATIVITY.

Under ordinary laboratory conditions the curvature of space-time is so extremely small that in most quantum experiments gravitational effects are completely negligible. Quantization in Minkowski space is then justified. Gravity is expected to play a significant role in quantum physics only at rather extreme conditions of strongly time-dependent fields, near or inside very dense matter. The scale of energies at which quantization of the metric itself becomes essential is given by $(\hbar c^5/G)^{1/2} \approx 10^{19}$ GeV, where G is the gravitational constant, \hbar is Planck's constant divided by 2π , and c is the velocity of light. Energies that can be reached in the laboratory or found in cosmic radiation are far below this order of magnitude. Only in the very early stages of the universe, within a proper time of the order of $(G\hbar/c^5)^{1/2} \approx 10^{-43}$ s after the big bang, would such energies have been produced. See BIG BANG THEORY.

Quantum fields in curved space-time. In most physical systems the metric is quasistationary over macroscopic distances so that its fluctuations can be ignored. A quantum description of fields in a curved space-time can then be given by treating the metric as a classical external field in interaction with the quantum fields. If the metric is asymptotically flat, asymptotic particle states can be defined associated with positive frequency-solutions of the freefield wave equation. "In" and "out" Hilbert spaces can then be constructed which will be related by a unitary transformation S. In particular, the vacuum states, $|0_{in}\rangle$ and $|0_{out}\rangle$, are related by $S|0_{out}\rangle = |0_{in}\rangle$, and in general, in a nonstationary background, $|0_{out}\rangle$ $\neq |0_{in}\rangle$. This means that particles are created by the nonstationary metric. In the special case of free quantum fields, an explicit form for the S-matrix can be obtained. It is found that particles are created in pairs of which one member is a particle and the other its antiparticle. One exception to particle creation is the case of a massless free field in interaction with a background Robertson-Walker metric which represents a spatially homogeneous and isotropic geometry. This property of space is assumed in the Friedmann models of an expanding universe, discussed below. The conformal invariance of these systems ensures that $|0_{out}\rangle = |0_{in}\rangle$. See SCATTERING MATRIX

Quantum effects of black holes. The most striking quantum effect in curved space-time is the emission of radiation by black holes. A black hole is an object that has undergone gravitational collapse. Classically this means that it becomes confined to a space-time region in which the metric has a singularity (the curvature becomes infinite). This region is bounded by a surface, called the horizon, such that any matter or radiation falling inside becomes trapped. Therefore, classically the mass of a black hole can only increase. However, this is no longer the case if quantum effects are taken into account. When, because of fluctuations of the quantum field, particle-antiparticle or photon pairs are created near the horizon of a black hole, one of the particles carrying negative energy may move toward the hole, being absorbed by it, while the other moves out with positive energy. Only these emitted particles can be observed; their number, in a given mode *j*, is the expectation value in the incoming vacuum $|0_{in}\rangle$ of the number operator n_{jout} associated with the outgoing Hilbert space. The index *j* stands for the frequency $\omega/2\pi$ and other quantum numbers. A careful analysis of the propagation of wave packets shows that the average number per unit frequency of particles emitted per unit time is given by Eq. (1), where $\Gamma(\omega)$ is the absorption

$$\langle n_{\omega} \rangle = \frac{\Gamma(\omega)}{\exp\left(2\pi\omega/\kappa - 1\right)}$$
 (1)

coefficient and κ is the surface gravity of the black hole. For a Schwarzschild black hole of mass M, $\kappa = 1/(4M)$, in geometric units chosen so that $G = \hbar = c = 1$. This result shows that a black hole emits radiation as a body in thermal equilibrium at temperature $T = \kappa/(2\pi k)$, where k is the Boltzmann constant. It lends support to the thermodynamic interpretation of the classical laws of black hole mechanics. *See* BLACK HOLE; GRAVITATIONAL COLLAPSE; HEAT RADI-ATION.

Back reaction on the metric. As energy is taken out of the black hole by the emission of radiation, its mass must decrease. This is an effect of the back reaction of the quantum field on the metric. As long as the mass of the black hole is very large compared to Planck's mass $(\hbar c/G)^{1/2} \approx 10^{-8}$ kg, a semiclassical quasistationary approach is valid and the rate of change of mass is given by the total energy emitted per unit time. If it is assumed that $\Gamma(\omega)$ is equal to the absorption coefficient of a blackbody of radius equal to the Schwarzschild radius, and is thus proportional to the area of the horizon, then from Eq. (1) it is found that the total rate of emission is inversely proportional to the square of the mass. As the mass decreases to very small values, the black hole evaporates with a burst of radiation. In this process, conservation of baryon number is violated. These conclusions rely on approximations that break down in the last stages of the black hole's evolution; hence they may not be entirely correct. For stellar black holes whose masses are of the order of a solar mass, the emission rate is negligibly small and unobservable. Only primordial black holes, of mass less than 10^{13} kg, formed very early in the quantum era of the universe, would have been small enough to produce quantum effects that could play any significant role in astrophysics or in cosmology.

Quantization of the metric. There are basically two approaches to the quantization of the metric, the canonical and the covariant quantization. Each method for quantizing gravity will be briefly discussed for the case in which other fields are absent. Einstein's field equations then reduce to Eq. (2), where the indices μ and ν take on the values 0, 1, 2, 3.

$$G_{\mu\nu} \equiv R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} = 0 \tag{2}$$

Canonical quantization. In the canonical quantization, a parametrization of the metric is used. There are

four nonpropagating fields, analogous to the timelike component A_0 of the vector potential A_μ in electrodynamics. The dynamical variables are the components g_{ii} (*i*, *j* = 1, 2, 3) of the three-geometry of a spacelike hypersurface and their conjugate momenta π^{ij} . Variation of the Einstein action with respect to the nonpropagating fields gives a set of constraint equations. At this stage it is possible either to use Dirac's method of quantization for constrained systems, or choose a coordinate system by imposing four gauge conditions $F_{\mu}(g,\pi) = 0$. Taking the time derivative of these gauge conditions and using the equations of motion yields further constraints $\dot{F}_{\mu} = 0$. This whole set of constraints is then used to eliminate the nonpropagation fields and four of the pairs of variables g_{ij} , π^{ij} , leaving only two pairs of dynamical variables, representing the two degrees of freedom of the system. These are then quantized by means of the usual canonical commutation rules. Solving the constraint is a major technical difficulty in this procedure. See HAMILTON'S EQUATIONS OF MOTION; NON-RELATIVISTIC QUANTUM THEORY; POTENTIALS; VARIA-TIONAL METHODS (PHYSICS).

An alternative is to follow J. A. Wheeler's superspace formulation of general relativity using the Hamilton-Jacobi formalism. The Einstein action is taken between two spacelike hypersurfaces where the metric of the three-geometry is given. The minimal of the action with the metric held fixed at the two hypersurfaces is the Hamilton principal function, which is a function of the two given threegeometries. If the second hypersurface is allowed to change, the Hamilton function gives the history of the geometries connecting the two hypersurfaces. The variation of the Hamilton action corresponding to a variation of the metric on the second hypersurface gives an expression for the conjugate momenta, π^{ij} . Replacement of the conjugate momenta with this expression in the constraint equations yields the full geometrodynamics of supersace at the classical level. See HAMILTON-JACOBI THEORY.

The theory can be quantized by replacing the dynamical variables g_{ij} and π^{ij} by operators satisfying canonical comutation relations, and imposing the constraint equations on the state vectors of the system. A remarkable feature in this procedure is that the "time" parameter no longer appears in the evolution equation of the system. Somehow it has to be defined as an observable in terms of the extra components of the metric g_{ij} . The question of defining a complete set of observable in the theory is still open. Attempts to address this problem have also been made using a new set of variables instead of g_{ij} . In these variables the constraints take on a simpler form.

Covariant quantization. In the covariant method the metric is split into a classical background field and a quantum operator. Since not all components of the quantum operator are independent, constraints must be imposed on the physical states, analogous to the constraints in the Gupta-Bleuler formalism in quantum electrodynamics. If the classical metric is chosen to be the Minkowski metric, it is possible to formally

write down Feynman rules in momentum space. One objection to this method is that spacelike surfaces are defined with respect to the classical metric, not the total metric. *See* QUANTUM ELECTRODYNAMICS.

Application of path integral representation. A third method, which can be derived from the first and is now most widely used, is based on the Feynman path integral representation for the vacuum-to-vacuum amplitude, which is the generator of Green's functions for the quantum theory. One important feature of this method is that since the topology of the manifold is not specified at the outset it is possible to include a sum over paths in different topologies. The outcome of this idea is that the vacuum would, at the level of the Planck length, $(G\hbar/c^3)^{1/2} \approx 10^{-35}$ m, acquire a foamlike structure. *See* FEYNMAN INTEGRAL; GREEN'S FUNCTION.

Problems. At present a complete, consistent theory of quantum gravity is still lacking. The formal theory fails to satisfy the power-counting criterion for renormalizability. In every order of the perturbation expansion, new divergences appear which could only be canceled by counterterms that do not exist in the original lagrangian. This may not be just a technical problem but the reflection of a conceptual difficulty stemming from the dual role, geometric and dynamic, played by the metric. *See* RENORMALIZA-TION.

Quantum cosmology. The existence of a singularity in generic solutions of Einstein's equation of general relativity was proven in 1970 under very general and realistic assumptions. Thus the big bang theory of the universe, based on classical general relativity, was given a solid theoretical foundation. However, near the singularity at the origin of the universe the classical description breaks down and the concepts of quantum mechanics are required. The universe should be described in terms of a wave function whose determination depends on what is assumed as initial boundary conditions. If the wave function is constructed by using R. P. Feynman's formalism of sum over histories, it is convenient to rotate the time in the complex plane to the imginary axis, turning "real time" into "imaginary time." The Minkowski metric of real space-time becomes a euclidean metric in the new manifold. Compact geometries of this four-dimensional euclidean manifold can then be considered.

Among the boundary conditions that can be imposed in such geometries is the condition of no boundary. This hypothesis has been advocated as appropriate for the construction of the wave function of the universe. Here not only would the singularity of the classical theory disappear in the quantum theory, but even the need to specify boundary conditions would be circumvented by the no-boundary hypothesis; in the euclidean construction of the wave function, all possible compact four-geometries should be included in the sum over histories without imposing any initial conditions at all. *See* BIG BANG THEORY; COSMOLOGY.

Problem of the cosmological constant. The cosmological constant is a parameter that can be introduced

in general relativity with the effect of giving a nonvanishing curvature to space-time even in the absence of matter. There are very stringent limits derived from astronomical observations on the value of the cosmological constant, which is either zero or extremely small. In a quantum theory, a cosmological constant will result from quantum fluctuations, even if it is absent in the classical action. It appears that a precise fine tuning of the boundary conditions at the Planck scale is required to obtain a vanishing cosmological constant. This is a major problem in quantum gravity whose resolution requires some mechanism to drive the cosmological constant to zero almost independently of the physics below the Planck scale, that is, for most boundary conditions on the scale.

Supergravity and superstrings. Supergravity is a geometric extension of general relativity which incorporates the principle of supersymmetry. Supersymmetry is a kind of symmetry, discovered in the 1970s, that allows for the transformation of fermions and bosons into each other. (Fermions carry half-integer spin while bosons carry integer spin; they also obey different statistics.) Supergravity can be formulated in space-time manifolds with a total of D = d + 1 dimensions, where *d*, the number of space dimensions, can be as large as 10. They constitute truly unified theories of all interactions including gravity.

One motivation for studying such theories was the hope that, because of the extra symmetry, they were better behaved when quantized, that is, that supersymmetry would make the divergencies of the quantum theory of gravity cancel out. This was expected, above all, of the D = 11 theory after reduction to four dimensions by some dynamical mechanism of compactification of the extra seven dimensions. All the efforts to achieve this goal proved unsuccessful. Moreover, it became apparent that the theories obtained in four dimensions by this process of dimensional reduction did not have some of the basic features of the physics of the real world at low energies. A possible exception might have been some theory derived from D = 10 supergravity with an additional gauge symmetry. See GAUGE THEORY.

In the early 1980s, some encouraging results were found with a theory based on the idea that the basic objects of nature are not pointlike but actually one-dimensional objects like strings, which can be open or closed. Consistency of the quantum theory of bosonic strings requires that they exist in a D = 26 space-time manifold, but the action principle for their motion is formulated in two dimensions. Conformal invariance of the action is a fundamental symmetry of the theory. It can be shown that the bosonic string theory is not realistic because it contains an unphysical tachyon. Incorporating supersymmetry into the theory leads to a critical dimension D = 10 and the removal of the tachyon. *See* TACHYON.

The relevance of superstring theories to quantum gravity is that they contain a zero-mass graviton (spin-2 state), and the equation of motion for a background metric in which the string is quantized reproduces (in some approximation) Einstein's equation.

In the approximation of neglecting string excitations, certain superstring models may be described in terms of local fields as a D = 10 supergravity theory. However, at least in every order of a certain perturbation expansion (a loop expansion) the quantized superstring theory is finite. At present, these are the only theories that both include gravity and can be consistently quantized. Although a superstring theory may eventually become the ultimate theory of all the interactions, there is still a very long way to go in making the connection between its fundamental fields and the fields representing the particles and their interactions as observed at low energies. See FUNDAMENTAL INTERACTIONS; GRAVITA-TION; SUPERGRAVITY; SUPERSTRING THEORY; SUPER-SYMMETRY. Samuel W. MacDowell

Bibliography. A. Ashketar and J. Stachel (eds.), Conceptual Problems of Quantum Gravity, 1991; S. W. Hawking, A Brief History of Time, 1988; S. W. Hawking and W. Israel (eds.), General Relativity: An Einstein Centenary Survey, 1979; C. J. Isham, R. Penrose, and D. W. Sciama (eds.), Quantum Gravity: An Oxford Symposium, 1975, and Quantum Gravity 2: A Second Oxford Symposium, 1981.

Quantum mechanics

The modern theory of matter, of electromagnetic radiation, and of the interaction between matter and radiation; also, the mechanics of phenomena to which this theory may be applied. Quantum mechanics, also termed wave mechanics, generalizes and supersedes the older classical mechanics and Maxwell's electromagnetic theory. Atomic and subatomic phenomena provide the most striking evidence for the correctness of quantum mechanics and best illustrate the differences between quantum mechanics and the older classical physical theories. Quantum mechanics is needed to explain many properties of bulk matter, for instance, the temperature dependence of the specific heats of solids.

The formalism of quantum mechanics is not the same in all domains of applicability. In approximate order of increasing conceptual difficulty, mathematical complexity, and likelihood of future fundamental revision, these domains are the following: (i) Nonrelativistic quantum mechanics, applicable to systems in which particles are neither created nor destroyed, and in which the particles are moving slowly compared to the velocity of light, as shown by relation (1). Here a particle is defined as a material

$$c \cong 3 \times 10^{10} \text{ cm/s} \tag{1}$$

entity having mass, whose internal structure either does not change or is irrelevant to the description of the system. (ii) Relativistic quantum mechanics, applicable in practice to a single relativistic particle (one whose speed equals or nearly equals *c*); here the particle may have zero rest mass, in which event, its speed must equal *c*. (iii) Quantum field theory, applicable to systems in which particle creation and destruction can occur; the particles may have zero or nonzero rest mass.

This article is concerned mainly with nonrelativistic quantum mechanics, which apparently applies to all atomic and molecular phenomena, with the exception of the finer details of atomic spectra. Nonrelativistic quantum mechanics also is well established in the realm of low-energy nuclear physics, meaning nuclear phenomena wherein the particles have kinetic energies less than about 10^8 electronvolts (1 eV = 1.6×10^{-12} erg, and is the energy gained by an electron in traversing a potential difference of 1 volt). Many quantum-mechanical predictions are not as quantitatively accurate for nuclei as for atomic and molecular systems, however, because nuclear forces are not accurately known. *See* ATOMIC STRUC-TURE AND SPECTRA.

The novel (from the standpoint of classical physics) features of nonrelativistic quantum mechanics are described in this article. Some of these features are retained, others modified, in the more complicated domains of relativistic quantum mechanics and quantum field theory. *See* QUAN-TUM ELECTRODYNAMICS; QUANTUM FIELD THEORY; RELATIVISTIC QUANTUM THEORY; SYMMETRY LAWS (PHYSICS).

For the formal mathematical structure of nonrelativistic quantum mechanics and justification for many of the assertions made in this article. *See* NON-RELATIVISTIC QUANTUM THEORY.

Planck's constant. The quantity 6.626×10^{-34} joule-second, first introduced into physical theory by Max Planck in 1901, is a basic ingredient of the formalism of quantum mechanics. Most of the fundamental quantum-mechanical relations, for example, Schrödinger's equation and Heisenberg's uncertainty principle, explicitly involve Planck's constant, as do many of the well-verified consequences of quantum mechanics, for example, the formula for the energy levels of atomic hydrogen. Planck's constant plays no role in the classical theories. Planck's constant commonly is denoted by the letter *b*; the notation $\hbar = b/2\pi$ also is standard.

Uncertainty principle. In classical physics the observables characterizing a given system are assumed to be simultaneously measurable (in principle) with arbitrarily small error. For instance, it is thought possible to observe the initial position and velocity of a particle and therewith, using Newton's laws, to predict exactly its future path in any assigned force field. According to the uncertainty principle (W. Heisenberg, 1927), accurate measurement of an observable quantity necessarily produces uncertainties in one's knowledge of the values of other observables. In particular, for a single particle relation (2*a*) holds,

$$\Delta x \Delta p_x > \hbar \tag{2a}$$

$$\Delta \phi \Delta l_z \gtrsim \hbar \tag{2b}$$

$$\Delta x \Delta E \gtrsim \frac{\hbar}{m} p_x \tag{2c}$$

(2d)

 $\Delta t \Delta E \gtrsim \hbar$

where Δx represents the uncertainty (error) in the location of the *x* coordinate of the particle at any instant, and Δp_x is the simultaneous uncertainty in the *x* component of the particle momentum. Relation (2*a*) asserts that under the best circumstances, the product $\Delta x \Delta p_x$ of the uncertainties cannot be less than about 10^{-27} erg-s; of course, with poor measurements, the product can be much greater than \hbar . On the other hand, there is no restriction on the simultaneous determination of position along *x* and momentum along *y*; that is, the product $\Delta x \Delta p_y$ may equal zero. Other typical uncertainty inequalities for a single particle are given by relations (2*b*) and (2*c*).

In relation (2b) the particle location is specified in spherical coordinates, with polar axis along z; Δ^{ϕ} is the uncertainty in azimuth angle; and Δl_z is the uncertainty in the *z* component of the orbital angular momentum. In relation (2c) ΔE is the uncertainty in energy and *m* is the particle mass. The uncertainty relation (2*d*) is derived and interpreted somewhat differently than relations (2*a*)–(2*c*); it asserts that for any system, an energy measurement with error ΔE must be performed in a time not less than $\Delta t \sim \hbar/\Delta E$. If a system endures for only Δt s, any measurement of its energy must be uncertain by at least $\Delta E \sim \hbar/\Delta t$ ergs.

Because the numerical value of \hbar is so small, and since $\Delta p_x = m \Delta v_x$, v denoting velocity, the restrictions implied by relations (2*a*)–(2*c*) are utterly inconsequential for macroscopic systems, wherein masses are of the order of grams. For an electron, however, whose mass is 9.1 × 10⁻²⁸ g, $\hbar/m \sim 1$ cm²/s, and the uncertainty principle cannot be ignored. Similarly, relation (2*d*) is unimportant for macroscopic systems, wherein energies are of the order of ergs, but is significant for atomic systems where $\Delta E = \hbar/\Delta t$ need not be negligible compared to the actual energy *E. See* UNCERTAINTY PRINCIPLE.

Wave-particle duality. It is natural to identify such fundamental constituents of matter as protons and electrons with the mass points or particles of classical mechanics. According to quantum mechanics, however, these particles, in fact all material systems, necessarily have wavelike properties. Conversely, the propagation of light, which, by Maxwell's electromagnetic theory, is understood to be a wave phenomenon, is associated in quantum mechanics with massless energetic and momentum-transporting particles called photons. The quantum mechanical synthesis of wave and particle concepts is embodied in the de Broglie relations, given by Eqs. (*3a*) and (*3b*). These give the wavelength λ and

$$\lambda = \frac{b}{p} \tag{3b}$$

$$f = \frac{E}{b} \tag{3b}$$

wave frequency f associated with a free particle (a particle moving freely under no forces) whose momentum is p and energy is E; the same relations give the photon momentum p and energy E associated with an electromagnetic wave in free space (that is,

in a vacuum) whose wavelength is λ and frequency is *f. See* PHOTON.

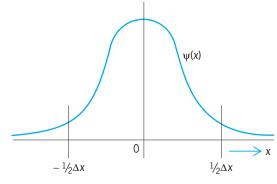
The wave properties of matter have been demonstrated conclusively for beams of electrons, neutrons, atoms (hydrogen, H, and helium, He), and molecules (H₂). When incident upon crystals, these beams are reflected into certain directions, forming diffraction patterns. Diffraction patterns are difficult to explain on a particle picture; they are readily understood on a wave picture, in which wavelets scattered from regularly spaced atoms in the crystal lattice interfere constructively along certain directions only. Moreover, the wavelengths of these "matter waves," as inferred from the diffraction patterns, agree with the values computed from Eq. (3a), as first demonstrated by C. J. Davisson and L. H. Germer in 1927. See ELEC-TRON DIFFRACTION; NEUTRON DIFFRACTION.

Photoelectric effect. The particle properties of light waves are observed in the photoelectric effect. When light of frequency f causes electrons to be emitted from a surface, all the electrons have very nearly the same maximum kinetic energy; the maximum kinetic energy is independent of the light intensity; the number of electrons emitted in unit time is proportional to the light intensity; as f is varied, the maximum electron kinetic energy W varies linearly with f, in fact, W = hf - C, W = hf - C, C being a constant characteristic of the emitting material. These observations are difficult to understand on the wave picture, wherein the magnitude of the electric field vector (which presumably exerts the force which ejects the electron) is proportional to the square root of the incident light intensity, and is not directly related to nor limited by, the incident light frequency f. The photoelectric effect is interpreted readily on the assumptions that the energy in the light beam is carried in quanta of energy E = bf; that emission of an electron results from absorption of a single quantum (a single photon); and that absorption of half or any fraction of a quantum is not possible, because the photons act as discrete indivisible entities. For additional information on the photoelectric effect see PHOTOEMISSION.

Compton effect. The particle properties of electromagnetic waves also are demonstrated in the Compton effect, wherein the wavelengths of x-rays are lengthened by scattering from free electrons. The change in wavelength is predicted quantitatively, assuming the scattering results from elastic collisions between photons and electrons, and using Eqs. (3a) and (3b) for the photon momentum and energy. The diffraction of x-rays by crystals was regarded, in the prequantum era, as conclusive proof that x-rays are waves and not "corpuscles." *See* COMPTON EFFECT; ELECTROMAGNETIC RADIATION; MAXWELL'S EQUA-TIONS; X-RAY DIFFRACTION.

Interference and diffraction. Wave propagation is distinguished from particle propagation by the phenomena of interference and diffraction. It is a general result of wave theories that interference and diffraction effects largely are confined to an angle (relative to the incident beam) which in radians equals about λ/d , where d is a characteristic dimension of the system causing the diffraction or interference, for example, the width of the slit diffracting the wave, or the distance between two interfering scattering centers. This fact and the magnitudes of λ inferred from Eq. (3a) are sufficient to explain why wave effects are not observed in the propagation of ordinary macroscopic bodies, but can be observed in the propagation of electrons, neutrons, and light atoms or molecules. For example, for a mass of 1 g moving at 1 cm/s, $\lambda = 6.6 \times 10^{-27}$ cm. But for a neutron or hydrogen atom (using p = Mv = 2ME, where $M = 1.66 \times 10^{-24}$ g) moving at velocity corresponding to room temperature (300 K), E = 3/2kT (k =Boltzmann's constant = 1.38×10^{-16} erg deg⁻¹), and λ turns out to equal 1.45 \times 10^{-8} cm. For an electron with an energy of 100 eV, $\lambda = 1.22 \times 10^{-8}$ cm. For a proton with an energy $10^6 \text{ eV} = 1 \text{ MeV}, \lambda =$ 2.88×10^{-12} cm. These numerical results and the discussion of this paragraph also explain the ability of crystals, wherein interatomic spacings are about 10^{-8} cm, to give a good demonstration of electron and molecular diffraction; suggest the need for quantum mechanics to "understand" atomic systems, wherein atomic dimensions are $\sim 10^{-8}$ cm and electron energies are ~ 10 eV; suggest the need for quantum mechanics to "understand" atomic nuclei, wherein nuclear dimensions are $\sim 10^{-13}$ cm and neutron or proton energies are ~ 10 MeV; and explain why quantum effects are more readily observed in H₂ and He than in heavier gases, and at low temperatures rather than high. See DIFFRACTION; INTERFER-ENCE OF WAVES.

Relationship to uncertainty principle. Wave-particle duality is intimately connected with the uncertainty principle in that the uncertainty inequalities can be derived from analyses of specific experiments. For a nonrelativistic particle the connection can be seen from the following argument, which contains the basic elements of the rigorous formal treatment. As explained later, the probability that the *x* coordinate of the particle will lie in the interval *x* to x + dxis $|\psi(x)|^2 dx$, where $\psi(x)$ is called the wave function. Suppose measurement has ascertained that the particle lies in an interval of width Δx centered at x = 0; that is, measurement has determined that the dependence of $\psi(x)$ on *x* is approximately as shown in **Fig. 1**. Because of wave-particle duality, the wave





packet of Fig. 1 can be looked upon as a superposition of waves. Since $\psi(x)$ rises from and falls to a very small value in an interval Δx , the packet must contain waves whose half wavelengths are as small as Δx ; furthermore, it can be proved that because $\psi(x)$ does not change sign, the packet must contain waves of very long wavelength. Thus, in the packet, the wavelengths λ run from about $2\Delta x$ to ∞ ; the reciprocal wavelengths λ^{-1} run from about 0 to $1/(2\Delta x)$. But from Eq. (3*a*), $\Delta p_x = b \Delta \lambda^{-1}$. Hence $\Delta p_x \Delta x$ cannot be less than about b/2. Considering the simplicity of the argument, this result is close enough to relation (2a). Similarly, relation (2d) can be understood from Eq. (3b) and from the fact that to transmit information in a time Δt , that is, to turn a measuring instrument on and off in an interval Δt , it is necessary to use frequencies higher than about $(2\Delta t)^{-1}$. Using relation (2a), the discussion of this paragraph suggests that in atoms, whose dimensions Δx are about 10^{-8} cm, one must expect to find electrons with energies as given by Eq. (4), which is of the order

$$E = \frac{p^2}{2m} = \frac{\hbar}{2m(\Delta x)^2} = 3.8 \text{ eV}$$
 (4)

of magnitude observed. Similarly, in atomic nuclei, whose dimensions Δx are about 10^{-13} cm, one must expect to find neutrons or protons with energies about 20 MeV, again of the order of magnitude observed. On the other hand, if atomic nuclei contained electrons, the electron energies would be as high as 200 MeV. In computing this value, it is necessary to use the relativistic relation $E = m^2 c^4 + c^4 c^4$ c^2p^2 connecting *E* and *p*. These energies are much too high to be explained by electrostatic forces between electrons and protons at separations of about 10^{-13} cm. Thus, the uncertainty principle leads to the inference that electrons are not contained in atomic nuclei, which inference is in accord with current theories of nuclear structure and beta decay. See NUCLEAR STRUCTURE; RADIOACTIVITY; RELATIVISTIC MECHANICS.

Complementarity. Wave-particle duality and the uncertainty principle are thought to be examples of the more profound principle of complementarity, first enunciated by Niels Bohr (1928). According to the principle of complementarity, nature has "complementary" aspects; an experiment which illuminates one of these aspects necessarily simultaneously obscures the complementary aspect. To put it differently, each experiment or sequence of experiments vields only a limited amount of information about the system under investigation; as this information is gained, other equally interesting information (which could have been obtained from another sequence of experiments) is lost. Of course, the experimenter does not forget the results of previous experiments, but at any instant, only a limited amount of information is usable for predicting the future course of the system.

The well-known double-slit experiment provides a good illustration of the principle of complemen-

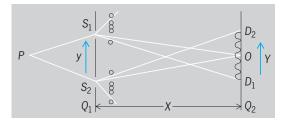


Fig. 2. Double-slit experiment.

tarity. Light from a monochromatic point source P(Fig. 2) is diffracted by the two slits S_1 and S_2 in the screen Q_1 . On the screen Q_2 , an interference pattern of alternating bright and dark bands is formed in the region D_1D_2 , where the two diffraction patterns (from the slits S_1 and S_2) overlap. Assuming that *P* is equidistant from the slits, and also that the slits are very narrow compared to the distances PS_1 or S_1O , it follows that interference maxima (bright bands) are observed whenever $S_2O - S_1O$ equals $n\lambda$, λ being the wavelength of the light and n an integer; when $S_2O - S_1O = (n + 1/2)\lambda$, interference minima (dark bands) are observed. In moving from any maxima to an adjacent maximum, the path difference $S_2O - S_1O$ changes by precisely one wavelength. Consequently, measurement of the distance Y between successive maxima, and knowledge of S_1S_2 and of the distance X between Q_1 and Q_2 , yields λ via the formula $\lambda = Yd/X$ (valid when $S_1S_2 = d$ is much smaller than X). Evidently, the double-slit experiment is understandable in terms of, and provides information concerning, the wave properties of light.

The double-slit experiment yields no information concerning the particle properties of light; in fact, introducing the particle picture leads only to conceptual difficulties. These difficulties appear with the recognition that reducing the source intensity does not modify the interference pattern; after a sufficiently long exposure, a photographic film at Q2 will show exactly the same interference pattern D_1D_2 as is observed by the eye using a more intense source. Since it is possible to make the source intensity so low that two photons almost never will be emitted during the very small time required for light to travel from P to Q_2 via either of the slits, it is necessary to conclude that the interference pattern is produced by independent individual photons, and not by interference between two or more different photons. On the other hand, the interference pattern is destroyed when either of the slits is closed. Thus, the question arises: How can a stream of independent photons, each of which presumably passes through only one of the slits, and half of which on the average pass through S_1 , produce an interference pattern that is destroyed by closing one of the slits? Or to put it differently, how can closing or opening a slit through which a photon does not pass affect the likelihood of that photon reaching any particular point on Q_2 ?

The principle of complementarity meets these difficulties with the assertion that the possibility of

demonstrating that the photons have well-defined trajectories through one or the other slit (a particlelike property) is complementary to the possibility of demonstrating the wavelike property of interference. In the double-slit experimental setup which has been described, until the photon is localized at Q_2 (by the visible evidence that a chemical effect has occurred in a photographic film), it is not possible to locate the photon at any particular point in space, nor is it legitimate to insist that the photon must have passed through only one of the slits. Moreover, according to the principle of complementarity, modifying the experimental setup so as to localize the photon at one of the slits, and thereby to determine through which slit the photon passes, necessarily destroys the interference pattern. This last assertion is supported by analysis of various photondetection schemes, recognizing that the proposed experiments perforce are entirely Gedanken (in the mind); an actual measurement of the slit through which the photon passes demands extreme precision, and this has not been attempted. It is concluded that quantum mechanics involves no inconsistencies of paradoxes. From the standpoint of the complementarity principle, the questions presented in the preceding paragraph, and other similar difficulties, rest always on the specious assumption of more information than actually is obtainable. See SUPERPOSI-TION PRINCIPLE.

The limitations on the experiments are imposed by the requirements of the uncertainty principle, relations (2), and wave-particle duality, Eqs. (3). To illustrate the analysis, suppose indicators (symbolized by circles), free to move in the vertical y direction, are placed immediately behind the slits S_1 and S_2 in Fig. 2; the recoil of an indicator signifies a collision with a photon and places the photon at the indicator. In order that the recoil establish the slit through which the photon has passed, the uncertainty in the vertical location of each indicator must be much less than d/2. According to relation (2*a*), this means that the vertical momentum p_y of each indicator will be uncertain by amount $\Delta p_{\gamma} \gg 2\hbar/d$. There can be no assurance that the indicator has recoiled unless δp_{ν} , the momentum transferred from photon to indicator, equals or exceeds Δp_{γ} . When the slits are narrow and $X \gg d$, a momentum transfer δp_y deflects the photon through an angle $\theta \sim \delta p_{\nu}/p$, and changes the point Y where the photon strikes the screen by an amount $\Delta Y = \theta X \cong X \delta p_{\nu}/p$, with $p = b/\lambda$, Eq. (3a). It follows that when there are observable recoils localizing the photon at one of the slits, then $\Delta Y \gg \lambda X/\pi d$; that is, $\Delta Y \gg Y/\pi$, where Y/2 = $\lambda X/2d$ is the distance between adjacent maxima and minima of the interference pattern. Thus, determining the slit through which the photon passes necessarily gives the photon an uncontrollable random vertical deflection (on the screen Q_2), which is very much larger than the distance between adjacent maxima and minima. The photons now spread with uniform intensity over vertical distances ΔY , which are large compared to the original widths of the light and dark interference bands; in other words, the interference pattern is destroyed.

Quantization. In classical physics the possible numerical values of each observable, meaning the possible results of exact measurement of the observable, generally form a continuous set. For example, the xcoordinate of the position of a particle may have any value between $-\infty$ and $+\infty$; similarly p_x , the x component of momentum, may have any value between $-\infty$ and $+\infty$; the kinetic energy T of a particle may have any value between 0 and $+\infty$; the total energy (kinetic plus potential) of an electron in the field of a proton may have any value between $-\infty$ and $+\infty$; the orbital angular momentum vector **l** of a particle moving in a central field, for example, an electron in a hydrogen atom, may have any magnitude between 0 and ∞ ; if the magnitude of **l** is known to be *l*, then, since the plane of the orbit may be arbitrarily oriented, the z component of **1** may have any value between -l and +l. In quantum mechanics the possible numerical values of an observable need not form a continuous set, however. For some observables, the possible results of exact measurement form a discrete set; for other observables, the possible numerical values are partly discrete, partly continuous; for example, the total energy of an electron in the field of a proton may have any positive value between 0 and $+\infty$, but may have only a discrete set of negative values, namely, -13.6, -13.6/4, -13.6/9, -13.6/16 eV, Such observables are said to be quantized; often there are simple quantization rules determining the quantum numbers which specify the allowable discrete values. For example L^2 , the square of the orbital angular momentum of a particle, must equal $l(l+1)\hbar^2$, where l is zero or a positive integer; the only allowed values of J^2 , the square of the total angular momentum (orbital angular momentum plus intrinsic angular momentum or spin), are $J^2 = j(j + 1)\hbar^2$ where j = 0, 1/2, 1, 3/2, 2, ...; that is, $j\hbar$ is an integral or half-integral multiple of $b/2\pi$. If the magnitude of the total angular momentum is given by some one of these values, $j\hbar = 3\hbar/2$, for example, then the only allowed values of the z component J_z of **J** are $-3\hbar/2$, $-\hbar/2$, $\hbar/2$, $3\hbar/2$; that is, for a given j, the allowed values of $J_z = -j, -j + 1, -j +$ 2,..., j - 1, j, in units of \hbar . On the other hand, the observables x, p_x , and T for a relativistic particle are not quantized, and these observables have precisely the same allowed values in quantum mechanics as they do classically. In the formal theory each observable is a linear operator, whose eigenvalues (characteristic values) are the allowed values of that observable. The set of eigenvalues is termed the spectrum of the operator, which spectrum may be discrete, continuous, or mixed. See QUANTUM NUMBERS.

The fact that nature is quantized has been amply verified experimentally. For instance, the quantization of energy and momentum in light waves is demonstrated in the photoelectric and Compton effects, described earlier.

Stern-Gerlach experiment. The quantization of angular momentum is strikingly exhibited in the

Stern-Gerlach experiment, wherein a beam of, for example, hydrogen atoms moving through a region of space containing an inhomogeneous magnetic field breaks up into two separate beams. Classically, the force on a hydrogen atom in an inhomogeneous magnetic field depends on the angle between the magnetic field and the plane of the electron's orbit. Thus, classically, the original beam, in which all orientations of the plane of the orbit are possible, is expected to spread out or defocus in the inhomogeneous magnetic field, but is not expected to form two distinct focused beams. Formation of two beams agrees, however, with the quantum mechanical prediction that the square of the total angular momentum of atomic hydrogen is $J^2 = 3\hbar^2/4 =$ $1/2(1/2 + 1)\hbar^2$, and therefore that the vector **J** (which classically always is perpendicular to the orbit plane) must be either parallel or antiparallel to the applied magnetic field (along z); for these two directions of J, and only for these two, the value of J_z equals one of the permitted values $J_z = -\hbar/2$ or $J_z = \hbar/2$. The Stern-Gerlach experiment also can be performed with beams of other atoms and molecules, as well as with neutron beams. In all cases the observations agree with quantummechanical expectation. Refinements of the experiment yield accurate measurements of the parameters characterizing the magnetic interactions of atoms, molecules, and atomic nuclei. See MAGNETIC RESONANCE; MOLECULAR BEAMS; NUCLEAR MOMENTS.

Atomic spectra. Spectroscopy, especially the study of atomic spectra, probably provides the most detailed quantitative confirmation of quantization. Granting that the energy in a light wave is quantized, it follows from conservation of energy and Eq. (3b) that when an atom emits a photon of frequency f, its initial energy E_i and final energy E_f after emission are related by Eq. (5*a*). Similarly, when a photon of frequency f is absorbed, the initial energy E'_i , and final energy E'_f after absorption, satisfy Eq. (5*b*).

$$E_i - E_f = bf \tag{5a}$$

$$E'_f - E'_i = bf \tag{5b}$$

In quantum mechanics, as in classical mechanics, an electron remains in the vicinity of a proton, that is, it is "bound" to the proton, when and only when the magnitude of its kinetic energy *T* is less than the magnitude of its negative potential energy *V*(*V* is set equal to zero at infinite separation). Thus, the total energy of a stable hydrogen atom necessarily is negative; in other words, the only allowed energies of atomic hydrogen are $-R/n^2$, where *n* is an integer and R = -13.6 eV. Consequently, the radiation emitted by atomic hydrogen must consist of a discrete set of frequencies, or lines, obeying the relation, from Eq. (5*a*), given by Eq. (6),

$$f = \frac{R}{b} \left(\frac{1}{n^2} - \frac{1}{m^2} \right) \tag{6}$$

where *m* and *n* are integers, m > n. This simple argument provides a convincing explanation of the observation that atomic hydrogen has a line spectrum and not a continuous spectrum; in other words, it radiates discrete frequencies rather than a continuous band of frequencies. The observed lines very accurately satisfy Eq. (6); moreover, except for small relativistic and quantum field theory effects, the nonrelativistic Schrödinger equation accurately predicts not merely the frequencies of the lines, but also their relative intensities and widths. Because on the average the levels E_i and E_f of Eq. (5*a*) endure only for a limited time, their energies are uncertain by an amount ΔE , as explained in the earlier discussion of the uncertainty principle; therefore, the observed lines have a width not less than $\Delta f \cong b^{-1}(\Delta E_i + \Delta E_f)$, according to Eq. (5a). Heavier atoms have more complicated line spectra than hydrogen, and the frequencies they emit cannot be described by formulas as simple as Eq. (6). For these atoms, the agreement between experiment and the predictions of the nonrelativistic Schrödinger equation, though always very good, is not as precise as for hydrogen. All the evidence indicates, however, that the discrepancies arise because the Schrödinger equation cannot be solved exactly in many-electron atoms, so that the theoretical predictions necessarily are approximate. If approximations were not necessary (for example, if perfect computing machines were available), there is every reason to think that the predictions of the Schrödinger equation would be exactly correct, except for the aforementioned small relativistic and field theory effects.

Probability considerations. The uncertainty and complementarity principles, which limit the experimenter's ability to describe a physical system, must limit equally the experimenter's ability to predict the results of measurement on that system. Suppose, for instance, that a very careful measurement determines that the x coordinate of a particle is precisely $x = x_0$. This is permissible in nonrelativistic quantum mechanics. Then formally, the particle is known to be in the eigenstate corresponding to the eigenvalue $x = x_0$ of the x operator. Under these circumstances, an immediate repetition of the position measurement again will indicate that the particle lies at $x = x_0$; if the particle is moving in a one-dimensional force field, described by the potential V(x), the particle's potential energy will be exactly $V(x_0)$. Knowing that the particle lies at $x = x_0$ makes the momentum p_x of the particle completely uncertain, however, according to relation (2a). A measurement of p_x immediately after the particle is located at x = x_0 could yield any value of p_x from $-\infty$ to $+\infty$ a measurement of $T = p_x^2/2m$ could yield any value from 0 to $+\infty$, and in fact the average or expectation value of T in these circumstances would be infinite.

More generally, suppose the system is known to be in the eigenstate corresponding to the eigenvalue α of the observable *A*. Then for any observable *B*, which is to some extent complementary to *A*, that is, for which an uncertainty relation of the form of relations (2) limits the accuracy with which *A* and *B* can simultaneously be measured, it is not possible to predict which of the many possible values $B = \beta$ will be observed. However, it is possible to predict the relative probabilities $P_{\alpha}(\beta)$ of immediately thereafter finding the observable *B* equal to β , that is, of finding the system in the eigenstate corresponding to the eigenvalue $B = \beta$. If the system is prepared in the eigenstate α of *A* a great many times, and each time the observable *B* is measured immediately thereafter, the average of these observed values of *B* will equal the expectation value of *B*, defined by Eq. (7), summed over all eigenvalues

$$\langle B \rangle = \sum \beta P_{\alpha}(\beta) \tag{7}$$

of *B*; when the spectrum is continuous, the summation sign is replaced by an integral. To the eigenvalues correspond eigenfunctions, in terms of which $P_{\alpha}(\beta)$ can be computed. In particular, when α is a discrete eigenvalue of *A*, and the operators depend only on *x* and p_x , the probability $P_{\alpha}(\beta)$ is postulated as in Eq. (8), where $u(x,\alpha)$ is the eigenfunction corre-

$$P_{\alpha}(\beta) = \left| \int_{-\infty}^{\infty} dx \, v^*(x,\beta) u(x,\alpha) \right|^2 \tag{8}$$

sponding to $A = \alpha$; $v(x,\beta)$ is the eigenfunction corresponding to $B = \beta$; and the asterisk denotes the complex conjugate. Since measurement of *A* in the state $A = \alpha$ must yield the result $A = \alpha$, it is necessary that the states $u(x,\alpha)$ satisfy the normalizing and orthogonalizing relation in Eq. (9), where $\delta_{\alpha\alpha'} = 0$

$$\left|\int_{-\infty}^{\infty} dx \, u^*(x,\alpha')u(x,\alpha)\right|^2 = \delta_{\alpha\alpha'} \tag{9}$$

when $\alpha \neq \alpha'$ (eigenfunctions normalized); the eigenfunctions $\upsilon(x,\beta)$ are similarly orthonormal. The integral in Eq. (8) is called the projection of $u(x,\alpha)$ on $\upsilon(x,\beta)$. The projection of the eigenfunction corresponding to $A = \alpha$ on the eigenfunctions of the *x* operator is $u(x,\alpha)$, and $|u(x,\alpha)|^2 dx$ is the probability that the system, known to be in the eigenstate $A = \alpha$, will be found in the interval *x* to x + dx. See EIGEN-FUNCTION; EIGENVALUE (QUANTUM MECHANICS).

The formalism just described embodies the essential feature that each measurement on an individual system, as it develops new information, necessarily loses or makes untrue some information gained in the past; in fact, this formalism leads to a rigorous derivation of uncertainty relations (2). For example, suppose B is measured with the system in the state $A = \alpha$, and it is found that $B = \beta$ exactly. The act of measurement necessarily and unavoidably disturbs the system, with the result that after the measurement, the system is in the eigenstate $v(\beta,x)$. After the measurement, therefore, it no longer is certain that $A = \alpha$; rather the probability of finding $A = \alpha$ is $P_{\beta}(\alpha)$, which by Eq. (8) equals $P_{\alpha}(\beta)$. Thus, after starting with $A = \alpha$, and then measuring $B = \beta$, it is possible to find that $A = \alpha' \neq \alpha$. Of course, as stressed previously, these considerations are unimportant for macroscopic systems, where the limitations imposed by the uncertainty principle are inconsequential. Even if all observations were extremely accurate by usual standards, when the momentum p_x of a 1-g mass at x is measured, the position x' immediately thereafter should be indistinguishable from x. Were x' and x distinguishable, the particle position seemingly would have changed discontinuously from x to x', contrary to all classical (macroscopic) experience.

This formalism yields predictions in excellent agreement with observation; furthermore, it can be seen that the formalism is internally consistent. Consequently, the following doctrine, embodied in a formalism, appears well established: Although it is possible to predict the average of a large number of observations on identical systems, the result of a measurement on a single (microscopic) system generally is unpredictable and largely a matter of chance. Nonetheless, some physicists have refused to accept this inherent indeterminancy of nature and believe that this doctrine is a serious deficiency of present physical theory. To put the problem in simplest terms, consider a gram of radium, containing approximately 10²¹ atoms. According to generally accepted theory, it is not possible to predict when any one atom will decay, but it is possible to predict very accurately the average number of atoms decaying every second. The objectors to this doctrine feel that it must be possible to predict the subsequent history of every individual atom; failure to do so represents, not an inherent indeterminism in nature, but rather a lack of obtainable information-and therefore a lack of understanding-concerning the mechanism of radioactive decay. To mention but one possible alternative, nonrelativistic quantum theory can be reinterpreted in terms of hidden variables, which in principle determine the precise behavior of an individual system but whose values are not ascertained in measurements of the type which now can be carried out. This alternative has not led to new predictions, however, and contains some unappealingly ad hoc features. See CAUSALITY; PROBABILITY (PHYSICS).

Wave function. When the system is known to be in the eigenstate corresponding to $A = \alpha$, the eigenfunction $u(x,\alpha)$ is the wave function; that is, it is the function whose projection on an eigenfunction $\upsilon(x,\beta)$ of any observable B gives the probability of measuring $B = \beta$. The wave function $\psi(x)$ may be known exactly; in other words, the state of the system may be known as exactly as possible (within the limitations of uncertainty and complementarity), even though $\psi(x)$ is not the eigenfunction of a known operator. This circumstance arises because the wave function obeys Schrödinger's wave equation. Knowing the value of $\psi(x)$ at time t = 0, the wave equation completely determines $\psi(x)$ at all future times. In general, however, if $\psi(x,0) = u(x,\alpha)$, that is, if $\psi(x,t)$ is an eigenfunction of A at t = 0, then $\psi(x,t)$ will not be an eigenfunction of A at later times t > 0. For example, suppose at t = 0 a free particle (a particle moving under no forces) is known to be in an eigenstate for which the uncertainty in *x* is $(\Delta x)_0$; $(\Delta x)_0$ is approximately the *x* interval within which $|\psi(x,0)|^2$ is not negligibly small. Suppose further that, at t = 0, the product of the uncertainties in position and momentum is as small as possible: $(\Delta x)_0(\Delta p_x)_0 \cong \hbar$, compare relation (2*a*). Then it can be proved that $(\Delta x)_t$, the uncertainty in *x* at time *t*, satisfies Eq. (10), where *m* is the particle mass.

$$(\Delta x)_t = \left[(\Delta x)_0^2 + \frac{t^2}{m^2} (\Delta p_x)_0^2 \right]^{1/2}$$
(10)

Equation (10) is readily interpreted. The root-meansquare spread at time *t* results from $(\Delta x)_0$ and from an uncertainty in the distance the particle has traveled; the latter uncertainty is $t(\Delta v_x)_0 = t(\Delta p_x)_0/m$. If $(\Delta x)_0 = 0$, meaning $\psi(x,0)$ is an eigenfunction of the x operator, $(\Delta x)_t$ is infinite, showing that $\psi(x,t)$ cannot be an eigenfunction of the x operator. When the particle is free, the projections of the wave function on the eigenfunctions of the momentum operator do not change with time, corresponding to the classical result that the momentum of a free particle does not change. Thus, the probability of measuring any value of the momentum does not change, $(\Delta p_x)_t =$ $(\Delta p_x)_0$, and Eq. (10) shows that $(\Delta x)_t (\Delta p_x)_t$ grows with time for a free particle, whatever the value of $(\Delta x)_0$, provided $(\Delta x)_0$ $(\Delta p_x)_0 \cong \hbar$. Nonetheless, $\psi(x,t)$ is known no less exactly than the initial wave function $\psi(x,0)$. The magnitude of $\Delta x \ \Delta p_x$ at any instant is no measure of the exactness with which the state of the system is known; increased uncertainties in position or momentum may be the price for increased certainty in the value of some other observable.

A system described by a wave function is said to be in a pure state. Not all systems are described by wave functions, however. Consider, for example, a beam of hydrogen atoms streaming in the *x* direction out of a small hole in a hydrogen discharge tube. According to the formal theory, if the beam were described by a wave function $\psi(x)$, then Eq. (11)

$$\psi(x) = C_{+}(x) u^{+}(z) + C_{-}(x) u^{-}(z)$$
(11)

holds, where $u^+(z)$ is the eigenfunction corresponding to finding a hydrogen atom with its z component J_z of total angular momentum equal to $\hbar/2$; $u^-(z)$ is the corresponding eigenfunction for finding $J_z =$ $-\hbar/2$; $|C_+(x)|^2$ is the probability of finding $J_z = \hbar/2$ at any point x along the beam; $|C_{-}(x)|^2$ is the corresponding probability of finding $J_z = -\hbar/2$. Since there are only two possibilities, $J_z = \pm \hbar/2$, $|C_+(x)|^2 +$ $|C_{-}(x)|^{2} = 1$, and since there seems no reason to favor either of these possibilities, it is reasonable to suppose that $|C_{+}(x)|^{2} = |C_{-}(x)|^{2} = \frac{1}{2}$. As Eq. (11) shows, however, to specify $\psi(x)$ it is necessary to know not merely the relative magnitudes of the complex numbers $C_{+}(x)$ and $C_{-}(x)$ but also their relative phase. It can be shown that each choice of relative magnitude and phase of C_+ and C_- corresponds to a direction γ for which there is probability 1 of finding $J_{\gamma} = \hbar/2$, and probability zero of finding $J_{\gamma} = -\hbar/2$.

Thus each choice of relative magnitude and phase of C_+ and C_- puts the system in an eigenfunction $u^+(\gamma)$ or $u^{-}(\gamma)$, that is, in a pure state. On the other hand, the discharge tube singles out no particular direction in space, so that in a Stern-Gerlach experiment the original beam must break up into two beams of equal intensity, whatever the direction of the external magnetic field. Consequently, the original beam is not in a pure state, but can be regarded as a statistical ensemble or mixture of pure states oriented with equal probability in all directions. Equivalently, Eq. (11) can be used for the original beam, provided calculations are averaged over all relative phases of C_+ and C_- . The distinction between mixtures and pure states is strongly analogous to the distinction between polarized and unpolarized light beams; consequently, beams of particles in pure spin states are termed polarized.

Schrödinger equation. Equation (12) describes a

$$\psi(x,t) = A(\lambda) \exp\left[2\pi i \left(\frac{x}{\lambda} - ft\right)\right]$$
 (12)

plane wave of frequency f, wavelength λ , and amplitude $A(\lambda)$, propagating in the positive x direction. The previous discussion concerning wave-particle duality suggests that this is the form of the wave function for a beam of free particles moving in the x direction with momentum $p = p_x$, with Eq. (3) specifying the connections between f, λ and E, p. Differentiating Eq. (12), it is seen that Eqs. (13) hold. Since for a

$$p_x \psi = \frac{h}{\lambda} \psi = \frac{h}{i} \frac{\partial \psi}{\partial x}$$
(13*a*)

$$E\psi = bf\psi = -\frac{\hbar}{i}\frac{\partial\psi}{\partial t}$$
(13b)

free particle $E = p^2/2m$, it follows also that Eq. (14) is valid.

$$\frac{-\hbar^2}{2m}\frac{\partial^2\psi}{\partial x^2} = -\frac{\hbar}{i}\frac{\partial\psi}{\partial t}$$
(14)

See WAVE MOTION.

Equation (14) holds for a plane wave of arbitrary λ , and therefore for any superposition of waves of arbitrary λ , that is, arbitrary p_x . Consequently, Eq. (14) should be the wave equation obeyed by the wave function of any particle moving under no forces, whatever the projections of the wave function on the eigenfunctions of p_x . Equations (13) and (14) further suggest that for a particle whose potential energy V(x) changes, in other words, for a particle in a conservative force field, $\psi(x,t)$ obeys Eq. (15).

$$\frac{-\hbar^2}{2m}\frac{\partial^2\psi}{\partial x^2} + V(x)\psi = -\frac{\hbar}{i}\frac{\partial\psi}{\partial t}$$
(15)

Equation (15) is the time-dependent Schrödinger equation for a one-dimensional (along x), spinless particle. Noting Eq. (13b), and observing that Eq. (15) has a solution of the form of Eq. (16), it is

$$\psi(x,t) = \psi(x) \exp\left(-iEt/\hbar\right)$$
(16)

inferred that $\psi(x)$ of Eq. (16) obeys the timeindependent Schrödinger equation (17).

$$\frac{-\hbar^2}{2m}\frac{\partial^2\psi}{\partial x^2} + V(x)\psi = E\psi$$
(17)

See FORCE.

Equation (17) is solved subject to reasonable boundary conditions, for example, that ψ must be continuous and must not become infinite as *x* approaches $\pm\infty$. These boundary conditions restrict the values of *E* for which there exist acceptable solutions $\psi(x)$ to Eq. (17), the allowed values of *E* depending on *V*(*x*). In this manner, the allowed energies of atomic hydrogen listed in the earlier discussion of quantization are obtained. When a $\psi(x)$ solving Eq. (17) exists, all probabilities inferred from the corresponding $\psi(x,t)$ are independent of time; see Eq. (8). Thus the allowed energies *E* of Eq. (17) are the energies of the stationary states (states not changing with time) of the system.

The forms of Eqs. (13*a*), (15), and (17) suggest that the classical observable p_x must be replaced by the operator $(\hbar/i) (\partial/\partial x)$. With this replacement, Eq. (18)

$$(xp_x - p_x x)\psi = i\hbar\psi \tag{18}$$

holds. In other words, whereas the classical canonically conjugate variables x and p_x are numbers, obeying the commutative law in Eq. (19a), the quantummechanical quantities x and p_x are noncommuting operators, obeying Eq. (19b).

$$xp_x - p_x x = 0 \tag{19a}$$

$$xp_x - p_x x = i\hbar \tag{19b}$$

In the formal theory the noncommutativity of x, p_x , leads directly to uncertainty relation (2*a*). For a derivation of relation (2*a*) from Eq. (19*b*), as well as for generalizations and more sophisticated derivations of Eqs. (13), (15), and (19*b*), *See* MATRIX MECHANICS; NONRELATIVIS-TIC QUANTUM THEORY; HAMILTON'S EQUATIONS OF MOTION.

Correspondence principle. Since classical mechanics and Maxwell's electromagnetic theory accurately describe macroscopic phenomena, quantum mechanics must have a classical limit in which it is equivalent to the older classical theories. Although there is no rigorous proof of this principle for arbitrarily complicated quantum-mechanical systems, its validity is well established by numerous illustrations, such as those mentioned in the preceding discussions of the uncertainty principle and wave-particle duality. In general, the classical limit is approached when (i) $b \rightarrow 0$; (ii) the mass becomes large; (iii) wavelengths become small; (iv) dimensions become large; (v) quantum numbers become large. (The notation $b \rightarrow 0$ refers to a mathematical operation in which one adheres to a fixed set of values of the other quantities involved and considers the effect of making *b* smaller and smaller.) These simple criteria must be employed cautiously, since they have not been stated in terms of dimensionless parameters; obviously the classical limit, in (ii) for instance, cannot depend on the unit of mass. Nonetheless, these criteria are useful guides; for example, Eq. (19*a*) is the limit of Eq. (19*b*) as $b \rightarrow 0$.

Before the introduction of the Schrödinger equation (1926) made possible exact determination of energy levels and related quantum numbers, the correspondence principle was very effectively employed as a heuristic means of arriving at the quantization rules. In particular, for periodic orbits there was evolved the rule in Eq. (20), when n is an integer,

$$\oint p \, dq = nb \tag{20}$$

q and p are canonically conjugate position and momentum variables, and the integration is performed along the orbit for one complete cycle. *See* HAMILTON-JACOBI THEORY; PERTURBATION (QUAN-TUM MECHANICS).

Equation (20) is not always exact, but nonetheless is useful, especially in the limit of high quantum numbers. In the case of a one-dimensional harmonic oscillator, for example, whose classical frequency is f and whose potential energy is $V(x) = 2\pi^2 m f^2 x^2$, Eq. (20) implies that the allowed energies are as defined by Eq. (21*a*), whereas the correct result, deduced from the Schrödinger equation, is given by Eq. (21*b*).

$$= nbf$$
 (21*a*)

$$E = (n+1/2)bf$$
 (21b)

See HARMONIC OSCILLATOR.

The existence of the zero-point energy $\frac{1}{2}bf$ in the ground state n = 0 has been confirmed in analyses of molecular spectra and has significance for many phenomena; for example, *see* IONIC CRYSTALS.

Ε

Quantum statistics; indistinguishability. Equation (20) means that the allowed orbits, plotted as functions of q and p, have area nb the area between two allowed orbits equals b. Since in classical statistical mechanics (i) all orbits are allowed, and (ii) the statistical weight of a volume dq dp of q, p (phase) space is proportional to dq dp, the correspondence principle suggests that in quantum statistics each allowed orbit replaces a set of classical orbits which cover an area b when plotted in the q, p plane. It follows that in quantum statistics: (i) each allowed quantized orbit should have the same statistical weight, which may be set equal to unity; and (ii) with unit weight for a quantized orbit, the weight of an unquantized volume dp dq of phase space must be (dp dq)/b. Because the average number of atoms in a state of energy E is proportional to exp (-E/kT), where k is Boltzmann's constant and T the absolute temperature, it is reasonable that quantum statistics yields different predictions than classical (Boltzmann) statistics when the energy-level spacing is large compared to kT. However, in the limit that the level spacing becomes small compared to kT, which corresponds to the limit $b \rightarrow 0$, quantum statistics must be equivalent to Boltzmann statistics. The quantum theory of specific heat and the theory of blackbody radiation illustrate and confirm the considerations of this paragraph. *See* BOLTZMANN STATISTICS; HEAT RADIATION; QUANTUM STATISTICS; SPECIFIC HEAT OF SOLIDS; STA-TISTICAL MECHANICS.

Quantum statistics is further complicated by the fact that identical particles are indistinguishable. Classically, if two He atoms in their ground states are placed in a box, the statistical weights of states are computed as if the atoms can be distinguished, that is, as if each atom had an identifying mark. Quantum mechanics rejects the possibility of this identification and simultaneously modifies the statistical weights which otherwise would be used in computing statistical averages. Formally, this modification is accomplished by insisting that for a system of identical particles, the wave function must be symmetric in the coordinates of all the particles, including spin coordinates specifying spin orientations, if the particles have integral spin, namely, 0, \hbar , $2\hbar$, ...; and antisymmetric in the coordinates of all the particles, including spin coordinates, if the particles have halfintegral spin, namely, $\hbar/2$, $3\hbar/2$, $5\hbar/2$, See BOSE-EINSTEIN STATISTICS; FERMI-DIRAC STATISTICS.

A symmetric wave function is unchanged under interchange of coordinates, for example, for two particles $\psi(x_1, x_2) = \cos(x_1 - x_2)$ is symmetric; an antisymmetric wave function changes sign under interchange of coordinates, for example, $\psi(x_1, x_2) =$ $\sin(x_1 - x_2)$ is antisymmetric. It is postulated that wave functions having other symmetry properties than these specified here never occur, and therefore must not be included in any enumeration of available states, even though without these symmetry restrictions, they might be acceptable solutions of the Schrödinger equation (for the many-particle system). These symmetry restrictions have profound consequences for macroscopic as well as microscopic systems, and are extremely well established for nuclei as well as atoms. The reason for the connection between spin and statistics is beyond the scope of this article, but appears to be a consequence of requirements imposed by the special theory of relativity. See EXCLUSION PRINCIPLE. Edward Gerjuoy

Bibliography. L. E. Ballentine, *Quantum Mechanics*, 2d ed., 1998; A. Bohm, *Quantum Mechanics: Foundations and Applications*, 3d ed., 1994; R. Eisberg and R. Resnick, *Quantum Physics of Atoms*, *Molecules, Solids, Nuclei, and Particles*, 2d ed., 1985; A. Messiah, *Quantum Mechanics*, vol. 1, 1961, vol. 2, 1962; H. Ohanian, *Principles of Quantum Mechanics*, 1989; R. Shankar, *Principles of Quantum Mechanics*, 2d ed., 1994.

Quantum numbers

The quantities, usually discrete with integer or halfinteger values, which are needed to characterize a physical system of one or more atomic or subatomic particles. Specification of the set of quantum numbers serves to define such a system or, in other words, to label the possible states the system may have. In general, quantum numbers are obtained from conserved quantities determinable by performing symmetry transformations consisting of arbitrary variations of the system which leave the system unchanged. For example, since the behavior of a set of particles should be independent of the location of the origin in space and time (that is, the symmetry operation is translation in space-time), it follows that momentum and energy are rigorously conserved. In similar fashion, the invariance of the system to arbitrary rotations of the three axes defining three-dimensional space leads to conservation of total angular momentum. Inversion of these three axes through their origin is likewise a symmetry operation and leads to conservation of parity. (Parity is not conserved in weak interactions and cannot be used as a quantum number to specify the state of a set of particles undergoing such interactions. For such a case, parity is not a "good" quantum number.) For a list of invariance or symmetry operations and the corresponding conserved quantities see SYMMETRY LAWS (PHYSICS)

Of these conserved quantities, the common ones used to label physical states are momentum, energy, angular momentum, charge, baryon number, lepton number, parity, charge parity, product of parity and charge parity, isotopic spin and third component thereof, hypercharge, isotopic parity (or G parity), and unitary spin. In general, each physical system must be studied individually to find the symmetry transformations, and thus the conserved quantities and possible quantum numbers. The quantum numbers themselves, that is, the actual state labels, are usually the eigenvalues of the physical operators corresponding to the conserved quantities for the system in question. See EIGENVALUE (QUANTUM MECHANICS); ELEMENTARY PARTICLE; NONRELATIVIS-TIC QUANTUM THEORY; PARITY (QUANTUM MECHAN-ICS).

It is not necessary that the conserved quantity be quantized in order to be regarded as a quantum number; for example, a free particle possesses energy and momentum, both of which can have values from a continuum but which are used to specify the state of the particle. For a particle at rest the momentum is zero, and the energy is given by the rest mass; the additional specification of values for the remaining (discrete) quantities in the list mentioned above completely identifies the particle.

The specification of the quantum numbers for a set of particles constitutes a spectroscopy. In the 1920s and 1930s the systematic study of the energy levels of atoms and molecules became known as atomic and molecular spectroscopy. In the 1960s and 1970s similar intensive studies of the quantum numbers of strongly interacting particles, mesons and baryons, was characterized as hadron spectroscopy. In the 1980s another spectroscopy

involving heavy quarks developed. The quarks, though still remaining unobserved particles postulated as having some 13 fractional quantum numbers, have yielded dramatically fruitful predictions for new experiments and new particles which have been observed. The special quantum numbers, known as flavors, which the quarks carry are labeled u, d, s, c, b, and t, where the first three are constituent labels of the older familiar particles of hadron spectroscopy, for example, protons, neutrons, and pi mesons. The (heavy) quark carrying the c (charm) quantum number was indirectly discovered in 1974 and that carrying the b (beauty) quantum number in 1977. The study of the levels of these quark-antiquark systems $c\overline{c}$ and $b\overline{b}$ is often characterized as quarkonium spectroscopy. The t quark is much heavier yet, and appears to have produced some experimental indications of its existence in a 1984 study at CERN, Geneva, Switzerland. A special quantum number given the name color is believed to be carried by the massless field exchanged between quarks which is responsible for their binding in particles. See COLOR (QUANTUM MECHANICS); QUANTUM CHROMODYNAM-ICS; QUARKS.

Some important numbers. Probably the most familiar quantum numbers are those associated with the restricted (or quantized) energy values which appear naturally in quantum mechanics whenever one or more particles are confined to a small region of space. The wave equation (Schrödinger equation in the nonrelativistic case) for the confined particle can be solved to express these allowed energy values in terms of one or more quantum numbers. The number of nodes, or zeros, that the solution of the wave equation possesses always enters such an expression and is usually called the principal quantum number n. In the case of a charged particle confined to the vicinity of an oppositely charged particle according to Coulomb's law (for example, the electron to the proton in a hydrogen atom), the Schrödinger equation yields possible energies, $-E_B/n^2$, where n = 1, 2, 3, ... and for hydrogen $E_B = 13.6$ eV. The magnitude of *n* is also directly related to the most probable distance from the center of the confining region where the electron may be found. See ATOMIC STRUCTURE AND SPECTRA; COULOMB'S LAW.

The wave equation for the confined particle also leads to quantized angular momentum, with possible values or quantum numbers L = 0, 1, 2, 3, ... in units of \hbar (Planck's constant divided by 2π). In addition, each particle, whether confined or not, possesses an intrinsic spin angular momentum *S* with possible values 0, $\frac{1}{2}$, 1, $\frac{3}{\sqrt{2}}$, ..., in units of \hbar . The electron, for example, has a spin of $\frac{1}{2}$. *See* SPIN (QUANTUM MECHANICS).

Since angular momentum is a quantized vector quantity (a direction in space is associated with it), it might be expected that the three space components would also be quantized observables. However, only one, the Z-axis projection, is observable and the associated quantum number m_L has possible values -L, $-L + 1, \ldots, L - 1, L$; correspondingly, the spin angular momentum M_S has possible values -S, -S $+ 1, \ldots, S - 1, S$. These are often called magnetic quantum numbers because they determine how the energy changes when a system with given n, L, S is placed in a magnetic field.

Determination of numbers. Thus the properties or state of a confined electron (for example, in the hydrogen atom) are known when the quantum numbers *n*, *L*, *M_L*, and *M_S* ($S = \frac{1}{2}\hbar$) are specified. Actually, when two or more angular momenta enter, as for the confined electron with an intrinsic $S = \frac{1}{2}\hbar$ in a state with orbital angular momentum L or especially for a system with more than one particle, the overall rotational symmetry of the system is best expressed by a total angular momentum quantum number J and a total magnetic quantum number M. The resultant J is obtained by vectorially adding all individual angular momenta, and M is obtained by summoning all the separate M_S 's and M_L 's. For a many-electron atom or many-nucleon nucleus, the preferred way to perform the vector addition is often indicated by the dynamics, that is, how the individual particle Land S values enter into interparticle interactions. In such a case an intermediate result in the determination of J may be a good quantum number; for example, the seniority quantum number is related to the number of particles whose spins are paired. The objective of atomic and nuclear spectroscopy is the determination of the quantum numbers for the states of many-electron and many-nucleon systems, respectively. Nucleon systems have the additional quantum numbers of isotopic spin and thirdaxis projection. See ELECTRON CONFIGURATION; NU-CLEAR STRUCTURE. Kenneth E. Lassila

Bibliography. J. P. Elliot and P. G. Dawber, Symmetry in Physics, 2 vols., 1979, paper 1985; D. Flamm and F. Schoberl, Quantum Numbers, Gauge Theories and Hadron Spectroscopy, 1982; W. Ludwig and C. Falter, Symmetries in Physics, 2d ed., 1996; U. Mosel, Fields, Symmetries, and Quarks, 2d ed., 1999; L. H. Ryder, Elementary Particles and Symmetries, 2d ed., 1986.

Quantum solids

Solids in which the quantum-mechanical wave functions of individual atoms overlap the wave functions of neighboring atoms in the quantum ground state of the system. The spatial extension of the wave functions is called zero-point motion since the location of the atoms cannot be determined within the width of the wave function even at the absolute zero of temperature. The zero-point motion becomes large when the constituent atoms have small mass and the attractive force between them is small. The isotopes of helium have the weakest interaction of any atom or molecule (and the smallest mass except for the much more strongly interacting

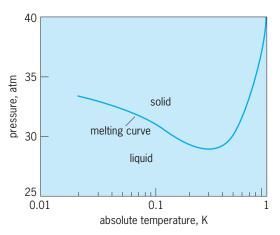


Fig. 1. Melting curve for solid ³He. (From data of E. R.Grilly, J. Low Temp. Phys., 11:33–52, 1973)

hydrogen molecules). The root-mean-square zeropoint motion in those solids is approximately 25% of the mean distance between atoms. This zeropoint motion results in some very unusual properties that are manifestations of the quantum statistical mechanics of many-particle systems. The properties of solid helium are currently an active subject of research with some fundamental questions remaining unanswered. *See* HELIUM; INTERMOLECULAR FORCES; NONRELATIVISTIC QUANTUM THEORY; QUANTUM ME-CHANICS; QUANTUM STATISTICS.

Melting pressure of ³He and ⁴He. Unlike all other solids, helium cannot be solidified by simply cooling liquid helium to low temperatures. The liquid must be compressed to at least 25 bar for ⁴He and 29 bar for ³He in order to freeze (1 bar = 10^5 Pa ≈ 1 atm). In further contrast with other solids, the pressure required to freeze the isotopes of helium increases with decreasing temperature below a certain temperature (**Fig. 1**). In that temperature range, a mixture of liquid and solid must be heated in order to freeze it. A general result of thermodynamics states that the pressure *P* and temperature *T* of any melting

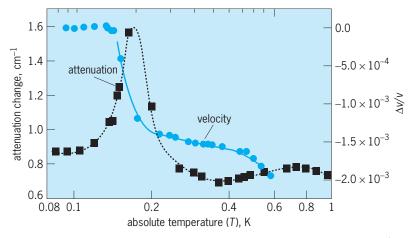


Fig. 2. Changes in the attenuation and speed of sound near 0.2 K for a sample of solid ⁴He with 27 ppm of ³He impurities. The attenuation is relative to an arbitrary reference value, and the velocity shift, $\Delta v/v$, is the fractional change relative to the velocity at the lowest temperature measured.

curve are related by the equation

$$dP/dT_{\text{melting}} = \frac{\Delta s}{\Delta u}$$

Here Δs is the liquid entropy (per mole) minus the solid entropy, and Δv is the liquid volume (per mole) minus the solid volume. When the slope of the melting curve, dP/dT, is negative, the entropy of the liquid is less than that of the solid, since the phase that exists at lower pressure must always have larger molar volume (Δv must always be positive). For all other materials, the solid state has lower entropy than the liquid due to the spatial order of the atoms in the crystal lattice. For helium, the entropy of the liquid is lower than the solid due to quantum statistics. *See* ENTROPY.

Quantum statistics becomes important when the wave functions of individual atoms or molecules overlap. In liquid helium at low temperatures, where thermal motion of the atoms becomes small, the overlap is so great that essentially all of the wave functions overlap each other. However, quantum statistics takes two different forms for the two fundamental types of particles so that a many-particle system consisting entirely of one type or the other will have very different properties at low temperatures. ³He atoms are "Fermi" particles (fermions) for which only one atom can occupy a given quantum state. The quantum states are specified by a momentum quantum number and a spin quantum number. The latter has only two possible values for ³He and represents the orientation of the magnetic moment of the atom with respect to some reference direction such as a magnetic field. Consequently, there are two atoms with each momentum quantum number, and their magnetic moments must be oriented antiparallel so that the bulk magnetization of liquid ³He decreases as temperature decreases. In solid ³He, the wave functions are much more localized so that the magnetization does not decrease until much lower temperatures. At temperatures higher than the ordering of the moments in the solid, the entropy of the liquid is lower than the solid. See FERMI-DIRAC STATISTICS.

⁴He atoms are "Bose" particles (bosons) in which any number of atoms can occupy the same quantum state. At low temperatures, this leads to Bose-Einstein condensation to a highly ordered, low-entropy state in which a large fraction of the atoms are in the same quantum state below a critical temperature. Bose-Einstein condensation is responsible for the remarkable states of superfluidity and superconductivity. In the solid phase, the greater localization suppresses Bose-Einstein condensation. *See* BOSE-EINSTEIN CONDENSATION; BOSE-EINSTEIN STATISTICS; LIQUID HELIUM; SUPERCONDUCTIVITY; SUPERFLUID-ITY.

Quantum statistics in solid phases. Although the wave functions overlap much less in the solid than in the liquid, there are still consequences of the quantum statistics.

As ³He atoms are forced closer together, the effect of the statistics is to tend toward some type of

antiparallel alignment of neighboring pairs of atoms. This effect of the statistics is called the quantum exchange interaction. It has no analog in Newtonian physics. Although there is an interaction between the magnetic dipole moments of the nuclei, it is smaller than the exchange interaction by a factor of 10^{-4} . The exchange interaction in ³He produces an unusual antiferromagnetic state below about 10^{-3} K. The magnetic moments in two adjacent planes of atoms are oriented in the same direction, and the moments in the next two planes are oriented in the opposite direction. *See* ANTIFERROMAGNETISM; EXCHANGE INTERACTION.

An analogous phenomenon in solid ⁴He would be some amount of Bose-Einstein condensation at much lower temperatures than the condensation in liquid ⁴He. If true, this would represent a peculiar state of matter which could display some of the properties of a superfluid as well as the spatial order of a solid. This possibility was proposed theoretically, and some experiments have found behavior suggestive of some type of a phase transition. Proof of its existence is an active topic of current research. The addition of as little as about 20 parts per million of ³He to solid ⁴He produced a sharp peak in the attenuation of sound and a shift in the speed of sound at about 0.2 K (Fig. 2). Some of the observed properties below 0.2 K were also produced by straining crystals without impurities. In a completely different experiment, the rotational moment of inertia of solid helium has been measured by placing the helium solid in a cylindrical container suspended on a torsion fiber. The period of the torsion pendulum was measured as a function of temperature and found to decrease below about 0.2 K (Fig. 3). See LOW-TEMPERATURE ACOUSTICS; MOMENT OF INERTIA; PHASE TRANSITIONS.

A decrease in the moment of inertia means that some of the mass of the solid is not moving with the container. This could be direct evidence that some of the solid becomes superfluid. However, the phenomenon appears to depend on the ³He impuritiy concentration and the strain of the solid since it becomes smaller and moves to lower temperatures for the pure solid and when the solid is annealed. A test for the flow of solid through small openings showed that, if there is supersolid, it does not flow in the same manner as superfluid. No flow was observed down to flow rates seven orders of magnitude smaller than would have corresponded to the observed velocities in the torsion pendulum. A measurement of the melting pressure in 2006 revealed unexpected behavior below 0.08 K, but it is not consistent with a supersolid phase transition.

Molecular quantum solids. The molecules H_2 , HD, D_2 , N_2 , and a few others also have sufficient zeropoint motion that some quantum solid properties might be expected. In the various isotopic combinations of solid hydrogen, these are manifested through the influence of quantum tunneling on nuclear magnetic resonance (NMR) relaxation times and on the orientational order of the molecules. *See* HYDROGEN;

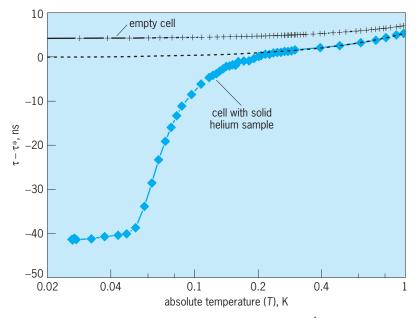


Fig. 3. Variation of the period of a torsion pendulum containing solid ⁴He as a function of temperature, indicating the decrease in the moment of inertia of the solid below about 0.2 K. The pendulum oscillation amplitude, for these data, was set so that the maximum velocity was 4 μ m/s. Temperature dependence of an empty cell was also measured in order to be certain that there was no effect intrinsic to the cell rather than to solid helium. Vertical scale is $\tau - \tau^*$, where τ is the period of the pendulum and τ^* is the value of the period at the temperature where the period starts to decrease.

NUCLEAR MAGNETIC RESONANCE (NMR); TUNNELING IN SOLIDS. John M. Goodkind

Bibliography. J. Day and J. Beamish, Pressure-driven flow of solid helium, *Pbys. Rev. Lett.*, 96:105304, 2006; P.-C. Ho, I. P. Bindloss, and J. M. Goodkind, A new anomaly in solid ⁴He, *J. Low Temp. Phys.*, 109:409421, 1997; E. Kim and M. H. W. Chan, Observation of superflow in solid helium, *Science*, 305:1941–1944, 2004; D. D. Osheroff, Nuclear magnetic order in solid ³He, *J. Low Temp. Phys.*, 87:297–342, 1992; I. A. Todoshchenko et al., Melting curve of ⁴He: No sign of a supersolid transition down to 10 mK, *Phys. Rev. Lett.*, 97:165302, 2006; J. S. Xia, W. Ni, and E. D. Adams, Melting pressure of solid ³He through the magnetic ordering transition, *Phys. Rev. Lett.*, 70:1481–1484, 1993.

Quantum statistics

The statistical description of particles or systems of particles whose behavior must be described by quantum mechanics rather than by classical mechanics. As in classical, that is, Boltzmann statistics, the interest centers on the construction of appropriate distribution functions. However, whereas these distribution functions in classical statistical mechanics describe the number of particles in given (in fact, finite) momentum and positional ranges, in quantum statistics the distribution functions give the number of particles in a group of discrete energy levels. In an individual energy level there may be, according to quantum mechanics, either a single particle or any number of particles. This is determined by the symmetry character of the wave functions. For antisymmetric wave functions only one particle (without spin) may occupy a state; for symmetric wave functions, any number is possible. Based on this distinction, there are two separate distributions, the Fermi-Dirac distribution for systems described by antisymmetric wave functions and the Bose-Einstein distribution for systems described by symmetric wave functions.

In relativistic quantum theory it is shown that particles having integer spin necessarily obey Bose-Einstein statistics, while those having half-integer spin necessarily obey Fermi-Dirac statistics. (Particles obeying Bose-Einstein statistics are often called bosons; particles obeying Fermi-Dirac statistics, fermions.) For sufficiently high temperatures, both forms of distribution functions go over into the familiar Boltzmann distribution, although strictly speaking no system is correctly described by this distribution. In practice, of course, the Boltzmann distribution gives an exceedingly good description of the experiments, but there are situations, such as those involving the behavior of electrons in metals and liquid helium, where the quantum description is essential. See BOLTZMANN STATISTICS; BOSE-EINSTEIN STATISTICS; EXCLUSION PRINCIPLE; FERMI-DIRAC STATISTICS; KINETIC THEORY OF MATTER; NONRELATIVISTIC QUANTUM THEORY; QUANTUM MECHANICS; RELATIVISTIC QUANTUM THE-ORY; SPIN (QUANTUM MECHANICS); STATISTICAL ME-CHANICS. Max Dresden

Quantum teleportation

A way to transfer the state of a quantum system over large distances by employing entanglement. Entanglement is a nonclassical connection between objects that Albert Einstein called "spooky."

To be able to travel from one place to another instantly and over arbitrary distances, or at least to move objects in this way, is an ancient dream. The concept of teleportation is frequently utilized in the literature of science fiction to overcome limitations imposed on space travel by the laws of physics.

In the standard science fiction approach, the sender, Alice, scans the object to be teleported in order to read out all the information needed to describe it. She then sends that information to the receiver, Bob, who uses this information to reconstitute the object, not necessarily from the same material as that of the original. However, according to quantum mechanics, it is impossible to succeed in this way. If only one individual object is at hand, it is impossible to determine its quantum state by measurement. The quantum state represents all that can be known about the object, that is, all possible (in general, probabilistic) predictions that can be made about future observations of the object.

In fact, it is quantum mechanics that comes to the rescue and makes quantum teleportation possible using a very deep feature of the theory, quantum entanglement. It is important to realize that there are significant differences between teleportation as



Fig. 1. Principle of quantum entanglement for two photons emitted by a source. Each photon travels to its own two-channel polarizer, each of which can be rotated around the respective beam direction. Independent of the orientation of the polarizer, each detector (H or V) has the same probability of registering a photon. If the two polarizers are oriented parallel, the two photons will always be registered in different detectors; that is, if one photon is registered in its H detector, the other is registered in its V detector, and vice versa. Yet neither photon carries any polarization before it is measured.

portrayed in science fiction and quantum teleportation as realized in the laboratory. In the experiments, what is teleported is not the substance an object is made of but the information it represents.

Quantum entanglement. Entangled quantum states as used in teleportation were introduced into the discussion of the foundations of quantum mechanics by Einstein, Boris Podolsky, and Nathan Rosen in 1935. In the same year, Erwin Schrödinger introduced the notion of entanglement, which he called the essence of quantum mechanics.

In order to discuss entanglement, one specific case will be considered, and the possible experimental results will be examined (Fig. 1). There are many possible sources that can create many different sorts of entangled states. The source under consideration will be assumed to be the one used in the first teleportation experiments, which produced photons in a singlet polarization state. This means that neither photon enjoys a state of well-defined polarization; each one of the photons on its own is maximally unpolarized. Yet, when one of the two photons is subject to a polarization measurement, it assumes one specific polarization. That specific experimental outcome is completely random. As a consequence of the two photons being in the entangled singlet state, the other photon is instantly projected into a state orthogonal to that of the first photon. The fact that the measurement result on the second photon can be perfectly predicted on the basis of the measurement result of the first photon, even as neither one carries a well-defined quantum state, is known as the Einstein-Podolsky-Rosen paradox. In 1964 John Bell showed that these perfect correlations cannot be understood on the basis of properties that the entangled photons carry individually before the measurement. The resulting conflict between the philosophical position of local realism and the predictions of quantum mechanics, which have been confirmed beyond reasonable doubt in experiment, is known as Bell's theorem. See HIDDEN VARIABLES; PHOTON; POLARIZATION OF WAVES.

From an information-theoretic point of view, the interesting feature of entanglement is that neither of the two photons carries any information on its own. All information is stored in joint properties.

Concept of quantum teleportation. It was first realized by Charles H. Bennett and his colleagues that entanglement can be utilized to make teleportation possible (**Fig. 2**). Alice, who is in possession of the

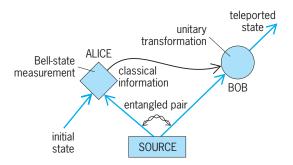


Fig. 2. Quantum teleportation procedure. Alice has an original particle in the initial state, and Alice and Bob also share an ancillary entangled pair. Alice then performs a Bell-state measurement and transmits the random result of that measurement in the form of two classical bits to Bob who, by a simple unitary transformation, can turn his ancillary photon into an exact replica of the original. (After D. Bouwmeester et al., Experimental quantum teleportation, Nature, 390:575-579, 1997)

original teleportee photon in a quantum state not known to her, and Bob initially share an ancillary pair of entangled photons, say in the singlet state described above. Alice then subjects her teleportee photon and her member of the ancillary pair to a Bell-state measurement. A Bell-state measurement is designed in such a way that it projects the two photons into an entangled state even if they were previously unentangled. This is a very tricky procedure both conceptually and experimentally. Conceptually, it means that the measurement must be performed in such a way that it is not possible, even in principle, to determine from the measurement result which photon was the teleportee and which was Alice's ancillary. They both have to lose their individuality. The result of the measurement must reveal only how the two photons relate to each other, and ignore individual properties. A Bell measurement has four possible results if the objects considered are defined in a twodimensional Hilbert space just as is done to describe the photon's polarization. One of the four states is the singlet state discussed above. The other three states also define specific relations between the two photons, though different ones than those for the singlet state.

By the Bell-state measurement, Alice now knows how the unknown state of the teleportee photon relates to her ancillary one. She also knows in which entangled state the two ancillaries were produced, that is, how these two relate to each other. Thus she finally knows precisely how the teleportee relates to Bob's photon. More formally speaking, as a result of Alice's measurement Bob's photon is projected into a state which is uniquely related to the original state; the specific relationship is expressed by which of the four Bell states Alice obtained. Alice therefore informs Bob of her measurement result via a classical communication channel, and he, by applying a simple unitary transformation on his photon, changes it into the original state.

In one of the four cases, Alice obtains the information that her two photons have been projected into the singlet state, the same state in which the ancillaries were produced. Then, she knows that Bob's photon is instantly projected into the original state; the transformation that Bob has to apply is an identity transformation, that is, one that makes no change to his photon. That Bob's photon then instantly becomes an exact replica of the original seems to violate relativity. Yet, while Alice knows instantly that Bob's photon, no matter how far away, is already an exact replica, she has to inform Bob of the Bell measurement result such that he knows that his photon is already in the correct state. That classical information can arrive only at the speed of light. This requirement is also true for the other possible Bellstate measurement results. Bob has to know them in order to apply the correct transformation to his photon.

The result of the Bell measurement is not related at all to any properties that the original photon carries. Thus, that measurement does not reveal any information about its state. Therefore, the operation that Bob has to apply is also completely independent of any properties of the original photon. The reason that quantum measurement succeeds is that entanglement makes it possible to completely transfer the information that an object carries without measuring this information.

Experimental realization. An experiment therefore faces a number of challenges. They include (1) how to produce the entangled photon pairs and (2) how to perform a Bell measurement for independent photons. In the experimental realization by D. Bouwmeester and his colleagues in 1997, the entangled photons were produced in the process of spontaneous parametric downconversion. This is a second-order nonlinear process where a suitable crystal, in the experiment beta barium borate (BBO), is pumped with a beam of ultraviolet radiation. A photon from that beam has a very small probability to decay spontaneously into two photons, which then are polarization-entangled in just the way necessary for the experiment. The more tricky part is the Bellstate measurement because, in essence, it requires that the two photons are registered such that all information about which was the teleportee photon and which the ancillary is irrevocably erased. This is a nontrivial requirement since the two photons are coming from different directions, they might arrive at different times, and so forth. See NONLINEAR OPTICS; OPTICAL MATERIALS.

In the experiment, the Bell-state measurement was performed using a semireflecting mirror, which acted as a 50/50 beam splitter. Two photons were incident on the beam splitter, one from its front side and one from its back, and each one had the same probability of 50% to be either reflected or transmitted. If each of the two detectors in the two outgoing beams, again one in the front and one in the back, registered a photon simultaneously, then no information existed as to which incoming photon was registered in which detector, and the two were projected into the entangled singlet state. Narrow-bandwidth filters in front of the detectors further served to erase any time information which could also serve to identify the photons. In this experiment, only one of the four possible Bell states could be identified, the singlet state. This certainly reduced the efficiency of the procedure, though in those cases in which the two detectors at the Bell-state analyzer registered, teleportation worked with a fidelity escaping all possible classical explanation.

In another experiment, also called entanglement swapping, it was even possible to teleport a photon that was still entangled to another one. That experiment started with two entangled pairs. A joint Bellstate measurement on one photon from each pair projected the other two photons onto an entangled state. In that way, two photons that neither came from the same source nor ever interacted with one another became entangled.

What all these experiments reveal is that the quantum state is really just a representation of the information that has been acquired. In the case of entanglement, it is only information on how objects relate to each other without any information on their individual properties. And in the case of teleportation, Alice's observation changes the quantum state that Bob observes. In other words, what can be said about the situation changes due to an observation by Alice. This gives very strong support to the Copenhagen interpretation of quantum mechanics. The first experiments were done with polarization-entangled photon pairs. Since then a number of experiments teleporting other properties such as continuous variables carried by the electromagnetic field of light, instead of the discrete ones discussed above, have been performed.

Prospects. While the teleportation distance in the first experiments was of the order of 1 m (3 ft), experiments in 2004 extended the distance to the order of 600 m (2000 ft), and there are plans to perform such experiments over much larger distances and even from a satellite down to laboratories on the ground. Other important experimental steps include the teleportation of quantum states of atoms (2004) and the teleportation of the quantum state of a photon onto that of an atomic cloud (2006).

Today quantum teleportation and entanglement swapping-the teleportation of an entangled stateare considered to be key building blocks of future quantum computer networks. At present there is intense research in the development of both quantum communication networks and quantum computers. Future quantum computers would use individual quantum states, for example those of atoms, to represent information in so-called quantum bits. They are expected to allow some algorithms to be performed with significantly higher speed than any existing computers. Quantum teleportation would allow the transfer of the quantum output of one quantum computer to the quantum input of another quantum computer. See NONREL-ATIVISTIC QUANTUM THEORY; QUANTUM COMPUTA-TION; QUANTUM MECHANICS; QUANTUM THEORY OF MEASUREMENT. Anton Zeilinger

Bibliography. M. D. Barrett et al., Deterministic quantum teleportation of atomic qubits, *Nature*,

429:737-739, 2004; C. H. Bennett et al., Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels, *Phys. Rev. Lett.*, 70:1895-1899, 1993; D. Bouwmeester et al., Experimental quantum teleportation, *Nature*, 390:575-579, 1997; A. Furasawa et al., Unconditional quantum teleportation, *Science*, 282:706-709, 1998; M. Riebe et al., Deterministic quantum teleportation with atoms, *Nature*, 429:734-737, 2004; J. F. Sherson et al., Quantum teleportation between light and matter, *Nature*, 443:557-560, 2006.

Quantum theory of matter

The microscopic explanation of the properties of condensed matter, that is, solids and liquids, based on the fundamental laws of quantum mechanics. Without the quantum theory, some properties of matter, such as magnetism and superconductivity, have no explanation at all, while for others only a phenomenological description can be obtained. With the theory, it is at least possible to comprehend what is needed to approach a complete understanding.

The theoretical problem of condensed matter, that is, large aggregates of elementary particles with mutual interactions, is that of quantum-statistical mechanics, or the quantum-mechanical many-body problem: an enormous number, of order 10²³, of constituent particles in the presence of a heat bath and interacting with each other according to quantummechanical laws. What makes the quantum physics of matter different from the traditional quantum theory of elementary particles is that the fundamental constituents (electrons and ions) and their interactions (Coulomb interactions) are known but the solutions of the appropriate quantum-mechanical equations are not. This situation is not due to the lack of a sufficiently large computer, but is caused by the fact that totally new structures, such as crystals, magnets, ferroelectrics, superconductors, liquid crystals, and glasses, appear out of the complexity of the interactions among the many constituents. The consequence is that entirely new conceptual approaches are required to construct predictive theories of matter. See CRYSTAL STRUC-TURE; FERROELECTRICS; GLASS; LIQUID CRYSTALS; SU-PERCONDUCTIVITY.

The usual technique for approaching the quantum many-body problem for a condensed-matter system is to try to reduce the huge number of variables (degrees of freedom) to a number which is more manageable but still can describe the essential physics of the phenomena being studied. In general, the fundamental laws of quantum mechanics give little or no guidance for this task. For example, while it seems natural to describe a solid by a collection of ions and valence electrons, it is not apparent from the elementary equations of motion that this makes sense. Here, the ions consist of the nuclei and the more tightly bound electrons; the valence electrons are the more loosely bound ones which participate in the chemical bonding. The problem remains extraordinarily complex because of the mutual Coulomb interactions among all these particles.

One-electron approach. In the several decades after the development of quantum theory and the understanding of the Pauli exclusion principle, enormous successes were achieved in developing the band theory of solids. The basic idea was to simplify the effects of all the interparticle interactions by approximating them by average potential fields. In this socalled one-electron approximation, in a crystalline lattice for example, the valence electrons move effectively independently of each other, under the influence of the average potential field arising from their mutual interactions and of the periodic electrostatic potential due to the ions which have arranged themselves in a periodic array. The quantum mechanics of this system then gives conclusions which can explain many observed properties of a large number of materials: the differences between a metal, a semiconductor, and an insulator; optical and thermodynamic properties; heat and electrical conductivity; thermoelectric effects; and some of the effects of impressed magnetic fields. None of these issues found satisfactory explanations before the quantum-mechanical band theory of solids. See BAND THEORY OF SOLIDS; ELECTRICAL CONDUCTIV-ITY OF METALS; FREE-ELECTRON THEORY OF METALS; SEMICONDUCTOR.

Even the motion of the ions must be described quantum-mechanically. In a very good approximation, the deviations of the ionic positions from their regular lattice points can be described in terms of a collection of masses whose near neighbors are coupled by springs. The quantum-mechanical description of the normal modes of these springs is in terms of elementary excitations called phonons whose thermodynamics gives the correct result for the contribution of ionic motion to the specific heat, which was previously unobtainable. *See* LATTICE VI-BRATIONS; PHONON; SPECIFIC HEAT OF SOLIDS.

While the band theory, based on the one-electron approximation described above, has been dramatically successful as a detailed microscopic explanation of a variety of properties of matter, it has failed for a number of others. It was realized that a more detailed consideration of the interelectron Coulomb interactions is essential to understand some of the deeply important properties. For example, the band theory cannot distinguish between a metal and a type of magnetic insulator, nor can it lead to a theory of superconductivity.

Many-body approach. The quantum theory of metals has been extended beyond the one-electron band theory to include some effects of the interelectron Coulomb interactions. It seems at first sight that this would be impossible since the interaction energies are not small compared to the characteristic oneelectron energy. However, the Pauli principle and the phenomenon of screening reduce the effectiveness of the interaction, and in many cases this allows a systematic treatment. In an elegant formulation of the many-body problem, the many interactions are reduced to a few simple parameters; the result is called a Fermi liquid. The Fermi liquid describes the behavior of the metallic electrons and certain collective excitations which arise as a consequence of the interactions. The same theory accounts for many of the properties of liquid helium-3. *See* LIQUID HELIUM.

Magnetism. One of the most dramatic of the phenomena which elude band theory is magnetism. The pioneers of the quantum theory lost no time in attacking the issue of magnetism, first on the level of atoms. Thus, P. A. M. Dirac's theory of the electron enabled an understanding of the magnetic properties of atoms in terms of the electron's orbital and intrinsic (spin) magnetic moments and the quantization of atomic energy levels. However, the character of the ground state of a given atom depends on the interplay between the Pauli exclusion principle and the Coulomb repulsion between the electrons. This was first accounted for qualitatively by Hund's rules, but the microscopic description requires a detailed treatment of the electron-electron interactions which goes beyond the Bohr one-electron theory for atoms. See ATOMIC STRUCTURE AND SPECTRA; ELEC-TRON SPIN; EXCLUSION PRINCIPLE.

It was quickly recognized by W. Heisenberg and Dirac that the interplay of the Pauli exclusion principle and the Coulomb interactions makes the behavior of a small group of electrons dependent on the relative orientations of their spins. This is called the exchange effect, and is expressed by the Heisenberg-Dirac exchange interaction. The latter is an example of the reduction of many degrees of freedom to a few: The complicated description of the behavior of all the electrons on, for example, two atoms is replaced by a simple expression involving only two variables, the scalar product of the net spin vectors on each atom. They are coupled with a strength which, in principle, is determined by the motions and interactions of all the underlying electrons but in practice is often taken as a phenomenological parameter. In solids, it is conceivable that the coupling parameter has either sign; that is, either spins parallel (ferromagnetic coupling) or spins antiparallel (antiferromagnetic coupling) could be the lowest energy configuration. This fundamental result, when applied to collections of atoms, is the basis for many theories of the magnetism in insulators. The foundation of these theories is the Heisenberg-Dirac exchange interaction with a phenomenologically determined coupling; the microscopic determination of the sign and magnitude of this coupling remains difficult. See ANTIFERROMAGNETISM; EXCHANGE INTER-ACTION; FERROMAGNETISM.

Modern many-body problem. Magnetism in metals such as iron and chromium poses even more complicated questions because, in contrast to insulators where the valence electrons are essentially localized, there are itinerant conduction electrons whose orbital motion is important for the magnetism. In such cases, the electron-electron interaction leads to correlations which play a central role in determining magnetic properties but for which there is no complete theoretical analysis. The same situation occurs even more dramatically in those magnetic insulators which within the one-electron band theory are predicted to be metals. Striking examples are found among the transition-metal oxides, for example, and the insulating magnetic parent compounds of the family of copper oxide high-temperature superconductors. In these cases, the electron-electron repulsions are so strong that not only is the oneelectron approach almost irrelevant but the systematics of the Fermi-liquid theory are no longer valid. This state of affairs also occurs frequently in systems of low dimensionality, which often have behavior dominated by strong interactions. These include the family of quasi-one-dimensional conductors formed by arrays of long chains of molecules, the planar copper oxide high-temperature superconductors, and the state of matter observed in two-dimensional electron systems in a magnetic field when in the regime of the fractional quantum Hall effect. See HALL EF-FECT; ORGANIC CONDUCTOR.

This situation defines the modern many-body problem in the quantum theory of matter. The properties of many materials are, in a fundamental way, primarily determined by the presence of strong interparticle interactions, and these are also responsible for the presence of magnetism in those materials which exhibit it. The modern challenge for the quantum theory of matter is to give a thorough analysis of the interparticle correlations which arise from the strong interactions and to describe how these correlations lead to the many effects which are observed in a wide class of materials. *See* NONRELA-TIVISTIC QUANTUM THEORY; QUANTUM MECHANICS; SOLID-STATE PHYSICS. Elihu Abrahams

Bibliography. N. W. Ashcroft and N. D. Mermin, *Solid State Physics*, 1976; P. M. Chaikin and T. C. Lubersky, *Principles of Condensed Matter Physics*, 1994; A. Ishihara, *Condensed Matter Physics*, 1991; G. D. Mahan, *Many-Particle Physics*, 3d ed., 2000; R. M. White and T. H. Geballe, *Long Range Order in Solids*, 1979.

Quantum theory of measurement

The attempt to reconcile the counterintuitive features of quantum mechanics with the hypothesis that it is in principle a complete description of the physical world, even at the level of everyday objects. A paradox arises because, at the atomic level where the quantum formalism has been directly tested, the most natural interpretation implies that where two or more different outcomes are possible it is not necessarily true that one or the other is actually realized, whereas at the everyday level such a state of affairs seems to conflict with direct experience.

Interference. It is useful to consider a group (ensemble) of identically prepared microscopic objects such as electrons or photons which start, one at a time, from some state A, and from there can proceed to one of two alternative states, B or C, and from there to E or some other state. (The various states might correspond to spatially separated positions,

as in Young's double-slit experiment, or to something more abstract, such as the different strangeness states of the neutral K meson.) Quantum mechanics describes the process of transition of the particles between the various states in terms of a set of probability amplitudes (wave functions) such that the probability of a given process occurring is proportional to the squared modulus of the corresponding amplitude. When a given result can be achieved by more than one process, and no measurement has been made to decide which occurred, the total amplitude for the result to occur is the algebraic sum of the amplitudes for the various contributing processes.

Thus, in the setup described, the probability (P_B) that a particle arrives at E via B can be determined; this is done by blocking alternative C (for example, in the Young's double-slit experiment, by simply obstructing the slit C). This probability is just the squared modulus ($|\psi_B|^2$) of the amplitude ψ_B for the process $A \rightarrow B \rightarrow E$. Similarly, by blocking alternative B, $P_C = |\psi_C|^2$ can be determined.

If both alternatives are now left open, the total amplitude ($\psi_{B \text{ or } C}$) to arrive at E is now the algebraic sum of ψ_{B} and ψ_{C} , and the probability of arriving at E is given by the expression in Eq. (1). The first two

$$P_{B \text{ or } C} = |\psi_{B \text{ or } C}|^{2} = |\psi_{B} + \psi_{C}|^{2}$$
$$= |\psi_{B}|^{2} + |\psi_{C}|^{2} + 2 \operatorname{Re} (\psi_{B}^{*}\psi_{C}) \quad (1)$$
$$= P_{B} + P_{C} + 2 \operatorname{Re} (\psi_{B}^{*}\psi_{C})$$

terms in this expression (which is very well verified at the atomic level) correspond to what would be obtained by assuming that each individual particle of the ensemble proceeded either via B or via C. However, the last term cannot be interpreted in this way; rather, it shows the effects of interference between the two possibilities, a characteristically quantummechanical phenomenon. In particular, if $\psi_{\rm C}$ happens to be equal to $\psi_{\rm B}$ in magnitude and opposite in sign, then $P_{\rm B}$ or $P_{\rm C}$ vanishes even though $P_{\rm B}$ and $P_{\rm C}$ are both individually nonzero. *See* COHERENCE; INTERFERENCE OF WAVES.

Thus, the properties of the complete ensemble of particles reaching E are not the same as the properties of a combination of the subensembles which are obtained by allowing only alternative B and only alternative C respectively. This strongly suggests that in this situation it is not the case that each individual particle of the ensemble realizes one or other of the two alternatives. Rather, each particle, as it were, keeps its options open. Although strenuous attempts have been made to avoid this conclusion, they all encounter severe difficulties. However, if a measuring device is set up to determine whether one or both paths are followed, each individual particle appears to choose either alternative B or alternative C, at random with a probability $|\psi_B|^2$ and $|\psi_C|^2$ respectively. But under these conditions no interference at E occurs. Thus, it seems that when no measurement is performed, a definite alternative is not realized, but

when a measurement is performed, it is. *See* HIDDEN VARIABLES.

Quantum measurement paradox. In the early days of quantum mechanics, the resolution of the apparent paradox was often located in the act of measurement itself, which was regarded as involving a supposedly unanalyzable interaction of the microscopic system with a macroscopic instrument which had to be described in classical terms. However, if one believes, as most physicists do, that quantum mechanics gives in principle a complete description of the physical world, it must be possible to describe the process of measurement itself, as well as the working of the measuring instrument, in quantum-mechanical language.

A collection of macroscopic counters can be set up to determine which of the paths, B or C, is followed by each individual particle of the ensemble. It may be supposed temporarily (and unrealistically, as discussed below) that whenever the particle in question is drawn from a subensemble described by ψ_B (alone), the final state of the counters, which registers the fact that path B was indeed followed, is some pure quantum state Ψ_B ; similarly, when the particle is drawn from a subensemble described by ψ_C , and is therefore guaranteed to follow path C, the final state of the counters is a pure state Ψ_C . Since it is necessary to read off which path was followed directly from an inspection of the counters, the states described by Ψ_B and Ψ_C must be macroscopically distinguishable.

If the particle in question is now drawn from an ensemble described by the so-called linearsuperposition wave function ψ_B or $\psi_C \equiv \psi_B + \psi_C$, then because quantum mechanics is by its nature a linear theory, it immediately follows that in this case the final state of the counters is given by Eq. (2).

$$\Psi_B \text{ or } \Psi_C \equiv \Psi_B + \Psi_C \tag{2}$$

That is, it is a linear superposition of macroscopically distinct states. If now it is supposed that no measurement has been made on the measuring instruments themselves, and the same interpretation of the quantum formalism is adopted as seemed to be required at the atomic level, it would seem to follow that no definite outcome of the experiment has been realized even at the everyday level of counters, and so forth—a conclusion which apparently conflicts violently with direct experience at this level. This is called the quantum measurement paradox.

Alleged resolution. The resolution of this paradox that is probably most favored by practicing physicists proceeds in two stages. At stage 1, it is pointed out that the above argument (and in particular the description of the final state of the counters by pure states Ψ_B and Ψ_C) is grossly oversimplified, and that a correct quantum-mechanical analysis shows that, quite generically, whenever the formalism appears to generate a superposition of macroscopically distinct states it is impossible to demonstrate the effects of interference between them. The reasons for this claim include the facts that the initial state of a macroscopic system is likely to be unknown in detail; its extreme sensitivity to random external noise; and most important merely by virtue of its macroscopic nature any such system will rapidly have its quantummechanical state correlated (entangled) with that of its environment in such a way that no measurement on the system alone (without a simultaneous measurement of the complete state of the environment) can demonstrate any interference between the two states in question-a result often known as decoherence. Thus, it is argued, the outcome of any possible experiment on the ensemble of macroscopic systems prepared in this way will be indistinguishable from that expected if each system had actually realized one or other of the two macroscopically distinct states in question. Stage 2 of the argument (often not stated explicitly) is to conclude that if this is indeed true, then it may be legitimately asserted that such realization of a definite macroscopic outcome has indeed taken place by this stage.

Most physicists agree with stage 1 of the argument. However, not all agree that the radical reinterpretation of the meaning of the quantum formalism which is implicit at stage 2 is legitimate; that is, that an interpretation in terms of realization, by each individual system, of one alternative or the other, forbidden at the atomic level by the observed phenomenon of interference, is allowed once, on going to the macroscopic level, the phenomenon disappears.

Alternative approaches. Consequently, various alternative interpretations have been developed. One, sometimes known as the statistical interpretation, holds that the symbols of the quantum-mechanical formalism correspond to nothing in the physical world at all, even at the macroscopic level, but are merely a recipe for calculating the probabilities of various directly observable outcomes of the experiment in question. In another direction, the Everett-Wheeler (or relative-state or many-worlds) interpretation holds (at least according to some expositions) that the subjective impression of having observed one and only one outcome of a given experiment is an illusion and that there exist other parallel universes, said to be equally real, in which the unknown outcomes are realized. Several other interpretations have also been proposed.

A quite different approach to the quantum measurement paradox is the conjecture that quantum mechanics is not, after all, a complete description of the physical world but must be supplemented by other laws of physics. Since the predictions of quantum mechanics appear to be very well verified at the atomic level, the corrections due to such new laws must be zero or very small in this regime; however, they must be effective enough at the macroscopic level to effect the realization of a definitive alternative in, at most, a time that is small compared to the minimum interval detectable by human observation. A concrete scenario along these lines was developed by G. C. Ghirardi, A. Rimini, T. Weber, and P. Pearle (the GRWP theory).

Experimental test. The existence, or not, of quantum superpositions of macroscopically distinct states

may be susceptible, to an extent, to experimental resolution. Obviously, to establish the existence of such superpositions it is necessary to observe the effects of interference between the two components (and to show that such effects could not follow from any theory in which realization had taken place by this stage), and thus a system is required where decoherence does not take place on the relevant time scale. Candidates for such systems include magnetic grains, various quantum-optical systems, and superconducting devices incorporating the Josephson effect. *See* JOSEPHSON EFFECT; NONRELATIVISTIC QUANTUM THE-ORY; QUANTUM MECHANICS; SQUEEZED QUANTUM STATES. Anthony J. Leggett

Bibliography. D. Z. Albert, *Quantum Mechanics* and Experience, 1992; L. E. Ballentine, The statistical interpretation of quantum mechanics, *Rev. Mod. Phys.*, 42:358-381, 1980; B. d'Espagnat, *Conceptual Foundations of Quantum Mechanics*, 2d ed., 1976, reprint 1989; A. I. M. Rae, *Quantum Physics: Illusion or Reality?*, 1987, paper 1994.

Quark-gluon plasma

A predicted state of nuclear matter containing deconfined quarks and gluons. According to the theory of strong interactions, called quantum chromodynamics, hadrons such as mesons and nucleons (the generic name for protons and neutrons) are bound states of more fundamental objects called quarks. The quarks are confined within the individual hadrons by the exchange of particles called gluons. However, calculations indicate that, at sufficiently high temperatures or densities, hadronic matter should evolve into a new phase of matter containing deconfined quarks and gluons, called a quark-gluon plasma when hot, or quark matter when

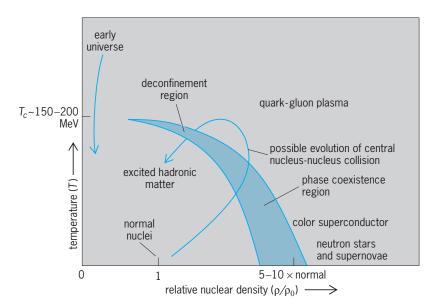


Fig. 1. Conjectured phase diagram of nuclear matter, temperature *T* versus relative nuclear density ρ/ρ_0 , where the normal nuclear density $\rho_0 = 0.16$ nucleon/fm³. Various regions of matter are shown with normal nuclei being at T = 0 and $\rho/\rho_0 = 1$. A possible path of the early universe and a possible trajectory followed by matter produced in a nucleus-nucleus collision are shown.

cold and dense. Such a state of matter is thought to have existed briefly in the very early universe until about 10 microseconds after the big bang, and its cold form might exist today inside the dense cores of neutron stars. One important prediction is that the mechanism, called spontaneous chiral symmetry breaking, that is thought to be responsible for 98% of the mass of the nucleon under normal conditions is not operative in the quark-gluon plasma. *See* BIG BANG THEORY; HADRON; NEUTRON STAR; QUANTUM CHROMODYNAMICS.

Thermodynamics of plasma. Since the quark-gluon plasma is made up of strongly interacting quarks and gluons at thermal equilibrium, an appropriate theoretical framework for describing its properties is statistical quantum chromodynamics. In this framework, calculations of quantities such as the energy density, entropy, pressure, and the equation of state of the matter, as a function of the temperature and chemical potential of the system, are carried out. (Temperature refers to the average energy of a particle in the system; the chemical potential denotes the energy required to add a nucleon to the system.) Present calculations, using computer simulations of statistical quantum chromodynamics on a discrete spatial grid (lattice gauge theory), estimate that the critical temperature T_c needed to heat matter to the point that the quarks and gluons become deconfined and form a quark-gluon plasma is about 150-200 MeV (Fig. 1), where 1 MeV is equivalent to about 10^{10} K. Such calculations assume that thermal and chemical equilibrium is achieved, and they involve extrapolations of volume and quark masses to realistic values. These assumptions need to be verified experimentally in attempts made to create a quark-gluon plasma in laboratory conditions, for example, in high-energy nuclear collisions. See STATISTICAL MECHANICS.

Dense but cold quark matter is predicted to exhibit the property of color superconductivity. This phenomenon occurs when quarks near the Fermi surface bind in pairs that act as bosons and form a Bose condensate, similar to the Cooper pairs of electrons in an electronic superconductor. Depending on the quantum numbers of the pairing quarks and the spatial structure of the Cooper-pair condensate, several different phases of color-superconducting quark matter can be distinguished. Which one of these phases would form in the core of a neutron star when the nuclear density is sufficiently high is still under investigation. *See* SUPERCONDUCTIVITY.

The nature of the phase transition, whether firstorder or higher, is an important issue. First-order transitions (such as the transformation of water to ice or steam) generally involve abrupt changes in experimental observables, such as the latent heat, whereas higher-order transitions are continuous. Present calculations indicate that the character of the hadronic matter to quark-gluon plasma transition is very sensitive to the masses of the quarks involved. For realistic masses of the up (u) and down (d) quarks, the constituents of the nucleon, and the strange (s) quark, the transition under the conditions of the early universe (high temperature and nearly zero nuclear density) appears to be very rapid, but not discontinuous. For cold but dense nuclear matter, the transition is conjectured to be of first order. The order of the transition may affect the ability of experiments to detect the plasma. *See* PHASE TRANSITIONS.

Production. The discovery and study of a new state of matter requires a means for producing it under controlled laboratory conditions. The extreme conditions of heat and compression required for the transition from the hadronic matter to the quark-gluon plasma phase can be achieved on Earth only by collisions of beams of heavy ions such as nuclei of gold or lead (although lighter nuclei can be used) with other heavy nuclei at extremely high energies (Fig. 1). To be most effective, these collisions must be as central as possible to bring to bear the full energy of the incident nuclei. Calculations using the lattice gauge model of quantum chromodynamics indicate that energy densities of at least 1-2 GeV/fm³ (1 fm = 1 femtometer = 10^{-15} m, 1 GeV = 1000 MeV), about 10 times that found in ordinary nuclear matter, must be produced in the collision for plasma formation to occur (Fig. 1). See NUCLEAR REACTION; RELATIVISTIC HEAVY-ION COLLISIONS.

Experimental signatures and results. At the earliest stage of a central collision, multiple nucleonnucleon collisions occur as the two nuclei interpenetrate, leading to a sudden formation of a highly excited region densely filled with gluons and quarkantiquark pairs. The interaction among these particles causes their fast equilibration and leads to a rapid increase of temperature and pressure (Fig. 1). On the assumption that sufficient energy density is deposited, the transition to a quark-gluon plasma may be initiated over a volume comparable to the overlap region of the colliding nuclei, which have radii of several femtometers, and on a time scale of the order of 10⁻²³ s. This hot compressed system then starts to expand under the action of its internal pressure and cools. As the temperature of the expanding matter drops below the critical value T_c needed to deconfine the quarks and gluons, they recombine to form hadrons and eventually restore the pure hadronic matter phase. Thus, no free quarks or gluons are expected at the end of the reaction

Although no single unambiguous signature for the quark-gluon plasma has been identified, several characteristic probes or observables have been suggested. Signatures are sought coming from the primary hot, dense plasma phase or from transition back into the hadronic phase. Photons and lepton pairs $(e^+e^- \text{ or } \mu^+\mu^-)$ are considered to be among the best probes of the quark-gluon phase owing to their relatively weak interaction with matter, which allows them to leave the production volume without modification of the information they carry about their primary production. *See* LEPTON.

Other sources of information about the plasma phase are hadronic fragments of very energetic quarks or gluons, which are scattered sideways during the original impact between the colliding nuclei. These quarks interact strongly with the medium and

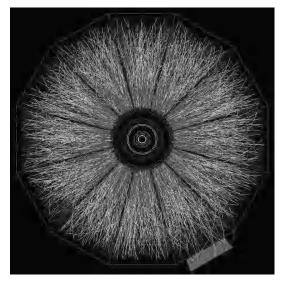


Fig. 2. Time projection chamber picture of a central gold-gold collision in which several thousand charged particles were produced. Each incident gold nucleus had a kinetic energy of 100 GeV per nucleon. The visible tracks of the particles can be analyzed to obtain information about the properties of the state from which the particles are emitted. This experiment was performed by the STAR collaboration at Brookhaven National Laboratory with the Relativistic Heavy Ion Collider (RHIC). (*Brookhaven National Laboratory*)

thereby can lose a significant part of their energy. The amount of energy lost is a measure of the transparency of the medium to quarks and gluons, which is expected to be low. (This property is analogous to the optical opacity of electromagnetic plasmas.) *See* PLASMA (PHYSICS).

The momentum spectrum of emitted hadrons is influenced by the collective flow pattern of the matter before its breakup into freely streaming particles. This flow is generated by the internal pressure of the expanding matter and reflects its equation of state as well as the initial conditions prevailing at the moment of formation of an equilibrated phase. The angular anisotropy of the flow pattern in not quite central collisions is predicted to be especially sensitive to the fluid properties of the matter.

Other probes of the properties of the produced matter involve heavier species of quarks, such as the strange (s) quarks and charm (c) or bottom (b) quarks. The reduction of the s quark mass and the ample availability of gluons in the quark-gluon plasma phase allow enough strange quarks to be produced to establish chemical equilibrium. As the system cools, the strange quarks later recombine with other, lighter (*u* and *d*) quarks to form strange hadrons, which can thus serve as vestiges of quark and gluon deconfinement during the high-density stage. Bound states of c or b quarks are predicted to dissolve at even higher temperatures due to the screening of the strong interaction in the plasma phase. A large suppression of mesons made exclusively of such heavy quarks would thus serve as evidence for the screening of forces that is characteristic of the plasma phase. See CHARM; MESON; STRANGE PARTICLES.

A systematic search for the quark-gluon plasma has been conducted at the Relativistic Heavy Ion Collider (RHIC) since 2000. Experiments at this accelerator facility colliding beams of gold nuclei with energies of up to 100 GeV per nucleon (Fig. 2) have found compelling evidence for the formation of a novel state of matter which exhibits many of the predicted properties of the quark-gluon plasma. The final state is found to be compatible with the assumption that particles are emitted from a thermally and chemically equilibrated source. The yield of energetic hadrons is strongly suppressed, indicating a high opacity of the medium to quarks and gluons. The intensity of the measured collective flow pattern among the emitted hadrons indicates that the produced matter is a nearly ideal fluid with an extremely low viscosity near the theoretical limit. Differences observed in the flow patterns of baryons and mesons suggest that the flow was generated in a phase in which quarks were not confined. Detailed studies of weakly interacting or rarely produced probes, such as lepton pairs and hadrons containing c quarks, are in progress. Experiments with nuclei of even higher energy are in preparation at the Large Hadron Collider at the European Center for Particle Physics (CERN). These experiments should provide critical tests of the theory of the strong interaction and illuminate the earliest moments of the universe. See ELEMENTARY PARTICLE; GLUONS; PARTICLE ACCELER-ATOR; QUARKS. Berndt Müller

Bibliography. M. G. Alford, Color superconducting quark matter, *Annu. Rev. Nucl. Part. Sci.*, 51:131-160, 2001; J. H. Harris and B. Müller, The quest for the quark-gluon plasma, *Annu. Rev. Nucl. Part. Sci.*, 46:71-107, 1996; T. Ludlam and L. McLerran, What have we learned from the relativistic heavy ion collider?, *Phys. Today*, 56(10):48-54, October 2003; B. Müller and J. L. Nagle, Results from the Relativistic Heavy Ion Collider, *Annu. Rev. Nucl. Part. Sci.*, 56:93-135, 2006; K. Yagi, T. Hatsuda, and Y. Miake, *Quark-Gluon Plasma: From Big Bang to Little Bang*, Cambridge University Press, 2005.

Quarks

The basic constituent particles of which elementary particles are understood to be composed. Theoretical models built on the quark concept have been very successful in understanding and predicting many phenomena in the physics of elementary particles.

Search for fundamental constituents. Physics research over the past two centuries has been probing progressively deeper into the structure of matter in order to seek, at every stage, the constituents of each previously "fundamental" entity. Thus, a hierarchy proceeding from crystals of bulk matter through molecules, atoms, and nuclei to nucleons (protons and neutrons) and mesons has been revealed. The energies required to dissociate each entity increase from thermal energies to gigaelectronvolts in proceeding from crystals to mesons. It is therefore only a logical extrapolation of past patterns to expect that hadrons—mesons and baryons (nucleons)—might be dissociated into more fundamental constituents if subjected to a sufficiently high energy. *See* ATOMIC PHYSICS; HADRON; NUCLEAR PHYSICS; NUCLEON.

Evidence supporting the quark model. The study of the elastic scattering of electrons on protons demonstrated that the proton has a finite form factor, that is, a finite radial extent of its electric charge and magnetic moment distributions. It was plausible that the charge cloud which constitutes the proton is a probability distribution of some smaller, perhaps pointlike constituents, just as the charge cloud of an atom was learned to be the probability distribution of electrons. Subsequent high-energy, deep inelastic scattering experiments of electrons on protons, leading to meson production, revealed form factors corresponding to pointlike constituents of the proton. These proton constituents, first referred to as partons, are now understood to include the constituent quarks of the proton. The quark structure of the proton was also invoked to explain the nature of secondary π -meson (pion) distributions produced in high-energy proton-proton inelastic collisions.

These high-energy collisions also produced an abundance of resonance states, equivalent to shortlived particles. The spectroscopy of these hadronic states revealed an order and symmetry among the observed hadrons that could be interpreted in terms of representations of the SU(3) symmetry group. This in turn is interpreted as a consequence of the grouping of elementary constituents of fractional electric charge in pairs and triplets to form the observed particles. The general features of the quark model of hadrons have withstood the tests of time, and the static properties of hadrons are consistent with predictions of this model. *See* SYMMETRY LAWS (PHYSICS); UNITARY SYMMETRY.

Thus, the proton and neutron are not fundamental constituents of matter, but each is composed of three quarks, very much as the nuclei of 3 H and 3 He are made of protons and neutrons, and the molecules of NO₂ and N₂O are made of oxygen and nitrogen atoms.

Kinds of quarks. There are two kinds (or "flavors") of quarks of very low mass of which the proton, neutron, and pions are composed, and a third, more massive quark which is a constituent of "strange" particles, including the *K* mesons and hyperons such as the Λ^0 . These are known as the up quark (*u*), the down quark (*d*), and the strange quark (*s*). Baryons are composed of three quarks, for example the proton (*uud*), neutron (*udd*), Λ^0 (*uds*), and Ξ^- (*dss*). Antiparticles such as antiprotons are formed by the antiquarks of those forming the particle, for example, the antiproton $\bar{p}(\bar{u} \ \bar{u} d)$. Mesons are composed of a quark-antiquark pair, such as the $\pi^+(u\bar{d})$, $\pi^-(\bar{u}d)$, $K^+(u\bar{s})$, and $K^-(\bar{u}s)$. See BARYON; HYPERON; MESON; STRANGE PARTICLES.

The quantum numbers of quarks are added to give the quantum numbers of the elementary particle which they form on combination. The unit of electrical charge of a quark is +2/3 or -1/3 of the charge

Flavor	Mass [†] , GeV/c ²	Electric charge ‡	Baryon number	Spin [§]	Isotopic spin	Strangeness	Charm
и	0.0015-0.005	+2/3	+1/3	1/2	1/2	0	0
d	0.003-0.009	-1/3	+1/3	1/2	1/2	0	0
С	1.1-1.4	+2/3	+1/3	1/2	0	0	+1
S	0.060-0.170	-1/3	+1/3	1/2	0	-1	0
t	174.3±5.1 [¶]	+2/3	+1/3	1/2	0	0	0
b	4.1-4.4	-1/3	+1/3	1/2	0	0	0

[§]Spin is in units of Planck's constant divided by 2π , written as \hbar .

The top quark mass is deduced from experimental measurements of its decay dynamics.

on a proton $(1.6 \times 10^{-19} \text{ coulomb})$, and the baryon number of each quark is +1/3 (**Table 1**). The charge, baryon number, and so forth, of each antiquark are just the negative of that for each quark.

The manner in which the u, d, and s quarks and their antiquarks may combine to form families of mesons may be indicated on two-dimensional plots for multiplets, such as the meson pseudoscalar nonet (**Fig. 1**). On such plots, hypercharge is a quantum number related to the quark strangeness and baryon number, and isospin (here the third component of isotopic spin) is a quantum number related to the u-d quark difference. See HYPERCHARGE; ISOBARIC PROCESS.

In the strong interactions of elementary particles, quark-antiquark pairs of the same flavor may be created if sufficient energy is present, but a quark does not transform into a different quark. The weak interaction, however, does permit quark transformations to occur, so that, for example, a d quark becomes a u quark in the radioactive decay of a neutron, and particles containing an s quark (strange parti-

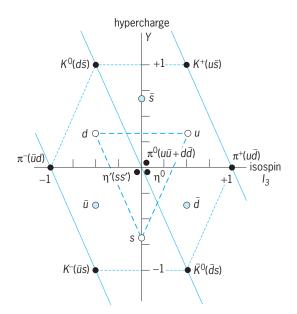


Fig. 1. Pseudoscalar meson nonet and the u, d, and s quarks and antiquarks as represented in isospin (l_3)-hypercharge space. The solid lines sloping downward to the right are lines of constant electric charge.

cles) decay to nonstrange particles only through the weak interaction. *See* FUNDAMENTAL INTERACTIONS; STRONG NUCLEAR INTERACTIONS; WEAK NUCLEAR IN-TERACTIONS.

A particularly interesting set of particles is the group of vector mesons composed of a quarkantiquark pair. An example is the ϕ meson, composed of an $s\bar{s}$ pair. It may decay to a K^+K^- meson pair but is inhibited from decaying to a $\pi^+\pi^-$, as the latter contain no *s* or \bar{s} quarks. The ϕ may also decay electromagnetically to a $\mu^+\mu^-$ or e^+e^- lepton pair; conversely, it may be produced by a photon or by the collision of an electron and positron of sufficient energy.

Such quark-antiquark systems are very analogous to the positronium atom, composed of a positron, e^+ , and an electron, e^- . The force binding this atom together is the electromagnetic (Coulomb) interaction, whereas the force binding the quark-antiquark system is the strong interaction. In both cases, the energy levels of excited states of the system are directly related to the details of the force holding the particles together. As with the quark-antiquark particles, positronium decays in a small fraction of a second through the annihilation of its particle-antiparticle constituents. *See* POSITRONIUM.

Charm, bottom, and top quarks. During the 1970s, experiments at electron-positron colliders and proton accelerators detected a relatively long-lived (that is, very narrow, in energy) resonant state of about 3.1 GeV total energy. This was interpreted as evidence for a new quark, the charm (c) quark, produced as a cc quark-antiquark resonance analogous to the ϕ . The discovery of this J/ ψ resonance was followed by the observation and study of meson systems, now labeled D mesons, containing a single cor \bar{c} quark (paired with an antiquark of another flavor), as well as baryon states containing these quarks. The two-dimensional diagrams (Fig. 1) may be elaborated by adding a third dimension corresponding to the charm quantum number, defining a polyhedron whose vertices correspond to identified charm (D), strange, and other mesons (Fig. 2). See CHARM; J/PSI PARTICLE; PARTICLE ACCELERATOR.

A few years later, experiments with higherenergy proton beams, studying the spectra of muonantimuon pairs at the Fermi National Accelerator Laboratory, discovered a more massive, narrow resonant

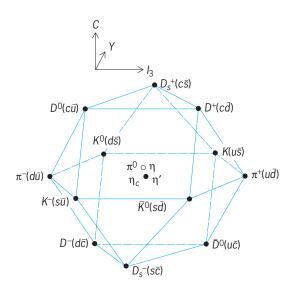


Fig. 2. Pseudoscalar mesons composed of *u*, *d*, *s*, and *c* quarks. The three orthogonal axes are the *z* component of isospin *I*₃, hypercharge *Y*, and charm *C*. (After *C*. Quigg, Lectures on charmed particles, Fermi National Accelerator Laboratory, Fermilab-Conf-78/37-THY, April 1978)

state at about 9.4 GeV, which was labeled the Υ (upsilon). This was interpreted as evidence for a more massive quark, the *b* (bottom) quark. Subsequent experiments at proton and electron accelerators confirmed the existence of the *b* quark and also observed a corresponding family of meson resonant states, now referred to as *B* mesons. A four-dimensional extension of the three-dimensional diagram of Fig. 2 can be imagined which would include *B* mesons containing the *b* quark, such as $b\bar{u}$, $b\bar{s}$, and $c\bar{b}$. See UPSILON PARTICLES.

During the 1990s, experiments observing collisions of protons and antiprotons at an energy of 1.8 TeV in the center of mass established the existence of the *t* (top) quark, primarily through analysis of its decay to a *B* meson and a *W* intermediate vector boson. The *t* mass of $174.3 \pm 5.1 \text{ GeV}/c^2$ (about the mass of a tungsten atom) is so great that its weak decay through this channel is very fast, and mesonic states of the *t* and \bar{t} quark (analogous to the Υ , the J/ψ , and the ϕ) are not observed, although the observed *t*'s are from the production of $t\bar{t}$ pairs. *See* INTERMEDIATE VECTOR BOSON.

Color. Quarks are understood to have a spin of 1/2; that is, their intrinsic angular momentum is $\hbar/2$ (where \hbar is Planck's constant *b* divided by 2π), just as for the electron and muon. Thus mesons (composed of a quark-antiquark pair) have integral spins, of 0,1, or greater (in units of \hbar). Baryons, made of three quarks, have spins of 1/2, 3/2, and so forth. A problem arose when the structure of observed baryons required two or, in some cases, three quarks of the same flavor in the same quantum state, a situation forbidden for spin-1/2 particles by the Pauli exclusion principle. In order to accommodate this contradiction, a new quantum variable, arbitrarily labeled color, was introduced; the idea is that each quark is red, green, or blue (and the antiquarks, anti-red, and

so forth). The color quantum number then breaks the degeneracy and allows up to three quarks of the same flavor to occupy a single quantum state. Confirmation of the color concept has been obtained from experiments with electron-positron storage rings, and the theory of quantum chromodynamics (QCD), based on this concept, has been developed. According to quantum chromodynamics, hadrons must be colorless; for example, baryons must consist of a red, a green, and a blue quark, and mesons of a quarkantiquark pair of the same color (for example, a red quark and an anti-red antiquark). *See* COLOR (QUAN-TUM MECHANICS); EXCLUSION PRINCIPLE; SPIN (QUAN-TUM MECHANICS).

The field quanta of quantum chromodynamics are gluons; massless, spin-1 quanta which interact with quarks. This is very analogous to the manner in which photons, the quanta of electromagnetic interaction, interact with particles containing electric charge and are responsible for electromagnetic forces. The QCD theory is part of the now widely accepted standard model of elementary particle interactions, together with the electroweak theory. Experiments have increasingly confirmed details of the standard model to the extent that most physicists are confident that it is fundamentally correct. *See* ELEC-TROWEAK INTERACTION; GLUONS; QUANTUM CHRO-MODYNAMICS; STANDARD MODEL.

Generations of particles. There are three sets, or "generations," of quarks and leptons. Each generation contains a charged lepton (electron, muon, or tau lepton); a correspoding neutrino; a charge -1/3 quark color triad; and a charge +2/3 quark triad. Evidence for a fourth kind of neutrino would signify a new massive lepton and a pair of very massive, new quarks. Detailed measurements of the decays of the Z^0 intermediate vector boson at electron-positron colliders have firmly established the number of neutrinos at three (2.984 \pm 0.008 in a 2000 compilation). This confirms the conclusion that the three generations of elementary particles now known are a complete set. *See* LEPTON; NEUTRINO.

Searches for free quarks. The QCD theory argues that free quarks do not exist, because the gluon-mediated force between quarks increases without limit proportional to the separation between single quarks. When two quarks are separated, beyond a certain point the energy required results in the creation of a quark-antiquark pair, resulting in a meson and a baryon, for example. During the 1960s and 1970s there were intensive searches for evidence of free quarks, at particle accelerators, in cosmic rays, and in stable matter.

The signature for a free quark should be clear. A fast-moving particle of charge $\pm 1/3$ or $\pm 2/3$ the charge on a proton or electron would produce ionization in passing through matter 1/9 or 4/9 as great as that of an integrally charged particle (such as an electron or muon). No confirmed evidence for free quarks was found. Upper limits to the possible production cross sections for free quarks from particle accelerator experiments are shown in **Fig. 3**. In spite

of the theory and the body of negative evidence, some physicists still seek evidence for the existence of free quarks. A Stanford University group has an ambitious program to study stable matter using a modernized, automated version of the Millikan oil drop method (first used in 1911 to establish the charge on an electron) to seek free quarks or at least to set much more stringent limits on their existence. *See* ELECTRON.

Unresolved questions. Quarks and the theory of quantum chromodynamics are now firmly established as cornerstones of the standard model of elementary particles (together with the electroweak theory, charged leptons, neutrinos, and so forth). However, unanswered questions remain.

The advanced string and M theories have the property of supersymmetry, which demands that every spin-1/2 quark and lepton must have a partner with integral spin. As of 2000, no experimental evidence for any of these supersymmetric (SUSY) particles had been found; if they exist, they must have rest masses greater than $30-100 \text{ GeV}/c^2$, depending on the candidate particle (or about 30-100 times the mass of a proton). As new accelerators enable searches at higher energies, the search for these particles will continue. *See* SUPERSTRING THEORY; SUPERSYMME-TRY.

Contemporary theories also predict that there exists one or more massive particles of integral spin, the Higgs paticles, responsible for the rest masses of the quarks and charged leptons. Experiments have precluded the existence of Higgs particles at masses below 100 GeV/ c^2 . In late 2000, however, groups working at the CERN electron-positron collider, LEP, reported possible evidence for a Higgs boson of a mass of about 115 GeV/ c^2 . The statistical evidence for this discovery is weak, and confirmation from future experiments is required before it can be included in tables of "fundamental particles" with confidence. *See* HIGGS BOSON.

Quarks may be permanently stable against decay via the weak interaction; however, it is also possible that quarks spontaneously decay to leptons. Intensive searches for the decay of the proton (into a neutral pion and a positron, for example) have been negative, setting a lower limit of over 10^{32} years for the proton lifetime. However, the apparent asymmetry of the universe between matter and antimatter (there is, at present, no evidence for primordial antimatter) suggests that antiprotons, for example, may spontaneously decay (or transform) more readily. Experiments have shown that the antiproton lifetime is at least 10^6 years; however, the age of the universe is over 10^{10} years. *See* ANTIMATTER; PROTON.

Some theories have postulated that quarks are composed of smaller constituents, just as other objects that were originally believed to be fundamental subsequently were found to have internal structure. Objects called preons by some theorists could make up the different flavors of quarks. Again, there is no evidence for such internal structures. So far, all ob-

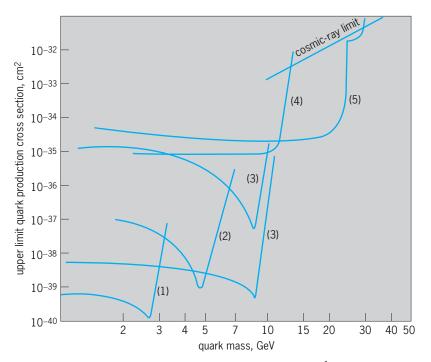


Fig. 3. Upper-limit cross section for production of quarks of charge $-\frac{1}{3}e$ as a function of quark mass from experiments at various particle accelerators and cosmic rays. The different experiments are indicated by numbers: (1) CERN 28-GeV proton synchrotron; (2) Serpukov 70-GeV proton synchrotron; (3, 4) Fermi National Accelerator Laboratory synchrotron; (5) CERN Intersecting Storage Rings. (*After L. W. Jones, A review of quark search experiments, Rev. Mod. Phys.*, 494:717–752, 1977)

servations are compatible with the quarks as point objects, like the electron. Lawrence W. Jones

Theory. The theory of the properties and interactions of quarks will now be discussed in detail.

Local gauge theories. The theory of the interactions of quarks is expressed, at a fundamental level, in terms of local gauge theories: theories which are invariant under independent symmetry transformations at each space-time point. Such a theory is characterized by its symmetry group, coupling constants, matter fields, and whether or not the symmetry is broken spontaneously. (Spontaneously broken symmetry is described below.) Local gauge theories first entered elementary particle physics in quantum electrodynamics (QED), the theory of the interactions of the electron-positron field with the Maxwell field of electromagnetism. For quantum electrodynamics, the symmetry group is U(1), the group of complex numbers of modulus one which can be identified with rotations of the complex plane; the coupling constant is the electric charge, the matter field is the electron-positron field, and the group is unbroken. See QUANTUM ELECTRODYNAM-ICS.

Local gauge theories were generalized to "nonabelian" theories by C. N. Yang and R. L. Mills. Nonabelian theories have a non-abelian group, that is, a group in which, in general, $g_2g_1 \neq g_1g_2$, where the g's are group elements. Typical non-abelian groups are the SU(N) and SO(N) groups. SU(N) is the group of N by N unitary (U) matrices which have determinant one (or are "special," S). Unitary matrices preserve the length in a complex space: they have the property $UU^{\dagger} = I$, where *I* is the unit matrix, and U^{\dagger}_{\dagger} is the complex conjugate transpose of *U*. SO(*N*) is the group of *N* by *N* orthogonal matrices. Orthogonal matrices preserve the length in a real space: they have the property $OO^{t} = I$, where O^{t} is the transpose of the matrix *O*. The group of rotations in three-dimensional space is the group SO(3). See GROUP THEORY; MATRIX THEORY.

The groups which are relevant for gauge theories are Lie groups, named after S. Lie. These groups are continuous, like the group of rotations in threedimensional space, and have further smoothness properties, which are also shared by SO(3). An arbitrary element of a Lie group near the unit or identity can be written as the identity plus a sum of a certain number of "generators" multiplied by real numbers. Each generator corresponds to a small or "infinitesimal" transformation away from the identity. For SO(3), the generators can be chosen to be the deviations from the identity associated with small rotations around the x, y, and z axes. See LIE GROUP.

In a local gauge theory, there must be a gauge field, whose associated gauge particle is a spin-1 boson, corresponding to each generator. For U(1), which has one generator, there is one gauge field. In quantum electrodynamics, this is the electromagnetic vector potential. For SU(*N*) and SO(*N*), there are $N^2 - 1$ and 1/2N(N - 1) generators and gauge fields, respectively. Gauge fields are the mediators of the interactions between the matter fields. When the local gauge symmetry is unbroken, the particles associated with the gauge fields are massless, corresponding to a long-range interaction.

For groups such as U(1), SU(N), or SO(N), the interactions are determined by a single constant. This fact, and the fact that the form of the interaction is completely fixed by the group and the matter fields, makes such gauge theories tightly constrained. There is a charge associated with each generator of a gauge theory. For quantum electrodynamics the charge is the electric charge of the electron. A crucial difference between non-abelian and abelian gauge theories is that the gauge fields themselves carry the charges in non-abelian theories, while in abelian theories the gauge fields, for example, the electromagnetic potential in quantum electrodynamics, are neutral (that is, do not carry the charges). Thus the gauge fields in a non-abelian theory can interact by exchanging other gauge fields, while the gauge field (or fields) in an abelian theory only interacts with each other indirectly via the matter fields.

Spontaneous symmetry breaking is the phenomenon in which the symmetry is broken by the vacuum state, although the lagrangian and equations of motion of the theory obey the symmetry. Spontaneous symmetry breaking in a local gauge theory gives rise to a mass for the gauge particles associated with the generators for which the symmetry is broken. The interactions mediated by massive gauge

Generation:	1	2	3
Quarks [*]	u _α	c_{lpha}	t_{lpha}
	d_{lpha}	S_{lpha}	b_c
Leptons	ν_{e}	$ u_{\mu}$	ν_{τ}
	e ⁻	μ	τ-
* The index α runs of	wer the three quark	colors. The three	uarke with

particles are short-ranged. *See* GAUGE THEORY; SYM-METRY BREAKING.

Electroweak interactions of quarks. The electromagnetic interactions of quarks can be described by using the electromagnetic current given by Eq. (1), where the

$$j_{\mu} = \sum_{i} \left({}^{2}\!/_{3} \bar{u}_{i} \gamma_{\mu} u_{i} - {}^{1}\!/_{3} \bar{d}_{i} \gamma_{\mu} d_{i} \right)$$
(1)

 γ_{μ} are the Lorentz covariant forms of the Dirac matrices, and the sum runs over the three presently known generations of quarks (**Table 2**). The u_i are the so-called up quarks with charge 2/3e (u, c, and t), and the d_i are the so-called down quarks with charge -1/3e (d, s, and b). The six quark species are called flavors. This form of the electromagnetic current says that the electromagnetic interaction does not change the flavor of the quark, and that quarks interact electromagnetically with a strength proportional to their electric charge. Electromagnetism is mediated by the photon, a neutral massless gauge particle.

The weak interactions of quarks can be described by using three weak currents: charge raising and lowering ones, and a neutral current. The prototype of charged-current weak interactions is beta decay, exemplified by the decay of the neutron, $n \rightarrow pe^{-}\bar{v}_{e}$. The charge raising current has the form of Eq. (2), where d_{i} stands for a superposition or

$$j_{\mu}^{+} = \sum_{i} \bar{u}_{i} \gamma_{\mu} d'_{i} \tag{2}$$

mixture of the three down quarks. N. Cabibbo introduced this mixing between the *d* and *s* quarks to take account of the fact that strangeness-conserving decays, such as neutron decay, go at a faster rate (when phase space factors are removed) than the analogous strangeness-violating decays, such as lambda decay, $\Lambda \rightarrow pe^-\nu_e$.

The theory with only three flavors of quarks leads to "strangeness-changing neutral-current weak decays," such as $K \rightarrow \mu \bar{\mu}$, which experimental data show occur at an extraordinarily low rate. To avoid these unseen decays, it was suggested that there should be a fourth flavor of quark, the charm quark *c*, whose presence in intermediate states would cancel the unobserved decays, provided the mass of the *c* quark was not too different from the mass of the *u* quark. The charm theory predicted new hadronic states containing the *c* quark. As discussed above, the discovery of the narrow J/ψ resonance at mass 3.1 GeV/ c^2 gave dramatic confirmation of the theory. The further narrow Υ resonance discovered at mass 9.4 GeV/ c^2 gave evidence for the *b*, the fifth flavor of quark; and the sixth flavor of quark, the *t*, was discovered in proton-antiproton collisions.

Until the work of S. Glashow, S. Weinberg, and A. Salam, who contributed theoretical developments which gave a partial unification of electromagnetism and weak interactions, the theory of weak interactions suffered from divergences in calculations beyond lowest order in the Fermi constant, $G_F \sim$ $10^{-5}M_P^{-2}$, where M_P is the proton mass. This partial unification used a local gauge theory based on the group $SU(2) \times U(1)$, which has four generators, three from the group SU(2) and one from U(1). The gauge fields W^+ and W^- associated with two of the SU(2) generators are identified as the mediators of the charge-changing weak interactions mentioned above. The electromagnetic vector potential is associated with a linear combination (or "mixture") of two of the remaining generators. A fourth gauge field Z^0 associated with the orthogonal mixture of the last two generators is required by the theory. Spontaneous symmetry breaking simultaneously produces the mixing of the generators, and gives masses to the W^{\pm} and Z^0 fields, as required by the short range of the weak interactions. Since the Z^0 is neutral, this partially unified theory predicts the existence of neutralcurrent weak processes, such as $v_{\mu}p \rightarrow v_{\mu}p$. Such processes were found in neutrino experiments, giving striking confirmation of the electroweak theory. See NEUTRAL CURRENTS.

The CP violation observed in the $K^0 - \bar{K}^0$ system is a puzzling phenomenon. M. Kobayashi and T. Maskawa extended the mixing of the d and s quarks due to Cabibbo to a three-dimensional mixing of the d, s, and b quarks. The CP violation is associated with complex elements of the Cabibbo-Kobayashi-Maskawa (CKM or KM) matrix. For fewer than six flavors of quarks, redefining the arbitrary phases of the quarks can always make the matrix real, so at least six flavors of quarks are needed for quark mixing to lead to CP violation. Experiments on weak decays of hadrons have given strong constraints on the elements of the CKM matrix. In general, the diagonal elements are close to one, which means that the mass eigenstates of quarks, which are relevant in strong interactions, are close to the weak eigenstates, which are relevant in weak processes. The elements displaced by one position from the diagonal are small, and the elements doubly displaced from the diagonal are very small or zero, which means that the main mixing is between neighboring generations.

Quantum chromodynamics. The apparently contradictory low-energy (massive, strongly bound, indeed permanently confined) and high-energy (almost massless, almost free) behaviors of quarks can be reconciled by using quantum chromodynamics, which is a local gauge theory based on the group SU(3), as the theory of the strong interactions. The eight gauge fields of quantum chromodynamics are called gluons.

The high-energy properties of quarks follow from the asymptotic freedom of quantum chromodynamics. The interquark potential has the form of Eq. (3).

V

$$V(r) = \frac{g_{\rm eff}^2(r)}{r}$$
(3)

Because of quantum corrections, the effective or running coupling constant g_{eff} of a quantum field theory depends on the distance scale r at which the coupling constant is measured. A theory is asymptotically free if g_{eff} goes to zero as r goes to zero so that perturbation theory is valid for small r. For large r, the potential grows linearly without bound, as given in Eq. (4), where σ is a constant, so that single quarks

$$V(r) \to \sigma r$$
 (4)

can never be free. This is the phenomenon of permanent quark confinement. Since confinement occurs when g_{eff} is large, nonperturbative methods are necessary to demonstrate confinement.

Lattice gauge theory. The best evidence for confinement comes from the Monte Carlo numerical approximation to lattice gauge theory on finite lattices. The properties of quantum chromodynamics and of quarks can be calculated by using this discrete formulation in which the dynamical variables are placed on the vertices and links of a lattice. This approach allows the treatment of nonperturbative features of quantum chromodynamics. Two types of problems have been studied: demonstration of confinement and finding its properties, and computation of observables such as hadron mass ratios and weak hadronic matrix elements. The results obtained are promising but not definitive. *See* MONTE CARLO METHOD.

Standard model. The local gauge theory based on the group $SU(3) \times SU(2) \times U(1)$, in which all fundamental interactions have a gauge theory origin, and SU(3) is unbroken while $SU(2) \times U(1)$ is spontaneously broken to the U(1) of electromagnetism, serves as a standard model for all elementary particle physics at energies below about 100 GeV.

Beyond the standard model. A major direction of theoretical research beyond the standard model is the search for a grand unified theory in which quantum chromodynamics and the electroweak interactions, and possibly gravity, are part of a larger scheme with a single coupling constant; however, the striking prediction of grand unified theories that the proton should decay has not been confirmed. Two other areas of such research are the exploration of supersymmetric theories in which all fermions have bosonic partners, and vice versa; and the study of models in which quarks, leptons, and other "elementary" particles are composites of still more basic particles. *See* GRAND UNIFICATION THEORIES.

The most ambitious departure concerns superstring theories, which combine the idea of supersymmetry with the use of a one-dimensional "string" as the object over which the quantum theory is defined to replace the zero-dimensional point over which conventional quantum field theories are defined. Superstring theories lead naturally to a quantum theory of gravity and, for the first time, seem likely to provide a well-defined theory, free of uncontrollable (technically, unrenormalizable) infinities. The hope that the entire content of elementary particle physics can be derived from a superstring theory has not yet been realized; nonetheless, it seems likely that these theories will play an important role in future developments. 0. W. Greenberg

Bibliography. P. C. Bhat, H. B. Prosper, and S. S. Snyder, Top quark physics at the Tevatron, Int. J. Mod. Phys. A, 13:5113-5218, 1998; S. L. Glashow, Quarks with color and flavor, Sci. Amer., 233(4):38-50, October 1975; D. E. Groom et al. (Particle Data Group), Review of particle physics, Eur. Phys. J. C, 15(1-4):1-878, 2000; N. M. Mar et al., Improved search for elementary particles with fractional electric charge, Phys. Rev. D, 53:6017-6032, 1996; Y. Nambu, Quarks, 1985; A. Pickering, Constructing Quarks, 1984; C. Rebbi, The lattice theory of quark confinement, Sci. Amer., 248(2):54-65, February 1983; R. F. Schwitters, Fundamental particles with charm, Sci. Amer., 237(4):56-70, October 1977; P. F. Smith, Searches for fractional electric charge in terrestrial materials, Annu. Rev. Nucl. Part. Sci., 39:73-111, 1989.

Quarrying

The process of extracting rock material from the outer portions of the Earth's crust. The shape of the excavated space is dependent on the geologic condition and orientation of the rock mass being quarried. This excavated space is either an open pit or an underground cavity that is accessed by tunnel or shaft. An underground quarry is usually called a mine, even though the process for making the excavation is usually the same as on the surface. *See* OPEN PIT MINING; ROCK; UNDERGROUND MINING.

The actual quarrying technique depends on the specific use of the excavated rock. Rock that is excavated for use as dimension stone—relatively flaw-less blocks or slabs of stone—is carefully extracted as large, 20-ton blocks (**Fig. 1**). These blocks are sawed into thin (3-cm or 1.2-in.) slabs, and then cut to size and polished as finished pieces (**Fig. 2**). Rock that is quarried for crushed stone is used either for construction aggregate or as an industrial mineral source. Crushed stone is broken in place by drilling and blasting into manageable-sized pieces for transport by truck or conveyor to a crushing plant, where it is crushed and screened into segregated, sized particles (**Fig. 3**). *See* STONE AND STONE PRODUCTS.

Dimension stone. An expanding construction market during the last decade of the twentieth century created an increased demand for dimension stone. This increased demand for stone over other building



Fig. 1. Pulling a 100-ton marble block in an underground quarry. This block was cut with a wire saw and will be cut into smaller (20-ton) blocks for transporting to the mill. (*Vermont Quarries, Inc.*)



Fig. 2. Marble slab sawed from a block of marble and polished, ready to be sawed to size. (Geomapping/ Technographics)

materials was due to its beauty and durability, as well as reduced costs for stone products. The reduction in cost was due to improved quarry and finishing techniques, as well as automated machinery. Dimension stone consumption is expected to increase, as more stone is used for countertops and floors in both new and renovated buildings.

The methods of quarrying large (20-ton) blocks, measuring approximately $3 \times 2 \times 1.5$ m ($10 \times 7 \times 5$ ft), have drastically changed. With the exception of some deep quarries, there are very few derricks used to lift dimension-stone blocks to the surface. Diesel-driven and hydraulically operated rubber-tired loaders are the predominant means for transporting blocks from the bottom of the quarry to the surface. Small, transportable electrically driven saws with segmented diamond wire or fixed bars with carbide-impregnated chains are used for cutting manageable-sized blocks from the quarry. Hydraulic jacks and pneumatic wedges have replaced mechanical wedges for splitting blocks.

In practice, the specific system that is selected for quarrying depends on the stone type. Slate is easily split along a major cleavage plane. Most slate, after being sized by circular diamond saws, is mechanically split into roofing tiles. Commercial marble deposits are more massive and do not split as readily as slate, but are easily sawed into blocks. Commercial granite deposits are also massive, but they tend to split in a preferred direction called the "easy way." A light explosive charge placed in equally spaced, drilled holes along this plane will separate individual blocks from the main stone mass. Blocks of soapstone, limestone, sandstone, and travertine are all quarried by using a combination of drill, saw, and various splitting techniques. See GRANITE; LIMESTONE; MARBLE; ROCK MECHANICS; SANDSTONE; SLATE; SOAP-STONE; TRAVERTINE.

Crushed stone. Construction aggregate is the stone that is used for making asphalt and concrete roadways. It also is the stone, sand, and gravel used for building foundations and for fill in landscaping. Industrial minerals are the minerals and rock quarried for agricultural and manufactured products. Toothpaste, paint, plastics, paper, and glass products all contain some minerals, including clay, titanium, wollastonite, talc, gypsum, as well as phosphate, limestone, and sandstone rock. Each person living in North America will consume or use in her or his lifetime more then 61 tons of these minerals, all of which have to be quarried from the Earth's crust. The method used for quarrying crushed stone has remained constant, while improvements in the technology and the size of the equipment have allowed a greater tonnage to be quarried, along with minimizing the environmental impact. See CON-CRETE; CRUSHING AND PULVERIZING; GRAVEL; MINER-ALS; SAND.

The main task in quarrying crushed stone is to break large rock masses into smaller pieces. Depending on the geology and uniformity of the rock being quarried, large areas are drilled for blasting into piles of smaller rock. The blast holes are 8-25 cm (3-10 in.) in diameter and usually 6-20 m (20-66 ft) deep. The drilled holes are filled with a blasting agent and detonated in a specific sequence, which breaks the rock mass into predeterminedsized pieces. These pieces are transported to a crushing and screening plant for further size reduction and segregation. Additional processing separates specific minerals for industrial use. Softer rock types, such as salt, gypsum, clay, and phosphate rock, are quarried without explosives. Various rock-grinding machines are used to cut, scrape, and remove the soft rock from the quarry floor. These machines produce small



Fig. 3. A 20-ton load of crushed marble being readied for transport to a plant facility to separate the pure calcium carbonate mineral from the rock. The calcium carbonate will be used as filler or extender in industrial products. (*Geomapping/Technographics*)

fragments, which are further processed into finished products. Lance Mead

Bibliography. D. D. Carr (ed.), *Industrial Minerals* and Rocks, 6th ed., Society for Mining Metallurgy and Exploration, 1994; H. L. Hartman (ed.), *SME Mining Engineering Handbook*, Society for Mining Metallurgy and Exploration, 2d ed., vol. 2, chaps. 14.1 and 14.2, 1992; *Mineral Yearbook*, U.S. Geological Survey, annually.

Quartz

The most common oxide on the Earth's surface, constituting 12% of the crust by volume. Quartz is a crystalline form of silicon dioxide (SiO₂). When cooling magmas solidify into igneous rocks and when preexisting rocks metamorphose into new mineral assemblages, the crystallization of quartz is favored by the sheer abundance of silica in the crustal environment. Among the igneous rocks, quartz is especially common within granites, granodiorites, pegmatites, and rhyolites. In addition, quartz can be observed in lowto high-grade metamorphic rocks, including phyllites, quartzites, schists, granulites, and eclogites. Because hydrothermal fluids are enriched in dissolved silica, the passage of fluids through rock fractures results in the emplacement of quartz veins. See GRAN-ITE; GRANODIORITE; IGNEOUS ROCKS; METAMORPHIC ROCKS; PEGMATITE; RHYOLITE.

Once quartz has formed, it persists through erosional reworking because of its low solubility in water (parts per million) and its high mechanical hardness (7 on Mohs scale). Consequently, quartz becomes increasingly concentrated in beach sands as they mature, and it is a major component of sandstone. In sedimentary environments, quartz also forms as the final crystallization product during

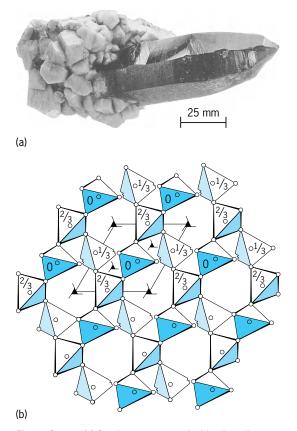


Fig. 1. Quartz. (a) Smoky quartz crystal with microcline, found in Florissant, Colorado (*American Museum of Natural History specimen*). (b) Projection of the atomic structure of α -quartz along the c axis. Each Si⁴⁺ cation is surrounded by an O²⁻ coordination tetrahedron (outlined). Numbers next to the Si atoms represent fractional coordinates along the c axis. Unit cell edge measures 0.491 nanometer.

silica diagenesis; amorphous silica on the sea floor that derives from the skeletons of diatoms, radiolarians, and sponges will transform to quartz upon prolonged exposure to increased temperatures (\leq 300°C or 572°F) and pressures (\leq 2 kilobars or 200 pascals) after burial. *See* DIAGENESIS; HARDNESS SCALES; SAND-STONE.

Crystal structure and chemistry. As with virtually all silicates, the atomic framework of the quartz structure consists of Si⁴⁺ cations that are tetrahedrally coordinated by oxygen anions (O^{2–}) [**Fig. 1**]. Every oxygen anion is bonded to two silicon cations, so that the tetrahedral units are corner-linked to form continuous chains. In low-temperature quartz (or α -quartz), two distinct tetrahedral chains spiral about the crystallographic c axis, creating ditrigonal tunnels along this direction. These chains wrap around the *c* axis in accordance with threefold screw symmetry, and α -quartz belongs to the trigonal crystal system. Prismatic crystals of quartz (Fig. 1*a*) are elongate parallel to these intertwining tetrahedral chains.

Although the silica tetrahedra can be depicted as spirals about the c axis in a left-handed sense, right-handed quartz crystals are found in nature as abundantly as are left-handed crystals. These enantiomorphic varieties are known as the Brazil twins of quartz, and they may be distinguished by crystal shape (corresponding crystal faces occur in different orientations) and by opposite optical activities. When polarized light passes through left-handed quartz, the plane of vibration is rotated counterclockwise as the light comes from the crystal toward the observer (levorotation); by contrast, right-handed quartz is dextrorotatory. *See* CRYSTAL OPTICS; POLAR-IZED LIGHT.

In addition to the Brazil twins, natural quartz crystals frequently contain structural domains called Dauphiné twins. These are regions within an α -quartz crystal that are rotated with respect to each other by 180° about the *c* axis. Usually Dauphiné twin domains are not evident to the unaided eye, but they are visible in reflected light after crystals are etched with hydrofluoric acid. Whereas the opposite handedness that describes the Brazil twins of quartz is a happenstance of crystal growth, Dauphiné twins generally form during the cooling of high-temperature β -quartz to α -quartz, and they may be induced when α -quartz crystals are stressed.

Impurity concentrations in natural α -quartz crystals usually fall below 1000 parts per million. Because the void space within the ditrigonal tunnels is fairly restricted, the most common interstitial cations are the monovalent species with small radii: sodium (Na⁺), lithium (Li⁺), and especially hydrogen (H⁺). Inclusion of even small amounts of hydrogen dramatically weakens quartz, and so-called wet quartz crystals respond to geologic pressures in a plastic rather than brittle fashion. Charge balance is achieved by the substitution of such trivalent cations as aluminum (AI^{3+}) and iron(III) [Fe³⁺] for silicon cations. Upon irradiation, crystals containing these impurities display characteristic color centers. The violet and yellow hues observed in amethyst and citrine are associated with Fe, and black smoky quartz contains Al. The white coloration of milky quartz reflects light scattering off minute fluid inclusions, and the pink tint in rose quartz is believed to arise from finescale intergrowths of a pegmatitic mineral called dumortierite [Al₂₇B₄Si₁₂O₆₉(OH)₃]. See AMETHYST; DU-MORTIERITE.

Silica system. When α -quartz is heated above 573°C (1063.4°F) at atmospheric pressure (1 bar or 100 kilopascals), the structure transforms to β quartz (Fig. 2). This transition requires the bending but not the breaking of primary Si-O bonds, and thus it is rapid and reversible. High-temperature β -quartz has a more open framework with hexagonal symmetry, and the density of β -quartz (2.53 g/cm³) is lower than that of α -quartz (2.65 g/cm³). With increasing temperature, β -quartz undergoes a radical restructuring that involves the rupture and reassembly of Si-O bonds to form the mineral tridymite. At even higher temperatures, tridymite in turn experiences a reconstructive transformation to the mineral cristobalite. At 1727°C (3141°F), cristobalite melts to silica liquid. Silica glass, cristobalite, and tridymite may be found naturally in volcanic rocks that have quenched upon eruption. See SILICA MINERALS; SILICATE MINER-ALS; SILICATE SYSTEMS.

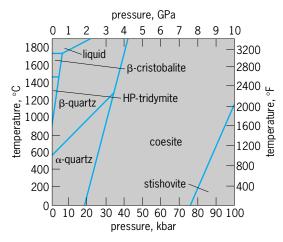


Fig. 2. Phase diagram for the silica system representing fields of thermodynamic stability for the various silica minerals as a function of pressure and temperature. HP indicates hexagonal-symmetry, primitive unit cell. (After C. Klein, and C. S. Hurlbut, Jr., Manual of Mineralogy, 21st ed., John Wiley and Sons, 1993)

When quartz is exposed to high pressure, it converts to the mineral coesite, which is structurally similar to feldspar. Above \sim 76 kbar (7.6 gigapascals), coesite transforms to stishovite, which is unique among the silica minerals in that each Si atom is coordinated by six rather than four oxygen atoms. Stishovite has the same structure as rutile (TiO₂). Coesite and stishovite occur at the sites of meteorite impacts on the surface of the Earth and may be major constituents of the Earth's mantle. *See* COESITE; COORDINATION CHEMISTRY; FELDSPAR; STISHOVITE.

Microcrystalline quartz commonly is intergrown with a metastable silica mineral called moganite. Microcrystalline varieties usually are classified by their textures: Fine-grained fibrous silica includes chalcedony, agate, and carnelian. By contrast, chert, flint, jasper, and onyx are equigranular varieties. These silica varieties have an extensive and sometimes ambiguous nomenclature arising from their historical role as gemstones. *See* AGATE; CHALCEDONY; CHERT; GEM; JASPER; ONYX.

Uses. Quartz is used predominantly by the construction industry as gravel and as aggregate in concrete. In addition, quartz is important in advanced technologies. Quartz is piezoelectric and has an extremely high quality factor. The high quality factor means that a bell made of quartz would resonate (ring) for a very long time. This property, combined with its piezoelectric behavior, makes quartz the perfect crystal for oscillators in watches.

Compression of α -quartz perpendicular to the *c* axis creates an electrostatic charge, and this property is exploited in oscillator plates in electronic components. Large flawless crystals of quartz are routinely synthesized for oscillators and for prisms in laser optic systems. Quartz also is employed in abrasives, fluxes, porcelains, and paints. *See* CONCRETE; OSCILLATOR; PIEZOELECTRICITY. Peter J. Heaney

Bibliography. C. Frondel, *The System of Mineral*ogy, 7th ed. vol. 3, 1962; P. J. Heaney, C. T. Prewitt, and G. V. Gibbs (eds.), *Silica: Physical Behavior, Geochemistry and Materials Applications*, 1994; M. O'Donoghue, *Quartz*, 1987; R. B. Sosman, *The Phases of Silica*, 1965.

Quartz clock

A clock that makes use of the piezoelectric property of a quartz crystal. When a quartz crystal vibrates, a difference of electric potential is produced between two of its faces. The actual displacement of the surface is generally only a few atomic diameters. The crystal has a natural frequency of vibration that depends on its size and shape. If it is placed in an oscillating electric circuit having nearly the same frequency as the crystal, it is caused to vibrate at its natural frequency, and the frequency of the entire circuit becomes the same as the natural frequency of the crystal. The first quartz clock was invented by Warren Marrison in 1927. *See* OSCILLATOR; PIEZO-ELECTRICITY.

The actual oscillation frequency depends not only on the quartz crystal but also on the oscillator circuit. In the quartz oscillator, this natural frequency may be used to produce other frequencies such as 1 or 5 MHz. A clock displaying the time of day can also be driven by using one of these frequencies.

The natural frequency of a quartz crystal is nearly constant if precautions are taken when it is cut and polished and it is maintained at nearly constant temperature and pressure. After a crystal has been placed in operation, its frequency usually varies slowly as a result of physical changes. If allowance is made for changes, laboratory quartz-crystal clocks may run for a year with accumulated errors of less than a few thousandths of a second. However, quartz crystals typically used in watches may accumulate errors of several tens of seconds in one year. Modern watches that utilize quartz crystals often use crystals that are chemically etched to shape from thin sheets of quartz. Accuracy depends on the precision of the cut and the degree of contamination of the quartz. See WATCH.

For comparison, clocks using rubidium as a frequency standard might be expected to have accumulated errors of less than a few ten-thousandths of a second, while those using cesium might be expected to have accumulated errors better than a few millionths of a second. *See* ATOMIC CLOCK.

The advantage of quartz clocks is that they are reliable, relatively accurate and inexpensive, and easy to use in various applications such as computers and microprocessors. Thus, despite their inaccuracy relative to some other types of clocks, they enjoy wide popularity, particularly in applications requiring accurate timekeeping over a relatively short time span. Natural and cultured quartz crystals are used in modern devices. In these applications, the rates and epochs of the quartz clocks may be readjusted periodically to account for possible accumulated errors. *See* CLOCK; COMPUTER; MICROPROCES-SOR; TIME. Dennis D. McCarthy

Quartzite

A metamorphic rock consisting largely or entirely of quartz. Most quartzites are formed by metamorphism of sandstone; but some have developed by metasomatic introduction of quartz, SiO₂, often accompanied by other chemical elements, for example, metals and sulfur (ore quartzites). The geological relations and the shape of quartzite bodies serve to distinguish between them (see **illus.**). The metasomatic quartzites are often found as contact products of intrusive bodies. *See* METASOMATISM; SAND-STONE.

The transition from sandstone to quartzite is gradational. All states of relic clastic structures are encountered. Some sandstones are soon completely metamorphosed. Others are very resistant, and in many highly metamorphic quartzites of the Precambrian, there are relic structures still to be observed.

Pure sandstones yield pure quartzites. Impure sandstones yield a variety of quartzite types. The cement of the original sandstone is in quartzite recrystallized into characteristic silicate minerals, whose composition often reflects the mode of development. Even the Precambrian quartzites correspond to types that are parallel to present-day deposits.

Carbonate cement reacts with silica to produce silicates during high-temperature metamorphism. However, ferric silicate has never been observed. The Fe_2O_3 pigment in deposits of desert sand resists any degree of metamorphism. Therefore, old Precambrian quartzites exhibit the same red color as the present-day sand of Sahara.

Under the conditions of regional metamorphism, cement composed of clay gives rise to sillimanite or kyanite, potash-rich cement yields potash feldspar or mica, lime and alumina yield plagioclase or epidote, dolomitic cement yields diopside or tremolite, and siderite cement yields gruenerite. Wollastonite will crystallize from pure calcite cement. Such quartzites



Waterloo quartzite, with schistosity developed by shearing on bedding planes. Dodge County, Wisconsin. (USGS)

occur as folded layers alternating with layers of other sedimentary rocks.

In some feldspathic quartzites, thin dark micaceous layers parallel to the foliation superficially suggest relic bedding. Many of the Moine schists of Scotland illustrate this. True bedding is marked, however, by thin strings of zircon, iron ore, or other inherited concentrations of the heavy minerals.

Under the condition of contract metamorphism, the cement of the original sandstone will recrystallize and minerals of the hornfels facies will develop. Quartz itself is usually stable. However, in very hot contacts, against basic intrusions, quartz may invert to tridymite, which again has reverted to quartz; or it may even melt. Such vitrified sandstones resulting from partial fusion of inclusions in volcanic rocks are called buchites. They often contain fritted feldspar fragments, corroded grains of quartz, and a matrix of slightly colored glass corresponding to the fused part of the sandstone. *See* HORNFELS; METAMORPHIC ROCKS; QUARTZ; SILICATE MINERALS. T. F. W. Barth

Quasar

An astronomical object that appears starlike on a photographic plate but possesses many other characteristics, such as a large redshift, that prove that it is not a star. The name quasar is a contraction of the term quasistellar object (QSO), which was originally applied to these objects for their photographic appearance. The objects appear starlike because their angular diameters are less than about 1 second of arc, which is the resolution limit of ground-based optical telescopes imposed by atmospheric effects. Stars also have angular diameters much less than this, and so they too appear unresolved or pointlike on a photograph.

Discovery. Quasars were discovered in 1961 when it was noticed that very strong radio emission was coming from a localized direction in the sky that coincided with the position of a starlike object. Prior to this time, it was believed that strong radio emission originating beyond the solar system came only from the direction of certain exploding galaxies (radio galaxies) or from previously exploded stars (supernova remnants). Two techniques were used by radio astronomers to measure within a few arc-seconds the direction of the radio emission. The more common technique, called interferometry, makes use of two (or more) antennas pointing in the direction of the source. After the signals are amplified at each antenna, they are combined via cables or a microwave radio link. Another method, very long baseline interferometry (VLBI), records the signals on magnetic tapes which are then synchronously played back so that the signals can be combined. The combined signals yield sinusoidally varying output voltages called fringes. Fringes represent the beating in and out of phase of the incoming radio waves arriving at different times at the two antennas. The phase of the fringes yields an accurate determination of the

direction of arrival of the radio waves. *See* INTERFER-OMETRY; SUPERNOVA.

The second technique makes use of the motion of the Moon through the sky in its orbit around the Earth. Occasionally the Moon will pass in front of a radio source and block its emission from reaching the Earth. The exact time of such an occultation is observed by using a radio telescope and is compared with the known position of the edge of the Moon at that instant. When the source reappears from behind the Moon several minutes later, the time of reappearance is also measured and compared with the location of the Moon's edge. These two fixes on the radio source are usually sufficient to permit its position to be measured accurately. *See* OCCULTATION.

When these techniques were used to make accurate position measurements of small-angulardiameter radio sources, the coincidence with starlike objects on optical photographs led to the discovery of a new, hitherto-unsuspected class of objects in the universe, the quasars. The full significance of the discovery was not appreciated until 1963, when it was noted that the hydrogen emission lines seen in the optical spectrum of the quasar 3C 273 were shifted by about 16% to the red from their normal laboratory wavelength. This redshift of the spectral lines is characteristic of galaxies whose spectra are redshifted because of the expansion of the universe and is not characteristic of stars in the Milky Way Galaxy, whose spectra are only slightly shifted to the red or blue from the Doppler effect of their small velocities relative to the Earth and Sun.

Optical characteristics. The color of quasars is generally much bluer than that of most stars with the exception of white dwarf stars. The optical continuum spectrum of stars has a blackbody distribution, while the energy distribution of quasars is generally a power law of the form f^{α} , where f is the frequency and α is the spectral index and is usually negative. The blueness of quasars as an identifying characteristic led to the discovery that many blue starlike objects have a large redshift and are therefore quasars. The quasistellar objects discovered this way turned out to emit little or no radio radiation and to be about 20 times more numerous than the radioemitting quasistellar radio sources (QSSs). Why some should be strong radio emitters and most others not is unknown at the present time. Several orbiting xray satellites have found that most quasars also emit strongly at x-ray frequencies. Gamma rays have also been observed in many quasars. See GAMMA-RAY AS-TRONOMY; X-RAY ASTRONOMY.

Visible and x-ray emission from some quasars is absorbed by nearby dust grains and is reradiated in the far-infrared near 100-micrometer wavelengths. Unfortunately, this emission is difficult to observe since it is absorbed by the Earth's atmosphere. Several quasars have been observed in the far-infrared from space by the *Infrared Astronomical Satellite* (*IRAS*). See INFRARED ASTRONOMY.

The emission from quasars also varies with time. The shortest time scale of variability ranges from years to months at short radio wavelengths, to days at optical wavelengths, to hours at x-ray wavelengths. These different time scales suggest that the emissions from the different bands originate from different regions in the quasar. Very highly active quasars are sometimes referred to as optically violent variables (OVVs), blazars, or BL Lac's, after the prototype BL Lacertae, a well-known variable "star" that turned out to be a quasar. The optically violent variables have no or very weak emission lines in their optical spectrum.

The rapid fluctuations indicate that there are some components in quasars that have diameters less than a light-hour or of the order of $10^9 \text{ km} (10^9 \text{ mi})$, the size of the solar system. The radiation mechanism responsible for the optical and x-ray continuum emission is a combination of thermal blackbody radiation, synchrotron emission, and inverse Compton radiation. *See* COMPTON EFFECT; SYNCHROTRON RADIATION.

Emission lines characteristic of a hot ionized gas are superimposed on the continuum. The emission lines are highly Doppler-broadened, indicating internal velocities of hundreds to thousands of kilometers per second. The most common permitted lines seen are those of H, He, C IV, Si IV, and Mg II, while forbidden lines originating from less dense gaseous regions in the quasar are emitted by the ions [C III], [O III], [Ne V], and others. *See* DOPPLER EFFECT.

In addition to the emission lines, many quasars, especially those with large redshifts, show very narrow hydrogen absorption lines. In contrast to the emission-line regions, the absorbing regions have internal velocities of about 10 km/s (6 mi/s). Frequently, many absorption-line systems are found with differing redshifts. The absorption-line redshifts are typically less than the emission-line redshifts by amounts corresponding to relative velocities of several thousand kilometers per second. In some quasars, there are absorption-line redshifts substantially less than the emission-line redshift. The most likely interpretation of the latter is that the light from the quasar is passing through and being absorbed by a distant galaxy that just happens to lie along the line of sight. The more numerous other lines could be absorptions by clouds in the vicinity of the quasar. Multiple images due to gravitational lensing by the mass of an intervening galaxy have been observed for a few quasars. See GRAVITATIONAL LENS.

For the few thousand quasars measured, the emission-line redshift *z* ranges from near 0 to 6.4. For quasars having a *z* greater than about 0.5, many normally ultraviolet spectral lines are seen that are redshifted into the visible region of the spectrum. These lines, such as the Lyman α line at 121.6 nanometers—seen when z > 1.6—cannot otherwise be seen in astronomical objects observed from the surface of the Earth because of the atmospheric absorption of all radiation less than about 310 nm.

Not all quasar images are entirely point sources of light. Quasar 3C 273 (**Fig. 1**) has an elongated jet of optical emission 20 arc-seconds from the quasar but pointing radially away from it as if the gas in the jet had been ejected from the quasar. The image of

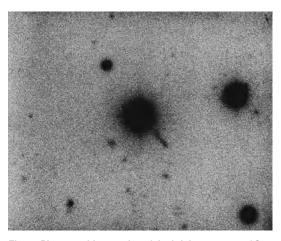


Fig. 1. Photographic negative of the brightest quasar, 3C 273. The cross-shape pattern is due to diffraction within the telescope. The jet extends toward the lower right. (*Palomar Observatory, California Institute of Technology*)

the quasar has been intentionally overexposed to reveal the ejected jet of material. Observations with the Hubble Space Telescope have shown that most quasars have faint nebulous emission surrounding the starlike image. The spectrum of this emission has been found in some cases to be an emission-line spectrum indicating the presence of heated gas extending thousands of light-years from the quasar. However, in most quasars the nebulous emission has a blackbody spectrum with superimposed absorption lines indicating the existence of a galaxylike distribution of stars surrounding the quasar. Furthermore, objects having quasar characteristics but with small redshifts are sometimes found surrounded by a faint distribution of stars. These are called N galaxies and may be similar to the Seyfert galaxies, which are spiral galaxies having a very bright quasarlike nucleus. The existence of these galaxy types and the starlike spectrum of the faint emission seen around some quasars strongly suggest that quasars are exceedingly luminous nuclei of galaxies. Thus, they are often referred to as active galactic nuclei (AGNs). See GALAXY, EX-TERNAL.

Distances of quasars. When the quasar redshift is plotted against the quasar's brightness or magnitude, as is done for galaxies, there is not a strong correlation between large z and faint magnitudes. Such a correlation is observed for galaxies and indicates that the fainter, more distant galaxies have progressively larger redshifts. This correlation, called Hubble's law, is considered to be proof that the universe is expanding. In the case of the quasars the scatter in the plot is so great that it does not prove that the redshift can be used as a distance indicator. This does not prove that the large-z quasars are not the most distant objects known, since the scatter in the diagram could be due to large differences in the intrinsic luminosities of the quasars. See HUBBLE CONSTANT; RED-SHIFT.

If the quasars exhibited the Hubble law relationship between their magnitude and redshift, then their redshifts would unambiguously show them to be the most distant known objects in the universe, ranging up to about 10^{10} light-years away. The statistics of the redshift number distribution could then be interpreted as showing a greater abundance of quasars at earlier epochs in the universe. However, the paucity of known quasars with z greater than 6 implies that they may have been absent at even earlier epochs when the density of the universe was hundreds of times greater than its present density.

A few astronomers have argued that the redshifts are not cosmological in origin and hence do not indicate the distances of quasars. One of these arguments rests on the observations that there are examples where two quasars of very different redshift are seen closer together than would be expected by chance if the two quasars were not physically related. On the other hand, the redshifts of quasistellar objects seen in nearby clusters of galaxies have been found to agree with the measured redshifts of the associated galaxies, implying that the redshifts are cosmological. Models postulating noncosmological redshifts, such as gravitational or Doppler redshifts, have met with difficulties. If the redshifts are not cosmological, then they are probably a result of some new unknown physical effect. See COSMOLOGY.

Radio characteristics. Some quasars emit a significant fraction of their radiated energy at radio frequencies ranging from about 30 MHz (wavelength, 10 m) to 300 GHz (wavelength, 1 mm). There appear to be two general categories of radio emitters. Those that emit predominantly at the lower radio frequencies (less than 1 GHz) have spectral indices α of about -0.8, showing that the flux of radio emission is rapidly decreasing with increasing frequency. Interferometers have shown that the radio emission from these types of quasistellar radio sources does not originate from the optical quasar but comes from two considerably distant regions symmetrically displaced on either side of the optical object. Such radio-frequency characteristics are nearly identical to those of radio galaxies, which are massive elliptical galaxies, suggesting that the two may be generically related.

The other, more interesting type of quasistellar radio source emits primarily at centimeter and millimeter wavelengths (1 GHz to 300 GHz) and originates in the same region as the optical quasar. The flux of emission of these sources is nearly independent of frequency ($\alpha = 0$) except that there is an abrupt drop-off in the emission spectrum at the lowfrequency end. There are a few examples of quasistellar radio sources that exhibit both sets of characteristics, suggesting that one type may evolve into another. In 1965 large variations in the flux of radio emission from the second type of quasistellar radio source were discovered, indicating that, since these types are active, they are young objects while those of the first type are probably decaying remnants. Sudden outbursts of emission are observed (Fig. 2), with enormous luminosities sometimes approaching 10³⁸ joules/s, nearly 10¹² times the luminosity of the Sun.

The nature of the origin of this energy is not known and represents one of the most important

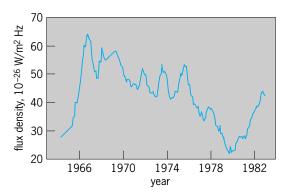


Fig. 2. Outbursts in radio emission which were observed in the nearby quasar 3C 273. Measurements were made at a frequency of 15.5 GHz (1.9 cm) with the 120-ft (37-m) radio telescope of the Haystack Observatory in Massachusetts. Sudden outbursts are evident, particularly the ones in 1966 and 1982.

unanswered questions concerning quasars. The polarization of the radio emission suggests that it is synchrotron radiation produced by electrons with velocities very near the speed of light that are spiraling in magnetic fields. The electrons can absorb their own photons at lower frequencies, producing the dropoff in their emission spectrum. From the observed frequency of this self-absorption, the magnetic field strengths can be estimated as 0.01 to 1 gauss (1 to 100 microtesla).

There have been VLBI observations of some of the active quasars. These observations show that very compact radio components with angular diameters of about 0.001 arc-second exist in quasars of the second type. Furthermore, successive observations seem to show that these compact components are expanding rapidly and separating at speeds greater than the speed of light, when cosmological distances are assumed. A number of theories have been advanced to explain the phenomena without requiring that the quasars be relatively nearby. The most likely explanation is that one of the separating radio components is a jet moving almost directly toward the Earth with a velocity close to the speed of light. At high velocities of approach, effects predicted by the special theory of relativity become important. Intervals of time in a rapidly approaching object would appear to be much shorter, giving the appearance that the object is moving much faster than it really is. This interpretation would also explain the high apparent radio luminosities. See RELATIVITY.

Theories of quasars. The many similarities of the observed characteristics of quasars with radio galaxies, Seyfert galaxies, and BL Lacertae objects strongly suggest that quasars are active nuclei of galaxies. There is good statistical evidence which shows that quasars with large redshifts are spatially much more numerous than those with small redshifts. Because high-redshift objects are very distant and emitted their radiation at an earlier epoch, quasars must have been much more common in the universe about 10^{10} years ago. Observations with the *Hubble Space Telescope* have shown that this is the same epoch when galaxies are observed to be forming. Thus it

is likely that quasars are associated with the birth of some galaxies.

More than 10^{53} J of energy are released in quasars over their approximately 10^{6} -year lifetime. Of the known energy sources, only gravitational potential energy associated with a mass about 10^{9} times the mass of the Sun can provide this energy, but it is unknown how this gravitational energy produces jets of particles that are accelerated to very near the speed of light.

Several theories have been proposed for quasars. However, the most favored interpretation is that quasars are massive black holes surrounded by rapidly spinning disks of gas in the nuclei of some galaxies. The hot gas in the disk emits the x-ray and optical continuum, a heated halo around the disk produces the emission lines, and the relativistic radio jets are ejected along the rotation axis of the spinning disk. *See* ASTRONOMICAL SPECTROSCOPY; AS-TROPHYSICS, HIGH-ENERGY; BLACK HOLE; INFRARED ASTRONOMY; NEUTRON STAR; RADIO ASTRONOMY. William Dent

Bibliography. G. Burbidge and M. Burbidge, *Quasi-Stellar Objects*, 1967; C. Hazard and S. Mitton (eds.), *Active Galactic Nuclei*, 1979; A. K. Kembhavi and J. V. Narlikar, *Quasars and Active Galactic Nuclei: An Introduction*, 1999; I. Robson, *Active Galactic Nuclei*, 1995; H. L. Shipman, *Black Holes, Quasars, and the Universe*, 1980; D. Weedman, *Quasar Astronomy*, 1986, paperback 1988.

Quasiatom

A structure in which the nuclei of two atoms approach each other closely and their electrons are arranged in atomic orbitals characteristic of a single atom of atomic number equal to the sum of the nuclear charges. Quasiatoms can be formed for short times in atom-atom and ion-atom collisions when the nuclei are much closer than the mean orbital radius of the innermost K-shell electrons. The electrons are then bound in the electric field of both nuclear charges Z_1 and Z_2 , which resembles the spherically symmetric $1/r^2$ Coulomb field of a single united atom having charge $Z_{ua} = Z_1 + Z_2$. See ATOMIC STRUCTURE AND SPECTRA.

Formation. The formation of transient quasiatoms requires that the electrons of the target atom and projectile ion smoothly rearrange their orbits in response to the changing electric field as the nuclei approach one another. This condition is fulfilled when the orbital velocity of the atomic electrons greatly exceeds the relative velocity of the colliding nuclei. Experimentally, this ratio of electron velocity to nuclear velocity has been achieved in collisions of heavy ions at energies near the nuclear Coulomb barrier. For example, collisions have been studied of 1.4-GeV uranium ions (Z = 92) with uranium atoms, in which the K-shell electron velocity v_e is approximately 0.6c, and the nuclear velocity v_{nuc} is approximately 0.1*c*, where c is the velocity of light. The orbits of at least the inner-shell electrons adjust nearly adiabatically

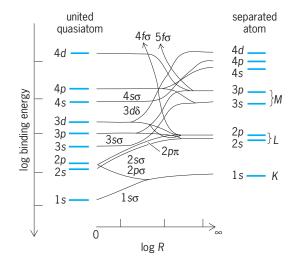


Fig. 1. Binding energies of quasimolecular states evolving from the separated-atom limit to the united-atom limit, as a function of the internuclear separation R.

to the changing nuclear charge configuration during the collision.

Initially, at very large internuclear distances, the electrons are bound to the incident projectile nucleus or the target nucleus, with wave functions characteristic of the separated atoms. As the approaching nuclei reach distances comparable to the size of the atomic electron clouds, the electrons respond to the two-centered Coulomb potential of both nuclei. At each moment the electron wave functions resemble those of a diatomic molecule with an interatomic spacing equal to the nuclear separation R. As a function of time, the electrons evolve through a continuous series of quasimolecular states, such as $1s\sigma$ or $2p\pi$ (Fig. 1), characterized by R(t). At the smallest internuclear distances of R = 20 femtometers $(1 \text{ fm} = 10^{-15} \text{ m})$ for 1.4-GeV uranium-uranium collisions, the nuclei are well within the orbit of the innermost K-shell, and for a short time (10^{-20} s) the electrons are governed not by the individual nuclei but by a single charge distribution with an effective radius of R/2. As the nuclei recede, the electrons pass again through the quasimolecular states to the separated-atom limit. See MOLECULAR STRUC-TURE AND SPECTRA.

Electron excitations. The quasimolecular picture presented above is somewhat idealized, because not all electrons can adjust adiabatically during the collision, especially those that are less tightly bound and have lower orbital velocities. These can be excited to higher quasimolecular orbitals or ejected from the collision system by the time variation of the Coulomb field. Certain electron excitations are enhanced at particular internuclear separations by a process called electron promotion, in which the principal quantum number of an electron is changed during the collision from its value in the separated atom. This occurs when quasimolecular orbitals associated with different principal quantum numbers in the separated-atom limit have nearly equal binding energies at a particular internuclear separation,

increasing the probability of induced excitations between these orbitals. In this process, an electron does not always return to its original state in the separated atom after the collision.

The feature of the collision that is responsible for the electron promotion, excitation, and ionization processes is the rapid variation of the two-centered Coulomb field during the approach. Nonadiabatic adjustment of the electrons to the radial component of the electric field as the internuclear distance changes produces a change in their principal quantum number. The rapid rotation of the axis between the nuclei in their plane of scattering as they pass one another can also couple angular momentum to the electrons, changing their orbital-angular-momentum quantum numbers. These excitation mechanisms, unique to the quasimolecular description of ion-atom collisions, explain several experimentally observed phenomena, such as very large excitation and ionization cross sections and the resonant production of K, L, and M x-rays in the separated atoms, which cannot be understood in terms of pure Coulomb excitation during the collisions. See X-RAYS.

The first direct experimental evidence for quasimolecule formation came from x-ray observations of medium-heavy collision systems, for example, nickel on nickel. In addition to the characteristic line radiation associated with transitions in the collisionally excited separated atoms, a continuum of x-rays with energies up to the united-atom limit (K x-rays characteristic of nuclear charge $Z_1 + Z_2$) was observed, consistent with the changing quasimolecular binding energies during the collision. A Doppler-shift analysis of this radiation showed that the x-rays were emitted from the center-of-mass system (that is, from the quasimolecule), and detected angular anisotropy was in agreement with the predicted effect of the rotation of the internuclear axis during the collision.

Superheavy quasiatoms. In addition to describing conventional atomic collision phenomena, quasimolecular collisions offer the prospect of forming short-lived quasiatoms at the smallest internuclear separations. This creates a new realm of atomic physics involving very high Z_{ua} atoms and making tests of quantum electrodynamics possible in extremely strong electromagnetic fields, far beyond those available in the heaviest artificial elements. Fundamentally new phenomena are expected to occur in atoms with $Z\alpha \ge 1$ ($\alpha = 1/137$, the finestructure constant). At Z = 137, the Dirac equation for a pointlike nucleus predicts a binding energy of the 1s electron equal to its rest-mass energy, $mc^2 =$ 511 keV (m = electron rest mass); and for Z above 137, the Dirac equation breaks down unless the finite size of the nucleus is taken into account. For Z >173, the 1s electron-binding energy is predicted to exceed $2mc^2 = 1.02$ MeV, which opens the interesting possibility of observing a phase change of quantum electrodynamics from a neutral vacuum state to a charged vacuum. In comparison, the heaviest stable atoms that have been studied (such as fermium with Z = 100) exhibit binding energies of only 140 keV. See RELATIVISTIC QUANTUM THEORY.

Wave functions. Theoretical studies of superheavy quasiatoms indicate that a very different pattern of atomic energy levels is expected in very high-Z systems (Fig. 2), in which the electron binding energies and wave functions are dominated by relativistic effects, unlike those encountered in atoms of lower Z. In light, nonrelativistic atoms, the atomic 1s electron binding energy scales as Z^2 , and the mean radius of the 1s orbit scales as 1/Z. As the atomic number increases past Z = 137, the binding energy increases more quickly than Z^2 , and the electron wave function becomes more tightly localized about the nucleus as relativistic effects become important. At the highest attainable atomic charges, such as $Z_{ua} = 188$ in collisions of uranium with curium, the 1s electron binding energy is almost three times larger than the nonrelativistic prediction, and the wave function is almost ten times more compressed about the nucleus than expected nonrelativistically.

The dominance of relativistic effects in the superheavy quasiatom drastically alters the character of the most tightly bound electron orbitals. For example, the $2p_{1/2}$ orbital, which is nearly degenerate with the $2s_{1/2}$ level in low-Z atoms, is more tightly bound than the $2s_{1/2}$ state by several hundred kiloelectronvolts. Furthermore, the fine-structure splitting between the $2p_{1/2}$ and $2p_{3/2}$ levels is on the order of mc^2 (= 511 keV), comparable to their total binding energies. These effects and the large density of the $2p_{1/2}$ state at the origin are understood in terms of the relativistic spin-orbit coupling, which mixes l = 0and l = 1 angular momentum states in the $2p_{1/2}$ wave function and gives it a close resemblance to the $1s_{1/2}$ state. On the other hand, the $2s_{1/2}$ binding is smaller because of the radial node in its wave function, and the $2p_{3/2}$ level because it is an admixture of l = 1and l = 2 states, which have low probability density at small radii where the Coulomb field is largest. See FINE STRUCTURE (SPECTRAL LINES); RELATIVITY.

Spontaneous decay of the vacuum. An interesting effect is associated with quasiatoms with Z > 173, in which the 1s binding energy is more than twice the electron rest mass, $E_{1s} > 2mc^2$. If a vacancy exists in this orbital, it is energetically favorable to create an electron-positron pair with the electron bound in this state. The positron would be repelled from the nucleus with kinetic energy equal to $E_{e^+} = |E_{1s}| 2mc^2$. In the Dirac hole picture, in which the vacuum consists of a negative energy continuum ($E < -mc^2$) filled with electrons, the 1s level is said to fall into the negative-energy Dirac sea as Z increases above the critical value, $Z_{cr} = 173$ (Fig. 2). A 1s hole (vacancy) becomes embedded in the negative continuum as an unstable resonance state that decays in a time of $\sim 10^{-19}$ s to a bound electron and a spontaneously emitted monoenergetic positron.

The quantum electrodynamic vacuum in the presence of a bare supercritical nuclear charge is therefore unstable and decays to a fundamentally new charged vacuum, which consists of the nucleus with two 1s electrons (from the two spin orientations). At higher values of Z_{ua} , as additional quasiatomic levels enter the negative continuum, the charge of

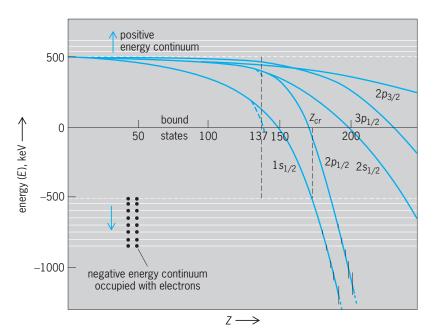


Fig. 2. Calculated quasiatomic binding energies as a function of nuclear charge Z. Broken curves represent levels in the field of a point charge; solid curves take into account finite size of nucleus. (After B. Mueller, Positron creation in superheavy quasi-molecules, Annu. Rev. Nucl. Sci., 26:351–383, 1976)

the quantum electrodynamic vacuum increases accordingly. If detected, spontaneous positron emission would represent the first observation of a phase transition in a gauge field theory. *See* ANTIMAT-TER; ELECTRON-POSITRON PAIR PRODUCTION; GAUGE THEORY; PHASE TRANSITIONS; POSITRON; QUANTUM ELECTRODYNAMICS; SUPERCRITICAL FIELDS.

Supercritical quasiatomic collisions. It was suggested in 1969 that quasiatomic collisions with Z_{ua} > 178 could provide the strong fields necessary to observe the spontaneous decay of the quantum electrodynamic vacuum experimentally. Figure 3 traces the time development of the internuclear distance and the inner-shell binding energies in a supercritical uranium-uranium collision at 1.4 GeV ($Z_{ua} = 184$). As the ions approach the united-atom limit (at time t = 0), the total energy of the $1s_{1/2}$ orbital falls rapidly until, for internuclear separations less than $R_{\min} =$ 27 fm, the 1s state drops into the continuum. On the incoming trajectory, electrons are excited to higher states or ejected to the positive energy continuum as delta electrons, as a result of the interplay between the increased binding and relativistic collapse of the electron wave functions mentioned above. Electron excitation or emission creates vacancies in the inner shells, which, if present in the $1s_{1/2}$ orbital, may decay spontaneously during the short time of supercritical binding ($t_{
m cr} \sim 2 \, imes \, 10^{-21}$ s in elastic collisions). Electron excitation and quasimolecular x-ray emission occur throughout the collision. See DELTA **ELECTRONS**

In addition to their creation by spontaneous decay of vacancies in the $1s_{1/2}$ state, positrons are produced by the so-called direct ionization of the negative-energy-continuum electrons to positive-continuum states, or by the so-called induced excitation of

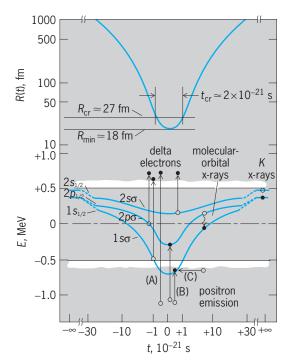


Fig. 3. Internuclear separation *R* and quasimolecular orbital binding energies *E* as a function of time *t* for 1.4-GeV uranium-uranium (²³⁸U-²³⁸U, *Z*_{ua} = 184) collisions. Delta-electron, x-ray, and (A) direct, (B) induced, and (C) spontaneous positron-emission processes are indicated. (After J. Eichler, I. V. Hertels, and N. Stalterfoht, eds., Electronic and Atomic Collisions, Elsevier Science, 1984)

negative-continuum electrons to quasiatomic bound states. This leaves holes in the Dirac sea, and this absence of negative-energy electrons is observed as positive-energy positrons. As opposed to spontaneous positron emission, an inherently static phenomenon occurring as a result of the static field strength, the production of these so-called dynamic positrons is a direct consequence of the rapidly timevarying Coulomb fields present during the quasiatomic collision.

Experimental verification. This picture of superheavy quasiatomic collisions has, to a large extent, been verified experimentally by measurements performed at the heavy-ion UNILAC accelerator in Darmstadt, Germany. In general, the salient feature of these collisions is that the sharply increased electron binding energies and the relativistic collapse of the electron wave functions about the origin with increasing effective nuclear charge combine to dominate the atomic excitation mechanisms. For example, a thousandfold enhancement was found in the K-shell ionization probability in uranium-lead collisions, as against nonrelativistic calculations.

The emission of very high-energy delta electrons, as shown in **Fig.** 4*a* for lead-lead collisions, also provides evidence for relativistic collapse of wave functions and increased binding energy. This follows from kinematic arguments that the maximum energy transferable to a lead K-shell electron from the incident nucleus is only ~200 keV, unless the electron becomes much more tightly bound. The measured energies of up to ~1 MeV, however, imply that the ejected electrons must have been bound with energies of up to $E_{1s} \sim 900$ keV, and localized in the quasiatom with mean radii of less than 100 fm, in agreement with the predictions of quasiatomic theory.

The production of dynamic positrons has also been studied over a large range of values of Z_{ua} . The total probability for emitting positrons in excess of those from nuclear pair-production backgrounds

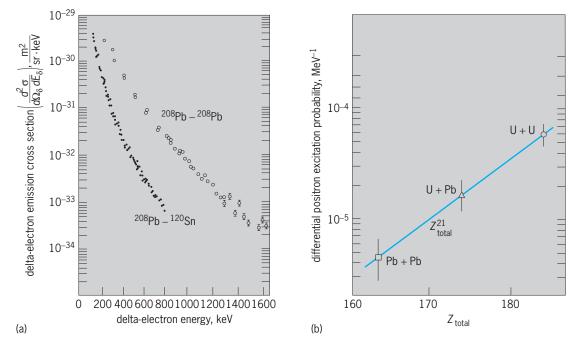


Fig. 4. Electron and positron emission in superheavy quasiatomic collisions. (a) Delta-electron energy distributions from lead-lead ($^{208}Pb-^{208}Pb, Z_{ua} = 164$) and lead-tin ($^{208}Pb-^{120}Sn, Z_{ua} = 132$) collisions at bombarding energy of 4.7 MeV/u. Presence of high-energy delta electrons reveals onset of relativistic effects in the higher- Z_{ua} quasiatom (*after S. Datz, ed., Electronic and Atomic Collisions, Elsevier Science, 1982*). (b) Total positron creation probability in quasiatomic collisions as a function of combined nuclear charge, $Z_{total} = Z_1 + Z_2$.

rises steeply with an unprecedented $(Z_1 + Z_2)^{20}$ power-law dependence on the combined nuclear charge (Fig. 4b), reflecting relativistic effects. The measured dependence of dynamic positron production on internuclear distance and on the incident beam energy, as well as its Z_{ua} dependence, are well reproduced by the theoretical calculations.

An important consequence of the rapid variation of the Coulomb field in quasiatomic collisions is that the distributions of emitted positron, electron, and x-ray energies are all dominated by dynamic effects and are therefore continuous and rather broad (typically \sim 1 MeV wide). This limits the amount of detailed spectroscopic information that can be experimentally extracted concerning specific quasiatomic energy levels.

Search for spontaneous positron emission. The short time scale of the collision also leads to several consequences that compound the difficulty of detecting spontaneously emitted positrons from quantum electrodynamic vacuum decay: (1) The short time during which the 1s state is supercritically bound $(2 \times 10^{-21} \text{ s})$ compared to the longer decay time ($\sim 10^{-19} \text{ s}$) implies that very few of the vacancies decay spontaneously. (2) By the Heisenberg uncertainty principle ($\Delta E \cdot \Delta t \ge \hbar/2$, where ΔE represents the energy spread, Δt is the time during which emission takes place, and \hbar is Planck's constant divided by 2π), the emitted positron energies would be broadened by 600 keV. (3) The amplitude for spontaneous emission adds coherently with dynamic positron pro-

duction. For these reasons, no observable signature of spontaneous positron emission is anticipated, and only continuous positron distributions are expected to be found. *See* UNCERTAINTY PRINCIPLE.

It came as a surprise, therefore, when a narrow peak was first discovered in the energy distribution of positrons emitted from uranium-curium collisions $(Z_{ua} = 188)$. Perhaps coincidentally, the energy of the positron peak ($E_{e^+} = 316$ keV) corresponds with the binding energy expected for the 1s state in the $Z_{ua} = 188$ quasiatom, and its width (~80 keV) implies a long-lived source moving with the center-ofmass velocity of the collision system. These observations encouraged speculation that the peak might be the signature of spontaneous positron emission. Indeed, it was demonstrated that the peak energy, intensity, width, and dependence of the production cross section on the incident beam energy could all be incorporated into the theory of spontaneous positron emission if, in a small fraction (10^{-3}) of the collisions, a dinuclear complex were formed with a lifetime of more than 10^{-19} s, an order of magnitude above the typical collision time. See NUCLEAR MOLECULE.

This attractive explanation was brought into question by measurements that probed the theoretically predicted $\sim Z_{ua}^{20}$ dependence of the emitted spontaneous positron energy on combined nuclear charge. Several supercritical collision systems were systematically studied, ranging from uranium-curium ($Z_{ua} = 188$) where the predicted spontaneous

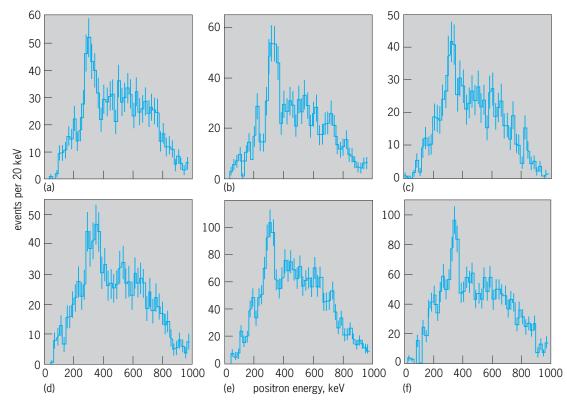


Fig. 5. Energy distributions of positrons emitted from collision systems. The kinematic constraints chosen prominently exhibit the narrow peak structures between 310 and 380 keV. (a) Uranium-curium, at bombarding energy of 6.07 MeV/u, $Z_{ua} = 188$. (b) Thorium-thorium at 6.02 MeV/u, $Z_{ua} = 186$. (c) Uranium-uranium at 5.8 MeV/u, $Z_{ua} = 184$. (d) Thorium-uranium at 5.8 MeV/u, $Z_{ua} = 182$. (e) Thorium-thorium at 5.75 MeV/u, $Z_{ua} = 180$. (f) Thorium-tantalum at 5.78 MeV/u, $Z_{ua} = 163$. (d) (for M. B. Gilbody et al., eds., Electronic and Atomic Collisions, Elsevier Science, 1988)

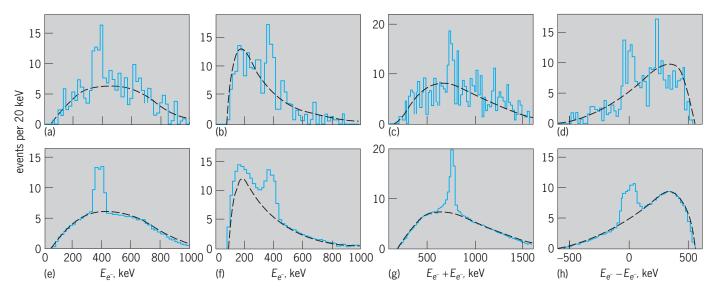


Fig. 6. Coincident positrons and electrons from 1.4-GeV uranium-thorium collisions. Spectra are shown of (a) energy E_{e^+} , (b) coincident electron energy E_{e^-} , (c) sum energy, $E_{e^+} + E_{e^-}$, and (d) difference energy, $E_{e^+} - E_{e^-}$ in coincident positron and electron events. Corresponding parts (e-h) present Monte Carlo simulation of expected yields, assuming production and decay of a 1.8 MeV/c² neutral object. (After W. Greiner, ed., Physics of Strong Fields, Plenum Publishing, 1987)

positron energy is $E_{\rm sp} = 300$ keV, and thoriumthorium ($Z_{ua} = 180$) where $E_{sp} = 80$ keV, to the subcritical thorium-tantalum system ($Z_{ua} = 163$), where the nuclear charge is clearly below the supercritical charge threshold and spontaneous positron emission is not expected. Each system studied gave narrow positron peaks between 310 and 380 keV with similar widths and cross sections (Fig. 5). No competing gamma-ray or internal conversion electron lines were observed that could explain the data in terms of nuclear-conversion processes, and no instrumental or background processes were found that could produce these structures. All attempts to explain the near independence of the peak energies on Z_{ua} or the presence of a line in the subcritical thoriumtantalum system within the context of spontaneous positron emission have failed.

Correlated electron and positron lines. The similar energies of the positron peaks imply a common origin not involving conventional nuclear or quantum electrodynamic processes. The narrowness of the lines suggest a two-body final state. An obvious possibility pursued experimentally was the two-body decay of a previously undetected neutral object into a positron-electron pair. The signature for this process would be the coincident detection of nearly monoenergetic positrons and electrons, correlated in such a way that the Doppler shifts of their energies cancel, indicating back-to-back emission in the rest frame of their source. *See* DOPPLER EFFECT.

Figure 6 shows the results of a first measurement of coincident positrons and electrons from 1.4-GeV uranium-thorium collisions. An electron peak at $E_{e^-} = 375 \text{ keV}$ (Fig. 6*b*) was detected in coincidence with the positron peak at $E_{e^+} = 380 \text{ keV}$ (Fig. 6*a*). The peak in the distribution of the sum of the positron and electron energies at 750 keV (Fig. 6*c*) is narrower than the individual lines. Each of these spectra, as well as the distribution of the difference of

positron and electron energies (Fig. 6d), is consistent with that expected from Monte Carlo simulations (Figs. 6e-b) of the production and decay of a neutral object of mass 1.8 MeV/ c^2 in a small fraction of the collisions. Additional measurements with improved electron energy resolution revealed two additional sets of correlated positron-electron structures with energy peaks at $(E_{e^+} + E_{e^-}) = 817$ keV and 612 keV. Explanations for these structures involving instrumental backgrounds, external or internal pair production, the uncorrelated emission of separate positron and electron lines, and multibody decay are all unable to describe several features of the experimental data, including the narrow sum-energy peak width and the absence of peak intensity in adjacent kinematic regions. Moreover, a kinematic analysis of the widths of the various correlated peaks suggests that the positron and electron are emitted with a 180° back-to-back angular correlation in the rest frame of their source.

These results represent a new source of positronelectron pair production in heavy-ion collisions, and appear to reflect the decay of at least three previously undetected neutral objects with masses of ~ 1.63 , 1.77, and 1.84 MeV/ c^2 , which are created with very low momenta in superheavy nuclear collisions. Theoretical attempts to explain these data in terms of new elementary particles have met with difficulties due to the negative results of experimental searches for new particles in high-precision quantum electrodynamic measurements, from decays of J/psi and upsilon mesons, in competition with nuclear transitions; from rare pion, kaon, and muon decays; and from electron brehmsstrahlung in beam dumps. Further experiments with superheavy quasiatoms, and attempts to form these objects resonantly in positron-electron scattering, should shed light on the nature of these anomalous correlated positronelectron pairs. Thomas E. Cowan

Bibliography. D. A. Bromley (ed.), *Treatise on Heavy-Ion Science*, vol. 5: *High Energy Atomic Physics*, 1985; W. Greiner (ed.), *Physics of Strong Fields*, 1987; W. Greiner, B. Muller, and J. Rafelski, *Quantum Electrodynamics of Strong Fields*, 1985.

Quasicrystal

A solid with conventional crystalline properties but exhibiting a point-group symmetry inconsistent with translational periodicity. Like crystals, quasicrystals display discrete diffraction patterns, crystallize into polyhedral forms, and have long-range orientational order, all of which indicate that their structure is not random. But the unusual symmetry and the finding that the discrete diffraction pattern does not fall on a reciprocal periodic lattice suggest a solid that is quasiperiodic. Their discovery in 1982 contradicted a long-held belief that all crystals would be periodic arrangements of atoms or molecules.

Since the eighteenth century, the idea that crystals are periodic arrangements of atoms or molecules gradually moved from a postulate to rationalize the limited number of different rotational symmetries seen in crystals; to a concept experimentally confirmed by the diffraction by crystals of x-rays, neutrons, and electrons; to a possible definition of a crystal. The conjecture that the lowest-energy arrangements of atoms or molecules would always be periodic was widely believed, although it was unprovable and is now known to be false. Early definitions of crystals invoked their polyhedral external form rather than their regular internal arrangements. It is easily shown that in two and three dimensions the possible rotations that superimpose an infinitely repeating periodic structure on itself are limited to angles that are $360^{\circ}/n$, where *n* can be only 1, 2, 3, 4, or 6. Various combinations of these rotations lead to only 32 point groups in three dimensions, and 230 space groups which are combinations of the 14 Bravais lattices that describe the periodic translations with the allowed rotations. Until the 1980s, all known crystals could be classified according to this limited set of symmetries allowed by periodicity. Periodic structures diffract only at discrete angles (Bragg's law) that can be described by a reciprocal lattice, in which the diffraction intensities fall on lattice points that, like all lattices, are by definition periodic, and which has a symmetry closely related to that of the structure. See CRYSTAL STRUCTURE; CRYSTALLOGRAPHY; X-RAY CRYSTALLOG-RAPHY: X-RAY DIFFRACTION.

Icosahedral quasicrystals were discovered in 1982 during a study of rapid solidification of molten alloys of aluminum with one or more transition elements, such as manganese, iron, and chromium. Since then, many different alloys of two or more metallic elements have led to quasicrystals with a variety of symmetries and structures. Quasicrystals are now made in a number of ways, including several methods of rapid solidification, such as powder and melt-spinning, solid-state transformations using thinfilm techniques, and annealing of foils after heavy

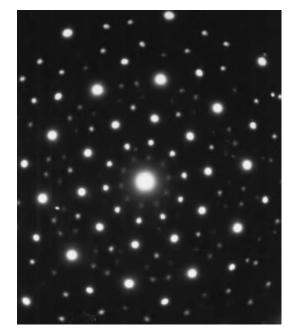


Fig. 1. Diffraction pattern, taken in an electron microscope along the fivefold axis of a rapidly solidified aluminummanganese alloy, that led to the discovery of icosahedral quasicrystals.

radiation damage and ion implantation. *See* ALLOY; ALLOY STRUCTURES; CRYSTAL GROWTH; ION IMPLAN-TATION; RADIATION DAMAGE TO MATERIALS.

The first quasicrystals were identified in the electron microscope, and their symmetry was established by electron diffraction (**Fig. 1**). There has been a lively debate about whether these early imperfect aperiodic structures would ever evolve toward a truly quasiperiodic structure, or whether the randomness was an intrinsic part of such structures. Because the perfection of the stable aluminum-copperiron quasicrystals approaches that of crystals, it is now believed that these structures can come as close to the ideal of quasiperiodicity as the skill of the crystal grower permits. **Figure 2** shows the external

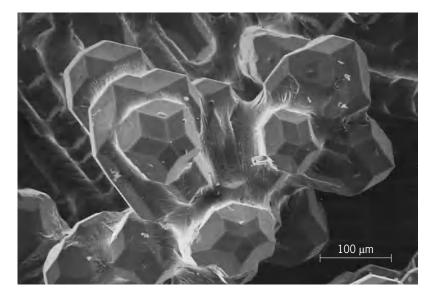


Fig. 2. Quasicrystals of an alloy of aluminum, copper, and iron, displaying an external form consistent with their icosahedral symmetry.

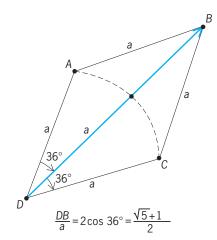


Fig. 3. Simple trigonometric construction showing that distances to successive spots in the diffraction pattern are not in simple rational ratios and therefore cannot be part of a periodic reciprocal lattice.

polyhedral form of an icosahedral aluminum-copperiron alloy. *See* ELECTRON DIFFRACTION; ELECTRON MI-CROSCOPE.

Crystallography. The diffraction patterns of quasicrystals violate several predictions resulting from periodicity. Quasicrystals have been found in which the quantity n is 5, 8, 10, and 12. In addition, most quasicrystals exhibit icosahedral symmetry in which there are six intersecting fivefold rotation axes. Furthermore, in the electron diffraction pattern (Fig. 1) the diffraction spots do not fall on a (periodic) lat-

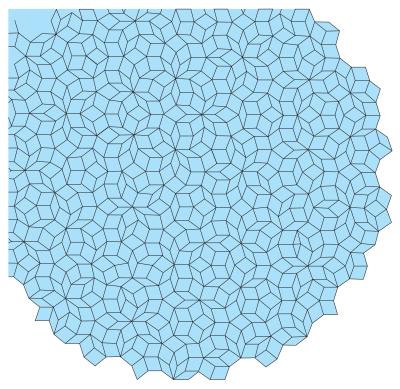


Fig. 4. Penrose tiling generated by projection of the facets of five-dimensional hypercubes in the neighborhood of a two-dimensional cut on an irrational plane. The corners of this quasiperiodic tiling display a tenfold diffraction pattern.

tice but on what has been called a quasilattice. The aperiodicity can be seen in Fig. 1 and in the construction in **Fig. 3** by noting that distances from the center to successive spots along a row are in the ratio $2 \cos 36^\circ$, which happens to be the algebraic number known as the golden mean, given in Eq. (1).

$$\frac{1+\sqrt{5}}{2} = 1.618034\dots$$
 (1)

Because this is not a rational number, the origin and two successive spots on a line cannot belong to a lattice. To locate all the spots in this two-dimensional section, a minimum of four basis vectors is required, rather than the two needed for a two-dimensional lattice. Instead of the usual three for a reciprocal lattice, six basis vectors are needed for indexing the diffraction pattern of this icosahedral alloy. If the diffraction patterns of a structure can be indexed with a finite basis, that is, each spot can be located as the sum of a finite number of vectors, then these patterns fall on a quasilattice and the structure is quasiperiodic. A quasilattice is not periodic, but has the property that any local arrangement is repeated infinitely many times in the infinite domain. A Fibonacci sequence of short and long lengths and corners of Penrose's aperiodic tiling of the plane with tenfold symmetry (Fig. 4) are one- and two-dimensional examples of quasilattices in which the golden mean plays a role. Quasilattices are quasiperiodic and lead to diffraction patterns whose peaks do not fall on the points of a lattice. See COMBINATORIAL THEORY; POLYGON

Quasiperiodicity. Strictly periodic structures give sharp diffraction. Deviations from strict periodicity commonly give rise to diffuse scattering. It has long been known and is easily shown theoretically, although not widely appreciated, that there are aperiodic structures that give rise to discrete diffraction patterns with no diffuse scattering. These are the structures whose densities can be described by almost periodic functions. Quasiperiodic functions are special cases of almost periodic functions. *See* FOURIER SERIES AND TRANSFORMS.

Among the interesting properties of quasiperiodic functions that are close to those of periodic functions is the existence, for any value of $\epsilon > 0$, of an infinite set of translation numbers *T*, shown in Eq. (2), such

$$|f(t+T) - f(t)| < \epsilon \tag{2}$$

that inequality (2) is satisfied for all *t*. For periodic functions, ϵ can be set equal to zero, and every *T* is then a multiple of the periodicity, that is, a one-dimensional lattice. For the quasiperiodic functions, the translation numbers *T* do not form a lattice.

An important concept in the mathematics of quasiperiodic functions is that such a function can be described as a planar cut on an irrational plane of a higher-dimensional periodic function. A periodic crystal cut on an irrational plane gives a two-dimensional quasiperiodic structure, which preserves some of the symmetries in the higherdimensional crystal. A rational cut of a cubic structure yields a periodic structure, which, depending on the orientation, could preserve the two-, three-, or fourfold axes. Since an *N*-dimensional isometric periodic structure always contains an (N + 1)-fold axis, quasiperiodic structures can have any integral value of *N*. The fivefold axes first appear in four-dimensional periodic structures, and the six intersecting fivefold axes needed for icosahedral symmetry first appear in six dimensions. Only irrational cuts of the higher-dimensional space can preserve the symmetries that fall outside those permitted for periodic crystals.

The relationship between a quasiperiodic function and a higher-dimensional periodic function can be easily demonstrated by considering the quasiperiodic function with two incommensurate lengths, 1 and $\sqrt{2}$, given in Eq. (3). The related two-

$$f(x) = \cos x + \cos \sqrt{2} x \tag{3}$$

dimensional function is given in Eq. (4). Along the

$$F(x, y) = \cos x + \cos y \tag{4}$$

line of Eq. (5) in the *xy* plane, the quasiperiodic function is recovered as shown by Eq. (6).

$$y = \sqrt{2} x \tag{5}$$

$$f(x) = F(x, \sqrt{2}x) \tag{6}$$

Structure determination. Although many methods of crystal structure determination depend heavily on periodicity, those that do not have been adapted for the structure determination of quasicrystals. The methods for analyzing liquid structures have also

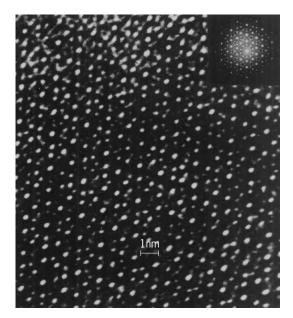


Fig. 5. High-resolution electron micrograph of a thin foil of an icosahedral (aluminum,silicon)-manganese quasicrystal seen in transmission along its fivefold axis. Inset shows electron diffraction pattern. (From R. Portier et al., High resolution electron microscopy of the icosahedral quasiperiodic structure in Al-Mn system, J. Microsc. Spectrosc. Electron., 10(2):107–116, 1985)

been used but have not been adapted to depict the order that is present in quasicrystals.

These aperiodic structures are best described as irrational cuts of periodic higher-dimensional structures that can be determined by standard crystallographic means from diffraction data. The structures examined so far all bear striking similarities to the complicated periodic structures found in intermetallic compounds, often containing hundreds of atoms per unit cell. These unit cells often contain atoms that pack into large icosahedral groupings common for the close packing of spheres of different sizes, but when these groupings pack periodically the fivefold axes are necessarily lost for the crystal as a whole. The quasicrystalline packings preserve these rotational symmetries in the diffraction at the expense of periodicity.

With modern high-resolution electron microscopy, quasicrystals can be seen in transmission (and thus in projection of specimens hundreds of atoms thick) with a resolution that is close to that of the size of atoms. **Figure 5** is an icosahedral example seen along one of the fivefold axes. Such pictures give information about the structure and clearly show the lack of periodicity.

Defects in quasicrystals. Defects can be defined only when the ideal defect-free structure is agreed upon. Not unexpectedly, the study of defects and their role in property determination has given rise to considerable debate, but the array of defects that are emerging is far more varied than those in ordinary periodic crystals. *See* CRYSTAL DEFECTS. John W. Cahn; Dan Shechtman

Bibliography. C. H. Janot, *Quasicrystals: A Primer*, 2d ed., 1997; M. V. Jaric (ed.), *Introduction to Quasicrystals*, 1988; K. H. Kuo and S. Takeuchi (eds.), *Quasicrystals and Imperfectly Ordered Crystals*, 1994; P. J. Steinhardt and S. Ostlund (eds.), *The Physics of Quasicrystals*, 1987; P. J. Steinhardt and D. Di Vincenzo (eds.), *Quasicrystals: The State of the Art*, 1991.

Quasielastic light scattering

Small frequency shifts or broadening from the frequency of the incident radiation in the light scattered from a liquid, gas, or solid. The term quasielastic arises from the fact that the frequency changes are usually so small that, without instrumentation specifically designed for their detection, they would not be observed and the scattering process would appear to occur with no frequency changes at all, that is, elastically. The technique is used by chemists, biologists, and physicists to study the dynamics of molecules in fluids, mainly liquids and liquid solutions. It is often identified by a variety of other names, the most common of which is dynamic light scattering (DLS).

Several distinct experimental techniques are grouped under the heading of quasielastic light scattering (QELS). Photon correlation spectroscopy (PCS) is the technique most often used to study such systems as macromolecules in solution, colloids, and critical phenomena where the molecular motions to be studied are rather slow. This technique, also known as intensity fluctuation spectroscopy and, less frequently, optical mixing spectroscopy, is used to measure the dynamical constants of processes with relaxation time scales slower than about 10^{-6} s. For faster processes, dynamical constants are obtained by utilizing techniques known as filter methods, which obtain direct measurements of the frequency changes of the scattered light by utilizing a monochromator or filter much as in Raman spectroscopy. *See* RAMAN EFFECT; SCATTERING OF ELEC-TROMAGNETIC RADIATION.

Static light scattering. If light is scattered by a collection of scatterers, the scattered intensity at a point far from the scattering volume is the result of interference between the wavelets scattered from each of the scatterers and, consequently, will depend on the relative positions and orientations of the scatterers, the scattering angle θ , the refractive index *n* of the medium, and the wavelength λ of the light used. The structure of scatterers in solution whose size is comparable to the reciprocal of the scattering vector length, q^{-1} , where $q \equiv [(4\pi n/\lambda) \sin(\theta/2)]$, may be studied by this technique. It is variously known as static light scattering, total-intensity scattering, integrated-intensity light scattering, Rayleigh scattering, or, in the older literature, simply light scattering. It was, in fact, developed in the 1940s and 1950s to measure equilibrium properties of polymers both in solution and in bulk. Molecular weights, radii of gyration, solution virial coefficients, molecular optical anisotropies, and sizes and structure of heterogeneities in bulk polymers are routinely obtained from this type of experiment. Static light scattering is a relatively mature field, although continued improvements in instrumentation (mainly by the use of lasers and associated techniques) are steadily increasing its reliability and range of application.

Both static and quasielastic light-scattering experiments may be performed with the use of polarizers to select the polarizations of both the incident and the scattered beams. The plane containing the incident and scattered beams is called the scattering plane. If an experiment is performed with polarizers selecting both the incident and final polarizations perpendicular to the scattering plane, the scattering is called polarized scattering. If the incident polarization is perpendicular to the scattering plane and the scattered polarization lies in that plane, the scattering is called depolarized scattering. Usually the intensity associated with the polarized scattering is much larger than that associated with the depolarized scattering. The depolarized scattering from relatively small objects is zero unless the scatterer is optically nonspherical.

Photon correlation spectroscopy. The average intensity of light scattered from a system at a given scattering angle depends, as stated above, on the relative positions and orientations of the scatterers. However, molecules are constantly in motion due to thermal forces, and are constantly translating, rotating and, for some molecules, undergoing internal rearrange-

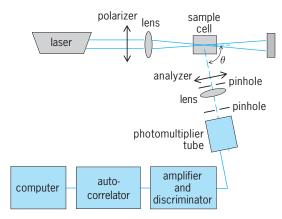


Fig. 1. Schematic diagram of a photon correlation spectroscopy apparatus.

ments. Because of these thermal fluctuations, the scattered light intensity will also fluctuate. The intensity will fluctuate on the same time scale as the molecular motion since they are proportional to each other.

Figure 1 shows a schematic diagram of a typical photon correlation apparatus. Light from a laser source traverses a polarizer to ensure a given polarization. It is then focused on a small volume of the sample cell. Light from the scattering volume at scattering angle θ is passed through an analyzer to select the polarization of the scattered light, and then through pinholes and lenses to the photomultiplier tube. The output of the photomultiplier is amplified, discriminated, and sent to a photon counter and then to a hard-wired computer called an autocorrelator, which computes the time autocorrelation function of the photocounts. The autocorrelator output is then sent to a computer for further data analysis.

The scattered light intensity as a function of time will resemble a noise signal. In order to facilitate interpretation of experimental data in terms of molecular motions, the time correlation function of the scattered intensity is usually computed by the autocorrelator. The autocorrelation function obtained in one of these experiments is often a single exponential decay, $C(t) = \exp(-t/\tau_r)$, where τ_r is the relaxation time.

The lower limit on decay times τ_r that can be measured by photon correlation spectroscopy is about a microsecond, although with special variations of the technique somewhat faster decay times may be measured. For times faster than this, filter experiments are usually performed by using a Fabry-Perot interferometer.

Fabry-Perot interferometry. Light scattered from scatterers which are moving exhibits Doppler shifts or broadening due to the motion. Thus, an initially monochromatic beam of light from a laser will be frequency-broadened by scattering from a liquid, gas, or solid, and the broadening will be a measure of the speed of the motion. For a dilute gas, the spectrum will usually be a gaussian distribution. For a liquid, however, the most common experiment of this type yields a single lorentzian line with its maximum

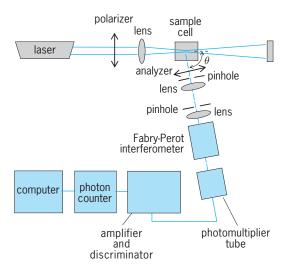


Fig. 2. Typical filter apparatus.

at the laser frequency $I(\omega) = A/\pi [(1/\tau_r)/(\omega^2 +$ $1/\tau_r^2$] $I(\omega) = A/\pi [(1/\tau_r)/(\omega^2 + 1/\tau_r^2)]$. Figure 2 shows a schematic of a typical Fabry-Perot interferometry apparatus. The Fabry-Perot interferometer acts as the monochromator and is placed between the scattering sample and the photomultiplier. Fabry-Perot interferometry measures the (average) scattered intensity as a function of frequency change from the laser frequency. This intensity is the frequency Fourier transform of the time correlation function of the scattered electric field. Photon correlation spectroscopy experiments utilizing an autocorrelator measure the time correlation function of the intensity (which equals the square of the scattered electric field). For scattered fields with gaussian amplitude distributions, the results of these two types of experiment are easily related. Sometimes photon correlation spectroscopy experiments are performed in what is called a heterodyne mode. In this case, some unscattered laser light is mixed with the scattered light on the surface of the photodetector. Photon correlation spectroscopy experiments in the heterodyne mode measure the frequency Fourier transform of the time correlation function of the scattered electric field. See INTERFEROMETRY.

Translational diffusion coefficients. The most widespread application of quasielastic light scattering is the measurement of translational diffusion coefficients of macromolecules and particles in solution. For particles in solution whose characteristic dimension *R* is small compared to q^{-1} , that is, qR < 1, it may be shown that the time correlation function measured in a polarized photon correlation spectroscopy experiment is a single exponential with relaxation time $1/\tau_r = 2q^2D$, where *D* is the particle translational diffusion coefficient. For rigid, spherical particles of any size a photon correlation spectroscopy experiment also provides a measure of the translational diffusion coefficient.

Translational diffusion coefficients of spherical particles in dilute solution may be used to obtain the particle radius R through use of the Stokes-Einstein

relation (1), where $k_{\rm B}$ is Boltzmann's constant, T the

$$D = \frac{k_B T}{6\pi \eta R} \tag{1}$$

absolute temperature, and η the solvent viscosity. If the particles are shaped like ellipsoids of revolution or long rods, relations known, respectively, as the Perrin and Broersma equations may be used to relate the translational diffusion coefficient to particle dimensions. For flexible macromolecules in solution and also for irregularly shaped rigid particles, the Stokes-Einstein relation is often used to define a hydrodynamic radius (R_H).

This technique is routinely used to study such systems as flexible coil macromolecules, proteins, micelles, vesicles, viruses, and latexes. Size changes such as occur, for instance, in protein denaturation may be followed by photon correlation spectroscopy studies of translational diffusion. In addition, the concentration and, in some cases, the ionic strength dependence of D are monitored to yield information on particle interactions and solution structure.

Photon correlation spectroscopy experiments are also used to obtain mutual diffusion coefficients of mixtures of small molecules (for example, benzenecarbon disulfide mixtures) and to measure the behavior of the mutual diffusion coefficient near the critical (consolute) point of binary liquid mixtures. Experiments of this type have proved to be very important in formulating theories of phase transitions.

Rotational diffusion coefficients. Rotational diffusion coefficients are most easily measured by depolarized quasielastic light scattering. The instantaneous depolarized intensity for a nonspherical scatterer depends upon the orientation of the scatterer. Rotation of the scatterer will then modulate the depolarized intensity. In a similar way, the frequency distribution of the depolarized scattered light will be broadened by the rotational motion of the molecules. Thus, for example, for dilute solutions of diffusing cylindrically symmetric scatterers, a depolarized photon correlation spectroscopy experiment will give an exponential intensity time correlation function with the decay constant containing a term dependent on the scatterer rotational diffusion coefficient D_R [Eq. (2)]. A depolarized filter experi-

$$\frac{1}{\tau_r} = 2(q^2 D + 6D_R)$$
(2)

ment on a similar system will give a single lorentzian with $1/\tau_r$ equal to one-half that given in Eq. (2). For small molecules (for example, benzene) and relatively small macromolecules (for example, proteins with molecular weight less than 30,000) in solution, filter experiments are used to determine rotational diffusion coefficients. In these cases, the contribution of the translational diffusion to τ_r is negligible. For larger, more slowly rotating macromolecules, depolarized photon correlation spectroscopy experiments are used to determine D_R .

Quasielastic light scattering is the major method of studying the rotation of small molecules in

solution. Studies of the concentration dependence, viscosity dependence, and anisotropy of the molecular rotational diffusion times have been performed on a wide variety of molecules in liquids as well as liquid crystals.

Rotational diffusion coefficients of large (\geq 100 nanometers) nonspherical particles may also be measured from polarized photon correlation spectroscopy experiments at high values of *q*. See DIFFU-SION.

Other applications. There are many variations on quasielastic light scattering experiments. For instance, polarized filter experiments on liquids also give a doublet symmetrically placed about the laser frequency. Known as the Brillouin doublet, it is separated from the incident laser frequency by $\pm C_s q$, where C_s is the hypersonic sound velocity in the scattering medium. Measurement of the doublet spacing then yields sound velocities. This technique is being extensively utilized in the study of bulk polymer systems as well as of simple liquids.

In a variation of the photon correlation spectroscopy technique, a static electric field is imposed upon the sample. If the sample contains charged particles, the molecule will acquire a drift velocity proportional to the electric field strength $v = \mu E$, where μ is known as the electrophoretic mobility. Light scattered from this system will experience a Doppler shift proportional to v. Thus, in addition to particle diffusion coefficients, quasielastic light scattering can be used to measure electrophoretic mobilites. *See* ELECTROPHORESIS.

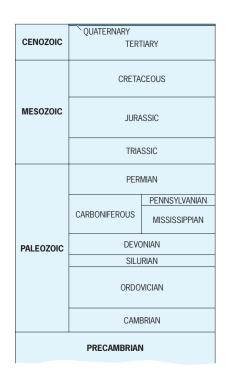
Quasielastic light scattering may also be used to study fluid flow and motile systems. Photon correlation spectroscopy, for instance, is a widely used technique to study the motility of microorganisms (such as sperm cells); it is also used to study blood flow. Robert Pecora

Bibliography. L. P. Bayvel and A. R. Jones, *Electro-magnetic Scattering and Its Applications*, 1981; B. J. Berne and R. Pecora, *Dynamic Light Scattering with Applications to Chemistry, Biology, and Physics*, 1976, reprint 1990; W. Brown (ed.), *Dynamic Light Scattering: The Method and Some Applications*, 1993; R. Pecora (ed.), *Dynamic Light Scattering: Applications of Photon Correlation Spectroscopy*, 1985.

Quaternary

A period that encompasses at least the last 3×10^6 years of the Cenozoic Era, and is concerned with major worldwide glaciations and their effect on land and sea, on worldwide climate, and on the plants and animals that lived then. The Quaternary is divided into the Pleistocene and Holocene. The term Pleistocene is gradually replacing Quaternary; Holocene involves the last 7000 years since the Pleistocene.

The Quaternary includes four principal glacial stages, each with subdivisions, in the western Rocky Mountains and Sierra Nevada and in the northcentral United States, matched by four correspond-

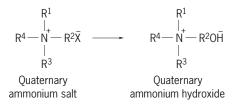


ing stages throughout northern Europe and the Alps. Interglacial stages, periods of mild climate during which ice sheets and alpine glaciers disappeared, intervened between the four main glacial stages. *See* CENOZOIC; HOLOCENE; PLEISTOCENE.

Sidney E. White

Quaternary ammonium salts

Analogs of ammonium salts in which organic radicals have been substituted for all four hydrogens of the original ammonium cation. Substituents may be alkyl, aryl, or aralkyl, or the nitrogen may be part of a ring system. Such compounds are usually prepared by treatment of an amine with an alkylating reagent under suitable conditions. They are typically crystalline solids which are soluble in water and are strong electrolytes. Treatment of the salts with silver oxide, potassium hydroxide, or an ion-exchange resin converts them to quaternary ammonium hydroxides, which are very strong bases, as shown in the reaction below.



A quaternary salt, methyllybenzulphenylammonium iodide, containing four different groups on nitrogen, proved to be resolvable (1899), clearly establishing that nitrogen is tetrahedral in these compounds. In general, optically active quaternary ammonium salts racemize readily, suggesting the possibility of an equilibrium between the tertiary amine and the alkyl compounds from which it is generated.

Following are some specific examples of useful quaternary ammonium salts and bases: Triton B (1)

$$(1) \qquad (2) \\ HOCH_2CH_2 N^{\dagger}(CH_3)_3 O\overline{H} \qquad CH_3(CH_2)_{15} N^{\dagger}(CH_3)_3 C\overline{I} \\ (3) \qquad (3) \\$$

is used in organic synthesis as a basic catalyst for aldo land Michael-type reactions. Benzyltriethylammonium chloride is used as a phase-transfer catalyst in organic synthesis to promote the reaction of water-insoluble reagents in an aqueous medium. Quaternary ammonium salts in which one of the alkyl groups is a long carbon chain, such as hexadecyltrimethylammonium chloride (2), are called invert soaps and have useful germicidal properties.

Choline (**3**) is a vitamin, since the human body makes it too slowly to meet all of its needs and a dietary source is required. Choline is a component of complex lipids such as lecithin, and acetylcholine serves in transmitting nerve signals. *d*-Tubocurarine chloride is a complex quaternary ammonium salt isolated from a tropical plant which is used in medicine as a skeletal muscle relaxant. *See* ACETYL-CHOLINE.

Other quaternary ammonium salts have found use as water repellants, fungicides, emulsifiers, paper softeners, antistatic agents, and corrosion inhibitors. *See* AMINE; AMMONIUM SALT; SURFACTANT.

Paul E. Fanta Bibliography. T. S. Eckert, Organic Chemistry, 1995; R. J. Fessenden and J. S. Fessenden, Organic Chemistry, 6th ed., 1998; J. C. Stowell, Intermediate Organic Chemistry, 2d ed., 1993.

Quaternions

An associative, noncommutative algebra based on four linearly independent units or basal elements. Quaternions were originated in Dublin, Ireland, on October 16, 1843, by W. R. Hamilton (1805-1865), who is famous because of his canonical functions and equations of motion which are important in both classical and quantum dynamics.

The four linearly independent units in quaternion algebra are commonly denoted by 1, i, j, k, where 1 commutes with i, j, k and is called the principal unit or modulus. These four units are assumed to have the multiplication table of Eqs. (1). The i, j, k do not

$$1^{2} = 1$$

$$i^{2} = j^{2} = k^{2} = ijk = -1$$

$$i(jk) = (ij)k = ijk$$

$$1i = i1$$

$$1j = j1$$

$$1k = k1$$
(1)

commute with each other in multiplication, that is, $ij \neq ji, jk \neq kj, ik \neq ki$, etc. But all real and complex numbers do commute with *i*, *j*, *k*; thus if *c* is a real number, then ic = ci, jc = cj, and kc = ck. On multiplying ijk = -1 on the left by *i*, so that iijk = i(-1) = -i, it is found, since $i^2 = -1$, that jk = i. Similarly jjk = ji = -k; when exhausted, this process leads to all the simple noncommutative relations for *i*, *j*, *k*, namely, Eqs. (2). More complicated products, for

$$ij = -ji = k$$

$$jk = -kj = i$$
 (2)

$$ki = -ik = j$$

example, jikjk = -kki = i, are evaluated by substituting for any adjoined pair the value given in the preceding series of relations and then proceeding similarly to any other adjoined pair in the new product, and so on until the product is reduced to ± 1 , $\pm i$, $\pm j$, or $\pm k$. Multiplication on the right is also permissible; thus from ij = k, one has ijj = kj, or -i = kj. Products such as jj and jjj may be written j^2 and j^3 .

All the laws and operations of ordinary algebra are assumed to be valid in the definition of quaternion algebra, except the commutative law of multiplication for the units *i*, *j*, *k*. Thus the associative and distributive laws of addition and multiplication apply without restriction throughout. Addition is also commutative, for example, i + j = j + i.

Now if *s*, *a*, *b*, *c* are real numbers, rational or irrational, then a real quaternion *q* and its conjugate q' are defined by Eqs. (3). In this case $qq' = q'q = s^2 + q' = q'q = s^2$

$$q = s + ia + jb + kc$$

$$q' = s - (ia + jb + kc)$$
(3)

 $a^2 + b^2 + c^2 = N$, and *N* is called the norm of *q*; the real quantity $T = \sqrt{N}$ is the tensor of *q*, and *s*, *a*, *b*, *c* are components (or coordinates) of *q*. The part $\gamma = ia + jb + kc$ is the vector of *q*, and it may be represented by a stroke or vector in a frame of cartesian coordinates, *a*, *b*, *c* being its components. Let now p = w + ix + jy + kz be another real quaternion; if pq = 0, either *p* or *q* or both are zero, which is called the product law. If, for example, p = 0, then all of its components w, x, y, z are zero. When p = q, that is, p - q = 0, then one must have w = s, x = a, y = b, z = c; otherwise Eq. (4) would constitute

$$(w-s) + i(x-a) + j(y-b) + k(z-c) = 0 \quad (4)$$

a linear relation between 1, *i*, *j*, *k*, which would be in conflict with their original definitions as linearly independent.

Multiplication. The product of two quaternions may be found by a straightforward process, and in full is given by Eqs. (5) and (6); and hence

Eq. (7) holds, so that pq - qp is zero only when

$$pq = ws - (ax + by + cz)$$

+ i(aw + sx + cy - bz)
+ j(bw + sy + az - cx)
+ k(cw + sz + bx - ay) (5)

$$qp = ws - (ax + by + cz)$$

+ i(aw + sx - cy + bz)
+ j(bw + sy - az + cx)
+ k(cw + sz - bx + ay) (6)

$$pq - qp = 2 [i(cy - bz) + j(az - cx) + k(bx - ay)]$$
(7)

cy - bz = az - cx = bx - ay = 0. This shows that $pq \neq qp$ except under special conditions; quaternion multiplication is not, in general, commutative.

In *q*, if any two of *a*, *b*, *c* are zero, one has, in effect, an ordinary complex number; if all of *a*, *b*, *c* are zero, then q = s and is an ordinary real number or scalar. Hence real quaternions include the real and ordinary complex numbers as special cases. It will be evident that real quaternions are a kind of extension of the ordinary complex numbers $z = x + \sqrt{-1y}$. *See* COMPLEX NUMBERS AND COMPLEX VARIABLES.

So far the case in which s, a, b, c are complex quantities has not been included, thus making q a complex quaternion, and for present purposes complex quaternions will be put aside.

It may be noted that the invention of vector analysis was inspired by Hamilton's quaternions; as early as 1846-1852 the Reverend M. O'Brien published papers in which he assumed $i^2 = j^2 = k^2 = 1$, and thus paved the way to the dot or inner product of vector analysis. Fundamentally, quaternion algebra provides much deeper concepts and consequences, and in some practical problems it presents clear advantages over vector analysis. *See* CALCULUS OF VEC-TORS.

Division. In the division of quaternions, reciprocals of the quaternion units i, j, k are easily found; thus Eqs. (8) hold. More complicated quotients may

$$i^{-1} = \frac{1}{i} = \frac{i}{ii} = -i \qquad j^{-1} = -j$$

$$k^{-1} = -k \qquad kk^{-1} = -k^2 = 1$$
(8)

be evaluated if care is taken to observe the defined conventions. Thus Eq. (9) holds, but so does Eq. (10).

$$\frac{ij}{j} = ijj^{-1} = i \tag{9}$$

$$\frac{ji}{j} = jif^{-1} = -kf^{-1} = kj = -i$$
(10)

It is best to write denominators with negative exponents and place them properly in the numerator to avoid errors. Similarly, the reciprocals and quotients of real quaternions yield unique results. If in q = s + ia + jb + kc, all of *s*, *a*, *b*, *c* are not zero,

then Eqs. (11) hold. Further, if p = w + ix + jy + kz

$$q^{-1} = \frac{1}{q} = \frac{q'}{qq'} = \frac{q'}{N}$$

$$N = s^2 + a^2 + b^2 + c^2$$
(11)

is a second real quaternion, then Eq. (12) holds.

$$\frac{p}{q} = \frac{pq'}{qq'} = \frac{pq'}{N} \tag{12}$$

Accordingly, real quaternions admit of division. If rq = p, then r = p/q = pq'/N; and if qr = p, $r = q^{-1}p = q'p/N$. Hence both right division and left division yield unique quotients. If, for example, s = a = 1, $b = \sqrt{-2}$, c = 0, then N = 1 + 1 - 2 + 0 = 0; this result is one reason why the discussion has been limited to real quaternions.

Hamilton adopted the name vectors for directed lines in space. If vectors are denoted with Greek letters, then $\alpha = ix + jy + kz$ is a vector whose components along a conventional right-handed rectangular cartesian coordinate system are x, y, z, respectively. If $\beta = ix' + jy' + kz'$ is another vector in the same coordinate system, their products are given by Eqs. (13) and (14). Both products yield a scalar u

$$\alpha\beta = -(xx' + yy' + zz') + i(yz' - y'z) + j(x'z - xz') + k(xy' - x'y) = -u + \gamma \quad (13)$$

$$\beta \alpha = -(xx' + yy' + zz') - [i(yz' - y'z) + j(x'z - xz') + k(xy' - x'y)] = -u - \gamma \quad (14)$$

and a vector γ as a sum, and such a sum is by definition a quaternion. Further, Eq. (15) holds. That is, in

$$\alpha\beta + \beta\alpha = -2(xx' + yy' + zz') \tag{15}$$

general, multiplication of nonparallel vectors is not commutative, which is a special case of the noncommutative multiplication of quaternions. If one sets x = x', y = y', z = z', then Eq. (16) results, which,

$$\alpha^2 = -(x^2 + y^2 + z^2) \tag{16}$$

with negative sign, is the square of the length of the vector α whose components are *x*, *y*, *z*. The quotient α/β emerges easily since Eq. (17) is valid. Hence α/β

$$\frac{\alpha}{\beta} = \frac{-\alpha\beta}{(x'^2 + y'^2 + z'^2)}$$
(17)

has a unique value if x', y', z' are all real and not all zero. The quotient of two vectors is a quaternion, and previously it was stated that a quaternion can be defined as the ratio of two vectors.

By multiplication a real q and its conjugate q' are found to commute, as in Eq. (18),to give a real posi-

$$qq' = q'q = s^2 + a^2 + b^2 + c^2 = T^2$$
(18)

tive scalar. Further, q + q' = 2s, and hence q and q' are the roots of a quadratic equation (19) with real

$$t^2 - 2st + T^2 = 0 \tag{19}$$

coefficients. When this equation is solved in the field

of ordinary complex numbers, Eq. (20) results. This

$$t = s \pm \sqrt{-(a^2 + b^2 + c^2)} \tag{20}$$

simple but important result emphasizes the fact that, in asking for a solution of a given algebraic equation, the field of the quantities for which a solution is desired must be specified.

Applications. Next, let r = s' + ia' + jb' + kc' be another real quaternion, then Eq. (21) holds, where S = ss' - (aa' + bb' + cc'), A = as' + a's - bc' - b'c, etc. for *B* and *C*. Also, Eqs. (22) are valid. But

$$qr = ss' - (aa' + bb' + cc') + i(as' + a's + bc' - b'c) + j(bs' + b's + a'c - ac') + k(cs' + c's + ab' - a'b) = S + iA + jB + kC$$
(21)

$$r'q' = S - (iA + jB + kC) = (qr)'$$

$$qr(qr)' = S^{2} + A^{2} + B^{2} + C^{2}$$
(22)

qr(qr)' = qrr'q' = qq'rr', since rr' is a scalar and commutes with q and q'. Therefore, one has Euler's famous result, Eq. (23), which is important in number

$$S^{2} + A^{2} + B^{2} + C^{2} = (s^{2} + a^{2} + b^{2} + c^{2})$$
$$\times (s'^{2} + a'^{2} + b'^{2} + c'^{2}) \quad (23)$$

theory since it is used in the proof that every positive rational integer can be represented as a sum of four squares. *See* NUMBER THEORY.

The quotient of real vectors α/β is the sum of a scalar and vector and hence is a quaternion, for example, q. If l_1 , m_1 , n_1 are the direction cosines of α in a rectangular cartesian coordinate frame, and l_2 , m_2 , n_2 are the direction cosines of β , then α and β can be expressed by Eqs. (24) and (25), and q is given by Eq. (26). Then $\alpha = q\beta$; that is, q is an op-

$$\alpha = a_0(il_1 + jm_1 + kn_1)$$

$$a_0 = \sqrt{x^2 + y^2 + z^2}$$
 (24)

$$\beta = b_0(il_2 + jm_2 + kn_2)$$

$$b_0 = \sqrt{x'^2 + y'^2 + z'^2}$$
(25)

$$q = \frac{\alpha}{\beta} = \frac{\alpha_0}{b_0} = [l_1 l_2 + m_1 m_2 + n_1 n_2 + i(m_2 n_1 - m_1 n_2) + j(l_1 n_2 - l_2 n_1) + k(l_2 m_1 - l_1 m_2)]$$

= $s + ia + jb + kc = s + \gamma$ (26)

erator which turns and stretches β to coincide with α . From analytical geometry $l_1l_2 + m_1m_2 + n_1n_2 = \cos \omega$, where ω is the angle ($<\pi$) between vectors α and β . Note that Eq. (27) holds, hence the vector

$$l_1(m_2n_1 - m_1n_2) + m_1(l_1n_2 - l_2n_1) + n_1(l_2m_1 - l_1m_2) = 0 \quad (27)$$

part γ of q is perpendicular to α , and the same is true for γ and β . This suggests the relationship of

Eq. (28), where
$$a_0^2/b_0^2 = s^2 + a^2 + b^2 + c^2 = T^2$$

$$q = \frac{a_0}{b_0}(\cos\omega + \epsilon\sin\omega) = TUq \qquad (28)$$

 $\cos \omega = s/T$, and ϵ is the unit vector $\epsilon = (ia + jb + kc)\sqrt{a^2 + b^2 + c^2}$, which is perpendicular to both α and β . Because $\epsilon^2 = -1$, the square of any real unit vector is -1. The factor $Uq = \cos \omega + \epsilon \sin \omega$ turns β through the angle ω to coincide with the direction of α and is called the versor of q. U is one of the symbols encountered in the grammar of earlier quaternion theory. If $U'q = \cos \omega - \epsilon \sin \omega$, then UqU'q = 1.

There is an important result in quaternion algebra which seems to be due to A. Cayley. With q as before, then its reciprocal is $q^{-1} = q'/T^2$, as before. If some other real quaternion is p = w + ix + jy + kz, then the expression qpq^{-1} is called the conical rotation (or transform) of p, and it finds important application in the specification of motions of rigid bodies. One has Eq. (29), where $T^2 = s^2 + a^2 + b^2 + c^2$. If $T^2 = 1$, w = w and X, Y, and Z are given by Eqs. (30), which is a linear transformation of X, Y, Z to x, y, z.

$$qpq^{-1} = \{wT^{2} + i[(s^{2} + a^{2} - b^{2} - c^{2})x + 2(ab - cs)y + 2(ac + bs)z] + j[2(ab + cs)x + (s^{2} - a^{2} + b^{2} - c^{2})y + 2(bc - as)z] + k[2(ac - bs)x + 2(as + bc)y + (s^{2} - a^{2} - b^{2} + c^{2})z]\}/T^{2}$$
$$= w + iX + jY + kZ = w + \tau$$
(29)

$$X = (s^{2} + a^{2} - b^{2} - c^{2})x + 2(ab - cs)y + 2(ac + bs)z Y = 2(ab + cs)x + (s^{2} - a^{2} + b^{2} - c^{2})y + 2(bc - as)z Z = 2(ac - bs)x + 2(as + bc)y + (s^{2} - a^{2} - b^{2} + c^{2})z$$
(30)

It is easy to show that, for example, Eqs. (31)

$$(s^{2} + a^{2} - b^{2} - c^{2})^{2} + 4(ab - cs)^{2} + 4(ac + bx)^{2}$$

= $T^{4} = 1$
 $2(ab + cs)(s^{2} + a^{2} - b^{2} - c^{2})$ (31)
 $+ 2(s^{2} - a^{2} + b^{2} - c^{2})(ab - cs)$
 $+ 4(bc - as)(ac + bs) = 0$

hold. Hence the coefficients of *x*, *y*, *z* are direction cosines of the frame *X*, *Y*, *Z* in the *x*, *y*, *z* frame of coordinates, and the transformation is orthogonal (unitary). The *a*, *b*, *c*, *s* are Euler's parameters (ξ, η, ζ, χ) . Clearly the angle ψ between $\alpha =$ ix + jy + kz and $\gamma = ia + jb + kc$ is given by $g(ax + by + kz) = \cos \psi$, with *g* a known quantity, and some easy algebra shows that the angle between the vector part τ of qpq^{-1} and γ of $q = s + \gamma$ is given by the same expression. Hence qpq^{-1} has rotated α conically about γ , and the magnitude of the rotation comes out, remarkably enough, to be just 2ω , where again $\cos \omega = s/T = s$. For example, if Eqs. (32) hold,

$$a = b = w = 0 \quad s = \cos \omega \quad c = \sin \omega \tag{32}$$

so that $q = \cos \omega + k \sin \omega$; then the formulas above yield easily Eqs. (33), which shows that p has been

$$X = x \cos 2\omega - y \sin 2\omega$$

$$Y = x \sin 2\omega + y \cos 2\omega$$
 (33)

$$Z = z$$

rotated conically about z and through the angle 2ω .

If λ , μ , ν are the angles which the vector $\gamma = ia + jb + kc$ makes with x, y, z, respectively, and if $T^2 = 1$, then the relations $a = \cos \lambda \sin \omega$, $b = \cos \mu \sin \omega$, $c = \cos \nu \sin \omega$, $s = \cos \omega$, and $a^2 + b^2 + c^2 = \sin^2 \omega$ are consistent with the above analysis. Another formulation for a, b, c, s is given by Eqs. (34). The angles ω , ϕ , ψ are the familiar eule-

$$a = \sin \frac{\theta}{2} \sin \frac{1}{2}(\psi - \phi)$$

$$b = \sin \frac{\theta}{2} \cos \frac{1}{2}(\psi - \phi)$$

$$c = \cos \frac{\theta}{2} \sin \frac{1}{2}(\psi + \phi)$$

$$s = \cos \frac{\theta}{2} \cos \frac{1}{2}(\psi + \phi)$$

(34)

rian angles relating a fixed cartesian frame to another with the same origin, the second being assumed, in kinematics, to be fixed in a rigid body moving about a fixed point, namely, the common origin of the two frames.

If α , β , γ , δ are defined by $\alpha = s + ic$, $\alpha = a + ib$, $\gamma = -a + ib$, $\delta = s - ic$, $i = \sqrt{-1}$, then in terms of the eulerian angles, they are given by Eqs. (35), and Eq. (36) holds. These are the Cayley-

$$\alpha = \cos \frac{\theta}{2} e^{i(\phi - \psi)/2} \qquad \beta = i \sin \frac{\theta}{2} e^{i(\phi - \psi)/2}$$
$$\gamma = i \sin \frac{\theta}{2} e^{i(\psi - \phi)/2} \qquad \delta = \cos \frac{\theta}{2} e^{i(-\psi - \phi)/2} \qquad (35)$$
$$i = \sqrt{-1}$$
$$\alpha \delta - \beta \gamma = s^2 + a^2 + b^2 + c^2 = 1 \qquad (36)$$

Klein parameters which lend themselves to homographic (bilinear) transformation, $z' = \alpha z + \beta/\gamma z + \delta$, in the complex *z* plane so that the motion of a solid body in space can be represented on a plane. The quantities $u = -i\gamma$, $v = i\delta$ are also components of a unit spinor, a function occurring in higher algebra and in the quantum theory (similarly for α , β). If $a_x = (1/2) (u^2 - v^2)$, $a_y = (-i/2) (u^2 + v^2)$, $a_z = uv$ are the components of a complex vector, then $a_x^2 + a_y^2 + a_z^2 = 0$ and the vector is of length zero. Such a vector is the starting point for some treatments of spinor theory. *See* CAYLEY-KLEIN PARAME-TERS; EULER ANGLES.

Returning now to the quaternion units 1, *i*, *j*, *k*, they may, as appears to have been first discovered by Sylvester, be represented by various matrices, one

 2×2 set being given by Eqs. (37), and the last three,

$$1 \rightleftharpoons \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad i \rightleftharpoons \begin{bmatrix} \sqrt{-1} & 0 \\ 0 & -\sqrt{-1} \end{bmatrix}$$

$$j \rightleftharpoons \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \quad k \rightleftharpoons \begin{bmatrix} 0 & \sqrt{-1} \\ \sqrt{-1} & 0 \end{bmatrix}$$
(37)

when multiplied by $\sqrt{-1}$, are the Pauli spin matrices occurring in the quantum theories of electron spin (spin = 1/2). It has been shown by S. Bochner that no set of 3 × 3 matrices have the multiplication table corresponding to 1, *i*, *j*, *k*. A set of 4 × 4 matrices, due to A. S. Eddington, which represents 1, *i*, *j*, *k* is given by Eqs. (38). These are four out of a group of

$$1 \rightleftharpoons \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad i \rightleftharpoons \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \quad k \rightleftharpoons \begin{bmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix}$$
(38)

sixteen 4×4 matrices used by Eddington in his fundamental theory; five of the sixteen, not including the above, are, when multiplied by $\sqrt{-1}$, the matrices occurring in Dirac's theory of the relativistic wave equation for the electron. Another set of four matrices for 1, *i*, *j*, *k* is given by Eqs. (39), and if, by

$$1 \rightleftharpoons \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad i \rightleftharpoons \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$
$$i \rightleftharpoons \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix} \quad k \rightleftharpoons \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{bmatrix}$$
(39)

ordinary matrix addition rule, x + ix, $+ jx_2 + kx_3 = q$ are formed, then Eq. (40) holds, and the determinant of q, |q|, has the form of Eq. (41), which is an

$$q = \begin{bmatrix} x & x_1 & x_2 & x_3 \\ -x_1 & x & -x_3 & x_2 \\ -x_2 & x_3 & x & -x_1 \\ -x_3 & -x_2 & x_1 & x \end{bmatrix}$$
(40)

$$|q| = (x^2 + x_1^2 + x_2^2 + x_3^2)^2$$
(41)

intriguing result. *See* MATRIX THEORY; RELATIVISTIC QUANTUM THEORY.

The basal quaternion elements 1, i, j, k, and their negatives, are the elements of a non-Abelian group of order eight, which is called the quaternion group. The group contains proper, self-conjugate (invariant) cyclical subgroups, one of the order 1 and, significantly, three of order 4; and there are five conjugate classes, three of which are of order 2 and the others each of order 1. *See* GROUP THEORY.

Hamilton's interest in analytical geometry led to many applications of quaternions in this field, especially to conics and quadric surfaces. In his hands and especially in those of P. G. Tait, P. Kelland, J. McAulay, C. J. Joly, and A. S. Hardy there have been many applications of quaternions to classical geometry and mathematical physics. These include the use of Hamilton's partial differential operator, Eq. (42).

$$\nabla = i\frac{\partial}{\partial x} + j\frac{\partial}{\partial y} + k\frac{\partial}{\partial z}$$
(42)

The great advances made in quantum theory, relativity, number theory, algebra, and group theory are associated with scholars who had an easy familiarity with quaternions. Don M. Yost

Bibliography. S. L. Adler, *Quaternionic Quantum Mechanics and Quantum Fields*, 1995; S. L. Altmann, *Rotations, Quaternions, and Double Groups*, 1986; E. Cartan, *The Theory of Spinors*, 1937, reprint 1980; H. Goldstein, C. E. Poole, and J. L. Safko, *Classical Mechanics*, 3d ed., 2002; W. R. Hamilton, *Elements of Quaternions*, 2 vols., 3d ed., ed. by C. J. Joly, 1901, reprint 1969; G. H. Hardy and E. M. Wright, *Introduction to the Theory of Numbers*, 5th ed., 1980; H. Jeffreys and B. Jeffreys, *Methods of Mathematical Physics*, 3d ed., 1972; L. D. Landau and E. M. Lifshitz, *Quantum Mechanics: Non-Relativistic Theory*, 3d ed., 1997; J. C. Slater, *Quantum Theory of Molecules and Solids*, 4 vols., 1963, 1965, 1967, 1974.

Quebracho

Any of a number of trees (see **illus**.) belonging to different genera but having similar qualities, all indigenous to South America and valuable for both wood and bark. The heartwood of one South American tree, *Schinopsis lorentzii* (family Anacardiaceae), is called quebracho (meaning axbreaker) in reference to the exceedingly hard wood, one of the hardest known. Quebracho is the world's most important source of tannin. The logs are chipped and the chips are cooked with steam in copper extractors. The liquor becomes highly concentrated with 40–60% of tannin. It is used either alone or in combination with



Quebracho, branch with fruit.

other chemicals for tanning all kinds of leather. *See* SAPINDALES. Perry D. Strausbaugh; Earl L. Core

Queueing theory

The mathematical theory of the formation and behavior of queues or waiting lines. The name is also applied loosely to the mathematical study of a wide variety of problems connected with traffic congestion and storage systems. Uneven flow through a service point, with fluctuating arrivals and service times, constitutes a major topic of operations research, and some current work in queueing theory goes under that title. For the mathematician, queueing theory is particularly interesting because it is concerned with relatively simple stochastic processes, which are in general nonmarkovian and possibly stationary. *See* OPERATIONS RESEARCH; STOCHASTIC PROCESS.

Origin. The principal pioneer of queueing theory was A. K. Erlang, who began in 1908 to study problems of telephone congestion for the Copenhagen Telephone Company. He was concerned with problems such as the following: A manually operated telephone exchange has a limited number (one or more) of operators. When a subscriber attempts to make a call, the subscriber must wait if all the operators are already busy making connections for other subscribers. It is of interest to study the waiting times of subscribers-for example, the average waiting time and the chance that a subscriber will obtain service immediately without waiting-and to examine how much the waiting times will be affected if the number of operators is altered, or conditions are changed in any other way. If there are more operators or if service can be speeded up, subscribers will be pleased because waiting will be reduced, but the improved facility will be more expensive to maintain; therefore, a reasonable balance must be struck.

Related problems in the use of automatic telephone exchanges and of long-distance lines able to carry only a limited number of messages simultaneously have resulted in much mathematical study of telephone traffic problems since Erlang's time. His name is commemorated by telephone engineers in the unit of traffic intensity of a channel, the erlang, which may be defined as the number of requests for service (during some period of time) that are actually made, divided by the number of requests that could have been satisfied if the channel had been used to full capacity the whole time.

Similar problems arise in other contexts. In a factory a number of machines, such as looms, may be under the care of one or more repair person. If a machine breaks down, it must stand idle until a repair person is free from repairing other machines. Machines here correspond to telephone subscribers, breakdown corresponds to attempts to make a call, and repair corresponds to connection. Other examples of congestion situations, mathematically similar or identical to the foregoing, are aircraft flying around in circles waiting to use an airport landing strip, automobiles lining up at a turnpike toll booth, and (perhaps the most familiar of all) customers lining up at the counter of a retail shop, waiting for service. Much of the literature of queueing theory uses terminology appropriate to the last example.

Theory. One of the simplest portions of queueing theory is the following presentation. Consider customers arriving at a shop counter. Suppose that they arrive singly in a purely random or haphazard fashion, specified by the following conditions: (1) The average number of customers arriving per unit time is constant, signified by λ ; (2) the numbers of customers arriving during any two nonoverlapping time intervals are independent; or (3) the chance that a customer will arrive during any specified time interval of infinitesimal duration dt is λdt . Expressed more exactly, condition (3) means that the chance of one new arrival during any short time interval of length δt is $\lambda \, \delta t + o(\delta t)$, and the chance of more than one new arrival is $o(\delta t)$, as $\delta t \to 0$. Events (arrivals) having this character are said to constitute a Poisson process (stochastic process). Suppose that there is just one server, and if that server is occupied in helping a customer and more customers arrive, the latter will line up to await their turn for service in the order of their arrival. Suppose that the time scale has been so chosen that the average duration of service of a customer is just 1 unit. It will be assumed that customers do not all require the same time for service but that their service times have what is known as an exponential distribution, such that the chance that a customer's service time will exceed any particular duration t is equal to e^{-t} . This service time distribution has the remarkable property that whether or not a customer now being served will have his or her service completed during the next infinitesimal time interval of duration dt is independent of how long the customer has already been served; the chance of completion is just dt. It is assumed that the service times of successive customers are independent and that the number of persons waiting has no effect on the speed of service. It may be noted that, because the mean service time of a customer is equal to 1 unit, λ is equal to the traffic intensity, measured in erlangs.

The queue length *L* can be defined as the total number of customers at the counter, so that L = 0 if there are no customers, and someone is waiting for service if L = 2; and *T* denotes the length of time that a customer has to wait before his service begins. Let $f_r(t)$ denote the chance that L = r at time *t*. During the infinitesimal time interval (t, t + dt) there is the chance λdt that a new customer will arrive and, provided that L = 1 at time *t*, there is the chance *dt* that the customer at the head of the queue will be served and depart. Hence, by comparing the states of the queue at times *t* and t + dt, it is easy to show that $f_r(t)$ satisfies the system of differential-difference equations (1).

$$f'_{0}(t) = f_{1}(t) - \lambda f_{0}(t)$$

$$f'_{r}(t) = f_{r+1}(t) - (1+\lambda)f_{r}(t) + \lambda f_{r-1}(t)$$

$$(r \ge 1)$$
(1)

It can be proved that if $\lambda < 1$, the process approaches an equilibrium or steady state, in which the chances $f_r(t)$ are independent of t, and the derivatives on the left-hand sides of Eqs. (1) vanish. The equilibrium values are easily found as shown by Eq. (2).

$$f_r = (1 - \lambda)\lambda^r \qquad (r \ge 0) \tag{2}$$

In particular, the chance that at any arbitrary instant the server is idle is shown by Eq. (3). For the

$$f_0 = 1 - \lambda \tag{3}$$

mean queue length one finds that Eq. (4) holds.

$$\varepsilon(L) = \sum_{r=0}^{\infty} \eta f_r = \frac{\lambda}{1-\lambda} \tag{4}$$

Now consider a newcomer to the queue. If there are *L* persons in the queue the instant before the newcomer arrives, the expected service time for each of these (including the one at the head) is 1, and so $\epsilon(T) = \epsilon(L)$. Thus, the mean waiting time is shown by Eq. (5). This will be large if λ is close to, but less

$$\varepsilon(T) = \frac{\lambda}{1 - \lambda} \tag{5}$$

than, 1. If $\lambda > 1$, it is obvious that the mean queue length increases indefinitely and no equilibrium is reached, and it can be shown that the same is true if $\lambda = 1$.

It has been assumed that the service times follow an exponential distribution. If instead it is assumed that every customer requires exactly the same time for service, namely 1 unit, a somewhat similar argument would lead to the result that, for equilibrium with $\lambda < 1$, Eq. (6) holds, which is just one-half the

$$\varepsilon(T) = \frac{\lambda}{2(1-\lambda)} \tag{6}$$

previous result. Both results are included in a formula derived by F. Polaczek (1930), as follows. If the service time distribution has mean equal to 1 and variance v and if all the other assumptions previously made are satisfied, then the mean waiting time of a customer, in the equilibrium condition for $\lambda < 1$, is shown by Eq. (7). This will tend to be large if either

$$\varepsilon(T) = \frac{(1+\nu)\lambda}{2(1-\lambda)} \tag{7}$$

 λ is close to 1 or v is large (service times are very variable).

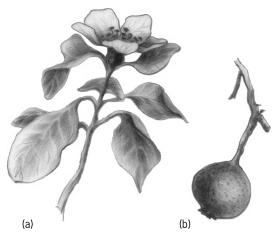
Applications. Many variants of the above problem have been studied. Arrivals need not constitute a Poisson process. For example, the arrival times of patients at a doctor's office can show a more even spacing than a Poisson process if they have been given equally spaced appointments, but their arrival times will not be exactly equally spaced. The rule of queue discipline, "first come, first served," will not always apply. Some customers, for example, may enjoy priority and move ahead of nonpriority customers. There may be more than one server, and if so, separate queues may be formed for each. There are then various possible rules to be considered for assigning a newcomer to a queue. It may happen that the frequency of arrivals or speed of service varies with the queue length; for example, if the queue is too long, new customers may be turned away. Compound queueing systems are formed when successive services have to be rendered to the same customers.

What is most interesting to investigate varies with the circumstances. Sometimes it is the mean waiting time of customers, sometimes the frequency with which the queue length exceeds a given limit, sometimes the proportion of the servers' time that is idle, and sometimes the average duration of a period during which a server is continuously occupied. In the study of stocking a warehouse or retail shop, known generally as the theory of inventories, the frequency with which the stock will be exhausted is considered under various reordering policies. Similar considerations apply in the theory of dams and water storage. *See* LINEAR PROGRAMMING; SYSTEMS ENGINEER-ING. Francis J. Anscombe

Bibliography. R. B. Cooper, *Introduction to Queueing Theory*, 2d ed., 1981, reprint 1990; J. W. Daigle, *Queueing Theory for Telecommunications*, 1992; B. Gnedenko and I. N. Kovalenko, *Introduction to Queueing Theory*, 2d ed., 1989; D. Gross and C. M. Harris, *Fundamentals of Queueing Theory*, 3d ed., 1997; L. Kleinrock, *Queueing Systems*, 2 vols., 1975, 1976.

Quince

The tree *Cydonia oblonga*, originally from Asia, grown for its edible fruit. The deciduous, crookedbranched tree attains a height of about 20 ft (6 m). It is cultivated in either bush or tree form. The undersides of the leaves are densely tomentose (hairy); the solitary flowers (**illus**. *a*), up to 2 in. (5 cm) across, are snowy white or pale pink; the fruit is a pearshaped or apple-shaped pome (illus. *b*), characteris-



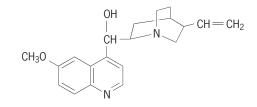
Quince. (a) Flower is borne on wood of the same season, not from an autumn fruit bud. (b) Short shoots start from the winter buds, and the flowers and fruits are produced on the ends of these shoots.

tically tomentose, up to 3 in. (7.5 cm) in diameter, aromatic, sour, astringent, and green, turning clear yellow at maturity. Used mostly for jam and jelly or as a stewed fruit, the fruit of the quince develops a pink color in cooking. *See* DECIDUOUS PLANTS; ROSALES.

The quince is propagated on hardwood cuttings or by budding on quince seedlings. It may be used as stock for dwarfing the pear. The tree is only slightly hardier than the peach, the wood being severely injured at -15 to -20° F (-26 to -29° C), and is subject to fire blight (Bacillus amylovorus), which is controlled by removal of diseased twigs and avoidance of overly vigorous growth. Trees are planted 15-20 ft (4.5-6 m) apart. Principal varieties are Orange, Champion, and Van Deman. Fruits keep for only a few weeks. There is but little commercial production of this fruit in North America. The dwarf Japanese quince (Chaenomeles japonica), with orange-scarlet flowers, is grown as an ornamental shrub. See BACTERIA; FRUIT, TREE; PEACH; PEAR; PLANT PATHOLOGY; PLANT PROPAGA-Harold B. Tukey TION.

Quinine

The chief alkaloid of the bark of the cinchona tree, which is indigenous to certain regions of South America. The structure of quinine is shown below.



The most important use of quinine has been in the treatment of malaria.

For almost two centuries cinchona bark was employed in medicine as a powder or extract. In 1820 P. Pelletier and J. Caventou isolated quinine and related alkaloids from cinchona bark, and the use of the alkaloids as such rapidly gained favor. Major credit is due to P. Rabe for the postulation of the correct structure of quinine. The difficult laboratory synthesis of quinine by R. Woodward and W. Doering in 1944, although economically unfeasible, corroborated Rabe's structure.

Until the 1920s quinine was the best chemotherapeutic agent for the treatment of malaria. Clinical studies definitely established the superiority of the synthetic antimalarials such as primaquine, chloroquine, and chloroguanide. *See* ALKALOID; CIN-CHONA; MALARIA. S. Morris Kupchan

Quinoa

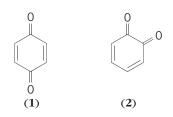
An annual herb, *Chenopodium quinoa* (family Chenopodiaceae), 4–6 ft (1.2–1.8 m) tall, a native of Peru, and the staple food of many people in South

America. These plants, grown at high altitudes, produce large quantities of highly nutritious seeds used whole in soups or ground into flour, which is made into bread or cakes. The seeds are also used as poultry feed, in medicine, and in making beer. The ash is mixed with coca leaves to flavor them as a masticatory. In the United States the leaves are sometimes used as a substitute for spinach. *See* CARYOPHYL-LALES; SPICE AND FLAVORING.

Perry D. Strausbaugh; Earl L. Core

Quinone

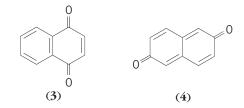
One of a class of aromatic diketones in which the carbon atoms of the carbonyl groups are part of the ring structure. The name quinone is applied to the whole group, but it is often used specifically to refer to *p*-benzoquinone (1). *o*-Benzoquinone (2) is



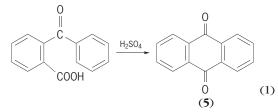
also known but the meta isomer does not exist.

Preparation. Quinones are prepared by oxidation of the corresponding aromatic ring systems containing amino (—NH₂) or hydroxyl (—OH) groups on one or both of the carbon atoms being converted to the carbonyl group. *p*-Benzoquinone is prepared by the oxidation of aniline with manganese dioxide, MnO₂, in the presence of sulfuric acid, H₂SO₄. Oxidation of phenol, *p*-aminophenol, hydroquinone, or *p*-phenylenediamine will also produce *p*-bezoquinone. *o*-Benzoquinone is prepared by oxidation of catechol with silver oxide, Ag₂O, in the absence of water. This quinone is much less stable and more reactive than the para isomer.

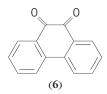
Three of the several quinones derived from naphthalene are known: 1,4-napthoquinone (3), 1,2-naphthoquinone, and 2,6-naphthoquinone (4).



The naphthoquinones are prepared by oxidation of the corresponding aminoaphthols. 9,10-Anthraquinone (5) is best prepared, as in reaction (1),

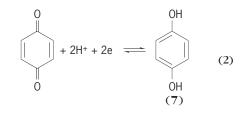


by dehydration of *o*-benzoylbenzoic acid which is prepared from Friedel-Crafts reaction of benzene and phthalic anhydride. Direct oxidation of phenanthrene with chromic acid yields 9,10-phenanthraquinone (6), the further oxidation of which



gives diphenyl-2,2'-dicarboxylic acid (diphenic acid). *See* FRIEDEL-CRAFTS REACTION.

Reactions. *p*-Benzoquinone is easily reduced, as in reaction (2), to hydroquinone (7) by a variety of

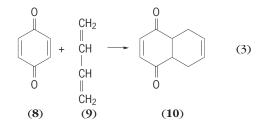


reagents. Reaction (2) is reversible, and the position of equilibrium can be made to depend on hydrogenion concentration and applied electrical potential.

This system ($E_0 = 0.699$ volt) has been useful for the measurement of hydrogen-ion concentration. The E_0 values for many other quinone-hydroquinone systems have been measured. An intermediate in the reduction of *p*-benzoquinone or in the oxidation of hydroquinone is quinhydrone, a 1:1 molecular complex of these two substances. *See* HYDROGEN ION.

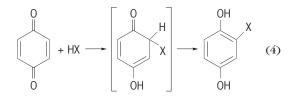
The most characteristic reactions of para quinones are those of the carbon to carbon double bonds and of the conjugated system

Reaction as a dienophile in the Diels-Alder process, reaction (3) [where *p*-benzoquinonel **(8)** and 1,3-butadiene **(9)** react to give 5,8,9,10-tetrahydro-1,4-napthoquinone **(10)**], is quite general and oc-

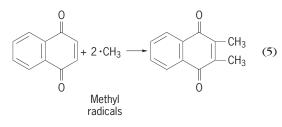


curs under mild conditions. The remaining \css\ bond of the quinone ring may also react in the same way. Halogen adds normally to the \css\ bond as in alkenes. Hydrogen halide, however, adds to the conjugated system

by 1,4 addition, and this is followed by enolization to a hydroquinone derivate, reaction (4). Malonic es-

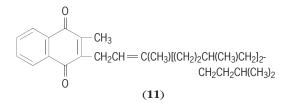


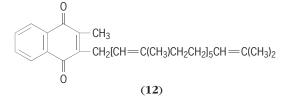
ter and acetoacetic ester react similarly with utilization of their active hydrogen atoms. Mercaptans and Grignard reagents give mixtures of normal adducts to the C=O group and 1,4 additions to the system. Simple quinones do not often undergo substitution reactions with the electrophilic reagents commonly used for aromatic systems. Free radicals from decomposition of acyl peroxides or lead tetraacetate substitute as in reaction (5).



See FREE RADICAL; GRIGNARD REACTION.

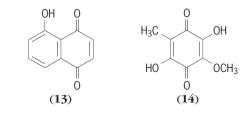
Important naturally occurring naphthoquinones are vitamins K_1 (11) and K_2 (12) which are found in





blood and are responsible for proper blood clotting reaction. The long aliphatic chain has been found unnecessary for the clotting reaction; its replacement by a hydrogen atom gives Menadione or 2-methyl-1,4-naphthoquinone, which is manufactured synthetically for medicinal use. *See* VITAMIN K.

A number of quinone pigments have been isolated from plants and animals. Illustrative of these are juglone (13) found in unripe walnut shells and spinulosin (14) from the mold *Penicillium spinu*-



losum. 9,10-Anthraquinone derivatives form an important class of dyes of which alizarin is the parent type. *p*-Benzoquinone is manufactured for use as a photographic developer. *See* AROMATIC HY-DROCARBON; DYE; KETONE; OXIDATION-REDUCTION. David A. Shirley

Bibliography. R. J. Fessenden and J. S. Fessenden, *Organic Chemistry*, 6th ed., 1998; S. Patai, *The Chemistry of Quinonoid Compounds*, 1974; Z. Rappoport (ed.), *The Chemistry of Quinones*, 1988; J. C. Stowell, *Intermediate Organic Chemistry*, 2d ed., 1993; R. H. Thompson, *Naturally Occurring Quinones*, 3d ed., 1987.