Contents of all Volumes

Volume 1 Natural Products Structural Diversity-I Secondary Metabolites: Organization and Biosynthesis

Volume 3 Development & Modification of Bioactivity

Volume 4 Chemical Ecology

Volume 5 Amino Acids, Peptides and Proteins

Volume 6 Carbohydrates, Nucleosides & Nucleic Acids

Volume 7 Cofactors

Volume 8 Enzymes and Enzyme Mechanisms

Volume 9 Modern Methods for the Investigation of Natural Products

9.01 Overview and Introduction

Lew Mander, Australia National University, Canberra, ACT, Australia

In the first edition of this comprehensive series, the chief editors chose to recognize that the field of natural products should encompass not only the traditional secondary metabolites, but molecules involved in fundamental metabolic cycles and the larger biomolecules (proteins, nucleic acids, polysaccharides, etc.) as well. Such a view brings together biologists, biochemists, chemists, and biophysicists, combining their skills and knowledge to bear on complex, challenging multidisciplinary problems. The aim of this volume is to provide a representative selection of recent methods employed in such projects with a view to providing an entry point for research workers to gain an appreciation of methods beyond their current experience and to cross traditional boundaries. Emphasis is given to the scope and limitation of the methods with valuable tips on the selection of the best procedures and traps to avoid. An archetypal example from the previous series is a chapter in Volume 7 on 'Cloning as a Tool for Organic Chemists' by Ben Miller, which is still very useful, but a subject that has been expertly updated by Cynthia Kinsland in Chapter 9.19 ('Bacterial Protein Overexpression: Systems and Strategies').

In most projects, it is essential to analyze and resolve mixtures (often on a sub-microscale) and gain access to pure products rapidly and efficiently, so the first chapter by Boysen and Hearn (Chapter 9.02) provides an appropriate beginning. Accordingly, an overview of the increasingly powerful techniques of high-performance chromatography, including normal phase, reverse phase, ion exchange, hydrophilic interaction, size exclusion, and affinity variants is provided. Also, multidimensional methods, including automation and off-line techniques are treated. The following chapters deal with methods for structure determination and here the gold standard is X-ray crystallography. It is the most commonly used technique for determining the atomic structure of biomacromolecules and underpins modern structural biology. Armed with the information provided by Oksanen and Goldman in Chapter 9.03, neophytes should gain an understanding of the scope and limitations of the technique sufficient to have an intelligent dialogue with crystallographers. Despite the enormous power of X-ray crystallography, it is constrained by the availability of suitable crystals and is largely limited in its ability to provide dynamic information. Fortunately, there are the very powerful supplementary spectroscopic techniques, the most important of which are addressed in the first half of this volume.

Three-dimensional molecular structures are usually chiral, that is, they have the same connectivity of atoms, but two arrangements are possible, possessing an object to mirror image relationship, as with left and right hands. These isomeric structures are termed 'enantiomers'. An important consequence of this situation is that the interaction of small molecules with a biomacromolecule will be different for each enantiomer. The most powerful technique to determine the absolute arrangement in space is circular dichroism (CD) spectroscopy, as described by Berova, Ellestad, and Harada in Chapter 9.04. The next chapter by Mori (Chapter 9.05) explores aspects of stereochemistry and absolute configuration further resorting, *inter alia*, to degradation and/or synthesis to finalize a structure, particularly where stereocenters are remote from each other.

The next four chapters are devoted to nuclear magnetic resonance (NMR), a technique of extraordinary power that has revolutionized the determination of molecular structure over the past five decades. In Chapter 9.06, Edison and Schroeder summarize the scope and limitation of the method and thread their way through the labyrinth of pulse sequences used to enhance what is an intrinsically insensitive technique, while extracting valuable structural and dynamic information. As well as addressing cutting edge technology, valuable advice is provided for dealing with mixtures, including metabolomics (cf. Chapter 9.15). The remaining three chapters deal with the application of the NMR technology to the major classes of biomacromolecules, namely, oligosaccharides and glycopeptides (Kover *et al.*, Chapter 9.07), to oligonucleotides and nucleic acids (James and Ulyanov, Chapter 9.08), and to peptides and proteins (King and Mobli, Chapter 9.09).

Equally important as NMR spectroscopy is the development of mass spectrometry where advances have been even more spectacular, partly due to its far greater sensitivity. A broad introduction is provided by Hocart (Chapter 9.10), who describes the range of available spectrometers and techniques. In the next chapter (Chapter 9.11), Dorrestein and his coauthors describe applications to the elucidation of biosynthetic pathways, focusing on nonribosomal peptides and polyketides, while Das provides a detailed account of the structure determination of proteins and peptides in Chapter 9.12. In yet another illustration of the power of mass spectrometry to analyze biomacromolecules, Li and Altman extend the technique to a study of bacterial polysaccharides (Chapter 9.13).

In Chapter 9.14, Karuso examines the interaction between natural products and receptors, involving the utility of chemical proteomics, reverse chemical proteomics, *in vitro* display technologies, and reverse genetics. Wang and her associates (Chapter 9.15) then provide a survey of bioinformatics and its applications to the treatment of biodiversity, gene expression, and signaling processes. Of particular value is a summary of software that is available for processing the enormous amounts of data that are generated, as in the next chapter (Chapter 9.16) by Trenerry and Rochfort on metabolomics, that is, the study of global metabolite profiles in a given system (e.g., cell, tissue, or organism).

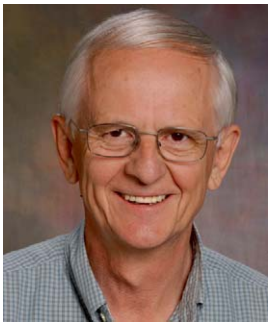
Genome sequencing projects have provided a wealth of information on gene identities in many prokaryotic and eukaryotic organisms. The daunting challenge for proteomic research is now to assign the molecular, cellular, and physiological functions for the full complement of proteins encoded by the genome. Rising to this challenge, a new powerful chemical proteomic strategy, termed activity-based protein profiling (‘ABPP’) has been systematically explored by Sieber and coauthors (Chapter 9.17) to characterize enzyme function and regulation directly in native biological systems. The principal concept of this technology is the covalent active site labeling of distinct enzyme classes by functionalized small molecules, which are used as chemical probes.

In Chapter 9.18, Eisenreich and Bacher outline a ‘retrobiosynthetic concept’, which is designed to improve the stringency of isotope incorporation studies. Advanced NMR and mass spectrometry methods enable the quantitative assessment of all major isotopologue species present in metabolites recovered from isotope incorporation experiments. Comparison of quantitatively determined isotopologue patterns of different metabolites resulting from a given experiment allow the reconstruction of labeling patterns of essential intermediates in the major pathways of central metabolism. On this basis, the building blocks that have contributed to assembling complex metabolites can be reconstructed.

As noted above, Cynthia Kinsland has provided an excellent guide to methods for the overexpression of proteins, a pivotal technique underpinning much of modern molecular biology. It may seem simple in principle, but in practice, with all its pitfalls, is another matter. This author, as described in Chapter 9.19, provides us with critical comment from the coal-face and steers us along the straight and narrow path to success.

In Chapter 9.20, Ollis and his coauthors provide a broad overview of directed molecular evolution with the intent of promoting interest in the field. They explain why we might want to evolve enzymes, how this might be done, and give some idea of the limitations of the available technology. They focus on those areas that have the most relevance to chemists, natural products chemists in particular. Also, the reader is directed toward reviews and important original studies that cover material that may become more important in the future.

Finally, in Chapter 9.21, Chen and Andoy provide an overview of single-molecule fluorescence methods in enzymology. The authors focus on the principle, features, generality, and experimental challenges illustrated by examples from the recent literature.



Lew Mander was born in Auckland, New Zealand, where he completed his BSc and MSc (hons.) degrees at the University of Auckland (the latter with R. C. Cambie). After moving to Australia, he obtained his Ph.D. in 1964 for his research on the structures of the Galbulimima alkaloids at the University of Sydney under the supervision of C. W. Shoppee, E. Ritchie, and W. C. Taylor. After 2 years of postdoctoral studies with R. E. Ireland, initially at the University of Michigan and then at the California Institute of Technology, he returned to Australia as a lecturer in organic chemistry at the University of Adelaide. He moved to the Australian National University in 1975 as a senior fellow in the Research School of Chemistry where he was subsequently appointed Professor in 1980, serving two periods as Dean (1981–85; 1992–95) and recently made emeritus. He was a Nuffield Fellow at Cambridge University in 1972 with A. R. Battersby, and a Fulbright Senior Scholar at the California Institute of Technology in 1977 and at Harvard University in 1986 (with D. A. Evans on both occasions), an Eminent Scientist of RIKEN, Saitama, Japan (1995–96), and a Distinguished Alumnus Professor, University of Auckland (1992). Also, he has been a visiting professor at the universities of Sydney, Cambridge, Alberta, Colorado, and Canterbury (New Zealand). He is a Fellow of the Australian Academy of Science and The Royal Society (London). His research interests are concerned with the development of methods and strategies for the assembly and manipulation of complex natural products with a special interest on the role of gibberellins in plant growth and development.

9.02 High Performance Liquid Chromatographic Separation Methods

Reinhard I. Boysen and Milton T. W. Hearn, Monash University, Melbourne, VIC, Australia

9.02.1 Introduction

The properties of natural products obtained from plants, fungi, animals, and other organisms of terrestrial or marine origin have intrigued mankind for millennia, in particular those products that exert an effect (i.e., have a particular bioactivity) on humans or other organisms. Such compounds have long been used as nutraceuticals, social elixirs, intoxicants, drugs of abuse, or as therapeutics.

At the highest level of classification, natural (= biological) products can be divided into two super classes, namely, primary and secondary metabolites. The primary class of natural products include nucleic acids, proteins, carbohydrates, lipids, and their precursors, all produced by metabolic pathways essential for life. Primary metabolic pathways therefore encompass all of the modes of synthesis, interconversion, or degradation associated with the production of primary metabolites.

The secondary metabolites are produced from key intermediates of the primary metabolism pathways. They are often found in limited quantity, can occur transiently in the cell cycle and can be unique for a particular group of organisms or even species. Secondary metabolites typically represent a chemically very diverse group of small molecules (molecular mass <2000 amu) and include (1) products from overflow metabolism as a consequence of nutrient limitation, (2) compounds for defense, (3) regulatory molecules, (4) signaling molecules, or (5) molecules that serve the requirements of evolutionary exploration within the physicochemical space available on this planet. Secondary metabolites can be grouped according to the primary metabolic pathway from which they are derived or in terms of their structural similarity.

The isolation of small quantities of unknown secondary metabolites from complex biological matrices and their unambiguous structural elucidation represent formidable analytical and purification challenges. Such tasks become even more demanding if they are used in conjunction with dereplication algorithms (i.e., using natural product libraries) to avoid time and resource allocation based on already known compounds.[1–4] These studies can, in some cases, be bioassay guided, which involves the testing of individual aliquots during a separation process in order to indicate which fractions are associated with a specific bioactivity.[5] Such methods increasingly involve high-throughput, robotic instrumentation and procedures.

The process of natural product isolation and identification can be split into several subprocesses of extraction, fractionation, isolation, and final purification to a level that enables structural elucidation and appropriate functional evaluation and discrimination. After bulk extraction (e.g., maceration, boiling, Soxhlet extraction, supercritical fluid extraction) from the biological material, a crude extract is usually fractionated by a variety of different chromatographic techniques, which take into account the chemical diversity and the low-molecular-mass ranges of the secondary plant metabolites. These fractionations can be performed with a variety of chromatographic techniques, including preparative thin-layer chromatography (PTLC), high-speed countercurrent chromatography (HSCC), open column chromatography (CC), flash chromatography (FC), or solid-phase extraction (SPE).[6] For further isolation, preparative or semipreparative high-performance liquid chromatography (HPLC) is usually applied, in conjunction with online ultraviolet (UV), evaporative light scattering, fluorescence, electrochemical detection (ED), or mass spectrometric detection. Nuclear magnetic resonance (NMR) spectroscopy, X-ray crystallography, Fourier transform infrared (FTIR) spectroscopy, or multiple stage mass spectrometry (MS^n) may then allow the unambiguous determination of the molecular structure.

The workflow integration of HPLC in the extraction, fractionation, isolation, and structural elucidation of natural products is illustrated in **Figure 1**. After the extraction of a natural product from a biological matrix (level 1), the crude extract is clarified (i.e., by filtration, centrifugation) to be free from particulate matter suitable for chromatography and in an appropriate (noninterfering) buffer compatible with the mobile phase(s) of the particular chromatographic mode used in the next steps. This is followed by an enrichment step (level 2), preferably with SPE or with restricted access materials (RAMs) in a step elution mode to eliminate the majority of macromolecular material and to drastically reduce the volume of the sample. The next purification step (level 3) comprises the intermediate purification of target compound(s) and a final chromatographic purification and uses a variety of high-performance (HP) chromatographic modes (i.e., reversed-phase chromatography (HP-RPC), normal-phase chromatography (HP-NPC), ion-exchange chromatography (HP-IEX), hydrophilic interaction chromatography (HP-HILIC), size exclusion chromatography (HP-SEC), or affinity chromatography (HP-AC)), to yield the desired compound(s) in the required amount and degree of purity. At the detection level (level 4), besides diode array detection (DAD), electrospray ionization (ESI) mass spectrometry (MS) can often be applied for the identification and – depending on the type of mass spectrometer – may allow quantification of the separated compound(s). Further structural elucidation can be performed with MS^n, that is, with an ion trap mass spectrometer, quadruple mass spectrometer, or FT mass spectrometer together with proton, carbon, or heteronuclear NMR spectroscopy.

Over the last two decades, HPLC has to a large extent superseded the classical modes of open column, thin-layer or paper chromatography previously used for natural product separation and has become an integral part of natural product analysis and preparative isolation. This can be attributed to various factors, including (1) availability of numerous chromatographic modes, robust high-resolution chromatographic materials and

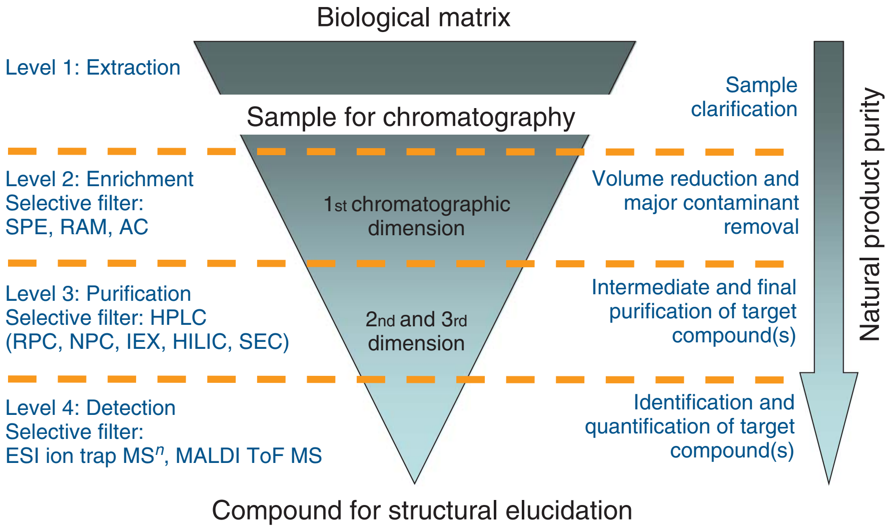

Biological matrix
Level 1: Extraction
Sample clarification
Sample for chromatography
Level 2: Enrichment
Selective filter:
SPE, RAM, AC
1st chromatographic dimension
Volume reduction and major contaminant removal
Level 3: Purification
Selective filter: HPLC
(RPC, NPC, IEX, HILIC, SEC)
2nd and 3rd dimension
Intermediate and final purification of target compound(s)
Level 4: Detection
Selective filter:
ESI ion trap MS^n, MALDI ToF MS
Identification and quantification of target compound(s)
Natural product purity
Compound for structural elucidation

Figure 1 Example of workflow in natural product isolation from a complex biological matrix using high-performance liquid chromatography for the target compound purification and identification. With successive application of several chromatographic modes of different selectivity (i.e., hydrophobicity/hydrophilicity, charge, molecular size) the chromatographic separation can become multidimensional.

reliable instrumentation, (2) availability of well-established optimization procedures, (3) ease of scalability from analytical to preparative separation, (4) ease of integration with spectroscopic methods (e.g., DAD, UV, fluorescence), (5) the ability to use fully automated high-throughput screening (HTS) modes, (6) the ease of in-line integration with MS, and (7) the possibility of hyphenation with NMR spectroscopy.[5] An illustrative example, where these chromatographic capabilities have been used with finesse, has been reported by Dugo *et al.*[7] to resolve a highly complex, crude extract of carotenoids present in red orange juice. In this example (**Figure 2**), the crude extract was partly separated by two-dimensional (2D) normal-phase × reversed-phase

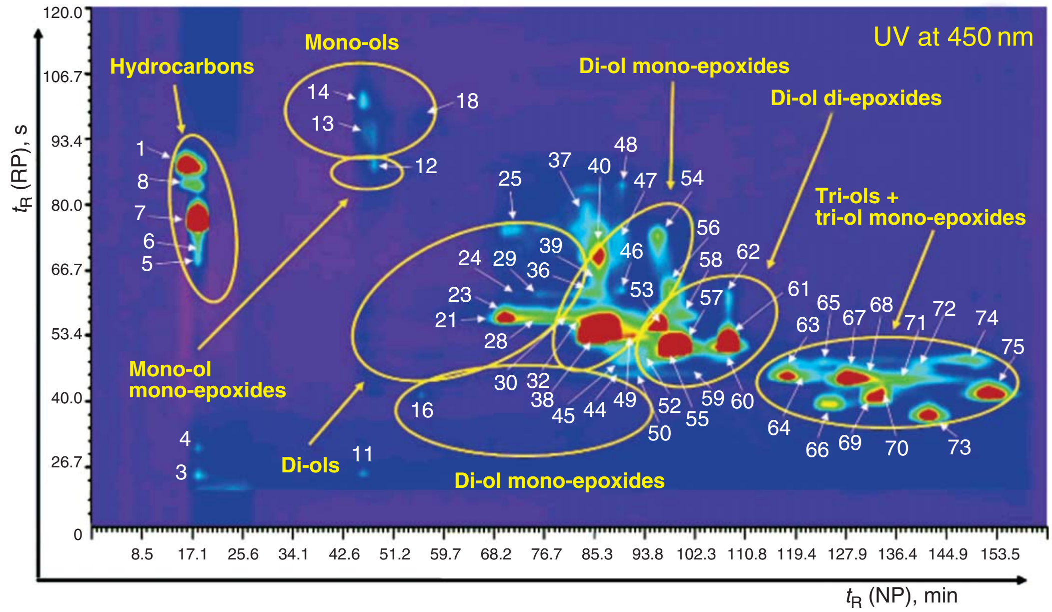


Figure 2 Contour plot of the comprehensive normal-phase × reversed-phase liquid chromatography analyses of carotenoids present in red orange juice with peaks and compound classes indicated. Reproduced from P. Dugo; V. Skerikova; T. Kumm; A. Trozzi; P. Jandera; L. Mondello, *Anal. Chem.* **2006**, *78*, 7743–7750.

liquid chromatography (LC), where the authors used a capillary 300 mm × 1.0 mm 5 μm normal-phase silica column in the first dimension with linear gradient elution (eluent A: *n*-hexane and eluent B: ethyl alcohol) at a flow rate of 10 μl min^{-1}. In the second dimension, they employed a monolithic 100 mm × 4.6 mm C_{18} column, with linear gradient elution (eluent A: 2-propanol and eluent B: 20% (v/v) water in acetonitrile (ACN)) at a flow rate of 4.7 ml min^{-1}. The incompatibility of the solvents that were used in the two dimensions (normal-phase chromatography (NPC) and reversed-phase chromatography (RPC)) and their effects on the separation were overcome by using a combination of a capillary column in the first dimension and a (larger) analytical monolithic column in the second dimension. Using flow splitting, the detection was performed in parallel using photodiode array and mass spectrometric detection. This approach then allowed establishment of a 2D contour plot, permitting group separations as well as identification of individual compounds.

The six chromatographic modes available in HPLC for the isolation and purification of natural products are described in more detail in Section 9.02.2. Since the groups of natural products differ in their molecular properties, certain chromatographic modes have been shown to work better with particular natural product groups. However, in order to take full advantage of a specific HPLC mode for a separation task and to effectively utilize time and resources, comprehensive method development should be performed. An example of such method development from the analytical to the preparative stage is described for reversed-phase chromatography (HP-RPC), the most frequently employed mode in natural product purification, in Section 9.02.3.

Since natural products are generally isolated from complex biological matrices, often more than one chromatographic step is needed for their purification. The successive application of several suitable different chromatographic modes must consider their applicability to the compound (class) of interest and the compatibility of the chromatographic modes to be integrated with the detection procedures. The design and implementation of a two- or higher-dimensional separation scheme for natural product purification is discussed in Section 9.02.4.

A variety of different approaches lead to successful natural product separation schemes. Representative (as well as some unusual) HPLC separation methods for selected natural product groups (isoprenoids, phenolics, and alkaloids) are reviewed in Section 9.02.5.

9.02.2 Modes of Separation by HPLC in Natural Product Isolation

There are six modes of HPLC currently in use for secondary metabolite analysis, namely, HP-RPC, HP-NPC, HP-IEX, HP-HILIC, HP-SEC, and HP-AC. The principles of these different modes are explained below. All of these various chromatographic modes can be operated under isocratic (i.e., fixed eluent composition), step gradient, or gradient elution conditions (variable step or continuous changes in eluent composition), except for SEC, which is usually performed under isocratic conditions. All modes can be used in analytical, semipreparative,[8] or preparative[9–14] situations.

9.02.2.1 High-Performance Reversed-Phase Chromatography

HP-RPC separates compounds according to their relative nonpolarity or hydrophobicity. In RPC, the polarity of the stationary and mobile phase is to the reverse of that used in NPC. HP-RPC is performed on porous or nonporous stationary phases with immobilized nonpolar polymers (i.e., *n*-alkylsilicas) or nonpolymer polymers (i.e., microparticulate polystyrenes). The most commonly accepted retention mechanism in RPC is based on the solvophobic theory, which describes the hydrophobic interaction between the nonpolar surface regions of the analytes and the nonpolar ligands/surfaces of the stationary phase.[15,16]

Typically, the nonpolar ligands are immobilized onto the surface of spherical, porous, or nonporous silica particles, although nonpolar polymeric sorbents (e.g., those derived from cross-linked polystyrene–divinylbenzene) can also be employed. Silica-based packing materials of 3–10 μm average particle diameter and 70–80 Å pore size, with *n*-butyl, *n*-octyl, or *n*-octadecyl ligands are widely used for the separation of natural products. Silica particles of 1 μm to over 65 μm have been developed in various size distributions and configurations, for example, spherical, irregular, with various pore geometries and pore connectivities; and in pellicular, fully porous, or monolithic structures by a variety of routes of manufacture and with different silica

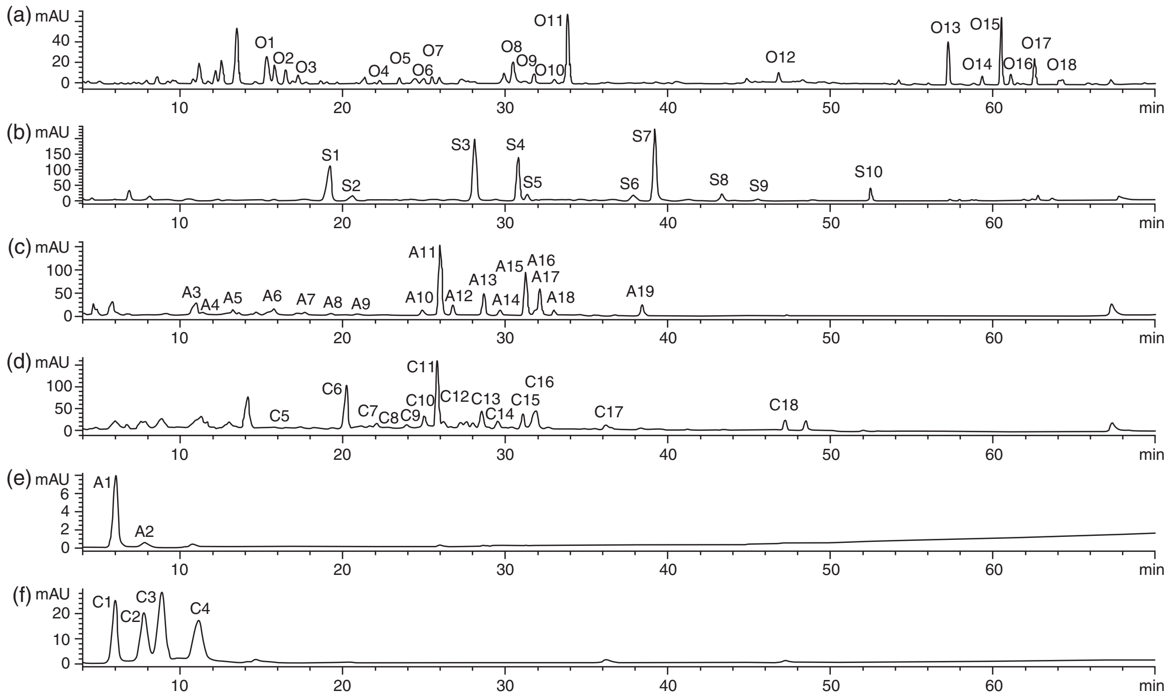


Figure 3 High-performance reversed-phase chromatography of glycosylated flavonoids and other phenolic compounds with UV absorption: (a) navel orange peel (350 nm), (b) soybean seeds (270 nm), (c) Fuji apple peel (270 nm), (d) cranberry (270 nm), (e) Fuji apple peel (520 nm), and (f) cranberry (520 nm). Reproduced from L.-Z. Lin; J. M. Harnly, *J. Agric. Food Chem.* **2007**, *55*, 1084–1096.

types, that is, whether they are based on type I, type II, or type III silica according to the classification of Unger.[17] For low-molecular-mass natural products (molecular mass <4000 Da) silica materials of 70–80 Å pore size and 3–5 μm average particle diameter are often used.

In HP-RPC, an organic solvent (i.e., methanol (MeOH), ethanol, ACN, *n*-propanol, tetrahydrofuran (THF)) is used as a surface tension modifier in the chromatographic eluent, which has a particular elution strength, viscosity, and UV cutoff. Mobile phase additives (i.e., acetic acid (AA), formic acid (FA), trifluoroacetic acid (TFA), and heptafluorobutyric acid (HBFA)) are used to obtain a particular pH value, typically at low pH (e.g., ~pH 2 for silica-based materials) with the exception of polymeric stationary phases, which have an extended pH range from pH 1 to 12. Some mobile phase additives may also function as ion pair reagents, which interact with the ionized analytes to form overall neutral eluting species and also suppress silanophilic interactions between free silanol groups on the silica surface and basic functional groups of the analytes. This property determines the suitability of a particular mobile phase additive for use with ESI MS, since strong ion pair interactions between analytes and mobile phase additives can suppress the ionization of the analytes. As noted above, HP-RPC can be operated in the isocratic, step gradient, or the continuous gradient elution mode, whereby the retaining mobile phase is aqueous and the eluting mobile phase is an organic solvent or an aqueous organic solvent mixture.

In terms of usage, due to its versatility and flexibility, HP-RPC techniques dominate the separation of secondary metabolites at the analytical, laboratory-scale, and preparative levels, since the majority of secondary metabolites possess some degree of hydrophobicity.[18,19] **Figure 3** shows the analysis of glycosylated flavonoids and other phenolic compounds using HP-RPC.[20]

9.02.2.2 High-Performance Normal-Phase Chromatography

HP-NPC can be performed on unmodified silica that separates analytes according to their intrinsic polarity. HP-NPC can also be operated in isocratic, step gradient, or gradient elution mode, where the retaining mobile

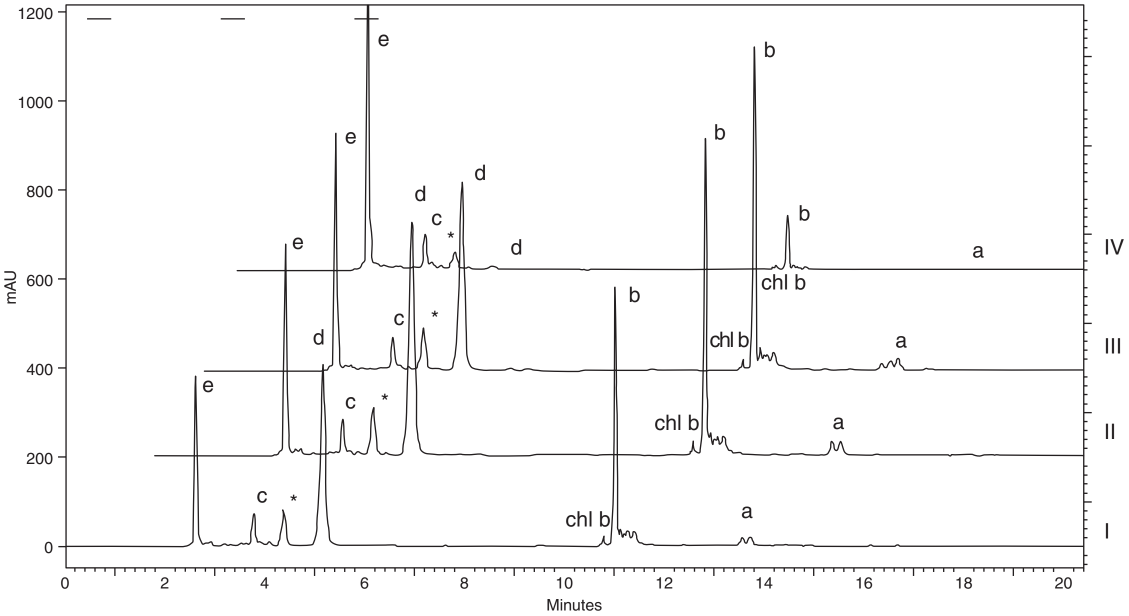


Figure 4 High-performance normal-phase chromatography using the four different solvent mixtures. I, 0h/100de; II, 20h/80de; III, 50h/50de; IV, 100h/0e (h, hexane; de, diethyl ether). a, Neoxanthin polar fraction; b, lutein polar fraction; c, internal standard; d, unknown (415, 435 nm); e, β-carotene; chl b, chlorophyll b (λ_{max} 435; 458 nm; 2nd D 458). Reproduced from M. M. Mendes-Pinto; A. C. S. Ferreira; M. B. P. P. Oliveira; P. Guedes de Pinho, *J. Agric. Food Chem.* **2004**, *52*, 3182–3188.

phase contains less polar organic solvents and the eluting mobile phase consists of more polar organic solvents and on occasions, water. **Figure 4** shows a separation of carotenoids in grapes using HP-NPC.[21]

9.02.2.3 High-Performance Ion-Exchange Chromatography

HP-IEX is performed on stationary phases with immobilized charged ligands and separates according to electrostatic interactions between the charged surface of the analyte(s) and the complementary charged surface of the sorbent. HP-IEX can be operated in isocratic, step gradient, or gradient elution mode. In high-performance anion-exchange chromatography (HP-AEX), analytes are separated according to their net negative charge, where the retaining mobile phase is aqueous, of high pH and low salt concentration and the eluting mobile phase is either aqueous, of high pH and high salt concentration, or aqueous, and low pH. In contrast, high-performance cation-exchange chromatography (HP-CEX) separates according to the net positive charge of the analytes, where the retaining mobile phase is aqueous, of low pH and low salt concentration and the eluting mobile phase is either aqueous, of low pH and high salt concentration, or aqueous, and high pH. With low-molecular-weight natural products, the anisotropy of the charge usually does not affect chromatographic resolution, although with conjugates and other metabolite derivatives charge distribution effects can influence the selectivity. **Figure 5** shows the isolation of ephedrine alkaloids and synephrine in dietary supplements using HP-IEX.[22]

9.02.2.4 High-Performance Hydrophilic Interaction Chromatography

HP-HILIC is performed on porous stationary phases with immobilized hydrophilic ligands and separates analytes according to their hydrophilicity. HP-HILIC can be operated in isocratic, step gradient, or gradient elution mode, where the retaining mobile phase is organic and the eluting mobile phase is aqueous. Being more suited to the isolation of polar substances, HP-HILIC, when linked to electrospray MS, has mainly found application for evaluation of polar compound mixtures for drug discovery. **Figure 6** shows an analysis of the polar components in a fermentation extract using HP-HILIC.

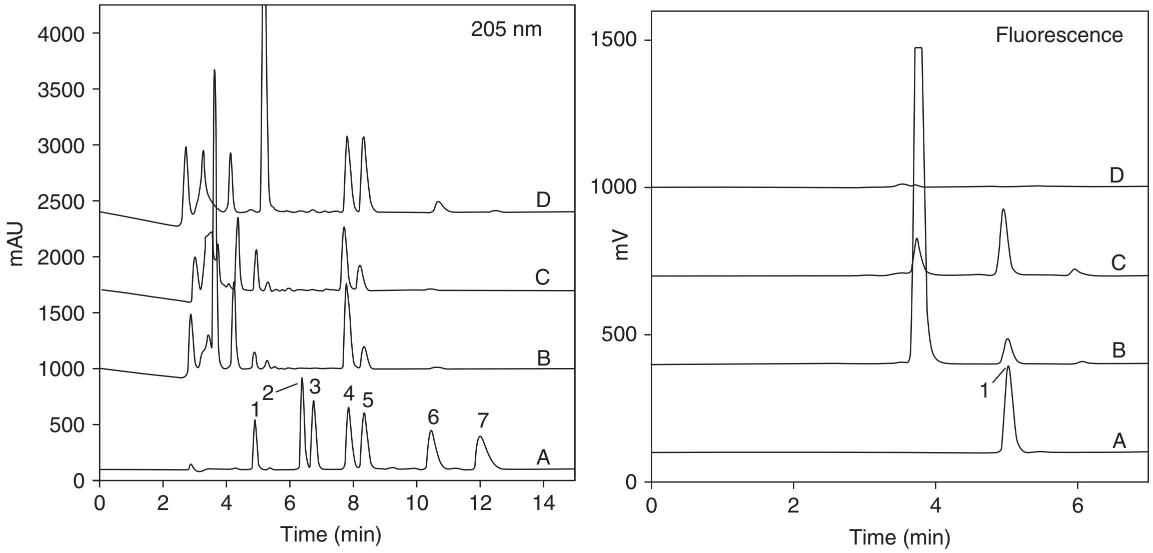


Figure 5 High-performance strong cation-exchange chromatography of ephedrine alkaloids and synephrine in dietary supplements. UV and fluorescence chromatograms of (A) standard solution (~3 μg per component, except 1.6 μg of component 1), (B) Xenadrine RFA-1, (C) Xetalean, and (D) Ultra Diet Pep. Analytes: 1, (±)-synephrine; 2, (−)-norephedrine; 3, (+)-norpseudoephedrine-HCl; 4, (−)-ephedrine-HCl; 5, (+)-pseudoephedrine; 6, (−)-*N*-methylephedrine; 7, (+)-*N*-methylpseudoephedrine. Reproduced from: R. A. Niemann; M. L. Gay, *J. Agric. Food Chem.* **2003**, *51*, 5630–5638.

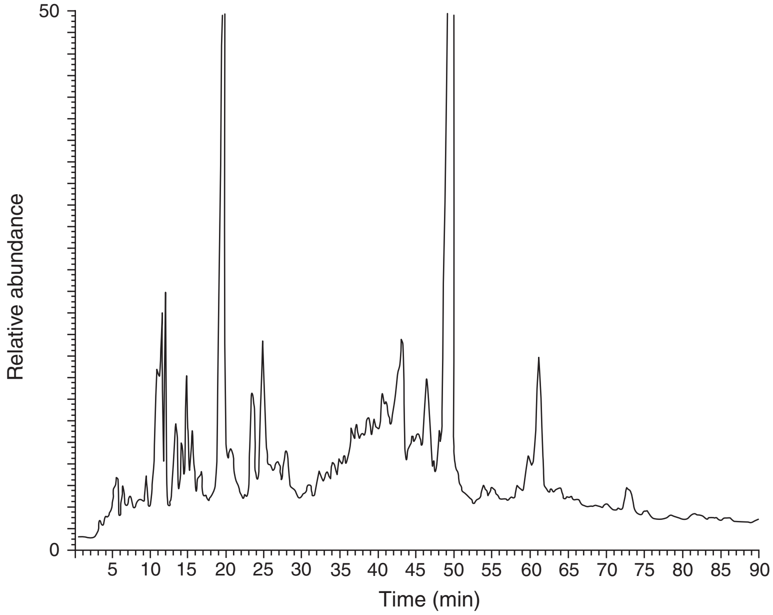


Figure 6 HILIC-ESI-MS separation of the polar components (unretained by reversed-phase solid-phase extraction pretreatment) of a fermentation extract, represented as a total ion chromatogram in positive ion ESI. The chromatogram was obtained using a TSKGel Amide 80 packing, 6.5 mmol l^{-1} ammonium acetate pH 5.5-buffered mobile phases, and a 90 min 10–40% aqueous gradient. Reproduced from M. A. Strege, *Anal. Chem.* **1998**, *70*, 2439–2445.

9.02.2.5 High-Performance Size Exclusion Chromatography

HP-SEC, also called high-performance gel-permeation chromatography (HP-GPC), is performed on porous stationary phases and separates analytes according to their molecular mass or their hydrodynamic volume. As a nonretentive separation mode, HP-SEC is usually operated with isocratic elution using aqueous low salt mobile

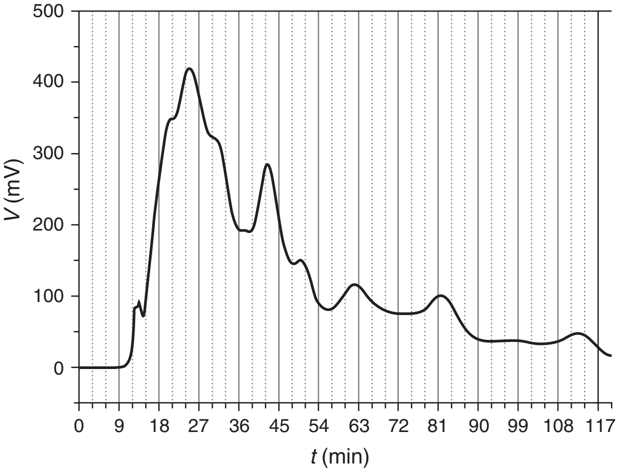


Figure 7 One-dimensional chromatogram of a concentrated Qingkailing injection. SEC; column: Toyopearl HW-40S, 300 mm × 8 mm ID; mobile phase: 0.05 mol l^{-1} Tris-HCl (pH 6.90); flow rate: 0.4 ml min^{-1}; injection volume: 20 μl; detection: UV 254 nm. Reproduced from S. Ma; L. Chen; G. Luo; K. Ren; J. Wu; Y. Wang, *J. Chromatogr. A* **2006**, *1127*, 207–213.

phases. The separation of analytes of different sizes is based on the concept that molecules of different hydrodynamic volume (Stokes radius) permeate to different extents into porous HP-SEC separation media and thus exhibit different permeation coefficients according to differences in their molecular masses/hydrodynamic volumes. Figure 7 shows an analysis of the complex Traditional Chinese Medicine (TCM), Qingkailing using HP-SEC.[23]

9.02.2.6 High-Performance Affinity Chromatography

HP-AC is performed on stationary phases containing immobilized biomimetic or biospecific ligands and separates according to principles of molecular recognition. Into this category can also be included the resolution of enantiomeric compounds by chiral HPLC. In HP-AC, analytes are usually eluted by step gradient or gradient elution, where the capture (loading) mobile phase is aqueous and of low ionic strength and the eluting mobile phase is aqueous and of higher ionic strength or different pH value or alternatively, contains a mobile phase additive that competes with the target compound for binding to the immobilized biospecific ligand. For chiral HPLC, organic solvent mixtures usually form the mobile phase eluents. In both HP-AC and chiral HPLC, separations can be performed with immobilized chemical or biological ligands, as well as with molecularly imprinted polymers (MIPs). In terms of achieving maximal selectivity and highest affinity for the interaction between the target substance(s) and the chromatographic sorbent, HP-AC provides the greatest gain as a trade-off between flexibility and versatility, that is, each affinity sorbent has to be tailored to the specific compound. Nevertheless, HP-AC has found application in natural product isolation. Figure 8 shows the isolation of scoparone (6,7-dimethoxycoumarin) and capillarisin, an extract from *Artemisia capillaris* using this technique.[24]

9.02.2.7 Summary

In order to achieve optimal selectivity and hence resolution of natural products in HP chromatographic separation, irrespective of whether the task at hand in analytical or preparative, the choice of the chromatographic mode must be guided by the properties of the analytes (i.e., their hydrophobicity/hydrophilicity, charge, molecular size). These attributes can often be rationalized in terms of structural features, solubility profiles, and source. A more detailed overview of existing methods that have been applied for seven classes of natural products is given in Section 9.02.5.

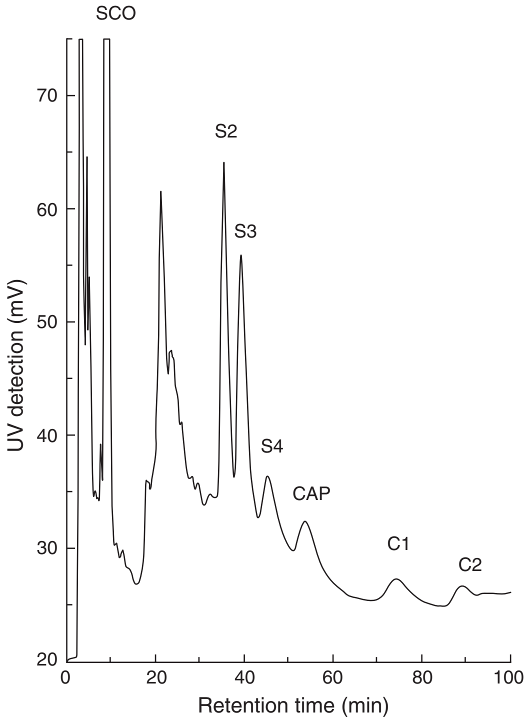


Figure 8 Affinity chromatography for the analysis of the methanol extract of *Artemisia capillaris* under optimized conditions. SCO, scoparone; CAP, capillarisin. Experimental conditions: the column used was of 150 mm × 4.6 mm ID packed with HSA immobilized on silica (7 μm), column temperature was 35 °C, flow rate was 0.8 ml min^{-1}, and UV detection wavelength was set to 238 nm. Initially, 10 min isocratic elution with mobile phase of 1.5% acetonitrile in 10 mmol l^{-1} phosphate buffer (pH 6.0); then, 5 min linear gradient elution from 1.5 to 12% acetonitrile in 10 mmol l^{-1} phosphate buffer (pH 6.0) with the elution of the latter mobile phase kept for an additional 45 min; and finally, another 45 min linear gradient elution from 12% acetonitrile in 10 mmol l^{-1} phosphate buffer (pH 6.0) to 15% acetonitrile in 10 mmol l^{-1} phosphate buffer (pH 7.4). Reproduced from H. L. Wang; H. F. Zou; J. Y. Ni; L. Kong; S. Gao; B. C. Guo, *J. Chromatogr. A* **2000**, *870*, 501–510.

9.02.3 From Analytical to Preparative Scale Illustrated for HP-RPC

As noted above, HP-RPC is currently the most frequently used HP liquid chromatographic mode for the analysis and preparative purification of secondary metabolites, in particular for applications that involve off-line or online ESI MS.

The development of a method for preparative HP-RPC purification for the purpose of isolation of one or more component(s) from a natural product sample (or alternatively the purification of a synthesized product from natural occurring precursors) is usually performed in four steps: (1) development, optimization, and validation of an analytical method, (2) scaling up of this method to a preparative chromatographic system, (3) application of the preparative method to the fractionation of the product, and (4) analysis of the individual fractions.

9.02.3.1 Development of Analytical Method

The development of an analytical method for the separation of a natural product encompasses the selection of the stationary and mobile phase taking into consideration the analyte properties (hydrophobicity/hydrophilicity, acid–base properties, charge, temperature stability, molecular size) and is followed by a systematic optimization of the (isocratic or gradient) separations, using either aliquots of the crude extract or, if available, analytical standards.

In the selection of the stationary and mobile phase, a variety of chemical and physical factors of the chromatographic system that may contribute to the variation in the resolution and recovery of natural products need to be considered. The stationary phase contributions relate to the ligand composition, ligand density, surface heterogeneity, surface area, particle size, particle size distribution, particle compressibility, pore diameter, and pore diameter distribution. The mobile phase contributions relate to the type of organic solvents, eluent composition, ionic strength, pH, temperature, loading concentration, and volume.

Typically, a particular HP-RPC material will be selected empirically as the starting point for the separation, taking into consideration its suitability for the separation task at hand, published procedures for similar types of natural products, availability of the stationary phase material for preparative chromatography, and if information is available, on the analyte properties.

Since the quality of a separation is determined by resolution of individual peak zones, method development always aims at optimization of the resolution. The resolution of adjacent peak zones for a two-analyte system can be defined as follows:

$$R_S = \frac{t_{R2} - t_{R1}}{(1/2)(w_1 + w_2)} \tag{1}$$

where t_{R1} and t_{R2} are the retention times, while w_1 and w_2 are the peak widths of two adjacent peaks corresponding to the analytes. To develop good resolution in the analytical separation of a complex mixture of natural products, method development always focuses on the least well-resolved peak pair(s) of interest.

In isocratic elution, resolution depends on the column efficiency or plate number N, the selectivity α, and the retention factor k, all of which can be experimentally influenced through systematic changes in individual chromatographic parameters. In the isocratic mode of separation, resolution is determined from

$$R_S = \left(\frac{1}{4}\right) N^{1/2} (\alpha - 1) \left(\frac{k}{(1 + k)}\right) \tag{2}$$

The plate number, N, is the efficiency of the column and is a measure of the column performance. The selectivity α describes the selectivity of a chromatographic system for a defined peak pair and is the ratio of the k values of the second peak to the first peak. The retention factor k is a dimensionless parameter and is defined as $k = (t_R - t_0)/t_0$, where t_R is the retention time of a particular peak and t_0 is the column void time. In this manner, normalization of the relative retention can be achieved for columns of different dimensions. While N and α change only slightly during the solute migration through the column, the value of k can be readily manipulated through changes in the elutropicity of the mobile phase by a factor of 10 or more. The best chromatographic separations for low- or mid-molecular-weight analytes are generally achieved with mobile phase–stationary phase combinations that result in a k value between 1 and 20.

In gradient elution, in contrast to isocratic elution, $\overline{N}$, $\overline{\alpha}$, and $\overline{k}$ are the median values for N, α, and k, since they change during the separation as the shape and duration of the gradient changes. The 'gradient' plate number $\overline{N}$ has no influence on the selectivity or the retention (except for temperature change). The selectivity $\overline{\alpha}$ and the retention factor $\overline{k}$ usually have only a minor influence on $\overline{N}$. While $\overline{N}$ and $\overline{\alpha}$ change only slightly during the solute migration through the column, the $\overline{k}$ value can change by a factor of 10 or more depending on the gradient steepness. Again, the best chromatographic separation is generally achieved with a $\overline{k}$ value between 1 and 20. Although resolution in isocratic and gradient elution is mainly influenced by the mobile phase variables α (or $\overline{\alpha}$) and k (or $\overline{k}$) and nearly independent of $\overline{N}$, for a given column, an optimization strategy should nevertheless start with appropriate selection of the stationary phase. This is because many initial choices (like column dimensions, choice of ligand) are determined by the overall strategy (i.e., separation optimization for quantification of several analytes or separation optimization for planned scaling up to preparative purification of specific target compounds) and by the purification goals. A number of computer-assisted, expert systems can be used to guide this selection, for example, for further insight into this field and the choice of different algorithmic expert systems approaches, see I *et al.*[25]

Based on these considerations and in view of the implication of Equation (2), the separation optimization for a natural product sample requires three steps to be performed: (1) the optimization of the column efficiency N, (2) the optimization of the selectivity α, and (3) the optimization of the retention factor $\overline{k}$ values.

Step 1: Optimization of the column efficiency N

The optimization of the peak efficiency, expressed as the theoretical plate number, N, requires an independent optimization of each of the contributing factors that influence the band broadening of the peak zones due to column and the extra-column effects. With a particular sorbent (ligand type, particle size, and pore size) and column configuration, this can be achieved through optimization of linear velocity (flow rate), the temperature, detector time constant, column packing characteristics, and by minimizing extra-column effects, by, for example, using zero dead volume tubing and connectors. The temperature of the column and the eluents should be thermostatically controlled in order to facilitate the reproducible determination of the various column parameters and to ensure resolution reproducibility. The flow rate (or alternatively the linear flow velocity) to achieve the minimum plate height, H, for a particular column can be taken from the literature or experimentally determined according to published procedures.

Step 2: Optimization of the selectivity α

Change in selectivity of the separation is the most effective way to influence resolution. This is mainly achieved by changing the chemical nature or concentration of the organic solvent modifier (ACN, MeOH, isopropanol, etc.) in conjunction with the appropriate choice of mobile phase additive(s). As noted above, this can be realized in both isocratic or gradient elution. Moreover, the interconversion of isocratic data to gradient data and vice versa can be achieved through the use of algorithms[25] based on linear and nonlinear solvent strength theory.

A good starting point is the solvent selectivity triangle approach (see **Figure 9**). Here, solvents are classified according to their relative dipole moment, basicity, or acidity. Combinations of three different solvents, plus water to provide an appropriate retention factor range, are selected to differ as much as possible in terms of their polar interactions. This selection permits the solvent combinations to mimic the selectivity that is possible for any given solvent, and defines the boundaries of the triangle.[26,27] At the same time, these solvents must be totally miscible with each other and with water. Three solvents, which best meet these requirements, are MeOH, ACN, and THF. Four solvent mobile phase optimization using water plus three organic solvents provides significant control over α-values in reversed-phase HPLC. However, if different organic solvents are used, different eluotropic strengths must be considered in order to allow elution of the analytes of the sample in the appropriate retention factor range.[28,29]

Once the selectivity parameter is fixed due to the initial choices of the mobile and stationary phase, further optimization should concentrate on resolution optimization via achieving the most appropriate retention factor for the different natural products in the mixture.

Step 3: Optimization of the retention factor $\overline{k}$ values

In the isocratic elution mode of HP-RPC, resolution optimization can take advantage of the relationship between the retention time of an analyte (expressed as the retention factor k) and the volume fraction of the

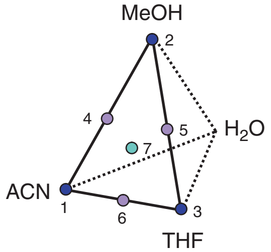


Figure 9 Solvent selectivity triangle approach for the selectivity optimization in HP-RPC. First, three initial experiments (1–3) with three mobile phases (binary mixtures of ACN/water, MeOH/water, and THF/water, respectively) are performed. If necessary, a further three experiments (4–6) with three mobile phase blends (ACN/MeOH/water, MeOH/THF/water, and THF/ACN/water) are performed. If necessary, a further experiment (7) with a mobile phase blend ACN/MeOH/THF/water is performed.

organic solvent modifier, φ. Although typically these dependencies are curvilinear, that is, not first order, for practical convenience they are often treated as linear relationships. Thus, the change in retention factor as a function of φ can be represented by

$$\ln k = \ln k_0 - S\varphi \tag{3}$$

where k_0 is the retention factor of the solute in the absence of the organic solvent modifier and S is the slope of the plot of $\ln k$ versus φ. The values of $\ln k_0$ and S can be calculated by linear regression analysis. Greater precision in the quality of fit of the experimental data, and thus improved reliability in the prediction of the retention behavior of analytes in HP-RPC systems for mobile phases of different solvent composition, can be achieved[30] through the use of an expanded form of Equation (3), that is,

$$\ln k = \ln k_0 - S\varphi + S'\varphi^2 - S''\varphi^3 + \cdots \tag{4}$$

Similarly, in gradient elution HP-RPC, resolution optimization can take advantage of the relationship between the gradient retention time of an analyte (expressed as the median retention factor $\overline{k}$) and the median volume fraction of the organic solvent modifier, $\overline{\varphi}$, in regular HP-RPC systems based on the concepts of the linear solvent strength theory,[15,31] such that

$$\ln \overline{k} = \ln k_0 - S\overline{\varphi} \tag{5}$$

A mapping of the dependence of analyte retention (expressed as the natural logarithm of the retention factor, k) on the mobile phase composition (expressed as the volume fraction of solvent in the mobile phase, φ) in isocratic elution (or as $\overline{k}$ versus $\overline{\varphi}$ in gradient elution) with a minimum of two initial experiments can be used to define the useful range of mobile phase conditions, and can indicate the mobile phase composition at which the band spacing is optimal (see **Figure 10**).

Irrespective of whether the data are obtained through isocratic or gradient elution, techniques employing two initial experiments (differing only by their mobile phase composition or gradient run times, respectively) with tracking and assignment of the peaks, a relative resolution map (RRM) can be established, which plots resolution R_S against the separation time (or gradient run time t_G). In the case of gradient elution, the RRM then allows determination of the optimal gradient run time (and gradient range). Such a procedure can be conveniently performed in any laboratory using Excel spreadsheets containing the relevant equations (see below) as macros or through software packages (e.g., DryLab, LabExpert). Such strategies greatly reduce the time to achieve an optimal separation, as well as saving on solvent, reagent, and analyte consumption. Moreover, as shown in various studies, the more sophisticated of these methods permit[24] instrumentation to be operated in a nearly fully automated, unattended fashion 24 h/7 days per week.

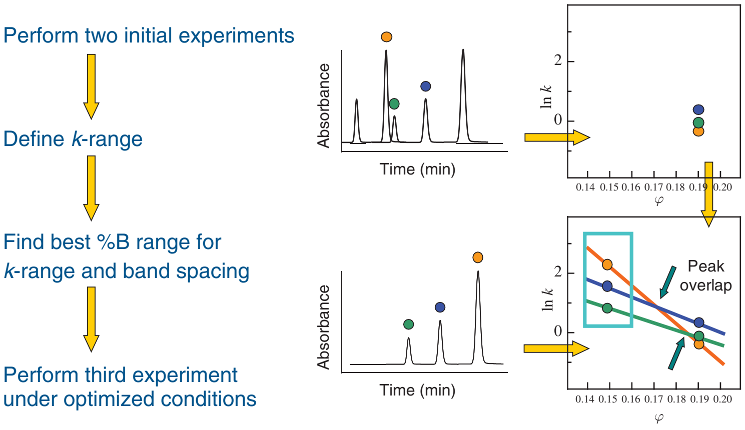


Figure 10 Optimization of isocratic elution. Two chromatograms obtained for 19 and 14% (v/v) of organic solvent modifier in the mobile phase (corresponding to $\varphi = 0.19$ and 0.14, respectively) can be used to plot the resulting retention factors versus the volume fraction of the organic solvent modifier in order to identify the mobile phase composition with optimal peak spacing.

In more advanced applications, optimization can be performed via computer simulation software (e.g., Simplex methods, multivariate factor analysis programs, DryLab G/plus, LabExpert). In such procedures, resolution of peak zones is optimized through systematic adjustment of mobile phase composition by successive change in the φ value (or equivalent parameters, such as the concentration of the ion pairing reagent employed). In gradient elution, advantage is taken of a strategy with the following eight steps: (1) performing of initial experiments, (2) peak tracking and assignment of the peaks, (3) calculation of ln k_0 and S values from initial chromatograms, (4) optimization of gradient run time t_G over the whole gradient range, (5) determination of new gradient range, (6) calculation of new gradient retention times t_g, (7) change of gradient shape (optional), and (8) verification of results.

(1) Initial experiments: In initial experiments, the natural product sample is separated using two linear gradients differing by a factor of three in their gradient run times (all other chromatographic parameters being held unchanged) to obtain the HP-RPC retention times for each of the natural product compounds.[32] Irrespective of what optimization strategy will be used, it is advisable to separate any sample with at least two different gradient run times, in order to identify overlapping peaks. For optimization of gradient shape and to achieve maximum resolution between adjacent peak zones, the ability to determine retention times of the natural compounds and to classify the parameters that reflect the contributions from the mobile phase composition and column dimensions is essential.[33–35] The determination of the volume, V_mix, is useful,[36] determination of dead volume and gradient delay are crucial.[31] With various inputs regarding stationary and mobile phase parameters, algorithms, for example, that of DryLab G/plus, can generate the RRM, based on calculation of corresponding S and k_0 values for each component. If no computer program is available the resolution information can be plotted directly from the distances of the individual peak zones of adjacent peak pairs ($R_\mathrm{S}= t_\mathrm{R2} - t_\mathrm{R1}$) against the gradient run time t_G.

(2) Peak tracking and assignment of the peaks: Complex chromatograms from reversed-phase gradient elution can often exhibit changes in peak order when the gradient steepness is changed. Before ln k_0 and S values are calculated, or computer simulation is used, the peaks from the two initial runs need to be correctly assigned. Several approaches to peak tracking have been described, using algorithms based on relative retention and peak areas,[37] or alternatively, based on DAD.[38,39]

(3) Calculation of ln k_0 and S values: The retention times t_g1 and t_g2 for a solute separated under conditions of two different gradient run times (t_G1 and t_G2, where $t_\mathrm{G1} < t_\mathrm{G2}$) can be given by the following equations:[33,40]

$$t_\mathrm{g1} = \left(\frac{t_0}{b_1}\right)\log\left(2.3k_0b_1\right) + t_0 + t_\mathrm{D} \tag{6a}$$

$$t_\mathrm{g2} = \left(\frac{t_0}{b_2}\right)\log\left(2.3k_0b_2\right) + t_0 + t_\mathrm{D} \tag{6b}$$

with

$$\frac{b_1}{b_2} = \frac{t_\mathrm{G2}}{t_\mathrm{G1}} = \beta \tag{7}$$

where t_G1,t_G2 are the gradient run time values of t_G for two different gradient runs, resulting in different values of $b(b_1,b_2)$, and $t_\mathrm{g}\left(t_\mathrm{g1},t_\mathrm{g2}\right)$ are the gradient retention times for a single solute in two different gradient runs; b_1,b_2 are the gradient steepness parameters for a single solute over the two differing gradient run times; k_0 is the solute retention factor at the initial mobile phase composition; β is the ratio of t_G2 and t_G1, which is equivalent to the ratio of b_1 and b_2; t_0 is the column dead time; and t_D is the gradient delay time. Steep gradients correspond to large b values and small $\overline{k}$ values.

For small molecules there is an explicit solution[40] for b and k_0, namely,

$$b_1 = \frac{t_0\log\beta}{\left[t_\mathrm{g1} - \left(\frac{t_\mathrm{g2}}{\beta}\right) + \left(t_0 + t_\mathrm{D}\right)\left(\frac{(1-\beta)}{\beta}\right)\right]} \tag{8}$$

$$\log k_0 = \left(\frac{b_1}{b_2}\right)\left(t_{g1} - t_0 + t_D\right) - \log(2.3b_1) \tag{9}$$

From the knowledge of b and k_0 the values of $\overline{k}$ and $\overline{\varphi}$ can be calculated.[31]

$$\bar{k} = \left(\frac{1}{1.15\, b_1}\right) \tag{10}$$

$$\overline{\varphi} = \frac{\left[t_{g1} - t_0 - t_D - \left(\frac{t_0}{b_1}\right)\log 2\right]}{t^0_{G1}} \tag{11}$$

where $\overline{k}$ is the value of k (retention factor) for a solute when it reaches the column midpoint during elution, φ the volume fraction of solvent in the mobile phase, $\Delta\varphi$ the change in φ for the mobile phase during gradient elution ($\Delta\varphi = 1$ for a 0–100% gradient), $\overline{\varphi}$ the effective value of φ during gradient elution and the value of φ at band center when the band is at the midpoint of column, and t^0_{G1} the normalized gradient time with $t^0_{G1} = t_{G1}/\Delta\varphi$.

By linear regression analysis, using $\overline{k}$ and $\overline{\varphi}$, the S value (empirically related to the hydrophobic contact area between solute and ligand) can be derived from the slope of the $\log \overline{k}$ versus $\overline{\varphi}$ plots, and $\ln k_0$ (empirically related to the affinity of the solute toward the ligand) as the y-intercept.[15]

$$S = \frac{\left(\ln k_0 - \ln \overline{k}\right)}{\overline{\varphi}} \tag{12}$$

(4) Optimization of the gradient time, t_G, over the entire gradient range: The retention factor $\overline{k}$ is a linear function of the gradient run time t_G if $\Delta\varphi$ is kept constant. Hence,

$$\frac{\overline{k}}{t_G} = \frac{0.87F}{V_m \times \Delta\varphi \times S} = \text{const.} = C \tag{13}$$

The optimized gradient run time t_{GRRM} can be obtained from the RRM or alternatively, from the plot of R_S versus t_G, and yields for each analyte the new values of $\overline{k}_{new}$ by t_{GRRM} being multiplied with C:

$$Ct_{GRRM} = \overline{k}_{new} \tag{14}$$

(5) Determination of the new gradient range: If the gradient run time t_{GRRM} is changed in relation to $\Delta\varphi$ with $t^0_{G1} =$ const., the $\overline{k}$ values do not change, as can be seen from the following equation:

$$t^0_{G1} = \frac{t_{GRRM}}{\Delta\varphi} = \frac{V_m S \overline{k}}{0.87 \times F} \tag{15}$$

where

$$\Delta\varphi_{opt} = \frac{t_{Gopt}}{t^0_{G1}} \tag{16}$$

and where the retention time t_g of the first peak is $> (t_0 + t_D)$ and the retention time t_g of the last peak is $< \Delta t_{Gopt}$.

(6) Calculation of the new gradient retention times t_g: Based on the knowledge of the S and the $\ln k_0$ values, new gradient retention times can then be calculated.

(7) Change of gradient shape (optional): Multisegmented gradients should only be used once the gradient delay has been measured. With multisegmented gradients, an error in the gradient delay will reoccur at the beginning and end of each gradient step. In addition, the effect of V_{mix} (which can be determined according to the procedures described in Ghrist *et al.*,[33] which modifies the composition of the gradient at the start and end (rounding of the gradient shape)) can lead to deviation of the experimentally determined retention times from the predicted 'ideal' values as derived, for example, with DryLab G/plus simulations.

(8) Verification of the results: After completion of the optimization process, the simulated chromatographic separation can now be verified experimentally using the predicted chromatographic conditions.

Examples where such systematic method development has been used for the analytical separation of natural products can be found in studies performed on flavones,[41] flavonoids,[42] and secoiridoids.[43] Computer-assisted method development was applied to anthranoids,[44] carotenoids,[45] coumarins,[46,47] flavonoids,[48,49] and other natural products.[50]

9.02.3.2 Scaling Up to Preparative Chromatography

While analytical HPLC aims at the quantification and/or identification of compounds (with the sample going from the detector to waste), preparative chromatography aims at the isolation of compounds (with the sample going to the fraction collector). For preparative separations, method development always focuses on the peaks of interest, and the two adjacent eluting peaks. In many cases, all other peaks can be viewed as superfluous, and directed to the waste. Optimization of the resolution of the peak of interest from the adjacent peaks has to take into account the sample size and the relative abundances of the three components that form the basis of the separation task. Once an analytical method is established, it can be scaled up[8] to a preparative separation by taking into consideration the operating ranges of the column (see **Table 1**) or used for scaling up by deliberate column overloading.

The concept of parity in scaling up or down implies that the performance features, selectivity behavior, and recyclability of the stationary phase material used for the analytical and the preparative separation are identical, with the exception of particle size. Both robust experimental methods as well as rules of thumb, acquired by experienced investigators, have been developed that enable such comparisons to be made. An extensive scientific literature is now available to indicate sound foundations for such scaling up strategies, coupled with suitable experimental methods for their validation. **Table 1** summarizes some of this information.

In order to obtain an equivalent elution profile, the flow rate needs to be adjusted for columns with different internal diameter, according to the following formula:

$$F_{\text{preparative}} = \left[\frac{r_{\text{preparative}}}{r_{\text{analytical}}}\right]^2 \times F_{\text{analytical}} \tag{17}$$

where F is the flow rate and r the column radius of the preparative or analytical column.

Estimates of the loading capacity of a particular column material can be usually obtained from the manufacturer. The mass loadability for a scaled up separation can be calculated with the following formula:

$$M_{\text{preparative}} = \left[\frac{r_{\text{preparative}}}{r_{\text{analytical}}}\right]^2 \times M_{\text{analytical}} \times C_{\text{L}} \tag{18}$$

where M is the mass, r the column radius of the preparative or analytical column, and C_{L} the column length ratio.

In many cases, despite some loss of resolution, column overloading is an economic and viable method for compound purification. In analytical LC, the ideal peak shape is a Gaussian curve. If under analytical conditions a higher amount of sample is injected, peak height and area change, but not peak shape or the retention factor. However, if more than the recommended amount of sample is injected onto the column, the adsorption isotherm becomes nonlinear. As a direct consequence, resolution decreases, and peak retention

Table 1 Operating ranges of column types in HPLC with sample amount, inner column diameter, column length, and flow rate range

Column type	*Sample amount*	*Column ID (mm)*	*Column length (mm)*	*Flow rate range (ml min^{-1})*
Preparative	mg–g	>4	15–250	5–20
Analytical	µg–mg	2–4	15–250	0.2–1
Capillary	µg	1	35–250	0.05–0.1
Nano	ng–µg	<1	50–150	<0.05
Chip	ng	<0.1	<50	<0.01

times and peak shapes may change. There are two methods of column overloading: concentration overloading and volume overloading. In concentration overloading, the volume of the injected sample is maintained, while the sample concentration is increased. The retention factors of the compound(s) decrease, and the peak shape may become triangular and fronting. The applicability of this method is limited by the solubility of the target compound(s) in the mobile phases employed. In volume overloading, the concentration of the sample is maintained, but the sample volume is increased. The retention factor of the compound(s) decrease(s), with broadened peak shape. Once a suitable method is established, it can be applied to the preparative purification of the target compound(s).

9.02.3.3 Fractionation

There are four types of fraction collection:[51] (1) manual, with a manually pressed button to start and stop collection, (2) time based, with a fraction collecting during fixed preprogrammed time intervals, (3) peak based, based on a chosen threshold of the up- and down-slope of a detector signal, and (4) mass based, with fraction collection occurring only if the specific mass of a trigger ion is detected by MS. In addition, a recovery collection can be performed, by which everything that is not collected as a fraction goes into a dedicated container where it can be easily recovered. Whatever the type of fraction collection, careful attention has to be given to the fraction collection delay times and a delay time measurement performed. For a peak with start time t_0 and end time t_{E}, fraction collection needs to be started when the start of the peak arrives at the diverter valve ($t_0 + t_{\mathrm{D1}}$) and ended when the end of the peak arrives at the needle tip ($t_{\mathrm{E}} + t_{\mathrm{D1}} + t_{\mathrm{D2}}$), where t_{D1} is the delay time between detector and valve and t_{D2} the delay time between valve and needle tip.

9.02.3.4 Analysis of the Quality of the Fractionation

In the absence of online MS, fractionation is usually accompanied by an off-line mode of quality analysis. The typical workflow comprises of (1) a prepreparative analysis of the unpurified material, (2) the purification/fractionation of the compound, and (3) a postpreparative analysis of the individual fractions. The pre- and postpreparative analysis can be performed with analytical HPLC, MS, and activity testing of the fractions, if an assay is available.

After the fractions have been collected, the solvent needs to be removed by using a freeze-dryer, rotary evaporator, or high-throughput parallel evaporator. Nonvolatile components can be removed with reversed-phase SPE procedures prior to solvent removal if the aqueous portion of the buffer is sufficiently large.

9.02.4 Multidimensional High-Performance Liquid Chromatography

Although HPLC is a powerful separation technique for the fractionation of natural products from complex biological mixtures, very often more then one chromatographic step is necessary to achieve a required degree of purity of the target compounds. In practice, this is achieved through a series of purification steps. As there are material losses associated with each purification step in these procedures (see **Figure 11**), the overall recovery of the product has to be optimized. This can be achieved if the number of employed purification steps is minimized. Thus, strategies and techniques that reduce the number of unit operations are to be preferred since they will lead to minimization of product loss(es), and save on capital costs for equipment or operational costs for staff, reagents, and other consumables.

After initial extraction, the enrichment and purification of the target compound can be typically achieved in two or three chromatographic dimensions using a combination of different chromatographic modes. For natural products, ion-exchange chromatography-reversed-phase chromatography (IEX-RPC), size exclusion chromatography-reversed-phase chromatography (SEC-RPC), normal-phase chromatography-reversed-phase chromatography (NPC-RPC), affinity chromatography-reversed-phase chromatography (AC-RPC) combinations are described in the literature.

Multidimensional (multistage, multicolumn) high-performance liquid chromatography (MD-HPLC) offers the possibility of cutting the elution profiles into consecutive fractions, where these fractions can be treated

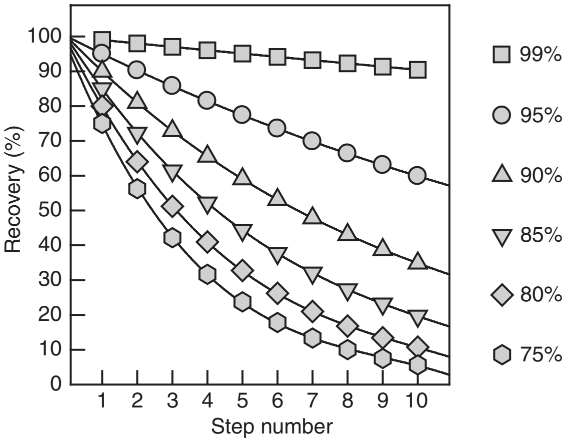


Figure 11 Effect on overall recovery at a fixed recovery per step value due to the additional steps in multidimensional chromatography.

independently from each other. One important consequence of this strategy is the gain in peak capacity (PC), defined as the number of peaks that can be accommodated between the first and the last peak in a separation of defined resolution.[52] Multidimensional LC has developed from column switching and related techniques[53] for a specific target or class of targets.[54] MD-HPLC has the potential of independent optimization of the separation conditions for each fraction and allows a relative enrichment/depletion/peak compression of components. An advanced conceptual framework of multidimensional LC has been developed for small molecule separations.[55–59]

MD-HPLC can be applied to the purification of a particular natural product or comprehensive fractionation of complex natural product mixtures.[60] **Figure 2** illustrated an example where such approaches have been employed. The two applications have some similarities in terms of basic separation science principles; however, they have fundamental differences in system design, optimization, and operation.

9.02.4.1 Purification of Natural Products by MD-HPLC Methods

To purify a particular natural product, it is often possible to select complementary chromatographic modes that allow the natural product to be obtained in high purity with only a few separation steps. In such noncomprehensive MD-HPLC, only a part of the analytes (as a single fraction) eluting from the first column is transferred to a second column for further purification (conventionally expressed by a hyphen, i.e., IEX-RPC). For such a 'heart-cutting' technique, knowledge of the retention properties of the analyte mixture in the first column is needed in order to choose the segment(s). Heart-cutting techniques are fast but not comprehensive, since the majority of analytes are not subjected to a separation in a second dimension. The main advantage of the technique is the improved resolution of compounds that coelute in the first dimension. A key requirement for such a purification scheme is that subsequent stages of the separation are orthogonal, with the two separation modes not correlated to each other in relation to their retention characteristics (i.e., selectivity).

For a single chromatographic dimension, the partly contradictory objectives of speed, resolution, capacity, and recovery usually cannot be maximized simultaneously (i.e., a high resolution can be achieved but at the expense of speed; a high-speed separation can reduce resolution) (see **Figure 12**).

MD-HPLC allows meeting the overall purification objectives, by placing the emphasis for each purification stage on a different pair of objectives and choosing a chromatographic mode that is particularly well suited for the task (see **Figure 13**). At the enrichment stage, the emphasis is on speed and capacity, employing, that is, strong cation exchange (SCX) or AC in the SPE format as a low-resolution step. At the intermediate purification stage, emphasis is placed on capacity and resolution, employing chromatographic modes with intermediate resolution, that is, ion-exchange chromatography (IEX) or SEC. At the final chromatographic polishing step, the emphasis is placed on resolution and recovery, employing high-resolution modes, that is, RPC.

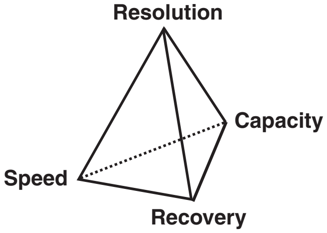


Figure 12 Optimization goals for a chromatographic purification and their interrelationship.

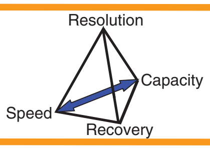
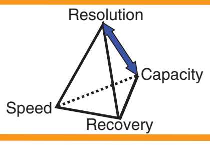
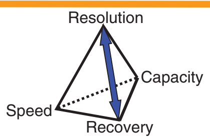
Purification stage	Separation mode	Priority
Enrichment (capture) stage: Rapid, high capacity, low resolution	IEX, SEC, AC	
Intermediate purification stage: High capacity, high resolution, intermediate speed and recovery	IEX, NPC, SEC	
Final purification stage: High resolution and recovery, low capacity and speed	RPC	

Figure 13 Optimization priorities at individual purification stages for a multidimensional fractionation exemplified for natural products and suitable chromatographic modes.

9.02.4.2 Fractionation of Complex Natural Product Mixtures by MD-HPLC Methods

If the objective of a purification scheme is the comprehensive fractionation of a complex, multicomponent natural product mixture, it is of advantage to use orthogonal chromatographic modes, but such extensive fraction collection requires additional, sometimes substantial, infrastructure. In comprehensive MD-HPLC, the entire analyte pool of the first column is transferred to the second column (expressed by a cross, i.e., IEX × RPC) as sequential aliquots, either successively onto one column or alternating onto two parallel columns. The resulting data can be represented as three-dimensional (3D) contour plots, with retention times of the second dimension plotted against retention times of the first dimension. The information content of such comprehensive 2D chromatograms is higher than the information content of individual one-dimensional chromatograms.

The first comprehensive 2D system was developed in the late 1970s by Erni and Frei, who applied IEX × RPC to the analysis of senna glycosides from plant extracts.[61] In the subsequent decades, comprehensive MD-HPLC methods have been further developed, mainly for peptides and proteins,[62,63] but also for separation of various natural products such as phenolic and flavone antioxidants[64] and carotenoids.[65] The theoretical aspects of MD-HPLC techniques have also been further developed.[66–68]

9.02.4.3 Operational Strategies for MD-HPLC Methods

From an operational perspective, multidimensional LC can be carried out off-line or online.[69] Regardless of which operational mode, off-line or online, is used, the compatibility of the mobile phases between successively employed chromatographic modes in a separation scheme needs to be considered (see Section 9.02.4.4.2). As a consequence, it may be necessary to process the fractions between two separation stages (e.g., through buffer exchange, concentration, or dilution) to enhance compatibility of eluent composition of fractions from the first chromatographic dimension with the retaining mobile phase of the second chromatographic dimension. If a nonretentive chromatographic mode such as SEC is employed in conjunction with a retentive chromatographic mode, such as RPC or IEX, it is usually performed first. This allows (1) relatively large eluent volumes stemming from isocratic elution in the nonretentive mode to be reduced through the capture of analytes under the retaining mobile phase conditions of the subsequent retentive chromatographic mode and (2) reduction of extra-column band broadening with resulting loss of resolution.

9.02.4.3.1 Off-line coupling mode for MD-HPLC methods

The off-line coupling mode in HPLC is comparable with that employed in conventional CC in natural product isolation. In the off-line mode, the eluent of the first column is collected as fractions, either manually or with an automated fraction collector, and reinjected onto the second column. Typical processing steps may include volume reduction by freeze-drying or automated high-throughput parallel evaporation systems taking into account the boiling point(s) or volatility of the target analyte(s) and organic solvent if these are contained within the eluates. The use of volatile mobile phase additives then allows a buffer exchange.

9.02.4.3.2 Online coupling mode for MD-HPLC methods

The online mode uses high-pressure, multiposition, multiport switching valves, which allow selection of pathways for single fractions from the first chromatographic dimension to subsequent column(s) of the second chromatographic dimension. The fractions from the first dimension are either transferred directly, or through one (or more) intermediate trapping columns for the purpose of concentration and automated buffer exchange. This approach requires complex instrumentation, results in increased optimization time and reduced system flexibility. It however has numerous advantages in terms of reproducibility, recovery, speed, and automation.

9.02.4.4 Design of an Effective MD-HPLC Scheme

MD-HPLC for natural products requires thoughtful selection of orthogonal and complementary separation modes, of the order of their utilization and independent optimization with respect to the chromatographic goals (speed, resolution, capacity, and recovery). Furthermore, besides the mobile phase composition of the employed chromatographic modes, the elution mode (isocratic, step, or gradient elution), flow rates, and mobile phase temperatures need to be considered.

9.02.4.4.1 Orthogonality of chromatographic modes

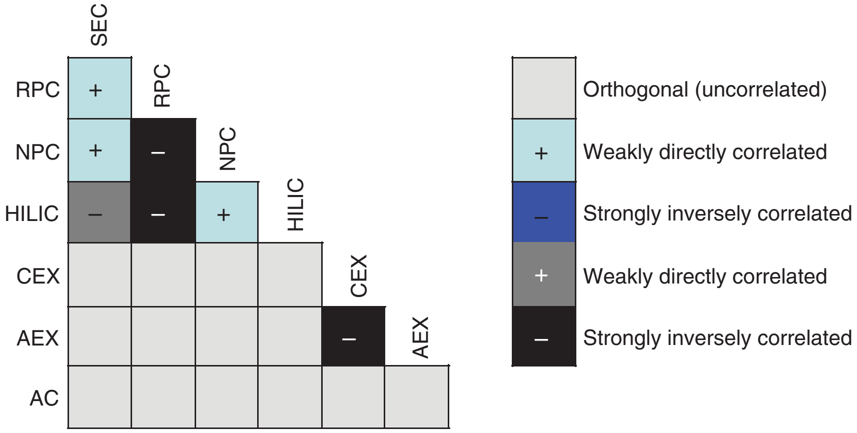
In order to exploit the full PC of a 2D system,[59] it is advantageous if the applied chromatographic modes are orthogonal. It is generally accepted that the dimensions in a 2D separation system are orthogonal, if the separation mechanism of the two dimensions are independent from each other causing the distribution of analytes in the first dimension to be uncorrelated to the distribution in the second dimension. An example of such orthogonality is liquid chromatography–capillary electrophoresis (LC–CE), where totally different separation mechanisms are used, that is, pressure-driven compared to electrically-driven separation.[70] In a similar manner, different separation modes in HPLC can be viewed as being orthogonal, for example, IEX (CEX or AEX) and RPC are orthogonal as they separate according to net charge or hydrophobicity, respectively. A very coarse classification of chromatographic modes commonly applied in the MD-HPLC of natural products according to their separation principles is depicted in **Figure 14**.

In MD-HPLC systems, combinations of chromatographic modes are usually designed to achieve analyte separation according to different characteristic analyte properties.[71]

For an ideal orthogonal, 2D separation, the overall PC is defined as the product of the PCs in each dimension:

$$\mathrm{PC}_{\mathrm{2D\ system}} = \mathrm{PC}_{\mathrm{first\ dimension}} \times \mathrm{PC}_{\mathrm{second\ dimension}} \tag{19}$$

However, if two nonidentical chromatographic modes with some degree of similarity are used in a 2D system, the increase in the PC and the total number of analytes that can be separated is much lower than the product of peak capacities of individual dimensions. The PC also depends on the elution mode. Gradient elution provides a higher PC than isocratic elution and is of advantage in 2D LC.



	SEC	RPC	NPC	HILIC	CEX	AEX
RPC	+					
NPC	+	–				
HILIC	–	–	+			
CEX						
AEX					–	
AC						

Orthogonal (uncorrelated)
+ Weakly directly correlated
– Strongly inversely correlated
+ Weakly directly correlated
– Strongly inversely correlated

Figure 14 Degree of orthogonality of chromatographic modes employed in the separation of natural products. Shading indicates the degree of correlation of the separation principles of paired modes.

It should be noted that since selectivity in chromatography depends not only on the stationary phase but on the mobile phase as well, orthogonal separations can be achieved through fine-tuning of the separation conditions, even if the principal separation mechanisms of both dimensions are similar. Such tuning removes the inaccessible area from the 2D retention plane and ensures that the remaining retention space is used efficiently.[66]

In addition, the structure of analytes has an effect on the PC. In many separation systems, the contribution of structural units, especially the repeating units, to the Gibbs free energy of association of the analytes with the immobilized chromatographic ligands are additive.[72] Such structural repeating units can be hydrophobic or polar. If one chromatographic system in a 2D LC has no selectivity for a structural element, then the first and the second dimension are noncorrelated (orthogonal) with respect to the repeating structural unit (see Figure 15(**a**)). In a completely correlated separation system, with correlated retention factors in two dimensions, the separation space is not utilized (see Figure 15(**b**)). Such 2D systems do not provide sufficient selectivity for the separation with respect to the structural property distribution of interest in either dimension and are generally not very useful in practice. In inversely correlated 2D LC × LC separation systems, the retention time increases in the first dimension, but decreases in the second dimension (see Figure 15(**c**)). Neither correlated nor inversely correlated 2D LC × LC increase the PC significantly. The selectivity of a 2D LC × LC with respect to hydrophobic or polar repeating units can determine the suitability of chromatographic modes employed in 2D separations and depends on the employed stationary as well as mobile phases. Orthogonal systems with noncorrelated selectivities provide the highest PC and therefore the highest number of resolved peaks. The PC in 2D LC × LC decreases with increasing correlation of the selectivity between the first and the second chromatographic dimension. In practice, however, 2D LC × LC systems are rarely fully orthogonal with respect to each structure distribution type (i.e., hydrophobic, polar).[67] Many partially orthogonal systems are using only part of the theoretically available 2D separation space but can be evaluated using analytes differing in the numbers of hydrophobic or polar structural units or by quantitative structure retention relationships (QSRR).

9.02.4.4.2 Compatibility matrix of chromatographic modes

In designing of 2D LC × LC systems, the selection of the mobile phase for each chromatographic dimension is of fundamental importance, in order to achieve maximal utilization of the 2D separation space. In contrast to off-line 2D LC procedures, where the collected fraction can be subjected to evaporation, dilution, or extraction, before injection onto the column of the second dimension, the compatibility of the mobile phases in online 2D LC × LC in terms of miscibility, solubility, viscosity, and eluotropic strength is much more important. The mobile phases used in SEC × RPC, SEC × NPC, RP × CEX, RP × AEX RP × CEX, NPC × HILIC, NPC × CEX, and NPC × AEX are compatible (see Figure 16).

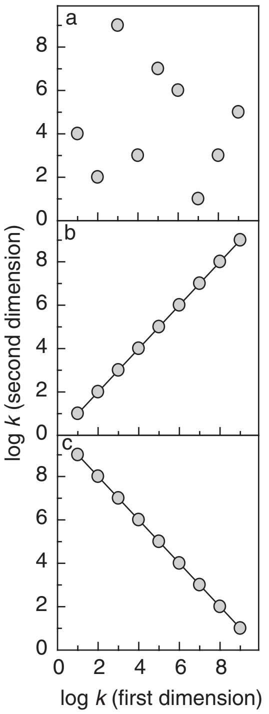
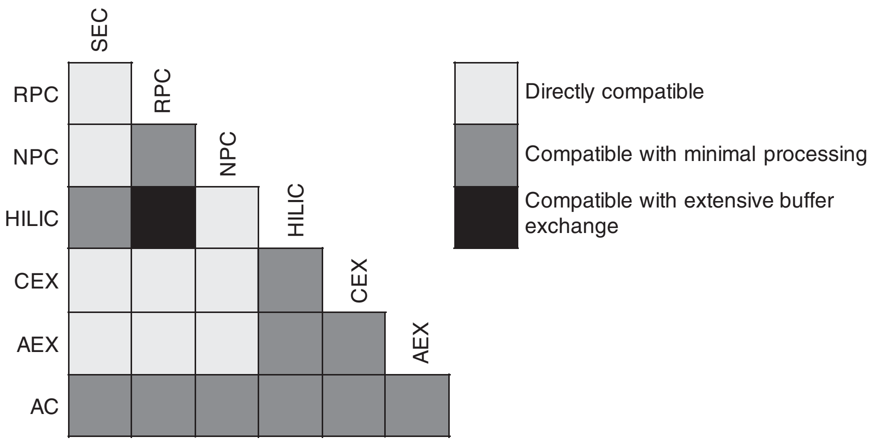


Figure 15 Two-dimensional separation space for a set of natural products utilizing separation systems that are (a) uncorrelated, (b) completely correlated, and (c) inversely correlated, where the retention factors obtained in the second dimension are plotted versus the retention factors obtained in the first dimension.



	SEC	RPC	NPC	HILIC	CEX	AEX
RPC	□					
NPC	□	▨				
HILIC	▨	■	□			
CEX	□	□	□	▨		
AEX	□	□	□	▨	▨	
AC	▨	▨	▨	▨	▨	▨

□	Directly compatible
▨	Compatible with minimal processing
■	Compatible with extensive buffer exchange

Figure 16 Pair-wise comparison of compatibility between common chromatographic modes. Compatibility is based on miscibility, solubility, and eluotropic strength for a particular class of natural products.

RPC and NPC are fully orthogonal separation modes, which can be useful for complex samples containing a mixture of analytes that are uncharged and possess different polarities or hydrophobicity. However, the use of RPC and NPC in 2D systems needs some consideration especially when NPC is performed with polar chemically bonded columns instead of an unmodified silica column in the first dimension and RPC is performed in the second dimension. When RPC is performed in the first dimension and NPC in the second dimension, the water in the aqueous–organic mobile phase of the first dimension may impact on the resolution of the following NPC separation on unmodified silica gel with nonaqueous mobiles phases.[67,73] When coupling

NPC with RPC, the immiscibility of the employed mobile phase can cause peak broadening in the second dimension. As recently demonstrated for the analysis of carotenoids, NPC × RPC can be performed with immiscible mobile phases, by adjusting the column dimensions and flow rates, for example, using a microbore column in the first dimension and a standard (4.6 mm ID (inner diameter)) analytical column in the second dimension in order to minimize the effects of band broadening.[65,73,74]

Mobile phases in RPC × RPC are usually compatible in terms of miscibility; however, big differences in the viscosities of the employed mobile phases may lead to flow instability at the mixing interface and should be therefore avoided.[75] Nevertheless, RPC × RPC can be performed when the mobile phase/stationary phases employed in the first and second dimension have considerable selectivity differences, as shown for phenolic antioxidants.[76]

A comprehensive AC × RPC approach has been employed for a natural compound extract used in TCM, based on an affinity column with human serum albumin (HSA) covalently linked to a silica column in the first dimension and a monolithic reversed-phase column in the second dimension.[77]

9.02.5 HPLC Separation of Natural Products

As documented in the scientific literature,[78,79] different approaches have attracted interest in the practical separation of natural products. Tabulation of the over 2500 citations for natural product chromatography and >4800 for HPLC purification of such compounds that are accessible through SciFinder Version 2007 using the keywords ‘natural product and high performance liquid chromatography’ is beyond the scope of this chapter. Rather, the following sections are intended to give an overview of the most frequently and significantly used procedures for the analysis and purification of several major classes of compounds that have long attracted the interest of natural products chemists. For these reasons, the following sections describe the application of HPLC to isoprenoids (with examples of mono-, sesqui-, di-, and triterpenes; iridoids and secoiridoids; carotenoids; saponins; and ecdysteroids), phenolics (with examples of coumarins as well as flavonoids and isoflavonoids), and alkaloids.

9.02.5.1 Mono-, Sesqui-, Di-, and Triterpenes

Terpenoids are the largest class of natural products. These compounds have a variety of roles in mediating beneficial (attracting) or antagonistic (deterring or defending) interactions between organisms.[80,81] Of the ~25 000 terpene structures reported so far, only a small proportion have been fully investigated, mainly those with high commercial (perfumes and flavors) or pharmaceutical value. Terpenoids are distinguished by their number of isoprene (C5) units. Several changes of the basic structures lead to methylations, ring closures, and ring openings, creating a plethora of diversity. Glycosylation is common and presents the molecule in a nonvolatile form. Under the influence of heat or enzymes (plant-derived glycosidases or, e.g., enzymes from yeast in winemaking), the bioactive or fragrant substances are frequently released.

Monoterpenes (C10) are typical ingredients of volatile plant oils and are widely used in the fragrance and flavor industry. However, they also have wide pharmaceutical application, for example, many are excellent bactericides. These plant oils also contain various other substances such as phenolics, lactones, esters, and alcohols, which may present problems during separation of the monoterpenes. Although gas chromatography (GC)–MS was the initial method of choice to separate volatile monoterpenes, the trend toward the use of LC–GC–MS has resulted in the reduction of artifacts and an easier interpretation of the resulting data.[82,83]

Furthermore, the chiral discrimination of monoterpenes has been recognized as one of the most important analytical techniques in flavor chemistry and pharmacology because the optically active stereoisomers have different sensory qualities and biological activities. HPLC offers powerful techniques for separation and quantification of enantiomers because of the progressive improvement of chiral chromatographic materials and chiral detectors such as optical rotatory dispersion (ORD) and circular dichroism (CD) detectors. In contrast, determination of chiral compounds by GC typically requires coinjection of the reference compound with known stereochemistry. An HPLC system equipped with a chiral detector, on the other hand, allows direct determination of the configuration of chiral compounds.[84]

Sesquiterpenes (15C) are a group of pentaprenyl terpenoid substances. Although they are a relatively small group of terpenoids compared to monoterpenes, their sources are widespread, having been isolated from terrestrial fungi, lichens, higher plants, insects, and various marine organisms, especially, sponges.[85] Some have pungent flavors as found in ginger (zingiberene) and pepper (rotundone)[86] and have been highly prized historically. Because of their commercial importance, fast analysis methods are required. In the group of 1000 different sesquiterpene lactones most are bitter (e.g., in absinth from *Artemisia*) and have anti-inflammatory, antibiotic, and in some cases cytotoxic qualities. When they possess exocyclic methyl groups they can react with sulfhydryl groups of, for example, skin or saliva proteins and this may lead to allergic reactions. Thus, sesquiterpenes often protect organisms. For example, in marine green algae (caulerpenyne) they are stored as polyacetates with the deterring sesquiterpene released only upon wounding. The plant hormone abscisic acid is a sesquiterpene and induces fruit ripening and leaf fall.

Diterpenes (C20) can be linear as in the phytol part of chlorophyll or mono- to tetracyclic.[87] Very well researched are the tricyclic taxanes of which taxol A from the *Taxus baccata* tree inhibits cell division and has found use in anticancer treatment. Certain labdanes inhibit blood platelet aggregation and kaurane norditerpene glycosyl ester atractylosides from the Mediterranean thistle block the ATP/ADP translocation – such diterpenes are very toxic when ingested. Ubiquitous plant-derived kaurane-type diterpenes are the plant hormones (gibberellins) or defense substances (phytoalexins).

Triterpenes (C30) are mainly polycyclic[88] and their basic structure is squalene, which was by itself only found in shark liver. Squalene is the building block for the sterane in steroid-like hormones, D vitamins, cholesterol, heart glycosides, and saponins. Modified triterpenes in plants, like the bitter hydroxylated sterol glycosides curcubitacins, are potent phytoalexins and very toxic to insects. They have been removed through breeding from edible parts of the cucumber family. However, investigation of new food sources from the widespread family of Curcubitaceae or Brassicaceae requires precise determination of undesirable triterpenes. As found for monoterpenes, triterpene enantiomers have different activities, for example, in cotton tissues the ratio of (−) to (+) gossypol determines its toxicity to ruminants or human cancer cells.[89] Cardenolides are triterpenes with unusual sugar conjugates and occur in 15 plant families, with heart-stimulating activity; some of the best known examples are the digitoxigenins from *Digitalis*. Saponins are glycosides of polycyclic triterpenes and amphiphilic, as their sterol part is lipophilic and the glycosylated end is hydrophilic. Thus, they can penetrate and interrupt biomembranes, which enables them to be good detergents (soaps) but also to be highly toxic to fish or mammals as they lead to membrane defects and the lysis of erythrocytes. A triterpene saponin of very low hemolytic activity is the very sweet glycyrrhetinic acid from the root of *Glycyrrhiza glabra* (licorice).

As described in more than 310 publications, mono-, sesqui-, di-, and triterpenes have been separated with HPLC, particularly by HP-RPC[89–97] and HP-NPC.[83,86,98]

A representative example of the separation of terpenes with HP-RPC has been described by Song *et al.*,[91] who applied a capillary HPLC–ESI-MS method for the determination of pinane monoterpene glycosides in TCM herbal extracts derived from *Paeonia suffruticosa*. In this study, the authors used a 3 mm × 150 mm 5 μm C_{18} silica column with linear gradient elution (eluent A: 0.1% aqueous FA and eluent B: ACN) at 5 μl min^{-1}. The monoterpenes paeoniflorin, albiflorin, oxypaeoniflorin, benzoyloxypaeoniflorin, galloylpaeoniflorin, and mudanpioside B and C were then detected with a single quadrupole mass spectrometer in negative ESI mode and single ion monitoring and quantified. Similar approaches, employing C_{18} phases with aqueous–organic mobile phases and volatile buffer additives are quite useful for analysis of terpenes by HP-RPC, since they allow rapid online MS.

A selection of examples from the scientific literature on the use of high-resolution chromatographic methods for the isolation and analysis of mono-, sesqui-, di-, and triterpenes (**Table 2**) have been included to demonstrate the commonality of the methodologies, despite the diversity of the structural types involved.

9.02.5.2 Iridoids and Secoiridoids

Iridoids are derivatives of monoterpenes and occur usually, but not invariably, as glycosides.[100,101] Structurally, they are cyclopentano [*c*] pyran monoterpenoids and they provide a biogenetical and chemotaxonomical link between terpenes and alkaloids. The cleavage of the cyclopentane ring of iridoids produces secoiridoids.[102]

Table 2 Mono-, sesqui-, di-, and triterpenes

Compound	*Natural product group*	*Organism/source*	*HPLC method*	*Detection*	*Reference*
α-Pinene, β-pinene, 3-carene, sabinene, limonene	Monoterpenes	Atmospheric aerosol samples	Micro RPC	ESI-MS	90
Paeoniflorin, oxypaeoniflorin, benzoylpaeoniflorin, benzoyloxypaeoniflorin, mudanpiosides B and C	Monoterpene Glycosides	*Paeonia suffruticosa*	Micro RPC	DAD UV, ESI-MS	91
α-Thujene, α-pinene, camphene, sabinene, β-pinene, myrcene, α-phellandrene , 6-3-carene, α-terpinene, β-cymene, limonene, (*Z*)-β-ocimene, (*E*)-β-ocimene, γ-terpinene and terpinolene	Monoterpenes	*Citrus*	Micro NPC	LC-GC-MS	83
Artemisinin	Sesquiterpene	*Artemisia annua*	Analytical RPC	ESI-MS	92
Caulerpenyne	Sesquiterpenes	Seaweed *Caulerpa taxifolia*	Analytical RPC	UV, ESI-MS	93
Euplotin C	Sesquiterpene	Marine ciliate *Euplotes crassus*	Analytical NPC	UV 215 nm, ESI-MS	98
Rotundone	Sesquiterpene	*Vitis vinifera*	Micro NPC	GC–MS–O	86
Various sesquiterpene lactones	Sesquiterpene lactones	*Lychnophora ericoides*	Analytical RPC	DAD UV, ESI-MS	94
Fifteen taxanes	Diterpenes	*Taxus baccata*	Analytical RPC with fluorinated and hydrocarbon phases	DAD UV	95
Bilobalide	Diterpene	*Ginkgo biloba*	Micro RPC	ELSD	99
Cucurbitacins, digitoxigenins	Triterpene sterols	Cucurbitaceae	Analytical RPC	DAD UV, ESI-MS	96
Gossypol	Terpenoid aldehyde	*Gossypium*	Analytical RPC	UV 272 nm	89
Glycyrrhizin, 18α-glycyrrhetinic acid, 18β-glycyrrhetinic acid and 18β-glycyrrhetinic acid methyl ester	Triterpene saponins	*Glycyrrhiza glabra*	Analytical RPC	UV 254 nm	97

Iridoids and secoiridoids are secondary metabolites found in terrestrial and marine flora and fauna. They are plant protectants and useful as markers of several gena in various plant families such as *Plantago* (Plantaginaceae), *Galium* (Rubiaceae), and *Scrophularia* (Scrophulariaceae). Many iridoids, for example, those from the genus *Gentiana*, taste bitter. They exhibit a wide range of bioactivities including cardiovascular, anti-inflammatory, antispasmodic, antitumor, antiviral, and immunomodulatory activities.

As shown from the more than 220 relevant references found with SciFinder Scholar or ISI Web of Knowledge, iridoids have been predominantly separated by HP-RPC[103–109] and to a lesser extent by HP-NPC.[110] Secoiridoids have been separated by HP-RPC in both analytical[43,107,109,111–116] and preparative scale.[113,117] Iridoids and secoiridoids have been almost exclusively separated with HP-RPC and detected with UV and MS. An example is the quantitative analysis of iridoids, secoiridoids, xanthones, and xanthone glycosides derived from roots of *Gentiana lutea* by Aberham *et al.*[107] These authors used an analytical 4.6 mm× 150 mm 5 μm C_{18} silica column thermostated to 30 °C with linear gradient elution (eluent A: 0.025% of TFA in water and eluent B: ACN and *n*-propanol, 50:50 (v/v)) at 1.0 ml min^{-1}. The LC system was directly coupled to an ESI mass spectrometer with flow splitting (split ratio 1:3) and the mass spectra were acquired in both positive and negative ionization mode. For the LC–MS method, eluent A was changed to a mixture of water, FA, and AA in a ratio of 99:0.9:0.1, other LC conditions remained unchanged. The described LC–MS method focused on the quantitative analysis of all major, currently known bioactive compounds in gentian roots, rather than on the iridoids (loganic acid), secoiridoids (swertiamarin, gentiopicroside, amarogentin, sweroside), xanthones (gentisin, isogentisin), and xanthone glycosides (gentiosides).

An example of the separation of iridoid glycosides with RPC-RPC chromatography systems has been published by Zhou *et al.*[109] for iridoid glycosides from *Gardenia jasminoides*, a plant used in TCM. For the separation in the first dimension, the authors have used a preparative 10 mm × 100 mm 5 μm C_{18} silica column and isocratic elution (eluent: 0.1 AA water/ACN (93:7) at 9.8 ml min^{-1}. In the second dimension they employed a 10 mm × 250 mm C_{18} silica column, with linear gradient elution (eluent: 0.1 AA water/ACN (93:7)) at 4 ml min^{-1}. Such a 2D column-switching system without sample loop trapping, enabled the isolation and purification of six iridoid glycosides including geniposide, gardenoside, shanzhiside, scandoside methyl ester, deacetylasperulosidic acid methyl ester, and genipin-1.

Additional examples as representative of approaches used for the HPLC separation are shown for iridoids in **Table 3** and for secoiridoids in **Table 4**.

9.02.5.3 Carotenoids

Carotenoids, which are tetraterpenes, are 40-carbon-atoms, aliphatic, conjugated double bond compounds with recurring isoprene units. They can be divided into carotenes (e.g., *β*-carotene, lycopene), which are formed when the end of the chain is transformed to a cyclic ionone ring, and into xanthophylls (e.g., lutein, cryptoxanthin), which contain hydroxyl-, oxo-, epoxy-, methoxy-, or carboxyl groups produced by oxidation. All carotenoids are very lipophilic and exhibit very low solubility in water. They are mainly formed in the chloro- and chromoplasts of plants and generate a diverse group of yellow-orange pigments, but also exist in other biological systems. They act as antioxidants, attractants, and UV attenuators. Due to their essential function in photosynthesis, they could also be regarded as part of the primary metabolism. They have received attention due to their health benefits and therapeutic value and in the food and perfume industries as they can be transformed into aroma molecules.

As described in more than 930 relevant publications identified by SciFinder, HPLC has been extensively used for the separation of carotenoids,[119–122] particularly in HP-RPC[123–126] and HP-NPC[21,45,127–131] modes. The primary tool of carotenoid separation is HP-RPC. For the separation of naturally occurring carotenoids as well as for carotenoid isomers, C_{30} bonded silica[132–134] exhibits a higher selectivity than the conventionally used C_8 and C_{18} materials.[135,136] The separation behavior of carotenoids on C_{30} silica phases is strongly temperature dependent, with best separations being obtained at lower temperatures.[137–139] The addition of antioxidants such as butylated hydroxytoluene (BHT) to the mobile phase (and to the extraction solvent) has been reported to avoid oxidization.[140] Carotenoids are colored compounds due to their conjugated double bonds, where the maxima in their absorption spectra, spanning from 380 to 550 nm, shifts to longer wavelengths with increasing number of conjugated double bonds. Besides UV and MS detection,[124,135,141–144] ED

Table 3 Iridoids

Compound	*Natural product group*	*Organism*	*HPLC method*	*Detection*[a]	*Other methods*	*Reference*
Ten iridoids	Iridoid glucosides	Genus *Galium* (Rubiaceae)	Analytical RPC	233 nm		105
Five iridoids	Iridoid glycosides	*Gardenia jasminoides*	Analytical RPC	UV at 240 nm, MS		106
Two iridoids	Iridoid glucosides	Noni (*Morinda citrifolia*)	Analytical RPC	UV at 240 nm, MS	HPTLC	108
Ten iridoids	Iridoid glucosides	Genus *Galium*	Analytical NPC	UV at 233 nm		110
Six iridoids	Iridoid glycosides	*Gardenia jasminoides*	RPC-RPC	DAD UV	MS, ^{1}H-NMR and ^{13}C-NMR analysis	109
Various iridoids	Iridoids	Various plants	SEC-RPC	DAD UV, MS		23

[a] If a particular technique, that is, mass spectrometry is listed in this column, it implies online coupling. Off-line application of mass spectrometry is listed in the column as other methods.

Table 4 Secoiridoids

Compound	*Natural product group*	*Organism*	*HPLC method*	*Detection*[a]	*Other methods*	*Reference*
Six secoiridoids	Acylated secoiridoid glycosides	*Gentiana* species (Gentianaceae)	Analytical and semipreparative RPC	DAD UV at 200–400 nm	MS, TLC, NMR spectroscopy	113
Oleuropein, ligustaloside A, ligustaloside B, ligstroside	Secoiridoids	*Ligustrum vulgare* (Oleaceae)	Analytical RPC	DAD UV at 190–450 nm, MS		114
Mangiferin, amarogentin, amaroswerin, sweroside, swertiamarin	Secoiridoid glycosides	*Swertia chirata* (Gentianaceae)	Analytical RPC	MS		115
Oleuropein, ligustroside, angustifolioside B	Secoiridoids	*Chionanthus virginicus* (Oleaceae)	Analytical RPC	MS	NMR spectroscopy	116
Several secoiridoids	Secoiridoids	*Fraxinus* species	Analytical RPC	DAD UV at 254 nm, MS		118

[a] If a particular technique, that is, mass spectrometry is listed in this column, it implies online coupling. Off-line application of mass spectrometry is listed in the column as other methods.

(array)[125,145–147] has been successfully applied for carotenoids. Recently, several comprehensive NPC × RPC methods for the analysis of carotenoids have been described.[7,65]

An example of the separation of free (i.e., not esterified) carotenoids as well as carotenoid esters present in mandarin essential oil with two separate comprehensive normal-phase × reversed-phase LC systems has been published by Dugo *et al.*[65] For the separation of the free carotenoids, the authors used in the first dimension a capillary normal-phase 1.0 mm × 300 mm 5 μm silica column with linear gradient elution (eluent A: *n*-hexane and eluent B: ethyl alcohol) at 10 μl min^{-1}. In the second dimension they employed a monolithic 4.6 mm × 100 mm C_{18} silica column with linear gradient elution (eluent A: 2-propanol and eluent B: 20% (v/v) water in ACN) at 4.7 ml min^{-1}. For the separation of the carotenoid esters, these authors used in the first dimension a capillary normal-phase 1.0 mm × 250 mm 5 μm cyanopropyl silica column with linear gradient elution (eluent A: *n*-hexane and eluent B: *n*-hexane/butyl acetate/acetone 80:15:5 (v/v/v) at 10 μl min^{-1}. The second dimension columns as well as the mobile phases were the same as those used for the free carotenoids analysis, with minor changes to the flow rate and gradient conditions. The incompatibility of the solvents that were used in the two dimensions (NPC and RPC) and their effects on the separation were overcome by using a combination of a capillary column in the first dimension and an analytical monolithic column in the second dimension. Using flow splitting, the detection was performed in parallel with photodiode array and mass spectrometric detection. This approach then allowed the establishment of two independent 2D contour plots, one for the carotenoids and one for the carotenoid esters, allowing group separations as well as identification of the individual compounds. With the first NPC × RPC system, predominantly targeting carotenoids, 19 compounds were identified and grouped separately into mono-ols, mono-epoxides, apo-carotenoids, di-ols, di-ol–mono-epoxides, di-ol–di-epoxides, and tri-ol–mono-epoxides. With the second NPC × RPC system, predominantly targeting carotenoid esters, 23 compounds were identified and grouped separately into hydrocarbons, β-cryptoxanthin esters, lutein esters, mutatoxoxanthin esters, isomeric esters, luteoxanthin esters, apocarotenoids, and free xanthophylls. In this study, the information was obtained from (1) analysis of carotenoids in both their free form as well as in their fatty acid esterified form, (2) group separation of the compounds in the respective 2D contour plots, (3) resolution enhancement brought about by the comprehensive NPC × RPC system, and (4) online analysis with MS. Such a comprehensive LC × LC approach demonstrates the paradigm shift from a combined application of open column/HPLC to HP multidimensional chromatographic separation methods in natural product purification.

Table 5 lists additional investigations that are representative of the approaches that have been used for the high-resolution chromatographic separation of carotenoids from diverse sources.

9.02.5.4 Saponins

Saponins consist of an aglycone with carbohydrate moieties. The aglycone can be a triterpene or a steroid and can have a number of different substituents (–H, –COOH, –CH_3). The number and type of carbohydrate moieties result in a considerable structural diversity of the saponins. Most carbohydrates in saponins are hexoses (i.e., glucose, galactose), 6-deoxyhexoses (rhamnose), pentoses (arabinose, xylose), uronic acid (glucoronic acid), or carbohydrates with amino functionality (glucosamine). Through the glycosylation of the hydrophobic aglycones, they can act as biological detergents and, when agitated in water, form foams, which gave rise to the name, saponin for this group of compounds. Saponins are widely spread throughout the plant kingdom,[148–150] and have been found in marine animals.[151] The amphiphilic nature of the saponins enables them to act as soaps and detergents as they can dissolve membranes; however, since the saponins can hemolyze erythrocytes they are highly toxic if they reach mammalian blood. Exceptions are the sweet triterpenesaponin glycirrhicin from *Glycirrhiza glabra*, and diosgenin from *Dioscorea* (yam), a steroidal saponin precursor for the synthesis of cortisone and progesterone.

As with many other natural product groups, HPLC has been increasingly applied to saponins, as shown from the 460 reference citations identified by SciFinder using the keywords 'saponins high performance liquid chromatography'. Saponins have predominantly been separated by HP-RPC[152–154] but separation by HP-NPC,[155,156] HP-IEX,[157,158] and HP-HILIC[159] has also been described. Saponins do not have a chromophore, which restricts their UV detection to 200–210 nm. This makes their detection in complex samples containing other analytes, which absorb in this wavelength range, challenging. Notable exceptions are

Table 5 Carotenoids

Compound	*Natural product group*	*Organism*	*HPLC method*	*Detection*	*Reference*
Various compounds	Carotenoids and carotenoid esters	*Capsicum annuum*	Analytical RPC with C_{30} ligand	MS	124
α- and β-Carotene, lycopene	Carotenoids	Vegetable oils	Analytical RPC with C_{30} ligand	ED	125
α- and β-Carotene, β-cryptoxanthin, lutein, zeaxanthin	Carotenoids	Grain: maize, oat, wheat, barley (Poaceae)	Analytical NPC	DAD UV at 350–500 nm	131
Various compounds	Carotenoids	*Citrus* products	NPC × RPC, first dimension microscale and second dimension analytical scale (monolithic C_{18})	DAD UV	7

DMMP (2,3-dihydro-2,5-dihydroxy-6-methyl-4-pyrone)-conjugated soya saponins that have UV absorption at 295 nm.[160] Besides UV detection, evaporative light scattering detection (ELSD) is applied as a validated analytical technique. As a consequence of the difficult UV detection, more recently, HPLC has been combined with online MS. Substantial work with LC–MS was performed on saponins from commercially significant plants, that is, black bean (*Vigna mungo*),[161] soybean (*Glycine max*),[162–164] and ginseng (*Panax notoginseng*).[165–167] In some investigations, a combination of ELSD and MS detection methods have been applied.[168–171]

An example of the separation of saponins with HP-RPC has been described by Sun *et al.*,[172] who identified triterpenoid saponins in crude extracts from nine species of *Clematis* with electrospray ionization multistage tandem mass spectrometry (HPLC/ESI-MS^n). These authors used an analytical 4.6 mm × 250 5 μm C_{18} silica column at 25 °C with linear gradient elution (eluent A: water with 0.05% FA (v/v) and eluent B: ACN) at 0.6 ml min^{-1}. Eight triterpenoid saponins, namely, hederacholichiside F, prosapogenin CP11, clematichinenoside B, huzhangoside D, clematiganoside A, clematichinenoside C, huzhangoside B, and HN saponin H, were isolated from the whole plant of *Clematis ganpiniana*, independently characterized and employed as reference compounds. These eight reference compounds were analyzed by HPLC/ESI-MS and HPLC/ESI-MS^n in negative ion mode to establish their MS^n fragmentation pathways ($n = 2$–4). This information provided structural insight into the carbohydrate sequence of the oligosaccharide chains and the mode of attachment of the aglycone of the saponins. This approach demonstrates the power of HP-RPC MS^n methods to allow the rapid discrimination of nonisomeric saponins. As a consequence, these authors have tentatively identified two new compounds hitherto unknown for the genus *Clematis*. This approach, employing C_{18} silica phases with aqueous–organic mobile phases and volatile buffer additives is typical of the procedures followed for the analysis of saponins by HP-RPC and provides, in conjunction with online MS, the methodological framework for the discovery of new compounds.

Table 6 lists examples that are representative of the approaches that have been used for the high-resolution chromatographic separation of saponins.

9.02.5.5 Ecdysteroids

The ecdysteroids are polar, polyhydroxylated steroids that function as the moulting hormones of insects and crustaceans. Ecdysteroids are also found in plants, often in high concentration, where it is thought they contribute to insect deterrence by acting as antifeedants, interfering in ecdysteroid metabolism or mode of action on ingestion by phytophagous insects. The chromatographic behavior of ecdysteroid glycosides is characteristic, as they appear much more polar than their corresponding free aglycones when analyzed by normal-phase HPLC, whereas the presence of glycosidic moieties has a very limited (if any) impact on polarity when using reversed-phase HPLC.[175]

As demonstrated from the 312 relevant SciFinder references, ecdysteroids have been isolated by RP-HPLC,[176–179] HP-NPC,[180–182] HP-IEX,[183,184] or HP-AC.[185] The dominant mode for the separation of ecdysteroids is HP-RPC, with numerous published applications including the use of superheated water as a mobile phase.[186]

An example for the separation of ecdysteroids with HP-RPC systems was published by Louden *et al.*[179] and related to the separation of an ecdysteroids extract from the plant *Lychnis flos-cuculi* (Caryopyllaceae). These authors used an analytical 4.6 mm × 100 mm 5 μm C_{18} silica column with isocratic elution (eluent ACN and D_2O 99.8% isotopic purity 20:80 v/v) at 1.0 ml min^{-1} and DAD between 190 and 360 nm. This study also explored the application of HPLC–NMR spectroscopy and HPLC–NMR spectroscopy–MS to these ecdysteroid-containing plant extracts, showing the advantages and limitations of the use of complex multiply hyphenated detection systems, which incorporate detectors of differing sensitivities.

Table 7 lists studies that are representative of the approaches that have been used for the high-resolution chromatographic separation of ecdysteroids.

9.02.5.6 Coumarins

Coumarins are 1,2-benzopyrones that are derived from the phenylpropanoid pathway; however, major details of their biosynthesis are still largely unknown.[187] Coumarins can also be produced, through the cleavage of meliotoside (the β-glucopyranoside of *o*-hydroxycinnamic acid) in dead plant material, or hay, where in the

Table 6 Saponins

Compound	*Saponin group*	*Organism*	*HPLC method*	*Detection*[a]	*Other methods*	*Reference*
Sixty saponins, differing in carbohydrate substructures	Glycosylated triterpenes	Tree *Quillaja saponaria* var. Molina	Preparative RPC	UV at 206 nm	ESI ion trap multiple stage tandem mass spectrometry	152
Twenty-eight saponins, differing in carbohydrate substructures	Glycosylated triterpenes	Tree *Quillaja saponaria* var. Molina	Preparative RPC	UV at 210 nm	SPE, NMR spectroscopy	154
Protodioscin, tribulosin, terrestrosin D	Steroidal saponins	*Tribulus terrestris* (Zygophyllaceae)	Analytical RPC	UV at 200 nm		173
Periandradulcin A, B and C	Glycosylated triterpenes	*Periandra dulcis* (Leguminosae)	Semipreparative NPC	UV at 207 nm, MS	RPC, MS, NMR spectroscopy	156
Various saponins	Glycosylated	Ginseng roots, *Sapindus mukurossi* and *Anemone rivularis*	Semipreparative IEX	UV		157
Elliptoside A–J	Glycosylated	*Archidendron ellipticum* (Leguminosae)	Preparative HILIC with aminopropyl ligand	DAD UV	SEC, RPC	159
Ginsenosides, notoginsenosides	Dammarane saponins	*Panax notoginseng*	Analytical RPC	ELSD, MS		168
Various saponins		Flos Lonicera, several genera (Caprifoliaceae)	Micro RPC	DAD UV, MS	SPE, IR, ^{1}H-, and ^{13}C-NMR	174

[a] If a particular technique, that is, mass spectrometry is listed in this column, it implies online coupling. Off-line application of mass spectrometry is listed in the column as other methods.

Table 7 Ecdysteroids

Compound	Natural product group	Organism	HPLC method	Detection[a]	Other methods	Reference
Various compounds	Ecdysteroids	*Lychnis flos-coculi* (Caryophyllaceae)	RPC	DAD UV, FT-infrared and ^{1}H-NMR spectroscopy, MS		179
Phyto-ecdysteroids	Ecdysteroids	*Silene* species (Caryophyllaceae)	Semipreparative NPC	UV	MS, bioassay	180
Various compounds	Ecdysteroids		Analytical IEX	UV		183
Various compounds	Ecdysteroids		AC	UV		185

[a] If a particular technique, that is, mass spectrometry is listed in this column, it implies online coupling. Off-line application of mass spectrometry is listed in the column as other methods.

presence of fungi, like *Penicillium* and *Aspergillus* sp., dicumarol, a vitamin K antagonist, can be formed leading to hemophilia in cattle. Coumarins may occur in the oils of some Apiaceae (Umbelliferae), Rubiaceae, and Poaceae and are best known as the bergamot oil from *Citrus bergamia*. In plants, coumarins contribute to the defense against phytopathogens, response to abiotic stresses, regulation of oxidative stress, and probably as signaling molecules.[187] Coumarins can be subclassified into simple coumarins (benzo-α-pyrones), 7-oxygenated coumarins (furanocoumarins), pyranocoumarins (benzodipyran-2-ones), and phenylcoumarins (benzo-benzopyrones). Some furano-coumarins are known to enhance the photosensitivity of human skin. In the presence of UV light, they produce radicals that block enzymes (and thus have found limited use therapeutically) but also lead to inflammation. Unfortunately, they are also incorporated into the cellular DNA via UV-mediated cycloaddition leading to mutations.

As evidenced by the over 360 references retrieved from SciFinder with the keywords 'coumarins high performance liquid chromatography', a variety of approaches have been used for the separation of coumarins. Coumarins have been separated by RP-HPLC,[46,47,188–192] HP-NPC,[193–198] HP-IEX,[199] and/or HP-AC.[24] The dominant mode for the separation of coumarins is HP-RPC and to a lesser extent HP-NPC, used in conjunction with UV and MS detection.

An example of the separation of coumarins with HP-RPC systems was published by Eeva *et al.*[47] for the 21 different coumarins and furanocoumarins isolated from the plants *Peucedanum palustre* and *Angelica archangelica*. In these studies the authors used an analytical 4.6 mm × 100 mm 3 μm C_{18} silica column with linear gradient elution at 1.0 ml min^{-1} and UV detection at 320 nm. A Turbo Method Development program was applied to optimize the mobile phase with two organic solvents (ACN and MeOH) and two aqueous solutions (1.0% FA and 10 mmol l^{-1} ammonium acetate). Optimization of the solvent gradients for the method was performed with the program DryLab, with the aid of coumarin standards. Once the LC methods were established, the techniques were transferred to an LC–MS system, employing a triple quadrupole mass spectrometer taking into consideration the gradient delay volume, and applied to the respective plant extracts containg >45 compounds of interest. Such an approach, employing computer-assisted method development considerably streamlined the method development phase of the separation when speed, high throughput, and method robustness are required.

An example of the separation of coumarins with comprehensive normal-phase × reversed-phase LC systems has been published by Dugo *et al.*[73] for coumarins and psoralens in cold-pressed lemon oil. For the separation in the first dimension, the authors used a capillary normal-phase 1.0 mm × 300 mm 5 μm silica column with isocratic elution (eluent: *n*-hexane:ACN (75:25)) at 20 μl min^{-1}. In the second dimension, a monolithic 4.6 mm × 25 mm C_{18} silica column (including a 4.6 mm × 5 mm guard column) was employed with linear gradient elution (eluent A: water and eluent B: ACN) at 4 ml min^{-1}. The interface between the first and the second dimension was a 10-port, 2-position valve equipped with two storage loops. The incompatibility of the solvents that were used in the two dimensions (NPC and RPC) and its effects in the separation were overcome by using a combination of a capillary column in the first dimension and an analytical monolithic column in the second dimension. With this NPC × RPC system, 11 heterocyclic compounds were analyzed and depicted in 2D contour plots.

Table 8 lists additional investigations that are representative of the approaches that have been used for the high-resolution chromatographic separation of coumarins.

9.02.5.7 Flavonoids and Isoflavonoids

Flavonoids are benzopyrane derivatives with a phenyl group in the second position. Flavonoids are polyphenols, which are biosynthetically derived from phenylalanine and can be O-glycosides, usually in position 3 or 7. They can be grouped into several subclasses, including the anthocyanins, catechins, flavones, isoflavones, flavonols, and chalcones. These subclasses, combined with glycosylation at multiple sites with a variety of different saccharides and further acylation of the saccharides, produce more than 5000 chemically distinguishable compounds.[200] Animals and humans need phenolics to build aromatic amino acids but cannot synthesize them, thus, they or their precursors must be ingested as food from plants. Flavonoids are the largest group of phenolics and contribute to color and oxidative stability of plant parts. They find medicinal uses as antioxidants and are known to increase blood flow and stabilize capillaries.

Table 8 Coumarins

Compound	*Natural product group*	*Organism*	*HPLC method*	*Detection*	*Reference*
Twenty-one different coumarin-type compounds	Coumarins and furanocoumarins	*Peucedanum palustre* and *Angelica archangelica*	Analytical RPC based on computer-assisted optimization	UV, MS	47
Nine coumarin compounds	Coumarins	*Angelica gigas*	Analytical RPC	DAD UV, MS	193
Osthol, corymbo-coumarin	Coumarins	*Seseli* species (Umbelliferae)	Analytical NPC	UV	198
6,7-Dimethoxycoumarin and capillarisin	Coumarins	*Artemisia capillaris*	AC	UV	24
Various compounds	Coumarins	*Citrus* products	NPC × RPC, first dimension microscale and second dimension analytical scale (monolithic C_{18})	DAD UV	73

As documented from the more than 1820 relevant SciFinder references, flavonoids have been separated by a variety of HPLC modes,[201–205] predominantly by RP-HPLC,[3,20,206–216] but also by HP-NPC,[217–221] HP-SEC,[222–225] HP-AC,[226] or AC with MIPs.[227–229] Separations are almost exclusively performed by HP-RPC using binary elution systems with an aqueous, acidified eluent A (AA, perchloric acid, phosphoric acid, or FA) and a less polar organic solvent eluent B such as MeOH or ACN, possibly acidified. Phenols absorb in the UV region. Two absorption bands are characteristic of flavonoids.[204] Band I, with a maximum in the 300–550 nm range, presumably arises from the B-ring. Band II, with a maximum in the 240–285 nm range, is believed to arise from the A-ring. Anthocyanins show band I and band II absorption maxima in the 465–560 and 265–275 nm regions, respectively.[205] Because there is little or no conjugation between the A- and B-rings, UV spectra of flavanones and isoflavones usually have an intense band II peak but a small band I peak.[230] This lack of conjugation also results in small band I peaks for the catechins. UV spectra of flavones and flavonols have a band I peak around 300–380 nm and a band II peak at around 240–280 nm.[230] Flavonoids have been identified using photo DAD,[231] fluorescence detection,[208] or ED.[216]

An example for the separation for flavonoids with HP-RPC is the screening method employed for the systematic identification of glycosylated flavonoids and other phenolic compounds in plant food materials by Lin *et al.*[20] These authors used an analytical 4.6 mm × 250 mm 5 μm C_{18} silica column at 25 °C with linear gradient elution (eluent A: (0.1% FA in water and eluent B: 0.1% FA in ACN) at 1.0 ml min^{-1}. DAD was performed at 270, 310, 350, and 520 nm to monitor the UV/VIS absorption. The LC system was directly coupled to an ESI mass spectrometer without flow splitting and the mass spectra acquired in the positive and negative ionization mode. The same analytical scheme (aqueous MeOH extraction, reversed-phase liquid chromatographic separation, and diode array and mass spectrometric detection) can be applied to a wide variety of samples and standards and therefore allows the cross-comparison of newly detected compounds in samples with standards and plant materials previously identified in the published literature.

As a further example, Prior *et al.*[221] have demonstrated how procyanidin oligomers can be separated using HP-NPC with UV detection (at 280 nm), fluorescence detection (the excitation and emission wavelengths were 276 and 316 nm), and MS detection. These authors used an analytical 4.6 mm × 250 mm ID 5 μm silica column at 37 °C with linear gradient elution using a ternary mobile phase (eluent A: dichloromethane, eluent B: MeOH, and eluent C: AA and water (1:1 v/v)) at 1.0 ml min^{-1}. For HPLC–MS analysis, 10 mmol l^{-1} ammonium acetate in MeOH was used as an ionization reagent and was added via a tee-junction at 0.1 ml min^{-1} into the eluent stream of the HPLC prior to the mass spectrometer by an auxiliary HPLC pump.

Table 9 lists additional examples that are representative of the approaches that have been used for the high-resolution chromatographic separation of flavonoids and isoflavonoids.

9.02.5.8 Alkaloids

Alkaloids are a very diverse class of secondary metabolites with more than 20 000 known compounds, encompassing a broad variety of chemical structures. Alkaloids are heterocyclic compounds, often possessing tertiary nitrogens, which gives them their basic properties. They are most commonly lipophilic; however, they can form water-soluble salts with acids. Many alkaloids are biosynthetically derived from the amino acids lysine, ornithine, phenylalanine, tryptophan, and tyrosine. Although in the past alkaloids were mostly isolated from plants, where they are sequestered as salts in special tissues to kill predators, they are also found in microorganisms, marine organisms, as well as insects and reptiles. Alkaloids have been classified according to their molecular structure (e.g., indole, isoquinoline, pyridine, pyrrolizidine, steroidal, tropane) or according to their botanical origin, for example, *Nicotina* alkaloids, *Papaver* alkaloids, and *Solanum* alkaloids. Most alkaloids are highly bioactive, and their effects in humans are a direct consequence of their structural similarities to neurotransmitters such as acetylcholine, dopamine, noradrenalin, or serotonin. Alkaloids are of considerable pharmaceutical interest (e.g. codeine, scopolamine, morphine, D-tubocurarine), others are drugs of abuse (mescaline, cocaine, and nicotine), stimulants (caffeine), or poisons employed for pest control (strychnine).

As evident in the literature of the last decade, with more than 1380 references relevant to the theme of this chapter are found within SciFinder Scholar or ISI Web of Knowledge, HPLC has developed into an important tool for the isolation and purification of alkaloids. Alkaloids have been isolated with a variety of HPLC modes,[232–234] including HP-RPC,[235–240] HP-NPC,[241–244] HP-IEX,[22,245–247] HP-HILIC,[248] or HP-AC.[249]

Table 9 Flavonoids and isoflavonoids

Compound	*Natural product group*	*Organism*	*HPLC method*	*Detection*[a]	*Other methods*	*Reference*
Various isoflavones	Isoflavones	*Glycine max* (Fabaceae)	Analytical RPC	MS	SPE	215
Anthocyanins compounds	Flavonoids	*Vitis vinifera*	Analytical RPC	ED		216
Glycosylated flavonoids	Various groups	Various plants	Analytical RPC	DAD UV, MS		20
Various compounds	Isoflavonoids	*Smirnowia iranica*	Analytical RPC	DAD UV	SPE, NMR spectroscopy	3
Various compounds	Procyanidins and anthocyanins	*Vaccinium* spp.	Analytical NPC	DAD UV, MS		221
Various compounds	Pyrano-anthocyanin–flavanols	*Vitis vinifera*	SEC, RPC	UV, MS		225
Quercetin	Flavanols	*Vitis vinifera*	AC with MIP	UV	SPE	228

[a] If a particular technique, that is, mass spectrometry is listed in this column, it implies online coupling. Off-line application of mass spectrometry is listed in the column as other methods.

Several LC–LC methods have been described including AC-SCX[250] where the HP-AC was performed with a MIP column. HP-RPC dominates the analytical and preparative HPLC application with alkaloids, notwithstanding the potential for secondary interaction of their basic primary, secondary, or tertiary amine moiety with residual silanol groups of the *n*-alkylsilica stationary phases, resulting in peak tailing. To overcome this problem, a variety of approaches have been developed, including the suppression of these interactions with mobile phase ion-pairing additives or the employment of end-capped stationary phases with low numbers of free silanol groups. Typically, mobile phases with a pH of 2–4 have been used to shift the equilibrium from the free base into the protonated form of the alkaloid. The detection of alkaloids has been mostly performed with UV detectors; however, fluorescence, electrochemical, or mass spectrometric detection has also been employed. Recently, MS has emerged as the primary tool for alkaloid identification, quantification, and structural elucidation, where structural analysis is usually performed by mass spectrometric fragmentation. Mobile phases suited for LC–MS applications contain preferably only volatile additives, for example, ammonium acetate, ammonium formate, or TFA.

A good example of the separation of alkaloids with HP-RPC is the separation of the protoberberine alkaloids (including berberine, palmatine, and jatrorrhizine) reported by Wu *et al.*[240] These authors used an analytical 4.6 mm × 250 mm 5 μm C_{18} silica column with linear gradient elution (eluent A: water with 0.0 034 mol l^{-1} ammonium acetate and 0.2% AA (v/v) and eluent B: ACN) at 0.5 ml min^{-1} at a temperature of 23 °C. A set of alkaloid standards, permitting external calibration, was employed permitting the quantitative analyses of the protoberberine alkaloids in herbs used in TCM, coupled with the multiple stage mass spectrometric fragmentation information obtained with ESI-FT-ICR-MS^n (Fourier transform ion cyclotron resonance multiple stage mass spectrometry) and ESI-MS^n in the positive ion mode. Such an approach, employing C_{18} silica phases with aqueous–organic mobile phases and volatile buffer additives can be considered typical of techniques now employed for the analysis of alkaloids by HP-RPC and provides, in conjunction with online MS, the methodological framework for fast herbal medicine authentication and the quantification of individual compounds.

The additional examples summarized in **Table 10** are representative of the approaches that have been used for the high-resolution chromatographic separation of alkaloids.

9.02.6 Conclusions

Due to the enormous growth in capability and separation power that has occurred over the past two decades, the benefits of HPLC in natural product chemistry may now seem obvious. However, as there is an immense choice of modes and procedures, further scope exists to improve the quality of such separations and achieve even higher resolutions based on even more efficient optimization procedures. For these reasons, a comprehensive overview of the principles and limitations of contemporary separation methods in various steps of purification and analysis of natural products has been presented at the beginning of this chapter.

In order to solve analytical problems for a particular compound or class of compounds, as well as to save time and resources, it is essential that systematic method development concepts are applied. Such methods then enable a successful scaling up to preparative purifications as well as the design and application of MD-HPLC purification schemes. Moreover, if used in conjunction with dereplication procedures, these advances in high-resolution chromatographic methods may lead to new discoveries that can be used to advance science or medicine and at the same time respect the environment through reduced solvent and reagent usage.

Finally, since many natural product compounds have been investigated with various chromatographic modes and detection techniques, a selection of examples has been summarized in this chapter. This information has been compiled in the form of tables for well-researched classes of secondary metabolites selected from the major subgroups of isoprenoids (mono-, sesqui-, di-, and triterpenes; iridoids and secoiridoids; carotenoids; saponins; and ecdysteroids), of phenolics (coumarins, flavonoids, and isoflavonoids), and of alkaloids.

Despite decades of research, there still remains a vast scope for new natural products to be discovered and isolated from microbial, marine, arthropod, or plant organisms and used as nutraceuticals or pharmaceuticals. Such tasks will require access to more sophisticated and better optimized separation and identification methodologies. To this end, it will also be the responsibility of future generations of natural product scientists

Table 10 Alkaloids

Compound	*Natural product group*	*Organism*	*HPLC method*	*Detection*[a]	*Other methods*	*Reference*
Twenty-two standards	Monoterpenoid indole alkaloids	*Rauvolfia serpentina* or *Rhazya stricta*	Analytical RPC			239
Berberine, palmatine, jatrorrhizine, coptisine	Protoberberine alkaloids	Ranunculaceae	Analytical RPC	DAD UV	FT-ICR-MS	240
Various alkaloids	Alkaloids		NPC	UV		244
Ephedrine alkaloids	Alkaloids	Ephedraceae	IEX	DAD UV and fluorescence		22
Ephedrine, atropine, theophylline, and nicotine	Alkaloids		HILIC	UV at 214 nm		248
Various alkaloids	Vinca alkaloids		AC	UV at 263 nm		249
Atropine and scopolamine	Tropane alkaloids		AC-SCX	UV at 210 nm		250

to ensure that such bioprospecting occurs responsibly and sustainably. It is therefore expected that increasingly the analysis of natural products will use the principles of green analytical chemistry, considering the issues of waste minimization and hazard reduction. Similar criteria will also apply to preparative and process developments. Thus, there is tremendous potential for investigators to pursue new aspects of method development, which hopefully has been encouraged by this chapter.

Abbreviations

AA	acetic acid
AC	affinity chromatography
ACN	acetonitrile
AEX	anion-exchange chromatography
BHT	butylated hydroxytoluene
CC	(open) column chromatography
CD	circular dichroism
CE	capillary electrophoresis
CEX	cation-exchange chromatography
DAD	diode array detection
DMMP	2,3-dihydro-2,5-dihydroxy-6-methyl-4-pyrone
ED	electrochemical detection
ELSD	evaporative light scattering detection
ESI	electrospray ionization
FA	formic acid
FC	flash chromatography
FT	Fourier transform
FT-ICR-MS	Fourier transform ion cyclotron resonance mass spectrometry
FTIR	Fourier transform infrared
GC	gas chromatography
GPC	gel-permeation chromatography
HBFA	heptafluorobutyric acid
HILIC	hydrophilic interaction chromatography
HP	high-performance
HP-AC	high-performance affinity chromatography
HP-AEX	high-performance anion-exchange chromatography
HP-CEX	high-performance cation-exchange chromatography
HP-GPC	high-performance gel-permeation chromatography
HP-HILIC	high-performance hydrophilic interaction chromatography
HP-IEX	high-performance ion-exchange chromatography
HPLC	high-performance liquid chromatography
HP-NPC	high-performance normal-phase chromatography
HP-RPC	high-performance reversed-phase chromatography
HP-SEC	high-performance size exclusion chromatography
HPTLC	high-performance thin-layer chromatography
HSA	human serum albumin
HSCC	high-speed countercurrent chromatography
HTS	high-throughput screening
ID	inner diameter
IEX	ion-exchange chromatography
LC	liquid chromatography
MALDI ToF	matrix-assisted laser desorption/ionization time-of-flight
MD-HPLC	multidimensional high-performance liquid chromatography
MeOH	methanol

MIP	molecularly imprinted polymer
MS	mass spectrometry
MS^n	multiple stage mass spectrometry
NMR	nuclear magnetic resonance
NPC	normal-phase chromatography
O	olfactometry
ORD	optical rotatory dispersion
PC	peak capacity
PTLC	preparative thin-layer chromatography
QSRR	quantitative structure retention relationship
RAM	restricted access material
RPC	reversed-phase chromatography
RRM	relative resolution map
SCX	strong cation exchange
SEC	size exclusion chromatography
SPE	solid-phase extraction
TCM	Traditional Chinese Medicine
TFA	trifluoroacetic acid
THF	tetrahydrofuran
2D	two-dimensional
3D	three-dimensional

References

1. H. L. Constant; K. Slowing; J. G. Graham; J. M. Pezzuto; G. A. Cordell; C. W. W. Beecher, *Phytochem. Anal.* **1997**, *8*, 176–180.
2. M. E. Hansen; J. Smedsgaard; T. O. Larsen, *Anal. Chem.* **2005**, *77*, 6805–6817.
3. M. Lambert; D. Strk; S. H. Hansen; M. Sairafianpour; J. W. Jaroszewski, *J. Nat. Prod.* **2005**, *68*, 1500–1509.
4. Y. Konishi; T. Kiyota; C. Draghici; J.-M. Gao; F. Yeboah; S. Acoca; S. Jarussophon; E. Purisima, *Anal. Chem.* **2007**, *79*, 1187–1197.
5. S. Urban; F. Separovic, *Front. Drug Des. Discov.* **2005**, *1*, 113–166.
6. O. Sticher, *Nat. Prod. Rep.* **2008**, *25*, 517–554.
7. P. Dugo; V. Skerikova; T. Kumm; A. Trozzi; P. Jandera; L. Mondello, *Anal. Chem.* **2006**, *78*, 7743–7750.
8. J. L. Mazzei; L. A. d'Avila, *J. Liq. Chromatogr. Relat. Technol.* **2003**, *26*, 177–193.
9. K. Hostettmann; M. Hostettmann; A. Marston, *Preparative Chromatography Techniques. Applications in Natural Product Isolation*; Springer-Verlag: London, UK, 1986.
10. B. A. Bidlingmeyer, Ed., *Journal of Chromatography Library, Vol. 38: Preparative Liquid Chromatography*; 1987.
11. E. Grushka, Ed., *Chromatographic Science Series, Vol. 46: Preparative-Scale Chromatography*; 1989.
12. K. K. Unger, Ed., *Handbook of HPLC, Part 2: Preparative Liquid Column Chromatography*; 1994.
13. K. Hostettmann; A. Marston; M. Hostettmann, *Preparative Chromatography Techniques: Applications in Natural Product Isolation*, 2nd ed.; 1997.
14. A. S. Rathore; A. Velayudhan, Eds., Scale-Up and Optimization in Preparative Chromatography: Principles and Biopharmaceutical Applications. In *Chromatographic Science Series.* 2003; Vol. 88.
15. C. Horvath; W. Melander; I. Molnar, *J. Chromatogr.* **1976**, *125*, 129–156.
16. C. Horvath; W. Melander; I. Molnar, *Anal. Chem.* **1977**, *49*, 142–154.
17. K. K. Unger, *Journal of Chromatography Library, Vol. 16: Porous Silica, Its Properties and Use as Support in Column Liquid Chromatography*; Elsevier: Amsterdam, The Netherlands, 1979; p 336.
18. P. Stead, *Methods Biotechnol.* **1998**, *4*, 165–208.
19. Z. Latif, *Methods Biotechnol.* **2005**, *20*, 213–232.
20. L.-Z. Lin; J. M. Harnly, *J. Agric. Food Chem.* **2007**, *55*, 1084–1096.
21. M. M. Mendes-Pinto; A. C. S. Ferreira; M. B. P. P. Oliveira; P. Guedes de Pinho, *J. Agric. Food Chem.* **2004**, *52*, 3182–3188.
22. R. A. Niemann; M. L. Gay, *J. Agric. Food Chem.* **2003**, *51*, 5630–5638.
23. S. Ma; L. Chen; G. Luo; K. Ren; J. Wu; Y. Wang, *J. Chromatogr. A* **2006**, *1127*, 207–213.
24. H. L. Wang; H. F. Zou; J. Y. Ni; L. Kong; S. Gao; B. C. Guo, *J. Chromatogr. A* **2000**, *870*, 501–510.
25. T.-P. I; R. Smith; S. Guhan; K. Taksen; M. Vavra; D. Myers; M. T. W. Hearn, *J. Chromatogr. A* **2002**, *972*, 27–43.
26. L. R. Snyder, *J. Chromatogr.* **1974**, *92*, 223–230.
27. L. R. Snyder, *J. Chromatogr. Sci.* **1978**, *16*, 223–234.
28. P. J. Schoenmakers; H. A. H. Biliet; L. D. Galan, *J. Chromatogr.* **1979**, *185*, 179–195.

29. H. B. Patel; T. M. Jefferies, *J. Chromatogr.* **1987**, *389*, 21–32.
30. M. T. W. Hearn, *Handbook of Bioseparations;* 2000; Vol. 2, pp 71–235.
31. L. R. Snyder, Gradient election. In *HPLC – Advances and Perspectives*; C. Horvath, Ed.; Academic Press: New York, 1980; Vol. 1, 208–316.
32. J. W. Dolan; D. C. Lommen; L. R. Snyder, *J. Chromatogr.* **1989**, *485*, 91–112.
33. B. F. D. Ghrist; B. S. Coopermann; L. R. Snyder, *J. Chromatogr.* **1988**, *459*, 1–23.
34. B. F. D. Ghrist; L. R. Snyder, *J. Chromatogr.* **1988**, *459*, 25–41.
35. B. F. D. Ghrist; L. R. Snyder, *J. Chromatogr.* **1988**, *459*, 43–63.
36. M. A. Stadalius; H. S. Gold; L. R. Snyder, *J. Chromatogr.* **1984**, *296*, 31–59.
37. E. P. Lankmayr; W. Wegscheider; K. W. Budna, *J. Liq. Chromatogr.* **1989**, *12*, 35–58.
38. J. C. Berridge, *Techniques for the Automated Optimization of HPLC Separations*; Wiley Interscience: Chichester, 1986.
39. J. K. Strasters; H. A. H. Billiet; L. de Galan; B. G. M. Vandeginste; G. Kateman, *J. Liq. Chromatogr.* **1989**, *12*, 3–22.
40. M. A. Quarry; R. L. Grob; L. R. Snyder, *Anal. Chem.* **1986**, *58*, 907–917.
41. J. Castillo; O. Behavente-Garcia; J. A. Del Rio, *J. Liq. Chromatogr.* **1994**, *17*, 1497–1523.
42. F. Dondi; Y. D. Kahie; G. Lodi; P. Reschiglian; C. Pietrogrande; C. Bighi; G. P. Cartoni, *Chromatographia* **1987**, *23*, 844–849.
43. F. Dondi; T. Gianferrara; P. Reschiglian; M. C. Pietrogrande; C. Ebert; P. Linda, *J. Chromatogr.* **1989**, *485*, 631–645.
44. W. Metzger; K. Reif, *J. Chromatogr. A* **1996**, *740*, 133–138.
45. A. S. Kester; R. E. Thompson, *J. Chromatogr.* **1984**, *310*, 372–378.
46. M. A. Hawryl; E. Soczewinski; T. H. Dzido, *J. Chromatogr. A* **2000**, *886*, 75–81.
47. M. Eeva; J.-P. Rauha; P. Vuorela; H. Vuorela, *Phytochem. Anal.* **2004**, *15*, 167–174.
48. T. H. Dzido; E. Soczewinski; J. Gudej, *J. Chromatogr.* **1991**, *550*, 71–76.
49. M. Krauze-Baranowska; T. Baczek; D. Glod; R. Kaliszan; E. Wollenweber, *Chromatographia* **2004**, *60*, 9–15.
50. T. Wennberg; K. Kreander; M. Laehdevuori; H. Vuorela; P. Vuorela, *J. Liq. Chromatogr. Relat. Technol.* **2004**, *27*, 2573–2592.
51. U. Rosentreter; U. Huber, *J. Comb. Chem.* **2004**, *6*, 159–164.
52. M. Martin, *Fresenius J. Anal. Chem.* **1995**, *352*, 625–632.
53. J. F. K. Huber; R. Van der Linden; E. Ecker; M. Oreans, *J. Chromatogr.* **1973**, *83*, 267–277.
54. L. Mondello; K. Bartle; A. Lewis, *Multidimensional Chromatography;* 2001.
55. J. F. K. Huber; E. Kenndler; G. Reich, *J. Chromatogr.* **1979**, *172*, 15–30.
56. J. C. Giddings, *Anal. Chem.* **1984**, *56*, 1258A–1260A, 1262A, 1264A, 1266A, 1268A, 1270A.
57. J. M. Davis; J. C. Giddings, *Anal. Chem.* **1985**, *57*, 2168–2177.
58. J. M. Davis; J. C. Giddings, *Anal. Chem.* **1985**, *57*, 2178–2182.
59. J. C. Giddings, *J. Chromatogr. A* **1995**, *703*, 3–15.
60. F. Regnier; G. Huang, *J. Chromatogr. A* **1996**, *750*, 3–10.
61. F. Erni; R. W. Frei, *J. Chromatogr.* **1978**, *149*, 561–569.
62. R. J. Simpson, Introduction to Chromatographic Methods for Protein and Peptide Purification. In *Purifying Proteins for Proteomics*; Cold Spring Habor Laboratory Press: Woodbury, N.Y., 2004; p 801.
63. K. K. Unger; K. Racaityte; K. Wagner; T. Miliotis; L. E. Edholm; R. Bischoff; G. Marko-Varga, *J. High Resolut. Chromatogr.* **2000**, *23*, 259–265.
64. F. Cacciola; P. Jandera; Z. Hajdu; P. Cesla; L. Mondello, *J. Chromatogr. A* **2007**, *1149*, 73–87.
65. P. Dugo; M. Herrero; T. Kumm; D. Giuffrida; G. Dugo; L. Mondello, *J. Chromatogr. A* **2008**, *1189*, 196–206.
66. Z. Liu; M. L. Lee, *J. Microcolumn Sep.* **2000**, *12*, 241–254.
67. P. Jandera, *J. Sep. Sci.* **2006**, *29*, 1763–1783.
68. P. Dugo; F. Cacciola; T. Kumm; G. Dugo; L. Mondello, *J. Chromatogr. A* **2008**, *1184*, 353–368.
69. R. E. Majors, *J. Chromatogr. Sci.* **1980**, *18*, 571–579.
70. M. T. W. Hearn, *Biologicals* **2001**, *29*, 159–178.
71. G. Guiochon; L. A. Beaver; M. F. Gonnord; A. M. Siouffi; M. Zakaria, *J. Chromatogr.* **1983**, *255*, 415–437.
72. A. J. P. Martin, *Ann. Rep. Progress Chem.* **1949**, *45*, 267–283.
73. P. Dugo; O. Favoino; R. Luppino; G. Dugo; L. Mondello, *Anal. Chem.* **2004**, *76*, 2525–2530.
74. P. Dugo; M. Herrero; D. Giuffrida; T. Kumm; G. Dugo; L. Mondello, *J. Agric. Food Chem.* **2008**, *56*, 3478–3485.
75. K. J. Mayfield; R. A. Shalliker; H. J. Catchpoole; A. P. Sweeney; V. Wong; G. Guiochon, *J. Chromatogr. A* **2005**, *1080*, 124–131.
76. E. Blahova; P. Jandera; F. Cacciola; L. Mondello, *J. Sep. Sci.* **2006**, *29*, 555–566.
77. L. Hu; X. Li; S. Feng; L. Kong; X. Su; X. Chen; F. Qin; M. Ye; H. Zou, *J. Sep. Sci.* **2006**, *29*, 881–888.
78. G. Cimpan; S. Gocan, *J. Liq. Chromatogr. Relat. Technol.* **2002**, *25*, 2225–2292.
79. M. Lambert; J.-L. Wolfender; D. Strk; S. B. Christensen; K. Hostettmann; J. W. Jaroszewski, *Anal. Chem.* **2007**, *79*, 727–735.
80. K.-H. Wagner; I. Elmadfa, *Ann. Nutr. Metab.* **2003**, *47*, 95–106.
81. A. Aharoni; M. A. Jongsma; H. J. Bouwmeester, *Trends Plant Sci.* **2005**, *10*, 594–602.
82. L. Mondello; K. D. Bartle; G. Dugo; P. Dugo, *J. High Resolut. Chromatogr.* **1994**, *17*, 312–314.
83. L. Mondello; P. Dugo; K. D. Bartle; G. Dugo; A. Cotroneo, *Flavour Fragr. J.* **1995**, *10*, 33–42.
84. J. Mookdasanit; H. Tamura, *Food Sci. Technol. Res.* **2002**, *8*, 367–372.
85. Y. Liu; L. Wang; J. H. Jung; S. Zhang, *Nat. Prod. Rep.* **2007**, *24*, 1401–1429.
86. C. Wood; T. E. Siebert; M. Parker; D. L. Capone; G. M. Elsey; A. P. Pollnitz; M. Eggers; M. Meier; T. Vossing; S. Widder; G. Krammer; M. A. Sefton; M. J. Herderich, *J. Agric. Food Chem.* **2008**, *56*, 3738–3744.
87. J. R. Hanson, *Nat. Prod. Rep.* **2007**, *24*, 1332–1341.
88. J. D. Connolly; R. A. Hill, *Nat. Prod. Rep.* **2007**, *24*, 465–486.
89. C. G. Benson; S. G. Wyllie; D. N. Leach; C. L. Mares; G. P. Fitt, *J. Agric. Food Chem.* **2001**, *49*, 2181–2184.
90. J. Warnke; R. Bandur; T. Hoffmann, *Anal. Bioanal. Chem.* **2006**, *385*, 34–45.
91. Y. Song; Q. He; P. Li; Y.-Y. Cheng, *J. Sep. Sci.* **2008**, *31*, 64–70.
92. P. Sahai; R. A. Vishwakarma; S. Bharel; A. Gulati; M. Z. Abdin; P. S. Srivastava; S. K. Jain, *Anal. Chem.* **1998**, *70*, 3084–3087.
93. A. Raffaelli; S. Pucci; F. Pietra, *Anal. Commun.* **1997**, *34*, 179–182.

94. L. Gobbo-Neto; N. P. Lopes, *J. Agric. Food Chem.* **2008**, *56*, 1193–1204.
95. R. Dolfinger; D. C. Locke, *Anal. Chem.* **2003**, *75*, 1355–1364.
96. S. Sturm; H. Stuppner, *Phytochem. Anal.* **2000**, *11*, 121–127.
97. Y.-C. Wang; Y.-S. Yang, *J. Chromatogr. B: Anal. Technol. Biomed. Life Sci.* **2007**, *850*, 392–399.
98. D. Savoia; C. Avanzini; T. Allice; E. Callone; G. Guella; F. Dini, *Antimicrob. Agents Chemother.* **2004**, *48*, 3828–3833.
99. M. J. Dubber; I. Kanfer, *J. Pharm. Biomed. Anal.* **2006**, *41*, 135–140.
100. B. Dinda; S. Debnath; Y. Harigaya, *Chem. Pharm. Bull. (Tokyo)* **2007**, *55*, 159–222.
101. B. Dinda; S. Debnath; Y. Harigaya, *Chem. Pharm. Bull. (Tokyo)* **2007**, *55*, 689–728.
102. S. Rodriguez; A. Marston; J. L. Wolfender; K. Hostettmann, *Curr. Org. Chem.* **1998**, *2*, 627–648.
103. B. Meier; O. Sticher, *J. Chromatogr.* **1977**, *138*, 453–457.
104. M. Willems, *Planta Med.* **1988**, *54*, 66–68.
105. K. Chervenkova; B. Nikolova-Damyanova, *J. Liq. Chromatogr. Relat. Technol.* **2000**, *23*, 741–753.
106. L. Ren; X. Xue; F. Zhang; Y. Wang; Y. Liu; C. Li; X. Liang, *Rapid Commun. Mass Spectrom.* **2007**, *21*, 3039–3050.
107. A. Aberham; S. Schwaiger; H. Stuppner; M. Ganzera, *J. Pharm. Biomed. Anal.* **2007**, *45*, 437–442.
108. O. Potterat; R. Von Felten; P. W. Dalsgaard; M. Hamburger, *J. Agric. Food Chem.* **2007**, *55*, 7489–7494.
109. T. Zhou; W. Zhao; G. Fan; Y. Chai; Y. Wu, *J. Chromatogr. B: Anal. Technol. Biomed. Life Sci.* **2007**, *858*, 296–301.
110. N. Dimov; K. Chervenkova; B. Nikolova-Damyanova, *J. Liq. Chromatogr. Relat. Technol.* **2000**, *23*, 935–947.
111. D. Schaufelberger; K. Hostettmann, *J. Chromatogr.* **1987**, *389*, 450–455.
112. A. Romani; P. Pinelli; N. Mulinacci; C. Galardi; F. F. Vincieri; L. Liberatore; A. Cichelli, *Chromatographia* **2001**, *53*, 279–284.
113. R.-W. Jiang; K.-L. Wong; Y.-M. Chan; H.-X. Xu; P. P.-H. But; P.-C. Shaw, *Phytochemistry* **2005**, *66*, 2674–2680.
114. A. Romani; P. Pinelli; N. Mulinacci; F. F. Vincieri; E. Gravano; M. Tattini, *J. Agric. Food Chem.* **2000**, *48*, 4091–4096.
115. S. Suryawanshi; N. Mehrotra; R. K. Asthana; R. C. Gupta, *Rapid Commun. Mass Spectrom.* **2006**, *20*, 3761–3768.
116. L. Boyer; R. Elias; K. Taoubi; L. Debrauwer; R. Faure; B. Baghdikian; G. Balansard, *Phytochem. Anal.* **2005**, *16*, 375–379.
117. D. Schaufelberger; K. Hostettmann, *J. Chromatogr.* **1985**, *346*, 396–400.
118. L. Zhou; J. Kang; L. Fan; X.-C. Ma; H.-Y. Zhao; J. Han; B.-r. Wang; D.-A. Guo, *J. Pharm. Biomed. Anal.* **2008**, *47*, 39–46.
119. M. Ruddat; O. H. Will, III, *Methods Enzymol.* **1985**, *111*, 189–200.
120. F. Khachik; G. R. Beecher; M. B. Goli; W. R. Lusby, *Methods Enzymol.* **1992**, *213*, 347–359.
121. N. E. Craft, *Methods Enzymol.* **1992**, *213*, 185–205.
122. H. Pfander; R. Riesen, *Carotenoids* **1995**, *1A*, 145–190.
123. R. Mateos; J. A. Garcia-Mesa, *Anal. Bioanal. Chem.* **2006**, *385*, 1247–1254.
124. U. Schweiggert; D. R. Kammerer; R. Carle; A. Schieber, *Rapid Commun. Mass Spectrom.* **2005**, *19*, 2617–2628.
125. N. L. Puspitasari-Nienaber; M. G. Ferruzzi; S. J. Schwartz, *J. Am. Oil Chem. Soc.* **2002**, *79*, 633–640.
126. A. B. Barua; J. A. Olson, *J. Chromatogr. B: Biomed. Sci. Appl.* **1998**, *707*, 69–79.
127. F. T. Gillan; R. B. Johns, *J. Chromatogr. Sci.* **1983**, *21*, 34–38.
128. S. H. Rhodes; A. G. Netting; B. V. Milborrow, *J. Chromatogr.* **1988**, *442*, 412–419.
129. L. Almela; J. M. Lopez-Roca; M. E. Candela; M. D. Alcazar, *J. Chromatogr.* **1990**, *502*, 95–106.
130. F. Khachik; G. R. Beecher; M. B. Goli; W. R. Lusby; J. C. Smith, Jr., *Anal. Chem.* **1992**, *64*, 2111–2122.
131. G. Panfili; A. Fratianni; M. Irano, *J. Agric. Food Chem.* **2004**, *52*, 6373–6377.
132. L. C. Sander; M. Pursch; B. Maerker; S. A. Wise, *Anal. Chem.* **1999**, *71*, 3477–3483.
133. L. C. Sander; K. E. Sharpless; M. Pursch, *J. Chromatogr. A* **2000**, *880*, 189–202.
134. L. Li; J. Qin; S.-a. Yin; G. Tang, *Chromatographia* **2007**, *65*, 91–94.
135. R. B. van Breemen, *Anal. Chem.* **1995**, *67*, 2004–2009.
136. Q. Su; K. G. Rowley; N. D. H. Balazs, *J. Chromatogr. B: Anal. Technol. Biomed. Life Sci.* **2002**, *781*, 393–418.
137. K. Albert, *Trends Anal. Chem.* **1998**, *17*, 648–658.
138. K. S. Epler; R. G. Ziegler; N. E. Craft, *J. Chromatogr., Biomed. Appl.* **1993**, *619*, 37–48.
139. M. Dachtler; T. Glaser; K. Kohler; K. Albert, *Anal. Chem.* **2001**, *73*, 667–674.
140. D. J. Hart; K. J. Scott, *Food Chem.* **1995**, *54*, 101–111.
141. R. B. Van Breemen; H. H. Schmitz; S. J. Schwartz, *Anal. Chem.* **1993**, *65*, 965–969.
142. R. B. van Breemen; C.-R. Huang; Y. Tan; L. C. Sander; A. B. Schilling, *J. Mass Spectrom.* **1996**, *31*, 975–981.
143. C. Rentel; S. Strohschein; K. Albert; E. Bayer, *Anal. Chem.* **1998**, *70*, 4394–4400.
144. P. A. Clarke; K. A. Barnes; J. R. Startin; F. I. Ibe; M. J. Shepherd, *Rapid Commun. Mass Spectrom.* **1996**, *10*, 1781–1785.
145. B. Finckh; A. Kontush; J. Commentz; C. Huebner; M. Burdelski; A. Kohlschuetter, *Anal. Biochem.* **1995**, *232*, 210–216.
146. B. Finckh; A. Kontush; J. Commentz; C. Hubner; M. Burdelski; A. Kohlschutter, *Methods Enzymol.* **1999**, *299*, 341–348.
147. F. Ladislav; P. Vera; S. Karel; V. Karel, *Curr. Anal. Chem.* **2005**, *1*, 93–102.
148. W. A. Oleszek, *J. Chromatogr. A* **2002**, *967*, 147–162.
149. W. Oleszek; Z. Bialy, *J. Chromatogr. A* **2006**, *1112*, 78–91.
150. C.-Z. Wang; E. McEntee; S. Wicks; J.-A. Wu; C.-S. Yuan, *J. Nat. Med.* **2006**, *60*, 97–106.
151. J. H. Gil; J. H. Jung; K.-J. Kim; M.-S. Kim; J. Hong, *Anal. Sci.* **2006**, *22*, 641–644.
152. D. C. Van Setten; G. J. Ten Hove; E. J. H. J. Wiertz; J. P. Kamerling; G. Van de Werken, *Anal. Chem.* **1998**, *70*, 4401–4409.
153. T. W. D. Chan; P. P. H. But; S. W. Cheng; I. M. Y. Kwok; F. W. Lau; H. X. Xu, *Anal. Chem.* **2000**, *72*, 1281–1287.
154. N. T. Nyberg; H. Baumann; L. Kenne, *Anal. Chem.* **2003**, *75*, 268–274.
155. R. Kasai; H. Yamaguchi; O. Tanaka, *J. Chromatogr.* **1987**, *407*, 205–210.
156. Y. Ikeda; M. Sugiura; C. Fukaya; K. Yokoyama; Y. Hashimoto; K. Kawanishi; M. Moriyasu, *Chem. Pharm. Bull. (Tokyo)* **1991**, *39*, 566–571.
157. H. Yamaguchi; H. Matsuura; R. Kasai; K. Mizutani; H. Fujino; K. Ohtani; T. Fuwa; O. Tanaka, *Chem. Pharm. Bull. (Tokyo)* **1986**, *34*, 2859–2867.
158. H. Frokiaer; S. A. J. Larsen; A. D. Sorensen; H. Sorensen; J. C. Sorensen; S. Sorensen, *Spec. Publ. – R. Soc. Chem.* **2001**, *269*, 74–76.
159. J. A. Beutler, *J. Liq. Chromatogr. Relat. Technol.* **1997**, *20*, 2415–2426.

160. K. Decroos; J. P. Vincken; L. Heng; R. Bakker; H. Gruppen; W. Verstraete, *J. Chromatogr. A* **2005**, *1072*, 185–193.
161. M.-R. Lee; C.-M. Chen; B.-H. Hwang; L.-M. Hsu, *J. Mass Spectrom.* **1999**, *34*, 804–812.
162. J. Lin; C. Wang, *J. Food Sci.* **2004**, *69*, C456–C462.
163. M. Jin; Y. Yang; B. Su; Q. Ren, *J. Chromatogr. A* **2006**, *1108*, 31–37.
164. M. A. Berhow; S. B. Kong; K. E. Vermillion; S. M. Duval, *J. Agric. Food Chem.* **2006**, *54*, 2035–2044.
165. A.-J. Lau; B.-H. Seo; S.-O. Woo; H.-L. Koh, *J. Chromatogr. A* **2004**, *1057*, 141–149.
166. D. Y. Q. Hong; A. J. Lau; C. L. Yeo; X. K. Liu; C. R. Yang; H. L. Koh; Y. Hong, *J. Agric. Food Chem.* **2005**, *53*, 8460–8467.
167. J.-B. Wan; P. Li; S. Li; Y. Wang; T. T.-X. Dong; K. W.-K. Tsim, *J. Sep. Sci.* **2006**, *29*, 2190–2196.
168. L. Li; T. Rong; J. Dou; F. Song; Z. Liu; S. Liu, *Anal. Chim. Acta* **2005**, *536*, 21–28.
169. Y. W. Ha; Y.-C. Na; J.-J. Seo; S.-N. Kim; R. J. Linhardt; Y. S. Kim, *J. Chromatogr. A* **2006**, *1135*, 27–35.
170. L. Yi; L.-W. Qi; P. Li; Y.-H. Ma; Y.-J. Luo; H.-Y. Li, *Anal. Bioanal. Chem.* **2007**, *389*, 571–580.
171. M. Liang; Z. Zheng; Y. Yuan; L. Kong; Y. Shen; R. Liu; C. Zhang; W. Zhang, *Phytochem. Anal.* **2007**, *18*, 428–435.
172. F. Sun; Q. He; P. Shi; P. Xiao; Y. Cheng, *Rapid Commun. Mass Spectrom.* **2007**, *21*, 3743–3750.
173. M. Ganzera; E. Bedir; I. A. Khan, *J. Pharm. Sci.* **2001**, *90*, 1752–1758.
174. J. Chen; Y. Song; P. Li, *J. Chromatogr. A* **2007**, *1157*, 217–226.
175. A. Maria; J.-P. Girault; Z. Saatov; J. Harmatha; L. Dinan; R. Lafont, *J. Chromatogr. Sci.* **2005**, *43*, 149–157.
176. M.-P. Marco; F. J. Sanchez-Baeza; F. Camps; J. Coll, *J. Chromatogr.* **1993**, *641*, 81–87.
177. M. DellaGreca; B. D'Abrosca; A. Fiorentino; L. Previtera; A. Zarrelli, *Chem. Biodiversity* **2005**, *2*, 457–462.
178. V. Rybin; E. Boltenkov; E. Novozhilova, *Nat. Prod. Commun.* **2007**, *2*, 1101–1104.
179. D. Louden; A. Handley; S. Taylor; E. Lenz; S. Miller; I. D. Wilson; A. Sage; R. Lafont, *J. Chromatogr. A* **2001**, *910*, 237–246.
180. Y. Meng; P. Whiting; L. Zibareva; G. Bertho; J.-P. Girault; R. Lafont; L. Dinan, *J. Chromatogr. A* **2001**, *935*, 309–319.
181. N. Kaouadji; R. Lafont, *J. Chromatogr.* **1990**, *505*, 408–412.
182. R. Ho; J.-P. Girault; P.-Y. Cousteau; J.-P. Bianchini; P. Raharivelomanana; R. Lafont, *J. Chromatogr. Sci.* **2008**, *46*, 102–110.
183. R. E. Isaac; N. P. Milner; H. H. Rees, *J. Chromatogr.* **1982**, *246*, 317–322.
184. T. M. Landon; B. A. Sage; B. J. Seeler; J. D. O'Connor, *J. Biol. Chem.* **1988**, *263*, 4693–4697.
185. V. D. Gildengorn, *J. Chromatogr. A* **1996**, *730*, 147–152.
186. D. Louden; A. Handley; R. Lafont; S. Taylor; I. Sinclair; E. Lenz; T. Orton; I. D. Wilson, *Anal. Chem.* **2002**, *74*, 288–294.
187. F. Bourgaud; A. Hehn; R. Larbat; S. Doerper; E. Gontier; S. Kellner; U. Matern, *Phytochem. Rev.* **2006**, *5*, 293–308.
188. K. Glowniak; M. L. Bieganowska, *J. Liq. Chromatogr.* **1985**, *8*, 2927–2947.
189. N. Nykolov; T. Iossifova; E. Vassileva; I. Kostova; G. Stoev, *Phytochem. Anal.* **1993**, *4*, 86–88.
190. T.-T. Wang; H. Jin; Q. Li; W.-M. Cheng; Q.-Q. Hu; X.-H. Chen; K.-S. Bi, *Chromatographia* **2007**, *65*, 477–481.
191. M.-J. Ahn; M. K. Lee; Y. C. Kim; S. H. Sung, *J. Pharm. Biomed. Anal.* **2008**, *46*, 258–266.
192. C. Sproll; W. Ruge; C. Andlauer; R. Godelmann; D. W. Lachenmeier, *Food Chem.* **2008**, *109*, 462–469.
193. H. J. Thompson; S. A. Brown, *J. Chromatogr.* **1984**, *314*, 323–336.
194. M. L. Bieganowska; K. Glowniak, *Chromatographia* **1988**, *25*, 111–116.
195. P. Harmala; H. Vuorela; E. L. Rahko; R. Hiltunen, *J. Chromatogr.* **1992**, *593*, 329–337.
196. P. Vuorela; E.-L. Rahko; R. Hiltunen; H. Vuorela, *J. Chromatogr. A* **1994**, *670*, 191–198.
197. P. Dugo; L. Mondello; E. Sebastiani; R. Ottana; G. Errante; G. Dugo, *J. Liq. Chromatogr. Relat. Technol.* **1999**, *22*, 2991–3005.
198. A. Tosun; M. Baba; T. Okuyama, *J. Nat. Med.* **2007**, *61*, 402–405.
199. K. Hunter; E. A. Sharp; A. Newton, *J. Chromatogr.* **1988**, *435*, 83–95.
200. K. Robards, *J. Chromatogr. A* **2003**, *1000*, 657–691.
201. D. J. Daigle; E. J. Conkerton, *J. Chromatogr.* **1982**, *240*, 202–205.
202. D. J. Daigle; E. J. Conkerton, *J. Liq. Chromatogr.* **1983**, *6*, 105–118.
203. D. J. Daigle; E. J. Conkerton, *J. Liq. Chromatogr.* **1988**, *11*, 309–325.
204. H. M. Merken; G. R. Beecher, *J. Agric. Food Chem.* **2000**, *48*, 577–599.
205. K. Robards; M. Antolovich, *Analyst* **1997**, *122*, 11R–34R.
206. E. Conde; E. Cadahia; M. C. Garcia-Vallejo, *Chromatographia* **1995**, *41*, 657–660.
207. J. L. Wolfender; S. Rodriguez; K. Hostettmann, *J. Chromatogr. A* **1998**, *794*, 299–316.
208. P. C. H. Hollman; J. M. P. Van Trijp; M. N. C. P. Buysman, *Anal. Chem.* **1996**, *68*, 3511–3515.
209. S. H. Hansen; A. G. Jensen; C. Cornett; I. Bjornsdottir; S. Taylor; B. Wright; I. D. Wilson, *Anal. Chem.* **1999**, *71*, 5235–5241.
210. J. Zhang; M. B. Satterfield; J. S. Brodbelt; S. J. Britz; B. Clevidence; J. A. Novotny, *Anal. Chem.* **2003**, *75*, 6401–6407.
211. B. D. Davis; J. S. Brodbelt, *Anal. Chem.* **2005**, *77*, 1883–1890.
212. P. Waridel; J. L. Wolfender; K. Ndjoko; K. R. Hobby; H. J. Major; K. Hostettmann, *J. Chromatogr. A* **2001**, *926*, 29–41.
213. G. Zgorka; A. Hajnos, *Chromatographia* **2003**, *57*, S/77–S/80.
214. M. V. Martinez-Ortega; M. C. Garcia-Parrilla; A. M. Troncoso, *Anal. Chim. Acta* **2004**, *502*, 49–55.
215. C. Cavaliere; F. Cucci; P. Foglia; C. Guarino; R. Samperi; A. Lagana, *Rapid Commun. Mass Spectrom.* **2007**, *21*, 2177–2187.
216. P. Kozminski; A. M. O. Brett, *Anal. Lett.* **2006**, *39*, 2687–2697.
217. M. C. Pietrogrande; C. Bighi; G. Blo; Y. D. Kahie; P. Reschiglian; F. Dondi, *Chromatographia* **1989**, *27*, 625–627.
218. M. V. Piretti; P. Doghieri, *J. Chromatogr.* **1990**, *514*, 334–342.
219. F. Buiarelli; G. P. Cartoni; F. Coccioli; E. Ravazzi, *Chromatographia* **1991**, *31*, 489–492.
220. J. C. Herrera; A. J. R. Romero; O. E. Crescente; M. Acosta; S. Pekerar, *J. Chromatogr. A* **1996**, *740*, 201–206.
221. R. L. Prior; S. A. Lazarus; G. Cao; H. Muccitelli; J. F. Hammerstone, *J. Agric. Food Chem.* **2001**, *49*, 1270–1276.
222. M. Lopez-Serrano; Ros A. Barcelo, *J. Chromatogr. A* **2001**, *919*, 267–273.
223. A. Yanagida; T. Shoji; Y. Shibusawa, *J. Biochem. Biophys. Methods* **2003**, *56*, 311–322.
224. K. Hashizume; S. Kida; T. Samuta, *J. Agric. Food Chem.* **1998**, *46*, 4382–4386.
225. J. He; C. Santos-Buelga; N. Mateus; V. de Freitas, *J. Chromatogr. A* **2006**, *1134*, 215–225.
226. D. Bongartz; A. Hesse, *J. Chromatogr. B. Biomed. Appl.* **1995**, *673*, 223–230.
227. J. Xie; L. Zhu; H. Luo; L. Zhou; C. Li; X. Xu, *J. Chromatogr. A* **2001**, *934*, 1–11.
228. R. Weiss; A. Molinelli; M. Jakusch; B. Mizaikoff, *Bioseparation* **2002**, *10*, 379–387.

229. G. Theodoridis; M. Lasakova; V. Skerikova; A. Tegou; N. Giantsiou; P. Jandera, *J. Sep. Sci.* **2006**, *29*, 2310–2321.
230. T. J. Mabry; K. R. Markham; M. B. Thomas, *The Systematic Identification of Flavonoids*; 1970.
231. A. P. Neilson; R. J. Green; K. V. Wood; M. G. Ferruzzi, *J. Chromatogr. A* **2006**, *1132*, 132–140.
232. R. Verpoorte; A. Baerheim Svendsen, *Journal of Chromatography Library, Vol. 23: Chromatography of Alkaloids, Part B: GasLiquid Chromatography and High-Performance Liquid Chromatography*; Elsevier: Amsterdam, The Netherlands, 1984; p 450.
233. J. Stockigt; Y. Sheludko; M. Unger; I. Gerasimenko; H. Warzecha; D. Stockigt, *J. Chromatogr. A* **2002**, *967*, 85–113.
234. P. J. Houghton, *J. Chromatogr. A* **2002**, *967*, 75–84.
235. P. Kulanthaivel; S. W. Pelletier, *J. Chromatogr.* **1987**, *402*, 366–370.
236. G. Bringmann; K. Messer; M. Wohlfarth; J. Kraus; K. Dumbuya; M. Rueckert, *Anal. Chem.* **1999**, *71*, 2678–2686.
237. X. Zhu; B. Chen; M. Ma; X. Luo; F. Zhang; S. Yao; Z. Wan; D. Yang; H. Hang, *J. Pharm. Biomed. Anal.* **2004**, *34*, 695–704.
238. H. Dong; Z. Liu; F. Song; Z. Yu; H. Li; S. Liu, *Rapid Commun. Mass Spectrom.* **2007**, *21*, 3193–3199.
239. I. Gerasimenko; Y. Sheludko; M. Unger; J. Stockigt, *Phytochem. Anal.* **2001**, *12*, 96–103.
240. W. Wu; F. Song; C. Yan; Z. Liu; S. Liu, *J. Pharm. Biomed. Anal.* **2005**, *37*, 437–446.
241. M. H. Stutz; S. Sass, *Anal. Chem.* **1973**, *45*, 2134–2136.
242. R. Verpoorte; T. Mulder-Krieger; M. J. Verzijl; J. M. Verzijl; A. Baerheim Svendsen, *J. Chromatogr.* **1983**, *261*, 172–175.
243. G. D. Manners; J. A. Pfister, *Phytochem. Anal.* **1993**, *4*, 14–18.
244. M. Waksmundzka-Hajnos; A. Petruczynik, *J. Liq. Chromatogr. Relat. Technol.* **2004**, *27*, 2247–2267.
245. J. M. Huen; J. P. Thevenin, *HRC & CC J. High Resolut. Chromatogr. Chromatogr. Commun.* **1979**, *2*, 154.
246. L. W. Doner; A. F. Hsu, *J. Chromatogr.* **1982**, *253*, 120–123.
247. R. M. Riggin; P. T. Kissinger, *Anal. Chem.* **1977**, *49*, 530–533.
248. Q.-W. Yu; B. Lin; Y.-Q. Feng; F.-P. Zou, *J. Liq. Chromatogr. Relat. Technol.* **2008**, *31*, 64–78.
249. I. Fitos; J. Visy; M. Simonyi; J. Hermansson, *J. Chromatogr.* **1992**, *609*, 163–171.
250. M. Nakamura; M. Ono; T. Nakajima; Y. Ito; T. Aketo; J. Haginaka, *J. Pharm. Biomed. Anal.* **2005**, *37*, 231–237.

Biographical Sketches

Reinhard I. Boysen holds a Ph.D. in natural sciences (Dr. rer. nat.) from the Faculty of Chemistry at the Freie Universität Berlin, Germany, where he later specialized in analytical biochemistry. Thereafter, he joined the Centre for Bioprocess Technology in the Department of Biochemistry and Molecular Biology of Monash University and worked in the fields of peptide synthesis, separation science, and thermodynamics of peptide/protein-immobilized–ligand interactions. He is currently a research fellow at the ARC Special Research Centre for Green Chemistry at Monash University working in the fields of separation science of chemical and biological molecules (multidimensional capillary-based separation methodologies including separations with molecularly imprinted polymers), mass spectrometry-based medical diagnostics, nanochemistry/nanotechnology (chemical nanoarrays), and in the investigation of interactions between proteins/DNA/cells and nonbiological surfaces.

Milton T. W. Hearn is currently professor of chemistry and director, ARC Special Research Centre for Green Chemistry, Monash University, Australia. He received his B.Sc. (Hons. First Class), Ph.D., and D.Sc. degrees from the University of Adelaide, Australia. Prior to joining Monash University, he held an NRC postdoctoral fellowship at the Department of Chemistry, University of British Columbia; ICI fellow at the Dyson Perrins Laboratory and research fellowships at Christchurch and Wolfson Colleges at Oxford University; MRCNZ senior research fellow at the University of Otago, and NHMRC Principal Research Fellow and McGauran Fellow at the St. Vincent's Institute of Medical Research, University of Melbourne. From 1986 to 2002 he was professor of biochemistry and director, Centre for Bioprocess Technology, Monash University. He has held distinguished professorships at Yale University, University of Paris, Johannes Gutenberg University, and Himeji Institute of Technology. His research interests focus on (1) the analysis, purification, and characterization of chemical and biochemical molecules from the nanoscale through to the process scale, including advanced technologies of importance to the chemical, pharmaceutical, and biotechnological industries, of which many of these developments have been successfully commercialized and (2) the structure/function of bioactive compounds, bioinspired synthesis, molecular imprinting, surface and combinatorial chemistry, and aspects of nanobiotechnology. Professor Hearn has authored 505 scientific publications, several books, and issued patents related to developments in chemistry/biochemistry, (bio)nanotechnology, protein purification, affinity chromatography, other 'downstream' aspects of biotechnology, as well as the development of several lead (bio)pharmaceutical compounds.

9.03 Introduction to Macromolecular X-Ray Crystallography

Esko Oksanen and Adrian Goldman, University of Helsinki, Helsinki, Finland

9.03.1 Introduction

Macromolecular crystallography is a powerful method for investigating the atomic structure of proteins and nucleic acids and thereby unravelling the molecular mechanisms of their functions. To visualize atoms, light must have a wavelength about the size of an atom, so that the atoms can diffract light, identical with cells and organelles that scatter visible light. For this purpose X-rays are required. However, since the interaction of X-rays with matter is weak, scattering is also weak unless there are multiple molecules that are ordered in the same way, which means we require crystals.

We therefore describe the basis of macromolecular crystallography and provide a summary of how to understand the results of a crystallographic experiment. We start with a mathematical description of what a crystal means in terms of symmetry; this applies to all crystals, whether macromolecular or not. Later, we describe how protein crystals grow by using the hanging drop and sitting drop vapor diffusion methods; this explains why protein crystals are so fragile and scatter X-rays very weakly.

The phenomenon of diffraction and its description as a Fourier transform (FT) is explained. The measured intensity of the diffracted X-rays related to the FT of the electron density, and the electron density – seen as an electron density map – is related to the (inverse) Fourier sum of the intensity of the diffracted X-rays. As we can only measure their intensity, we do not know the phases of the diffracted X-rays; we have to determine them to solve the structure. Therefore, three principal methods are used, two experimental approaches (isomorphous replacement and anomalous scattering) and one based on known structures (molecular replacement).

Next, we describe how, the 'electron density map' is improved and interpreted in terms of the atoms of a structure. The inconsistency between this structure and the experimental data is later minimized to make sure it is as accurate as possible. Finally, we provide an example of a crystal structure from recent literature and summarize the various statistics reported in papers on crystal structure.

9.03.2 Why Crystallography?

The standard approach to study minute details is microscopy, in which light scattered by the specimen is focussed onto the image plane by a lens. The smallest observable detail, however, is limited to half of the wavelength of light (∼300–700 nm). The inter-atomic distances in organic molecules are ∼0.1–0.2 nm; therefore, we cannot observe atoms under a light microscope but require light of a shorter wavelength. Photons of this wavelength are X-rays, and so an ‘X-ray microscope’, if it existed, would allow us to visualize atoms. Unfortunately, the refractive index of X-rays is so small that the lenses required to focus X-rays in an X-ray microscope are impossible to make. The scattered radiation, however, still contains the information about the structure of a molecule. It is not practical to image a single molecule because >99% of the X-rays pass straight through, hence to obtain any appreciable signal a macroscopic sample is required.

This imposes a further condition: the molecules must line up with each other in a well-defined spatial arrangement; that is, they must form a crystal. From such a crystal, the structure of a macromolecule may be determined by using single-crystal X-ray diffraction. The method is essentially the same for all biological macromolecules or complexes. However, as most of the structures determined are of proteins, we often refer this method as protein X-ray crystallography.

What topics can be addressed by X-ray crystallography? In recent years, the structures of important biological systems have been solved, for example, β_2-adrenergic receptor, which upon binding to adrenaline or nor-adrenaline causes the ‘fight-or-flight’ response[1] and the structure of the ribosome, which translates the messenger RNA in the cell into protein.[2] The ribosome is also an important drug target; the macrolide antibiotics like erythromycin bind to the 50S ribosomal subunit, and the structure of the ribosome explains how they work, and how mutations in the bacterial ribosome lead to antibiotic resistance.

The understanding of the degradation of natural products such as camphor has been greatly enhanced by understanding the catalytic cycle of the cytochrome P-450 enzyme P-450cam in structural detail.[3,4] These enzymes catalyze the addition of O_2 to nonactivated hydrocarbons at room temperatures and pressures – a reaction that requires high temperature to proceed in the absence of a catalyst. *O*-Methyltransferases are central to the secondary metabolic pathway of phenylpropanoid biosynthesis. The structural basis of the diverse substrate specificities of such enzymes has been studied by solving the crystal structures of chalcone *O*-methyltransferase and isoflavone *O*-methyltransferase complexed with the reaction products.[5] Structures of these and other enzymes are obviously important for the development of biomimetic and thus environmentally more friendly approaches to natural product synthesis.

9.03.3 Protein Crystals

What is a crystal? We need a mathematical description of a crystal in order to understand, even in a qualitative way, how crystals diffract X-rays and why we see the patterns we do?

Let us start from a crystal of salt, sugar, diamond – or even the enzyme inorganic pyrophosphatase (**Figure 1**). Why do they have sharp edges and regular faces? The reason is that crystals are macroscopic objects in which the constituent atoms or molecules arrange themselves in the same manner. This means that an ideal crystal consists of a series of repeated units (the unit cell, see below), with each unit containing the same arrangement of atoms inside it – known as the motif (**Figure 2**). Exact motions – translations – relate the atoms to each other; clearly they form an array – a lattice of indistinguishable points – and we can move from one point (O in **Figure 2**) to an indistinguishable point (P in **Figure 2**) along a straight line. Such a vector between two equivalent points is known as a lattice vector. Three noncoplanar lattice vectors, called the basis vectors[1] **a**, **b**, and **c** define a coordinate system. Any point in the crystal can then be referred to by a vector ($x\mathbf{a} + y\mathbf{b} + z\mathbf{c}$), where x, y, z are the coefficients in the **a**, **b**, and **c** directions – just as in a three-dimensional Cartesian coordinate system. Unlike a normal Cartesian system, the basis vectors are neither necessarily perpendicular to each other nor of the same length, and so they do not enclose a cuboid, but a general prismoid shape.

[1] Vectors are marked with bold type so **a** is a vector and a is a scalar.

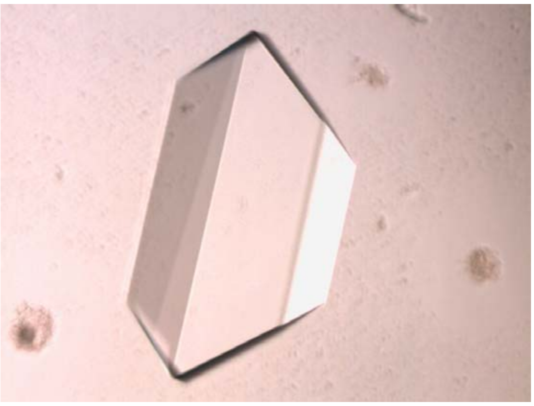


Figure 1 A crystal of yeast inorganic pyrophosphatase grown by temperature-controlled batch crystallization.[23] The longest dimension of the crystal is ~700 μm.

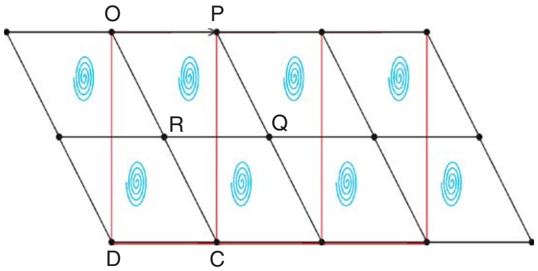


Figure 2 A schematic representation of a protein crystal in which the motif (protein molecule) is shown as a spiral in cyan. The smallest unit cell (OPQR) is shown in black; it is a rhombus: lozenge-shaped with all sides equal. A larger rectangular unit cell (OPCD) is shown in red. This cell leaves one lattice point in the middle and is known as *centered*.

This prismoid box is called the unit cell, and the entire crystal can be constructed by translating the box parallel to its edges. It is possible to choose many different unit cells, as shown in **Figure 2**, but we normally choose the smallest unit cell. However, in some cases a more convenient, but larger, cell is chosen as seen in **Figure 2**, where the rectangular centered cell (red) would be chosen. Such choices make cells easier to understand. Clearly, a rectangle picked up and rotated 180° looks the same; this is equivalent to the statement that a rectangular unit cell has twofold symmetry: that is, 90° angles and unequal edges. The same is true of the lozenge-shape in **Figure 2** but it is not as obvious.

Up to this point, our description has been general; it applies to crystals of rhenium chloride as accurately as to crystals of the ribosome. In addition to the translational symmetry (periodicity) that is inherent in the definition of a crystal, other symmetry can occur, but the kinds that can occur are restricted to crystals of biological macromolecules. Because the molecules are chiral, the symmetry operations in crystals must not change the handedness of the molecule, and so mirror planes, inversions, and 'glide planes' (sliding mirror planes) do not occur. This leaves only rotations and 'screws' (helical-type symmetry, sliding rotations).

Rotations of 60°, 90°, 120°, or 180° are the only ones allowed,[2] corresponding to six-, four-, three-, and twofold rotations. In addition, screw axes can occur, where the molecule is rotated by the same angles – 60°, 90°, 120°, 180° – and translated by a fraction of one of the lattice vectors **a**, **b**, or **c**. These have symbols like 2_1 (a 180° rotation followed by a translation of 1/2 of a lattice vector) or 4_3 (a 90° rotation followed by a translation of 3/4 of a lattice vector). These symmetry operators, lined up through the entire crystal, are the crystallographic symmetry operators.

[2] Other rotations are not compatible with the translational symmetry, which is the essence of a crystal. For a detailed explanation, see, for example, Giacovazzo *et al.*[6]

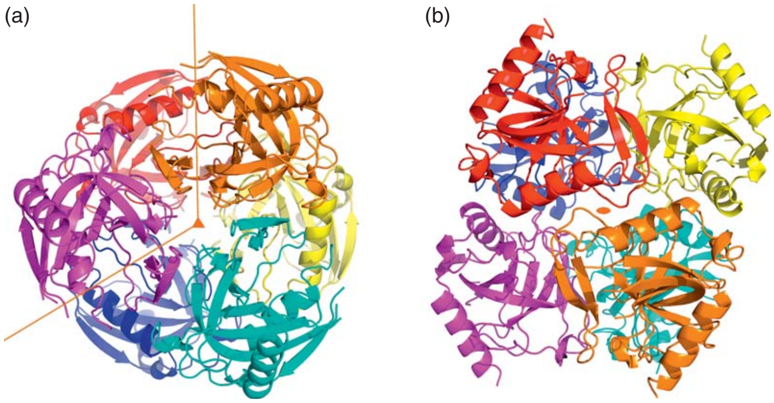


Figure 3 The *Escherichia coli* pyrophosphatase hexamer in a standard protein representation: the spirals are α-helices and the arrows are β-strands. The unit cell axes are marked with orange lines and each monomer has a different color (a) viewed along one of the twofold axes. The twofold rotation axis, marked with an ellipse, relates the orange and red monomers, and the yellow and purple monomers (and the hard-to-see cyan and blue monomers) to each other. (b) A view along the threefold axis, marked with a triangle. The threefold axis relates the orange, purple, and cyan monomers to each other, as well as the yellow, red, and blue ones to each other.

The presence of rotational or screw symmetry means that the unit cell has internal symmetry. Therefore, only part of the unit cell, known as the asymmetric unit, is needed to uniquely define the unit cell. (The asymmetric unit may also contain more than one molecule, related by movements – symmetry operations – that are not part of the crystal symmetry – noncrystallographic symmetry operators. This can be very important in determining the protein structure, as discussed in Section 9.03.9.3).

The convention we have described implies a hierarchy. We can generate the unit cell from the asymmetric unit by applying the various additional crystallographic symmetries: rotations, screws; and we can generate the entire crystal by translating the unit cell parallel to its edges – by its lattice vectors. Indeed, multimeric proteins sometimes crystallize so that the asymmetric unit contains only one monomer and the other monomers in the biological multimer are related by crystallographic (rotational) symmetry operators. One such example is hexameric *Escherichia coli* pyrophosphatase,[7] which has D_3 (32) point group symmetry[3] and crystallizes in space group R32, where each lattice point also has D_3 symmetry. The point group describes the rotational symmetries of an object – be it a molecule or something else. For *E. coli* pyrophosphatase, the point group 32 means there is a threefold (120°) rotation perpendicular to a twofold (180°) rotation. In the crystal, the center of the hexamer and a lattice point with the same symmetry coincide and therefore the asymmetric unit of the crystal contains a monomer. Crystal symmetry then relates the monomers in the hexamer (**Figure 3**).

The combination of rotational and translational symmetry defines the space group of the crystal. It is shown that 235 space groups exist, but only 65 allow the handedness of the molecule to be preserved, and so only 65 can occur in macromolecular crystallography. The space groups are numbered, but are commonly referred to by their symbols, such as $P2_12_12_1$. The most common in macromolecular crystallography are $P2_12_12_1$, P1, $P2_1$, and C2.

9.03.4 Obtaining Protein Crystals

A crystal is a very precisely ordered aggregate that represents the thermodynamically most favorable state under the conditions of crystallization. Since attractive forces between protein molecules are not very specific, successful crystallization requires both a very pure protein sample (typically ≈ 99% pure) and a careful search for the right conditions. In addition to favorable thermodynamics, crystal growth must also be kinetically favored over nonspecific aggregation. This often means a relatively slow process and while some protein crystals grow in hours, most take weeks to months to form.

[3] The point group, which is D_3 in the Schönflies notation used for example in molecular spectroscopy, is called 32 in the International (or Hermann–Mauguin) notation used by crystallographers.

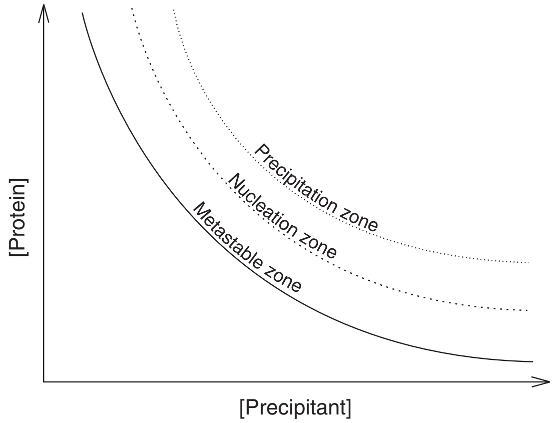


Figure 4 Phase diagram for a protein solution. In the undersaturation (soluble) zone crystals do not grow but dissolve; the first line marks the saturation limit. Above that, the solution is supersaturated *and* metastable with respect to the crystals; existing crystals will grow, but no spontaneous nucleation occurs. In the nucleation zone, new crystals form on their own, and in the precipitation zone nonspecific aggregation dominates.

Proteins, like any other molecule, have a certain solubility limit, above which an aggregated state (either amorphous or crystalline) is thermodynamically favored. The solubility of a protein as a function of some variable like temperature, pH, or the concentration of a precipitant such as salt or polyethylene glycol (PEG) may be represented with a phase diagram (Figure 4).

There is, however, an activation energy associated with the formation of a crystal from homogeneous solution. Therefore, immediately above the solubility limit the nucleation of crystals do not occur, but existing crystals grow. The solution is metastable. Above this metastable zone spontaneous nucleation occurs, resulting in a large number of small crystals or a polycrystalline precipitate. The objective of protein crystallization is to produce a limited number of nuclei and allow them to grow to sufficient size under supersaturating conditions. This may be achieved by slowly changing the concentration of the protein, the precipitant, or both. Variables such as temperature or pH are more difficult to change in a gradual way and hence are less often used. By far the most popular crystallization technique is vapor diffusion (Figure 5) because it can easily be done in a multi-well format and because it normally increases both protein and precipitant concentration at the same time. The protein solution – typically at a concentration of around 10 mg ml^{-1}[4] – is mixed with a precipitant

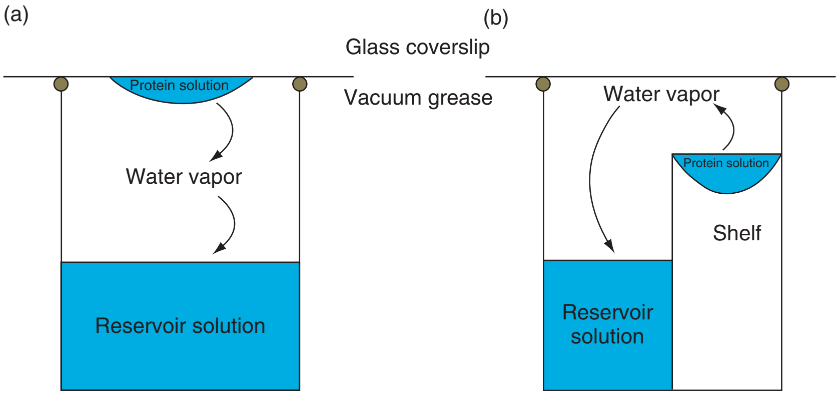


Figure 5 Setups for vapor diffusion. (a) Hanging drop vapor diffusion. (b) Sitting drop vapor diffusion in a modern 96-well plate. In both cases, a greased coverslip seals the well from the outside atmosphere, allowing equilibration *via* the vapor phase.

[4] Protein crystallographers usually measure protein concentration in mg ml^{-1} instead of molar units; for a 10 kDa protein 10 mg ml^{-1} would be 10 mmol^{-1}.

solution in a drop and placed in a gastight chamber with a reservoir of a more concentrated precipitant solution. The activity of the volatile components in the drop and the well – usually just water, but sometimes low-molecular weight alcohols too – equilibrate through the vapor phase. If, as is usual, the activity of the water in the drop is lower than in the well, there will be slow evaporation of water from the drop. The concentrations of both the protein and the precipitant in the drop increase, corresponding to a diagonal movement on the phase diagram (**Figure 6**).

There are other techniques, however, including microbatch crystallization, where the protein and precipitant are just mixed at the final supersaturation concentration. Free interface diffusion is similar to microbatch but the two components have to diffuse toward each other; the concentrations of both protein and precipitant therefore vary with distance from the original interface. In microdialysis, the precipitant solution is allowed to equilibrate with the protein solution through a semipermeable membrane, which permits passage of the precipitant but not the protein (**Figure 7**). Of these techniques, the first two also lend themselves to automation.

Inasmuch as the right conditions for crystallization cannot be predicted, a large number of conditions (precipitant, pH, temperature, protein concentration, additives, etc.) need to be screened to produce a crystal suitable for data collection. To minimize the amount of precious protein material used in these preliminary experiments and avoid the large amount of manual labor involved, automation is becoming very common.[8] When setups are done by hand, the pipetting is usually done using standard air-displacement pipettes, which are extremely inaccurate under 1 µl, making this the minimal practical protein volume per experiment. All of the various robotic systems can use sample volumes as low as 50 nl, and some as little as 1 nl.

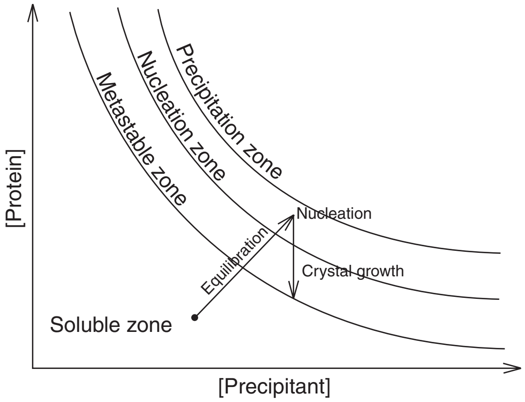


Figure 6 Movement on the phase diagram in a vapor diffusion experiment. The simultaneous increase of both precipitant and protein concentrations corresponds to a diagonal movement in the phase diagram. Once crystal nucleation occurs, the growing crystals consume the protein in the solution, until the solution is no longer supersaturated.

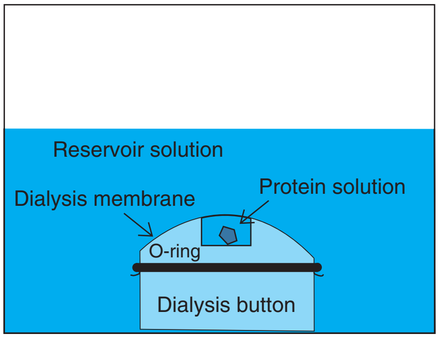


Figure 7 A schematic representation of a crystallization setup using a microdialysis button. The reservoir solution contains the precipitant, which slowly diffuses to the small depression in the button through the semi-permeable membrane.

Even though protein crystals are technically solids, the properties of protein crystals and ionic and molecular solids are very different. Strong covalent or ionic bonds hold crystals of ionic and molecular solids together. Even sugar crystals ($C_{12}H_{22}O_{11}.6H_2O$) are stabilized by seven hydrogen bonds per molecule of 23 nonhydrogen atoms. Protein crystals, on the other hand, are held together by relatively few noncovalent interactions, mainly hydrogen bonds and electrostatic interactions. For example, in yeast inorganic pyrophosphatase there are just 24 hydrogen bonds per 4600 nonhydrogen atoms connecting one asymmetric unit to its neighbors; that is, holding the crystal together. The forces due to these few interactions are thus at least 10 times weaker than the forces that maintain proteins in a folded conformation. Besides, in contrast to ionic and molecular crystals, disordered solvent typically comprises 30–70% of the volume of protein crystals. Although these properties make protein crystals difficult to grow and fragile to handle, they also mean that the structure the protein adopts in the crystal is generally the same as that in solution.[9] Comparisons of crystal structures to solution structures determined by nuclear magnetic resonance (NMR) spectroscopy[10] have shown that the differences between the crystal structure and the NMR structure are indeed smaller than the error margin. Artifacts may occur when multiple conformations of the protein exist in solution, but only one conformation forms crystals. Also the side chain conformations at crystal contacts (the points where molecules related by crystallographic symmetry interact) may be artifacts in the sense that there is no specific conformation in solution, but one or two specific conformations are seen in the crystal structure. In all cases, however, the X-ray structure represents one of the lowest-lying equilibrium states of the molecule, in both its overall shape and in the conformations of individual side chains. It is also worth noting that the total concentration of proteins in the cell is much closer to the protein concentration of a crystal than to the typical protein concentration in an NMR sample tube. In other words, the natural environment of at least intracellular proteins is a very concentrated solution.

9.03.5 Principles of Diffraction

We now briefly explain why the diffraction pattern looks the way it does (Figure 8). To start with: why are there spots? After all, the atoms in the crystals are not points. Why is there – at least in some representations of the diffraction pattern – clear symmetry (Figure 8(a)) and what, in outline, determines the distinct pattern of lighter and darker spots (Figure 8)? We start with the classical Bragg description of X-ray diffraction, which only explains where the spots will be, not the pattern of dark and light. We then outline how the same phenomenon can be described using FTs. It turns out that the pattern of dark and light spots – the intensities of the diffraction spots – that we can measure is related to the electron density in the crystals (which interests us).

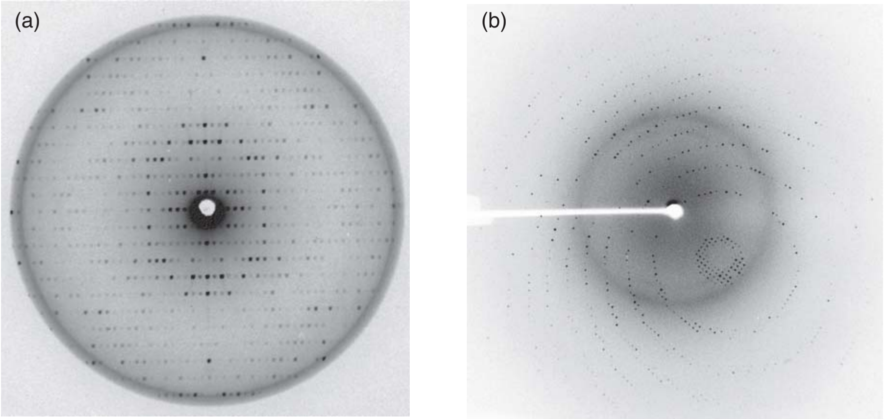


Figure 8 X-ray diffraction images. (a) A precession photograph of muconate lactonizing enzyme. The fourfold symmetry in the diffraction pattern is clearly visible. This gives an undistorted view of the reciprocal lattice but are no longer used because they are not as efficient as rotation images. (b) A rotation image of hen's egg white lysozyme. This easily obtainable image gives a distorted projection of the reciprocal lattice, but this is no obstacle for modern programs.

X-rays are scattered mainly by electrons and the elastic, coherent scattering process that gives rise to diffraction is called Thompson scattering. The intensity of the scattering decreases as the angle between the scattered beam and the original beam, called 2θ, increases. The incident photons will experience an identical environment at equivalent points, that is, points related through a lattice vector, and so will scatter in the same way at each of these points. Generally, the photons scattered by different points in the crystal interfere destructively and cancel each other out, but at certain values of 2θ constructive interference occurs and diffraction is observed. Why should this be so?

The regular array of lattice points forms planes, in the same way that trees planted at regular intervals form different rows as seen from a passing car (**Figure 9**). If we take such a plane of equivalent positions at an angle θ relative to the incident beam, we can think of the scattering by that plane as if it were a reflection from a mirror – that is, the angle of incidence (θ) and the angle of reflection must be equal (**Figure 10**).This means that the angle between them will be 2θ. If we now add another parallel plane at a perpendicular distance d from the first one, we can see that constructive interference between the beams reflected by the two planes only occurs if the path difference between the two beams is an integral number of wavelengths, that is, at $n\lambda$ where n is an integer and λ is the wavelength. The path length difference in **Figure 10** (AB + BC) is $2d\sin\theta$, which leads to Equation (1), known as Bragg's law.

$$n\lambda = 2d\sin\theta \qquad (1)$$

The closer the two planes are to each other, the higher the scattering angle 2θ.

The crystal is a three-dimensional object and the different Bragg planes have unique orientations in space. Since the reflection angle θ is defined with respect to the plane, the orientation of that plane (and hence the orientation of the crystal) with respect to the incoming X-ray beam will determine the actual direction to which the X-rays are diffracted. The X-rays diffracted in a given direction give rise to one diffraction spot (**Figure 8**), also known as a reflection.

Figure 9 Rows of trees seen from a passing car. Note that the row seen from each point is a different one.

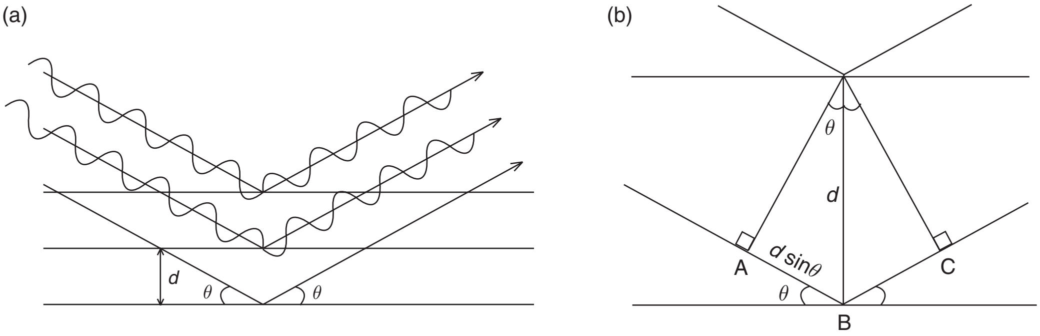


Figure 10 (a) The origin of Bragg's law. The X-ray waves of wavelength λ are 'reflected' at an angle θ from successive planes of equivalent atoms, separated by a lattice repeat d. (b) For constructive interference to take place between the waves reflected the path length difference AB + BC needs to be a multiple of the wavelength. This is true when $2d\sin\theta = n\lambda$.

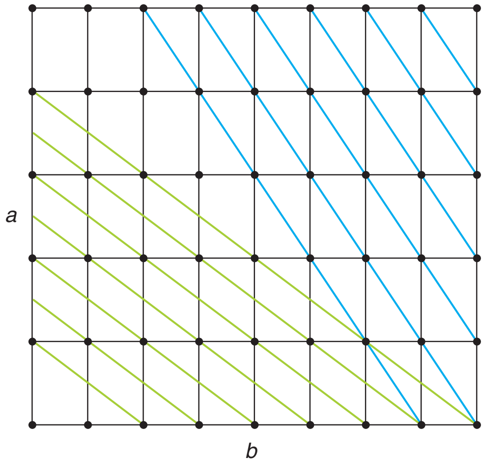


Figure 11 Two sets of Bragg planes viewed parallel to the *c* axis. The cyan set of planes intersects the *a* axis once, so $h = 1$ and *b* axis once, so $k = 1$. It is parallel to the *c* axis and therefore intersects it zero times, so $l = 0$. By a similar logic the green set of planes has indices $h = 2$, $k = 1$, and $l = 0$.

The standard way to describe the orientation of a plane is by a vector **d** perpendicular (normal) to it. An equivalent description for Bragg planes is in terms of how many times they intersect each of the three unit cell axes in one lattice repeat (**Figure 11**). These Miller indices *h* (for axis ***a***), *k* (for axis ***b***), and *l* (for axis ***c***) uniquely define the plane and its X-ray reflection; for instance, 1 (1,0,0) plane intersects the *x*-axis once, a (2,1,0) plane, the *x*-axis once, and the *y*-axis twice, and so on. In principle, it is possible to calculate the vector $\mathbf{d}_{hkl}$ knowing the Miller indices *h*,*k*,*l* and the unit cell vectors **a**, **b**, and **c**. In practice, this may not be easy. As we want to have a simple description of the normal vectors $\mathbf{d}_{hkl}$ (which determine when Bragg's law will hold) we adopt a different set of basis vectors ($\mathbf{a}^*$, $\mathbf{b}^*$, $\mathbf{c}^*$), called the reciprocal lattice and the space they define is called reciprocal space. Each plane can be described by a vector:

$$\mathbf{d}^*_{hkl} = h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^* \tag{2}$$

which is parallel to $\mathbf{d}_{hkl}$, but with the length $d^*_{hkl} = 1/d_{hkl}$.[5]

How can we then relate $\mathbf{d}^*$ to a diffracted beam in some particular direction? A very useful, if somewhat abstract, way of describing the diffraction geometry is the Ewald construction. Here the wavevectors (a wavevector is a vector parallel to the propagation of the beam with a length of λ^{-1}) of the incident ($\mathbf{s}_0$) and scattered ($\mathbf{s}$) beams are drawn within a circle of radius λ^{-1} (**Figure 12**). (Physically, this corresponds to a crystal located at the origin O of the circle.) The reflecting plane in the Bragg picture bisects the angle between **s** and $\mathbf{s}_0$. The difference vector $\mathbf{s}-\mathbf{s}_0$, known as the scattering vector, is perpendicular to the Bragg plane. From Bragg's law (Equation (1)) $n/d = (2\sin\,\theta)/\lambda$, so $|\mathbf{s}–\mathbf{s}_0|$ must be n/d, or nd^* for diffraction to occur. From **Figure 12**, it is clear that since the vector **s** and the reflected beam in **Figure 10** are parallel, the vector $|\mathbf{s}-\mathbf{s}_0|$ is perpendicular to the Bragg plane, just like *d* in **Figure 10**. Since $s = s_0 = \lambda^{-1}$, we can calculate the length of the scattering vector $\mathbf{s}-\mathbf{s}_0$ by simple trigonometry; $|\mathbf{s}–\mathbf{s}_0| = 2\sin\theta\,\lambda^{-1}$. By rearranging Bragg's law we know that the length of this vector has to be n/d (or nd^*) for diffraction to occur.

From Equation (2), we deduce that diffraction is observed only when the indices *h*, *k*, *l* in $\mathbf{d}^*$ take integral values. These reciprocal space vectors form a lattice, the reciprocal lattice, and the mathematical relationship between the real and reciprocal lattices (and between other aspects of the diffraction pattern) is a FT, as we will explain below. The interpretation of the Ewald construction is that diffraction is observed when the scattering vector $\mathbf{s}-\mathbf{s}_0$ is equal to a reciprocal space vector $\mathbf{d}^*_{hkl}$ with integral indices *h*, *k*, *l*. This occurs whenever such a

[5] There is a complicated relationship between **a, b, c** and $\mathbf{a}^*$, $\mathbf{b}^*$, $\mathbf{c}^*$ of the form $\mathbf{a}^* = (\mathbf{b}\times\mathbf{c})/(\mathbf{a}\cdot\mathbf{b}\times\mathbf{c})$. This is simple if the axes are orthogonal, then $a = 1/a$, and the two are parallel.

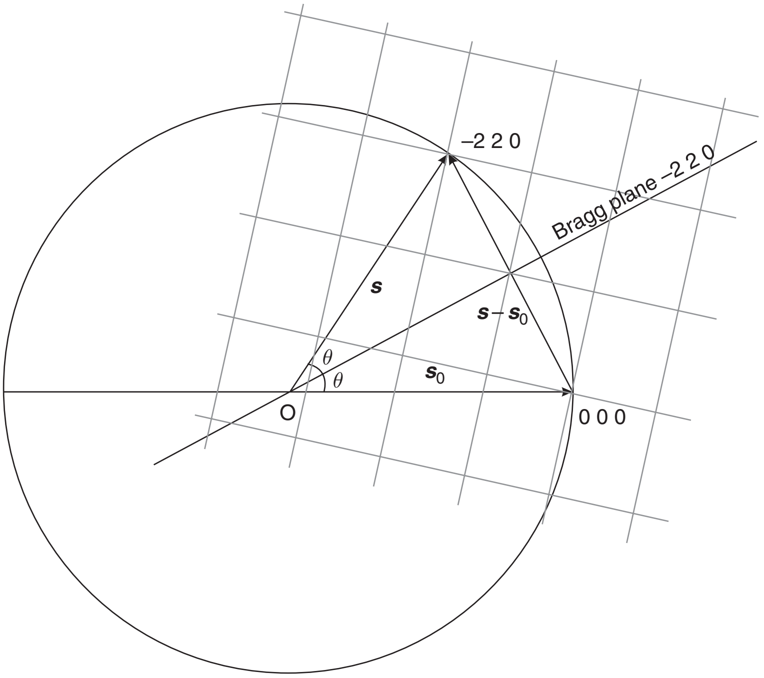


Figure 12 The Ewald construction drawn for the reflection (–2 2 0). The crystal is located at the origin O and the endpoint of the vector ***s*** lies at a lattice point of the reciprocal lattice (gray). The radius of the circle is λ^{-1}.

reciprocal space vector intersects the Ewald sphere. This is useful in predicting the reflections that are in diffracting position at a given orientation of the crystal as we discuss in more detail in Section 9.03.8.

To summarize, each spot on the diffraction pattern (**Figure 8**) can be uniquely referred to by integral indices h, k, and l, which are multipliers of the reciprocal lattice basis vectors $\mathbf{a}^*$, $\mathbf{b}^*$, and $\mathbf{c}^*$. The position of each spot in this reciprocal space may then be expressed as a vector $(h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*)$ analogously to the positions of the atoms in the crystals, which can be defined by a vector $(x\mathbf{a} + y\mathbf{b} + z\mathbf{c})$ in real space. Although the positions of the diffraction spots are defined by the FT of the crystal lattice, their intensities are defined by the FT of the contents of the unit cell.

9.03.6 Fourier Transforms

Given the above, understanding the concept of a FT is very useful in understanding and describing diffraction, hence we devote a section to their properties. A more detailed account can be found in textbooks such as Rhodes[9] or Blow.[11] A periodic function[6] in x, $f(x)$, can be approximated either as a sum of sine waves $F_h \sin 2\pi(hx + \alpha)$, cosine waves $F_h \cos 2\pi(hx + \alpha)$, or both characterized by an amplitude F_h at harmonic number h (the harmonic number determines the frequency and thereby the wavelength) and phase α. Using the cosine function as a basis we can write:

$$f(x) = \sum_{h=0}^{n} F_h \sin 2\pi(hx + \alpha_h) \qquad (3)$$

The more terms are used, the better the approximation, because with increasing h the frequency of the wave increases, contributing to finer and finer detail in the function being approximated. This is known as a Fourier sum; a one-dimensional example is shown in **Figure 13**. When n tends to infinity we have a perfect description of the function, known as a Fourier series.

[6] With certain limitations of little practical importance for crystallography; for details see for example Bracewell.[12]

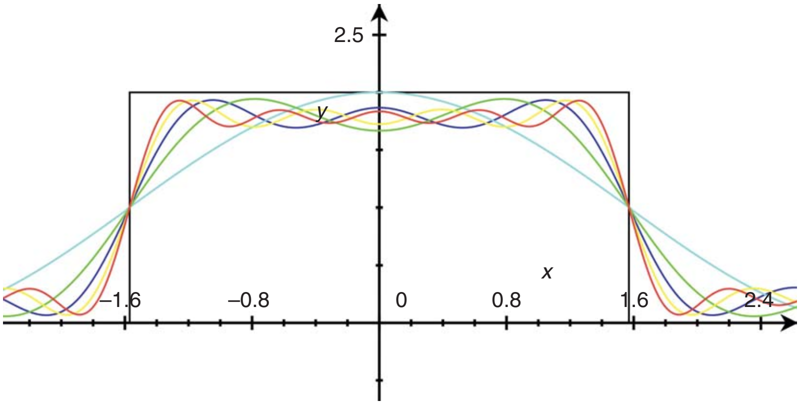


Figure 13 A ‘top hat’ function (black) approximated by different numbers of sine waves. Cyan: 1 + sin *x*, green: 1 + sin *x* + 1/3 sin 3*x*, blue: 1 + sin *x* +1/3 sin 3*x* + 1/5 sin 5*x*, yellow: 1 + sin *x* +1/3 sin 3*x* + 1/5 sin 5*x* +1/7 sin 7*x*, red: 1 + sin *x* +1/3 sin 3*x* + 1/5 sin 5*x* +1/7 sin 7*x* + 1/9 sin 9*x*. The approximation gets better with the increasing number of sine functions.

In principle, we could sum up any set of periodic functions – use them as the basis set – but a particularly convenient choice is the linear combination of sine and cosine waves:

$$f(x) = \sum_{h=0}^{n} (\cos 2\pi(hx) + \mathrm{i}\sin 2\pi(hx)) \tag{4}$$

where *i* is the square root of -1.

This is because by using the Euler formula

$$\cos\theta + \mathrm{i}\sin\theta = \mathrm{e}^{\mathrm{i}\theta} \tag{5}$$

we can express the Fourier sum as the exponential

$$f(x) = \sum_{h=0}^{n} F_h\, \mathrm{e}^{2\pi \mathrm{i} hx} \tag{6}$$

The information about the phase α_h is then contained in the exponent. Consequently, we can represent each term with the Argand diagram (**Figure 14**), where the imaginary component specifies the phase. Each term is then characterized by two parameters; the amplitude $|F_h|$ and the phase ϕ.

We have seen that any waveform can be described as a sum of waves, but how then can we find the values of F_h – the amplitudes of the waves? We can do so by taking the FT of the function *f*(*x*); multiply *f*(*x*) by $\mathrm{e}^{-2\pi \mathrm{i} hx}$ and integrate over *x*:

$$F_h = \int_{-\pi}^{\pi} f(x)\mathrm{e}^{-2\pi \mathrm{i} hx}\mathrm{d}x^7 \tag{7}$$

The concept presented here in one dimension can be relatively and easily extended to two or three dimensions, in which case the terms in the Fourier sum have extra indices *k* and *l*. The sum would be:

$$f(x,y,z) = \sum_h \sum_k \sum_l F_{h,k,l}\mathrm{e}^{2\pi(hx+ky+lz)} \tag{8}$$

and the transform:

$$F_{h,k,l} = \int_x \int_y \int_z f(x,y,z)\mathrm{e}^{-2\pi \mathrm{i}(hx+ky+lz)}\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z \tag{9}$$

[7] The exponent needs to be dimensionless, so Fourier space ends up having inverse dimensions compared to real space, such as m^{-1} for length or s^{-1} (i.e., frequency) for time.

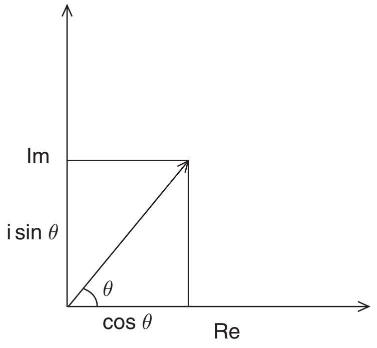


Figure 14 A complex number represented on an Argand diagram. The real (cos θ) and imaginary (i sin θ) components sum up to a vector in the complex plane.

where the integration is over the period 2π of the exponential functions, and $\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$ is the volume element $\mathrm{d}V$. Clearly, these two equations have a similar form, and it is indeed correct to say that $F_{h,k,l}$ is the FT of the function $f(x,y,z)$. There is also an inverse FT of the function $F_{h,k,l}$, where the summation in Equation (8) is replaced by an integration over a continuous variable. For all practical purposes $f(x,y,z)$ can be considered to be the inverse FT of $F_{h,k,l}$. Indeed, FTs are calculated most often as discrete summations using the fast Fourier transform (FFT) algorithm. The difference between the FT and the inverse FT is the sign of the exponent.

The physical interpretation is that each Fourier term $F_{h,k,l}$ is a wave in a plane defined by h, k, and l. By summing these waves in different directions, we can approximate any three-dimensional function, just like in the one-dimensional case above. This is the inverse FT, also known as Fourier synthesis, as the function is being 'built' from component waves. The normal FT is called Fourier analysis, since the function is being 'broken down' to its component waves.

What we will describe but not prove below is that when we convert a three-dimensional object in xyz by FT into a wave description, we end up describing the X-ray diffraction pattern of that three-dimensional object.

9.03.7 Diffraction as a Fourier Series

First, we need to know what is meant by a periodic function. The crystal contains a periodic arrangement – a regular array – of atoms but, as mentioned above, X-rays scatter electrons. Therefore it is more convenient to think about the crystal and thus the unit cell in terms of its electron density: not $f(x,y,z)$ where f describes the 'scattering factor' of the atoms, but $\rho(x,y,z)$, where $\rho(x,y,z)$ is the electron density at point x,y,z. As the atoms are periodically arranged, so also is their electron density; $\rho(x,y,z)$ is a periodic function. We can therefore approximate it with a Fourier series just as above. If we know the electron density function, we can use a FT to calculate the individual coefficients $F_{h,k,l}$. However this is completely useless; the shape of the electron density, that is, the arrangement of the atoms in the crystal: its structure – is precisely what we want to find out. In order to achieve this we need to do quite the opposite – calculate the electron density from the diffraction pattern. Before we consider how, we will try to build up a physical picture of what the FT of the electron density means.

Let us return, for a moment, to **Figure 10**, the Bragg's law description of X-ray diffraction. X-rays are reflected by planes of lattice points, uniquely described by the three indices h, k, l. These three indices form the basis of another lattice, which we called the reciprocal lattice, where the distance from the origin to each point hkl was $1/d_{hkl}$, where d_{hkl} was the distance between the Bragg planes. Each Bragg plane can be defined by its normal, which turns out to be a multiple of the reciprocal space basis vectors $\mathbf{a}^*$, $\mathbf{b}^*$, $\mathbf{c}^*$. We can then refer to this plane, as well as to the Fourier term associated with it, by a reciprocal lattice vector $\mathbf{d}^*_{hkl} = (h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*)$. Rewriting Equation (9) in terms of electron density, we get

$$\mathbf{F}_{h,k,l} = \frac{1}{V} \int_x \int_y \int_z \rho(x,y,z) \mathrm{e}^{-2\pi \mathrm{i}(hx+ky+lz)} \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z \tag{10}$$

where $1/V$ corrects for the change in volume between real and reciprocal space.

Can we connect this equation to the Bragg's law picture? Each $\mathbf{F}_{h,k,l}$ is an individual Bragg reflection hence each Bragg reflection describes part of the entire electron density distribution. The Fourier sum (or in other words inverse FT) describes the entire electron density. More accurately, the reflections sample, or are caused by, the direction- and distance-dependent properties of the electron density in the unit cell. The way it varies periodically in space can be imagined as an electron density 'wave'. This is perhaps possible if there is no variation in y and z. In this case, Equation (10) would reduce to the one-dimensional FT (analogous to Equation (7): $\mathbf{F}_{h,k,l} = 1/V \int_x \rho(x)\mathrm{e}^{-2\pi \mathrm{i} hx}\mathrm{d}x$. The value of $F_{h,k,l}$ is independent of k or l, and we sample only the variation along the x-axis.

It turns out that the diffraction pattern can be considered as the FT of the unit cell sampled at the reciprocal lattice points or, equivalently, the FT of the crystal is the reciprocal lattice multiplied by the FT of the unit cell. This follows from the convolution theorem; see Blow[11] for a detailed explanation.

The terms in the Fourier sum of the electron density are known as structure factors

$$\mathbf{F}_{h,k,l} = \frac{1}{V}\int_V \rho(x,y,z)\mathrm{e}^{-2\pi \mathrm{i}(hx+ky+lz)}\mathrm{d}V \qquad (11)$$

where V is the volume of the unit cell. The concept of the electron density as a Fourier series also gives a direct meaning to crystallographic resolution. It is simply the largest reciprocal space vector, or, in other words, the smallest Bragg spacing observed. In real space, it is the same as the optical resolution of a microscope: the closest that two objects can be and still be distinguishable from each other. The resolution of a diffraction pattern for a crystal of a given unit cell determines the number of structure factors (Fourier components) we can use to calculate the electron density. (In a larger unit cell there are more reflections, but they must also account for a larger volume in real space.)

In addition to frequency (or wavelength) and amplitude, a wave is characterized by a phase: the location of the first maximum with respect to the origin. This also applies to Fourier terms, which are waves. For waves of electron density in a crystal unit cell the relative phases of different waves determine where constructive interference produces peaks that may be identified as atoms. Unfortunately, the phase information is contained in the imaginary exponent and cannot be directly measured. The physical quantity that we observe is the intensity of the scattered waves, which is proportional to the square of the structure factor $\mathbf{F}_{h,k,l}$ (or mathematically $\mathbf{FF}^*$, where $\mathbf{F}^*$ is the complex conjugate of $\mathbf{F}$). As a result, the imaginary exponential in Equation (11) becomes zero and since $e^0 = 1$ the phase factor vanishes. In terms of the Argand diagram we only observe the length of the structure factor (or actually its square), not its angle with respect to x, or real, axis. This is a crystallographic phase problem, for which there is no general solution in macromolecular crystallography.

Worse still, the phases, not the structure factor amplitudes $|\mathbf{F}|$, dominate the shape of the calculated electron density. If, as in **Figure 15**, we FT a picture of a duck and a cat, and then calculate the inverse transform using duck amplitudes and cat phases, only the shape of the cat is discernible. This is exactly what happens to electron density when the phases of the structure factors are incorrect. The features of the electron density, largely determined by the phases, will also be wrong. This is known as model bias (see Section 9.03.10.2).

9.03.8 The Diffraction Experiment in Practice

What kind of experimental setup is required to record the diffraction patterns from the macromolecular crystals and what kind of information is usually reported about the experiment? **Figure 16** shows a typical X-ray setup. The main parts are the X-ray source, including the optics to focus the parallel radiation onto the crystal, a device called a goniostat, and the detector that records the diffracted radiation. X-ray sources fall into two categories: conventional generators and synchrotron sources. Most data these days are collected at synchrotrons.

In conventional generators, high voltage is applied between a tungsten cathode filament and a metal anode target. The electron bombardment excites a transition of the core electrons in the metal, which

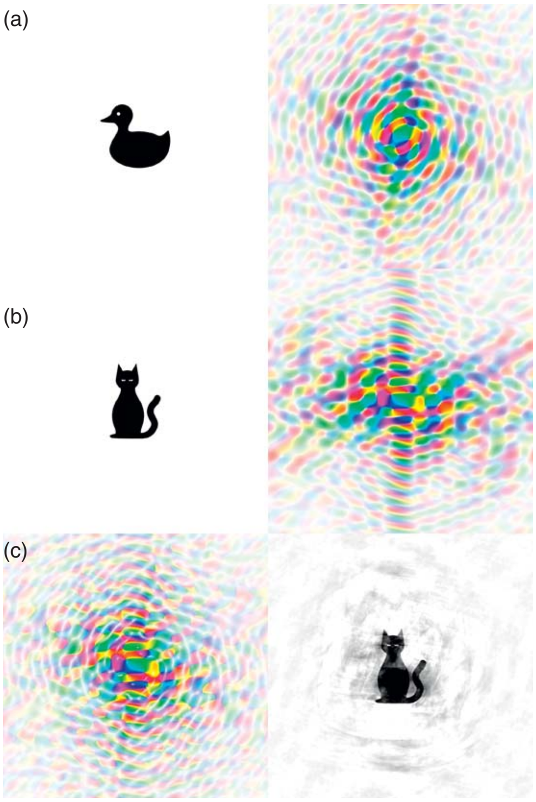


Figure 15 FTs in two dimensions. (a) A drawing of a duck and its FT. The amplitude is represented by brightness and the phase by color. (b) A drawing of a cat and its FT, represented as above. (c) A combination of the amplitudes from the duck transform and the phases from the cat transform. The inverse transform shows the features of the cat. Reproduced from the web-based book ‘Book of Fourier’, University of York, UK, with permission from Kewin Cowtan.

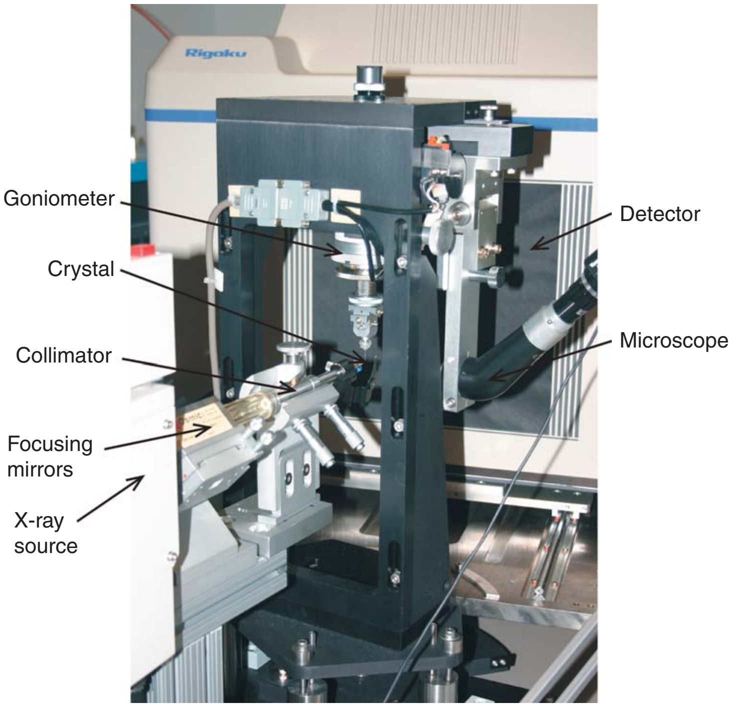


Figure 16 A typical laboratory X-ray diffraction setup and its main components. The cryostream maintaining the crystal at 100 K was removed for clarity.

relaxes and emits X-rays. The wavelength of the emitted radiation is determined by the electronic transition in the metal, so the only way to change the wavelength λ is to change the anode. Even then one is limited to the emission lines of existing metals. In addition, the main problem with conventional generators is that most of the electron energy is released as heat, limiting the achievable intensity of the X-ray beam. In addition, the X-rays from an X-ray generator are divergent, again limiting the maximal X-ray intensity on the sample.

Electrons accelerated in an external field, for instance around an evacuated circle, produce synchrotron radiation. The phenomenon is similar to the emission of radio waves from an antenna but, to produce X-rays, the electrons have to move much faster, almost at the speed of light. This requires large particle accelerators, which are built as national or international facilities, such as the European Synchrotron Radiation Facility (ESRF) in Grenoble, France (http://www.esrf.eu, **Figure 17**), the Advanced Photon Source (APS) at Argonne National Laboratories in the USA (http://www.aps.anl.gov) or Diamond, in Didcot (near Oxford) in England (http://www.diamond.ac.uk).The highly intense X-ray beams produced at such facilities can be 10^{12} times more brilliant than conventional laboratory sources. An additional advantage is the wavelength

Figure 17 The European Synchrotron Radiation Facility (ESRF) in Grenoble, France. The circumference of the storage ring is 844.4 m. Reproduced from the ESRF web press site, with permission from ESRF.

tuneability of synchrotron X-rays – typically between energies E from 6 and 20 keV or wavelengths[8] λ between 2.066 and 0.620 Å. Tuneable sources are particularly important for solving the phase problem, as described in Section 9.03.9.2. Synchrotron sources are thus indispensable for modern macromolecular X-ray crystallography: they are much brighter, they are tuneable, and the beams are essentially parallel; they have a very low angular divergence.

The X-ray beam from the source is monochromated, focused, and collimated to deliver a parallel beam of defined size and wavelength to the crystal. Because of the intrinsically superior optical qualities of synchrotron beams, the radiation delivered to the crystal is also superior to that from conventional sources. The crystal is mounted on a goniostat, which allows the crystal to be rotated. The crystal is usually flash-cooled to a temperature of 100 K by a cold stream of nitrogen gas to reduce radiation damage. X-rays are ionizing radiation and the free radicals produced as they pass through the protein destroy the crystal. Without flash cooling, protein crystals last only seconds on a synchrotron beamline.

The radiation diffracted from the crystal is measured by an electronic area detector, which allows the measurement of a large number of Bragg reflections in a single exposure. The reciprocal lattice is three-dimensional just like the real lattice, only a fraction of the reciprocal lattice points (reflections) are in diffracting position at any given orientation of the crystal (**Figure 18**). Therefore, the crystal is also rotated through an angle of 0.1–1° during the exposure to bring more reflections to diffracting position. Exposures at different orientations of the crystal are required to cover all of reciprocal space – to measure all the terms of the Fourier series up to the resolution limit. The reciprocal lattice has the same rotational symmetry as the real lattice, so just as points within the unit cell are related by internal symmetry and may have the same atoms and electron density, some reciprocal space vectors are the same – equally intense – due to symmetry (**Figure 8**). Crystals with higher symmetry thus require fewer diffraction images to cover the entire reciprocal lattice.

So how do we know the unit cell of the crystal and its orientation? The first step in the collection of crystallographic data consists of taking one or two test images, from which the spot positions are determined. Each diffraction spot is then assigned indices h,k,l based on its position on the detector. This is called indexing and the unit cell parameters and crystal orientation are determined here. Once the diffraction pattern is indexed, we can use the Ewald construction to predict where spots should be observed. The prediction is important, since some of the spots may be so faint that detection would be impossible unless we knew where to expect them.

If the cell has internal symmetry only one of the symmetry-related reflections needs to be measured. Once we know the unit cell and the crystal orientation, we can plan the actual data collection in such a way that all the unique reflections are measured at least once.

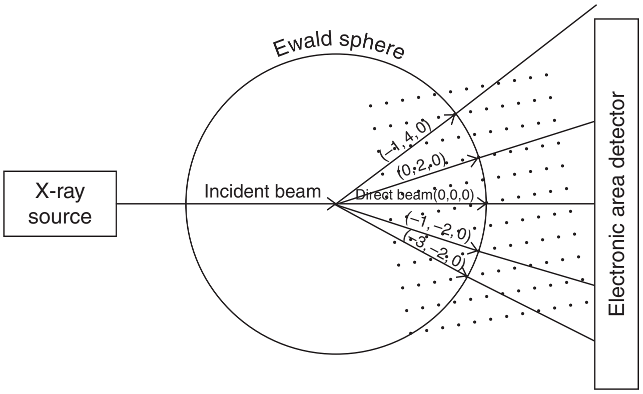


Figure 18 A schematic representation of data collection with an electronic area detector. The reciprocal lattice plane $l = 0$ is shown as black dots. The direct beam and four scattered beams with their respective indices are shown; three of them produce diffraction spots on the detector, while the fourth (–1,4,0) falls outside the detector area.

[8] The relationship is $\lambda = 12.398/E$ with E in keV and λ in Å.

Table 1 Typical values of data collection statistics for data collected at a synchrotron

Quantity	*Reasonable value (all data)*	*Reasonable value (highest resolution shell)*
Completeness	95%	90%
Overall R_{sym}	0.10	0.45
I/σ[a]	30	2

[a] I/σ tends to vary a lot due to unit cell size and other factors.

Papers usually report this information as the completeness: the ratio of observed reflections to this theoretical maximum, expressed as a percentage (see **Table 1** for expected values). So how do we decide what reflections we have actually observed? Even where a spot is not visible to the eye, the diffraction pattern may nonetheless yield useful information. The maximum resolution to which data are used is decided in a somewhat subjective manner.

First, the intensities of all the predicted spots on the detector are measured and the errors in the intensities are estimated. Owing to the diffraction geometry and other experimental factors, different measurements of the same reflection are not directly comparable to each other, and a computational procedure known as relative scaling must be used to bring them on a common scale. In all cases some reflections are related by symmetry, therefore, we have multiple observations of the same reflection. This gives us additional information on the experimental errors, because even though the intensities of symmetry-related reflections should be equal, they differ due to experimental error. The differences in the related intensities can be quantified by the residual

$$R_{\mathrm{sym}} = \frac{\sum_{h,k,l} \left| I_{h,k,l} - \langle I_{h,k,l} \rangle \right|}{\sum_{h,k,l} I_{h,k,l}} \tag{12}$$

where $<I_{h,k,l}>$ is the mean of the symmetry-related reflection intensities. It may be used to judge the overall reliability of the data, and, calculated for the highest resolution shell, the resolution limit (**Table 1**). However, R_{sym} inherently depends on the redundancy (see below) of the data,[13] so for incomplete data the quality is overestimated and for highly redundant data it is underestimated. A redundancy-independent R_{meas} has been devised,[13] but R_{sym} is still often reported in publications.

By scaling together symmetry equivalent reflections, we also obtain better estimates of the standard deviations (σ) and hence we can calculate signal-to-noise values (I/σ). Therefore it is often desirable to collect redundant data (collecting many symmetry equivalent reflections) in order to obtain more accurate estimates of the true intensities. This is particularly important for phasing methods relying on anomalous scattering (see Section 9.03.9.2). Redundancy is usually reported in the table of data processing statistics. The average signal-to-noise ratio is usually reported for the highest resolution shell as well, providing another criterion in addition to R_{sym} for determining the resolution limit; **Table 1** collects values to be expected for good data collected at a synchrotron.

We then have a list of indices h, k, l, the associated intensities, and their standard deviations. All we now need are the phases to reconstruct the electron density by a Fourier transformation.

9.03.9 Phasing Methods

The central problem in crystallography lies in obtaining the phase for every observed structure factor amplitude. We judge how correct a given set of phases is by the result: does the electron density map make chemical sense? For small molecules, very accurate data is usually available to high resolution (1 Å or better), which allows the use of 'direct methods'[9] to obtain the phases rapidly and correctly. The approach uses statistical relationships between the phases of certain reflections. Unfortunately, direct methods are not easily

[9] The Nobel Prize in Chemistry in 1985 was awarded to Herbert A. Hauptman and Jerome Karle 'for their outstanding achievements in the development of direct methods for the determination of crystal structures'.

applicable to macromolecular crystallography because the individual atoms need to be resolved into clear peaks, which is much harder when there are more than about 1000 atoms. (Macromolecular structures have 2000–100 000 nonhydrogen atoms.) There are three commonly used phasing methods: isomorphous replacement, anomalous dispersion techniques, and molecular replacement. The first two are experimental phasing methods, since no prior structural knowledge of the macromolecule is required. In these two methods, as we discuss below, we 'bootstrap' our way from a structure of a few heavy atoms to a structure for the entire protein. Direct methods are very useful in the first part of this process; as they are very efficient at solving heavy atom substructures. They have replaced old-fashioned trial-and-error approaches.

9.03.9.1 Isomorphous Replacement

The classical method for solving the phase problem in macromolecular crystal structures, known as isomorphous replacement, dates back to the earliest days of protein crystallography.[10,16] The concept is simple enough: we introduce into the protein crystal an atom or atoms heavy enough to affect the diffraction pattern measurably. We aim to figure out first where those atoms are (the heavy atom substructure) by subtracting away the protein component, and then 'bootstrap' – use the phases based on the heavy atom substructure to solve – the structure of the protein.

The first step is to introduce heavy atoms into the protein crystal. This is usually done by soaking the crystals in a solution containing 0.1–10 mmol l^{-1} of the heavy atom compound (Hg, Pt, Au, U compounds are often used) but sometimes the macromolecule is also co-crystallized with the heavy atom compound. As discussed in Section 9.03.4, protein crystals contain large solvent channels, which allow the diffusion of small molecules within the crystal. An important *caveat* is that the binding of the heavy atom compound must not distort the crystal appreciably: neither the overall unit cell dimensions nor the conformation of the macromolecule. If it does, the underlying assumption that we can subtract away the protein component is false. In other words, the native (no heavy atom) and derivative (with heavy atom) must be isomorphous, and the techniques are called in general isomorphous replacement.

To return to the crystallographic experiment itself: the addition of such a heavy atom must result in a measurable change in the structure factors $\mathbf{F}_{h,k,l}$. If we denote the structure factors in the absence of the heavy atom as $\mathbf{F}_\mathrm{P}$ (the protein **F**s) and those in its presence as $\mathbf{F}_\mathrm{PH}$ (the protein-and-heavy-atom **F**s), the difference $\mathbf{F}_\mathrm{PH}-\mathbf{F}_\mathrm{P}$ is $\mathbf{F}_\mathrm{H}$, the contribution of the heavy atom(s) alone. As the structure factors are complex, the subtraction must be represented in an Argand diagram as a vector difference (**Figure 19**).

Unfortunately, we can only measure the amplitudes $|\mathbf{F}_\mathrm{P}|$ and $|\mathbf{F}_\mathrm{PH}|$, but if we make the assumption that $\mathbf{F}_\mathrm{H}$ is much smaller than $\mathbf{F}_\mathrm{PH}$, the phase difference between $\mathbf{F}_\mathrm{PH}$ and $\mathbf{F}_\mathrm{P}$ will also be small. We can then write the

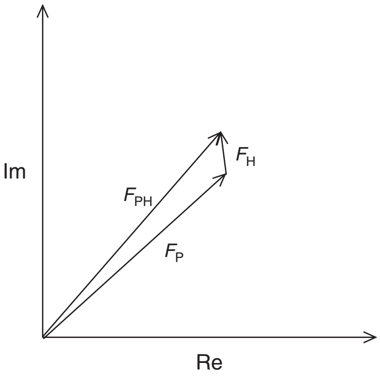


Figure 19 Vector diagram in the Argand plane of $\mathbf{F}_\mathrm{PH}=\mathbf{F}_\mathrm{P}+\mathbf{F}_\mathrm{H}$. As can be seen, the maximum angular difference between $\mathbf{F}_\mathrm{PH}$ and $\mathbf{F}_\mathrm{P}$ occurs when $\mathbf{F}_\mathrm{H}$ is perpendicular to $\mathbf{F}_\mathrm{P}$. When $|\mathbf{F}_\mathrm{H}|$ is much smaller than $|\mathbf{F}_\mathrm{P}|$, this angular difference is small. However, not all $|\mathbf{F}_\mathrm{P}|$s are large; not all $|\mathbf{F}_\mathrm{H}|$ will be much smaller than $|\mathbf{F}_\mathrm{P}|$ even when the averages are very different.

[10] For a historical account, see Rossmann[14] or Rossmann[15].

approximation: $|\mathbf{F}_\mathrm{H}| \approx |\mathbf{F}_\mathrm{PH}| - |\mathbf{F}_\mathrm{P}|$ (**Figure 19**). These structure factor amplitudes should then contain information only about the position of the heavy atom in the unit cell, that is, the substructure. Compared to the carbon, nitrogen, and oxygen atoms of the macromolecule the heavy atoms are few, far between, and electron dense. Thus, the substructure essentially resembles a small molecule structure and the methods for small molecule structure determination, such as direct methods, can be used to solve it – to find the positions of the heavy atoms. We can then calculate the phases for the heavy atom structure factors $\mathbf{F}_\mathrm{H}$ using Equation (11). (The above also explains why isomorphicity is so important; if the heavy atom derivative is not isomorphous, the change in $\mathbf{F}_\mathrm{PH}$ is not just the added heavy atoms $\mathbf{F}_\mathrm{H}$, but is due to changes in the protein induced by the added atoms. The approximation breaks down.)

What use is the substructure? Let us return to the Argand diagram representation of the structure factors; as long as we only have the measured amplitudes, the only thing we know about reflection *h*, *k*, *l* is that its structure factor vector lies somewhere on the circle of radius $|\mathbf{F}_\mathrm{P}|$. However since we also know that $\mathbf{F}_\mathrm{P} + \mathbf{F}_\mathrm{H} = \mathbf{F}_\mathrm{PH}$, we can draw two circles, one with radius $|\mathbf{F}_\mathrm{PH}|$ centered at the origin and the other with radius $|\mathbf{F}_\mathrm{P}|$ centered at the end of the only vector we know, $\mathbf{F}_\mathrm{H}$.[11] This Harker construction (**Figure 20**) then gives us two possible values of the phase for $\mathbf{F}_\mathrm{P}$, since the condition $\mathbf{F}_\mathrm{P} + \mathbf{F}_\mathrm{H} = \mathbf{F}_\mathrm{PH}$ is true only when the two circles intersect. Which one then is the correct choice? This phase ambiguity in single isomorphous replacement (SIR) can be resolved by using another derivative for which a circle of radius $|\mathbf{F}_\mathrm{PH2}|$ centered at the end of $\mathbf{F}_\mathrm{H2}$ can be drawn (**Figure 21**). The three circles only intersect at one point, giving the phase of $\mathbf{F}_\mathrm{P}$. The same may be repeated for all reflections $\mathbf{F}_{h,k,l}$. This is known as multiple isomorphous replacement (MIR). Another way to break the phase ambiguity is to exploit a phenomenon known as anomalous scattering, which is discussed below.

Whenever the three circles are drawn, they tend not to intersect precisely at the same point. This lack of closure gives, instead of a single unambiguous value, a probability distribution. The centroid[12] of this

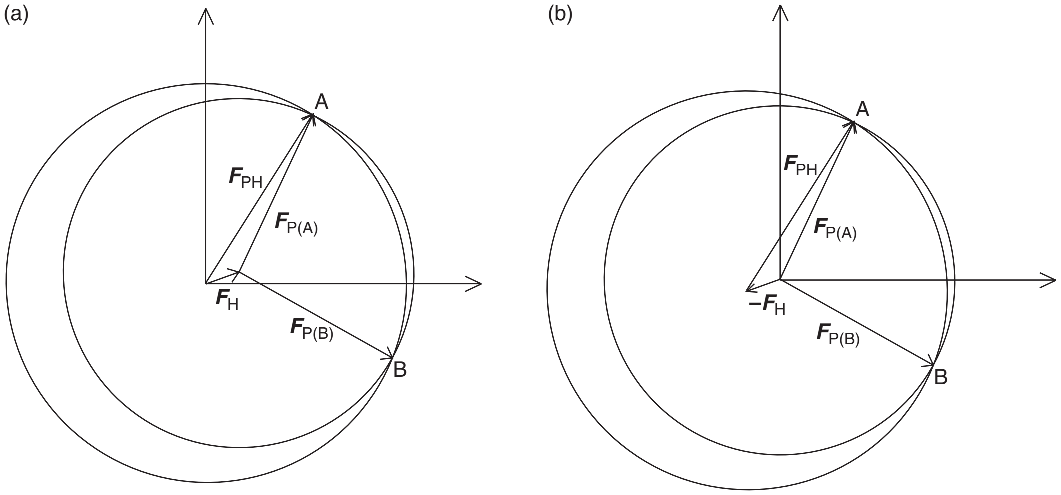


Figure 20 (a) The Harker construction for the SIR method. The circles intersect at two places, A and B, leading to phase ambiguity. (b) An alternative way of drawing the Harker construction with $\mathbf{F}_\mathrm{P}$ centered at the origin is often more convenient for more complicated phasing schemes, such as SIRAS (**Figure 24**).

[11] The vector sum of $\mathbf{F}_\mathrm{P}$ and $\mathbf{F}_\mathrm{H}$ gives $\mathbf{F}_\mathrm{PH}$, but we only know the *lengths* $|\mathbf{F}_\mathrm{P}|$ and $|\mathbf{F}_\mathrm{PH}|$; the phase is unknown. We can therefore only draw circles. The centers of the two circles are related by the only vector we know, $\mathbf{F}_\mathrm{H}$. We therefore draw the vector $\mathbf{F}_\mathrm{H}$ out from the origin. At the end of that vector, we draw a circle of radius $\mathbf{F}_\mathrm{P}$. The end of the vector $\mathbf{F}_\mathrm{P}$ must lie somewhere on this circle. We also draw a circle centered at the origin of radius $\mathbf{F}_\mathrm{PH}$; again the end of vector $\mathbf{F}_\mathrm{PH}$ lies somewhere on that circle. As you can see from , these two circles intersect at just two points, and those two points are the only ones that satisfy the vector equation $\mathbf{F}_\mathrm{P} + \mathbf{F}_\mathrm{H} = \mathbf{F}_\mathrm{PH}$.

[12] The centroid in this context is the point at the geometric center of the area on the Argand diagram where the end of the vector could *possibly* be.

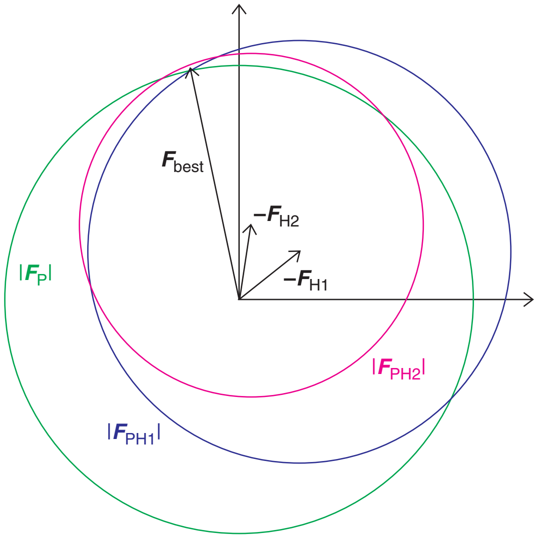


Figure 21 The MIR method, illustrating the problem of lack of closure due to experimental error. There are two heavy atom substructures, giving vectors $\mathbf{F}_{H1}$ and $\mathbf{F}_{H2}$. By plotting out two SIR constructs as in Figure 20, we get the construct shown here. The structure factor with the most probable phase, F_{best}, is at the centroid of the area limited by the three circles ($|\mathbf{F}_P|$ (green), $|\mathbf{F}_{PH1}|$ (blue), and $|\mathbf{F}_{PH2}|$ (magenta)). It therefore has a different length than $|\mathbf{F}_P|$.

distribution corresponds to the structure factor with the most probable phase $|\mathbf{F}_{best}|$. As can be seen from Figure 21, this is different from $|\mathbf{F}_P|$ due to the lack of closure. The figure of merit

$$m = \frac{|\mathbf{F}_{best}|}{|\mathbf{F}_P|} \tag{13}$$

is the cosine of the phase error and hence a measure of the reliability of an individual phase. It is often used as a weighting factor in the calculation of electron density maps.

9.03.9.2 Anomalous Dispersion

The phenomenon of anomalous scattering is extensively used in modern macromolecular crystallography to solve the phase problem. To understand how this is done, we need to return to the simple picture of X-rays reflecting from Bragg planes, where it makes no difference which side of the plane is the reflecting 'surface'. This leads to two structure factors $\mathbf{F}_{h,k,l}$ and $\mathbf{F}_{-h,-k,-l}$ differing only in the sign of their phase. The phase – a complex number – drops out because we measure intensities ($I = \mathbf{F}^2$; see above); and $I_{h,k,l}$ and $I_{-h,-k,-l}$ are equal. This is known as Friedel's law and the pairs of related reflections $\mathbf{F}_{h,k,l}$ and $\mathbf{F}_{-h,-k,-l}$ are called Friedel pairs.

For the third row (K–Kr) elements, there are electronic transitions in energy ranges from 3.6 to 14.3 keV (3.4–0.86 Å), which is close to or within the energy window of the available X-rays (6–20 keV or 2.0–0.62 Å). When the X-ray energy is sufficient to excite such a transition, resonance between the electrons and the X-rays occurs. Some photons are actually absorbed and re-emitted by fluorescence at a lower energy. However, not all resonance events lead to absorption; some X-ray photons are scattered in a process known as anomalous scattering. It is called anomalous because it does not obey Friedel's law; $|\mathbf{F}_{h,k,l}|$ no longer equals $|\mathbf{F}_{-h,-k,-l}|$ and the reflections are called a Bijvoet pair. Furthermore, the difference between $|\mathbf{F}_{h,k,l}|$ and $|\mathbf{F}_{-h,-k,-l}|$ is wavelength (i.e., energy) dependent. The relative size of the anomalous contribution is maximal at the absorption energy and depends on the number of electrons resonating at that energy. The anomalous scattering is therefore stronger from heavier atoms like Hg.

Why is this important? It gives us another way to solve the phase problem, because we again have two measurements of related reflections (this time h, k, l and $-h$, $-k$, $-l$) with a difference in their **F**s, much as for isomorphous replacement. Again, the simplest way to understand the problem is through an Argand diagram.

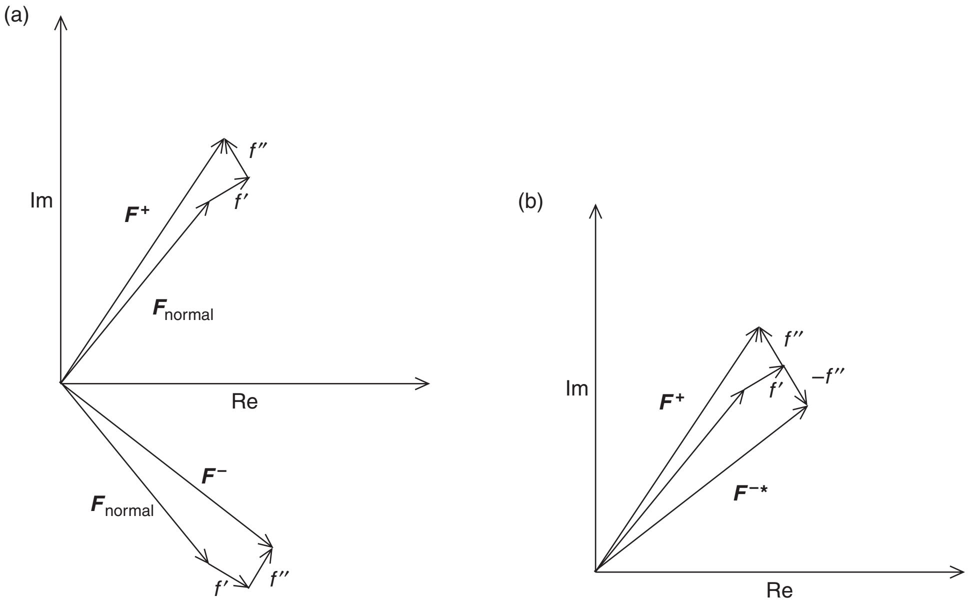


Figure 22 (a) The separation of the 'normal' and anomalous components of a Bijvoet pair of structure factors presented on an Argand diagram. The anomalous contribution can be further divided into real (f') and imaginary (f'') components that are perpendicular to each other. (b) The Bijvoet pairs are often shown in the same quadrant, so that $\boldsymbol{F}^{-*}$, the complex conjugate of $\boldsymbol{F}^{-}$ is drawn. Because f' is the same for both (at one wavelength), $\boldsymbol{F}^{+}$ and $\boldsymbol{F}^{-*}$differ only by direction of the f'' vector.

Each structure factor $\mathbf{F}_{h,k,l}$ can be separated into the normal contribution and the anomalous contribution (Figure 22). The anomalous contribution can be further separated into a dispersive term f' and an anomalous term f'', which are perpendicular to each other on the Argand diagram (Figure 22). In comparison to the overall scattering from the unit cell of a macromolecular crystal, which typically contains thousands of light atoms, the anomalous component is very small – smaller, even, than the differences when a heavy atom is added. In anomalous scattering, the differences between $I_{h,k,l}$ and $I_{-h,-k,-l}$ originate only from the anomalously scattering atoms – the anomalous substructure. As these atoms are few and far apart, we may again use small molecule 'direct methods' for solving the substructure. Once we know that, we can calculate the phase of the anomalous component. In the Argand diagram this corresponds to the difference between the structure factors $F_{h,k,l}$ and $F_{-h,-k,-h}$ so we can draw a Harker construction very similar to that in SIR (Figure 23). Again the circles intersect at two points, leaving a phase ambiguity. This is Single wavelength Anomalous Dispersion or SAD. The phase information from SAD experiments is frequently sufficient for solving the structure, but in those cases it critically depends on methods of density modification, which we will discuss later.

As an aside, let us return for a moment to SIR. How can anomalous scattering be used to break the phase ambiguity of SIR if both methods have similar Harker constructions? Fortunately, the information from the isomorphous differences and the anomalous differences is not the same, but complementary. If, as is usually the case, the heavy atom is the only anomalous scatterer, the substructure is the same, that is, we can use the same F_{H} for a reflection h, k, l. When the anomalous difference f'' and its inverse are added to $\mathbf{F_H}$, we can draw two circles of radii $|\mathbf{F}_{\mathrm{PH}}{}^{+}|$ and $|\mathbf{F}_{\mathrm{PH}}{}^{-}|$ centered at the ends of each vector (Figure 24). With the $|\mathbf{F}_{\mathrm{P}}|$-circle centered at the origin, these three circles only intersect at one point, which defines the phase of $\mathbf{F}_{\mathrm{P}}$. This method is Single Isomorphous Replacement with Anomalous Scattering – or SIRAS.

The contribution of anomalous scattering changes with wavelength. The differences between measurements made at different wavelengths are called dispersive differences. They are due entirely to the heavy atom(s) present as the light atoms do not scatter anomalously. The wavelength dependence can be used to resolve the phase ambiguity in SAD. This method is known as multiple wavelength anomalous dispersion (MAD). The phenomenon of anomalous scattering is related to absorption and indeed the anomalous differences (f'') are maximal at the absorption edge (Figure 25). At energies above the edge, they slowly decrease. The dispersive

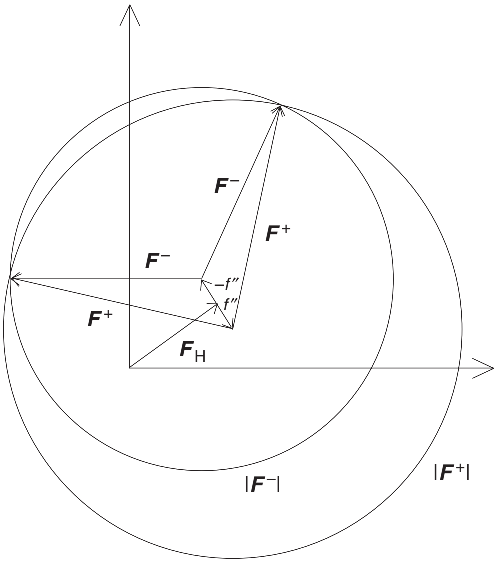


Figure 23 The Harker construction for the SAD method. The anomalous differences *f*″ are drawn from the end of $\boldsymbol{F}_\mathrm{H}$, giving rise to two circles. These two circles intersect at two points (just like in SIR), leaving the phase of $\boldsymbol{F}_\mathrm{P}$ (not shown for clarity) ambiguous.

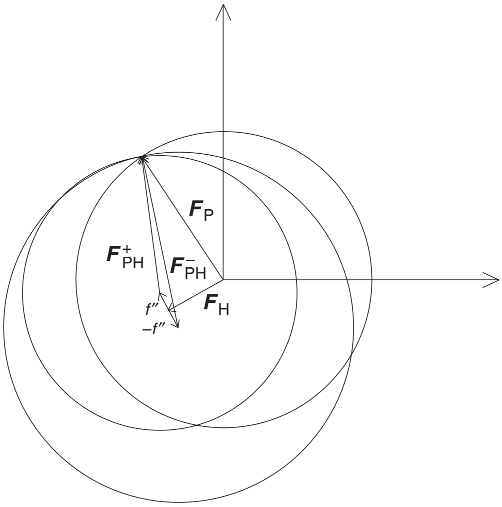


Figure 24 The Harker construction for the SIRAS method. The phase ambiguity in SIR is broken by drawing two circles $|\mathbf{F}^+{}_\mathrm{PH}|$ and $|\mathbf{F}^-{}_\mathrm{PH}|$ centered at the ends of the vectors $\mathbf{F}^+$ and $\mathbf{F}^-$ respectively. These circles have only one intersection with the circle $|\mathbf{F}_\mathrm{P}|$ centered at the origin, just like in MIR (**Figure 21**), leaving only one possible value for the phase of $\mathbf{F}_\mathrm{P}$.

component (f'), however, has a far steeper energy dependence. It is at a minimum at the inflection point of the absorption curve. MAD phasing involves collecting two or more data sets at different wavelengths (energies). In order to have the maximal phasing information, one data set is usually collected precisely at the edge where the anomalous differences (f'') are maximal, but the dispersive component (f') is small (**Figure 25**). (This is often called the peak data set and it can be used for SAD.) Another data set is then collected at the inflection point of the absorption curve, where the f' is at a minimum and therefore the difference to the peak data (and hence dispersive signal) is greatest (**Figure 25**). The exact energy of the absorption edge and inflection point in the protein crystal are determined experimentally by scanning the fluorescence as a function of energy. For a more detailed explanation of the method (see, e.g., Hendrickson and Ogata[17]).

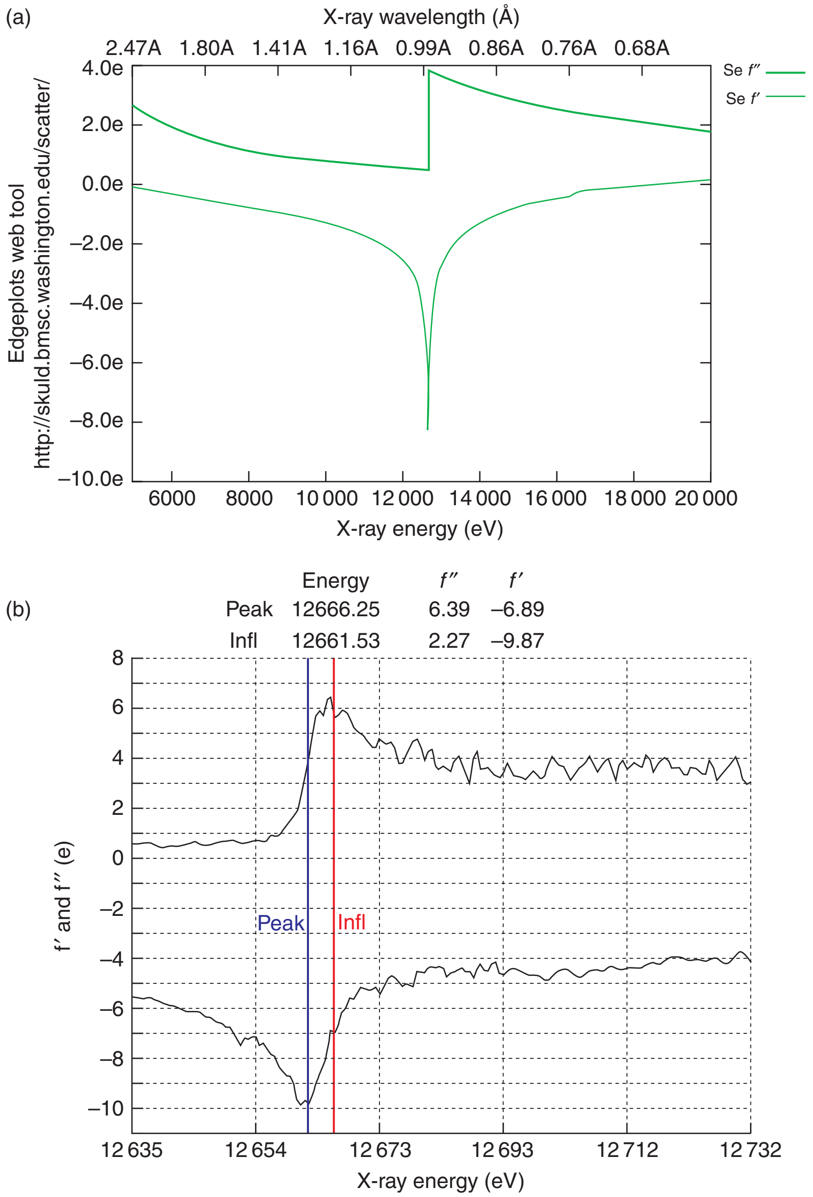


Figure 25 A plot of *f*′ and *f*″ for selenium as a function of energy. (a) A theoretical plot from 5 to 20 keV. Reproduced from Ethan Merritt's website, with permission from Ethan Merritt. (b) A fluorescence scan of the selenomethionyl derivative of the CBS domain of *Clostridium perfringens* inorganic pyrophosphatase (H. Tuominen, unpublished) around the absorption edge showing energies where peak (red) and inflection point (blue) data were collected. The values of *f*″ and *f*′ were fitted by the program CHOOCH.[27]

As mentioned at the beginning of this section, anomalous scattering based methods (SAD and MAD) have become very important. This is due to two reasons. First, unlike in the isomorphous replacement methods, the methods do not require multiple crystals, one with and one without a heavy atom. All the measurements are made from one crystal, so there is no problem with isomorphicity; a crystal is by definition isomorphous with itself. As long as the protein has a 3rd transition row or heavier element bound, MAD and SAD experiments can be performed. Second, modern techniques have made the method extremely easy to use; in most cases, it is not even necessary to introduce a heavy atom into the protein after crystallization. Selenium has a K-edge at 12.6578 keV, or 0.9795 Å, a very good energy for data collection at a synchrotron and cells can be grown on selenomethionine as a substitute for methionine. This is true for *E. coli*, yeast (*Saccharomyces cerevisiae* and *Pichia pastoris*), and even insect cells. This introduces an anomalously scattering atom that interferes neither with the fold nor function of the protein.

However, these two advantages would count for nothing without easy access to synchrotrons. Anomalous scattering is only observed near the resonance condition:

$E_{\text{X-ray}} \approx E_{\text{electronic transition}}$, so tuneable synchrotron radiation is essential for using these methods. MAD measurements in particular need to be made in a narrow wavelength window near the absorption peak where the anomalous scattering contribution changes rapidly. From the above discussion it means that selenomethionine MAD has become the most popular method for obtaining phases experimentally. Typically, one selenomethionine provides enough signal to phase a protein of about 17 kDa.

9.03.9.3 Molecular Replacement

Experimental phasing does not require any prior knowledge about the structure of the macromolecule in the crystal, but multiple, often complicated diffraction experiments have to be performed. What if we already know something about the molecule, such as the structure of a closely related molecule? Can we use such prior knowledge?

We have shown above that if the electron density is known, the structure factors, including their phases, may be calculated by an FT. The electron density clearly depends on the coordinates of the constituent atoms; if we know the atomic coordinates we can calculate the structure factors using an inverse FT which, in principle, can be done over atoms (cf. Equation (14)):

$$\boldsymbol{F}_{hkl} = \sum_{j=1}^{n} f_j \, \mathrm{e}^{2\pi \mathrm{i}(hx_j + ky_j + lz_j)} \tag{14}$$

where f_j is the atomic scattering factor of atom j and x_j, y_j, and z_j are its coordinates in the unit cell. (As a reminder, $\boldsymbol{F}_{hkl}$ is a complex number, or can be expressed as $\boldsymbol{F}_{hkl} = |\boldsymbol{F}_{hkl}|\alpha_{hkl}$ where α_{hkl} is the phase in an Argand diagram.) Given a pre-existing model, we can calculate the diffraction pattern to any resolution we want. Resolution, as we mentioned in Section 9.03.7, is the ability to distinguish two points from each other. Consequently, at low resolution we see only the broad outlines of our structure, not the fine detail (**Figure 26**). The information about the overall shape is present in the low-resolution terms in the FT of the atomic model. Two molecules of similar shape will thus have a very similar FT at low resolution and differ only at higher resolution. Furthermore, if we knew the atomic coordinates of a related molecule of similar shape, then we could use the phases from that model as initial phases for our unknown structure. There is one problem: the structure factors are affected not only by the atomic positions with respect to each other but also with respect to the origin of the unit cell; look at the exponent in Equation (14).

As we do not know where our model is in the unit cell, any set of structure factors $\boldsymbol{F}_{hkl}$ that we calculate (called $\boldsymbol{F}_{\text{calc}}$ with phase α_{calc}) will be wrong. However, if the model (which we have) and the unknown structure are the same, we do not actually need to find x_j, y_j, and z_j for each and every atom of our model in the unit cell of the unknown structure. We need to find 'where and how' the model should sit in the unit cell of the unknown structure. Once we know that, the model diffraction pattern and the unknown structure diffraction pattern (for which we have measured structure factors, called $\boldsymbol{F}_{\text{obs}}$, but not phases) will be the same – to the extent that the model and the unknown structure are the same.

What is the 'where and how' of the model in the unit cell? 'Where' is the x, y, and z that put the center of the model in the correct place in the unit cell: where the center of the unknown structure is; and 'how' is the three rotation angles (α, β, γ)[13] that rotate the model so that it is in the same orientation as the unknown structure in the unit cell.

[13] The orientation of any object can be uniquely described by rotations around three angles. There are multiple conventions for how these angles are chosen; the Euler angles (usually α, β, γ) and spherical polar angles (usually ϕ, ψ, κ) in different variants are the most common. For definitions, see, for example,. Giacovazzo *et al.*[6] A useful web resource for converting different conventions is the CSB rotation server (http://seal.csb.ki.se/xray/convrot.html).

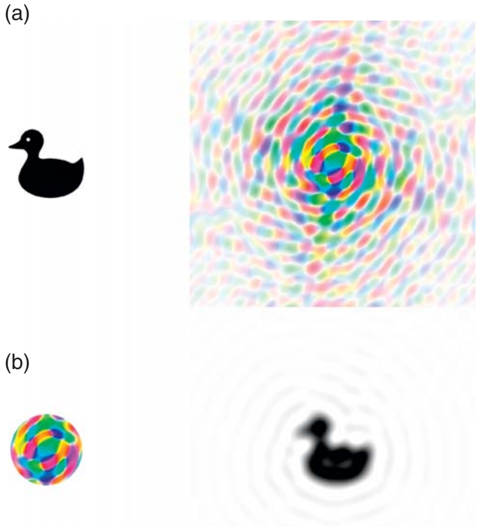


Figure 26 The effect of resolution on FTs. (a) A picture of a duck and its FT. (b) A back transform of the duck using only the low-resolution terms of the FT. Only the general shape of the duck is discernible; the details are lost. Reproduced from the web-based book 'Book of Fourier', University of York, UK, with permission from Kewin Cowtan; reimplemented by I. Karonen.

This is a six-dimensional search, which is time-consuming even on modern computers. Fortunately, the orientation and position searches can be done separately – with the orientation search first. So how do we actually know when we have found the correct position and orientation? As we rotate and move the model in the unit cell, the calculated structure factors F_{calc} change. We can calculate the correlation between the observed intensities (I_{obs}) and the 'calculated intensities' which are calculated as $\boldsymbol{F}_{\text{calc}}\boldsymbol{F}^{*}_{\text{calc}}$. The maximal correlation should correspond to the correct position. When the model is correctly positioned, we then have not only structure factors $|\boldsymbol{F}_{\text{calc}}|$ but also phases for each reflection α_{calc} using Equation (14). Using these phases, which are not quite the correct phases for the structure being investigated (because the model is not correct in detail, cf **Figure 15**), but close, we can calculate an electron density map. This map has at least some of the features of the real molecule.

9.03.10 The Electron Density Map

Once we have measured the structure factor amplitudes in a diffraction experiment and obtained a phase for each, we can calculate the electron density function using a formula like this:

$$\rho(x,y,z) = \sum_{h}\sum_{k}\sum_{l} \boldsymbol{F}_{hkl}\, \mathrm{e}^{\mathrm{i}\alpha_{hkl}}\, \mathrm{e}^{2\pi \mathrm{i}(hx+ky+lz)} \tag{15}$$

How $\boldsymbol{F}_{hkl}$ and α_{hkl} are derived will depend on how the phases were derived: experimentally or by molecular replacement. This three-dimensional function has a value everywhere in the unit cell, but for practical reasons its value is calculated at selected grid points, and is usually represented as an isocontour surface[14] at a given value. **Figure 27** shows such a surface represented by a chicken wire model contoured at a level of 0.39 electrons Å^{-3} or one standard deviation.

[14] An isocontour surface is the collection of points with the same value of the function in question. A two-dimensional example is a weather map, where the points on the isobars all have the same value of pressure. The electron density function is three-dimensional, so we have surfaces instead of curves.

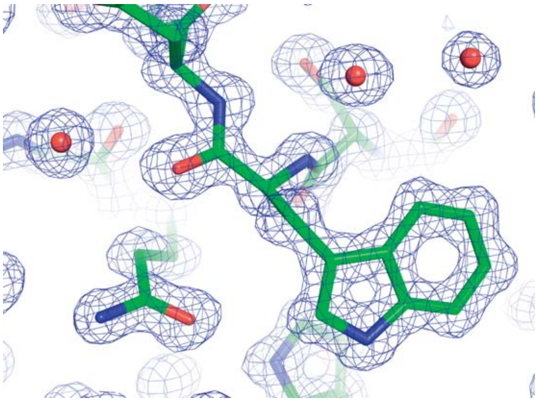


Figure 27 A $2F_{\mathrm{obs}}–F_{\mathrm{calc}}$ electron density map of *Aspergillus flavus* urate oxidase contoured at the 1σ level (M. Spano, unpublished). The absolute value of the electron density is actually not useful in macromolecular crystallography; what matters are the relative values. Electron density values are commonly given as multiples of the standard deviation σ, and where there are electrons, there are atoms. The protein atoms are shown in sticks and the red spheres represent ordered water molecules.

9.03.10.1 Modifying the Electron Density Map

Inasmuch as the phases dominate the appearance of the electron density map, errors in the phases will make the map much more difficult to interpret. Unfortunately, the initial phases obtained from the methods outlined above always contain errors. In experimental methods, very small differences in amplitudes are measured and the resulting phases are the statistical best estimates: This does not mean they are right; they may not yield the most interpretable map. Even before constructing a model to interpret the density, various forms of density modification are used to improve the map.

The most powerful form is noncrystallographic symmetry averaging, if such symmetry exists. Recalling Section 9.03.3 we mentioned the concept of crystallographic and noncrystallographic symmetry. If the protein exists as a multimer in the asymmetric unit, the different copies of the protein are often very similar to each other. However, the initial electron density usually does not display that similarity, and errors in the phases are the culprit. The errors in phases can thus translate into artifactual differences between regions in the electron density map that should be the same. We can decipher how the different regions of the electron density map are related to each other through a process similar to the one we used to calculate how to rotate and translate a known model into the unknown unit cell. This will give us one or more rotation–translation operator(s) that relate different regions of the electron density in the asymmetric unit to each other, and we can then use the relationship(s) to average the different regions together. The averaged electron density map describes the true electron density of the molecule better than the unaveraged map as long as two conditions hold. First, the rotation–translation operator(s) used must be correct and, second, the molecules must have the same conformation in detail. Averaging can be combined with the other density modification techniques described below.

The high solvent content of macromolecular crystals leads to another way to modify the electron density. The electron density map is the space average of all the unit cells in the crystal, so atoms that are in random positions (as in liquid water) in different unit cells will not show up as peaks. Why? They do not obey the periodicity of the crystal (remember that the FTs concerned periodic functions) and so are called disordered. The crystal then consists of ordered molecules, where the electron density is the same in each unit cell, and so visible, and disordered solvent, where the electron density averages to zero. In order to make physical sense, the electron density also has to be positive; a property not imposed by the FT. Consequently, electron density peaks outside the macromolecule are noise and can be got rid of, as can negative electron density inside the macromolecule. We can therefore apply these conditions: modify the initial electron density so that it is zero outside the molecules and positive within them. This, as for noncrystallographic averaging, alters the electron density map to conform to what must be true, and the map is thus a better representation than the initial map.

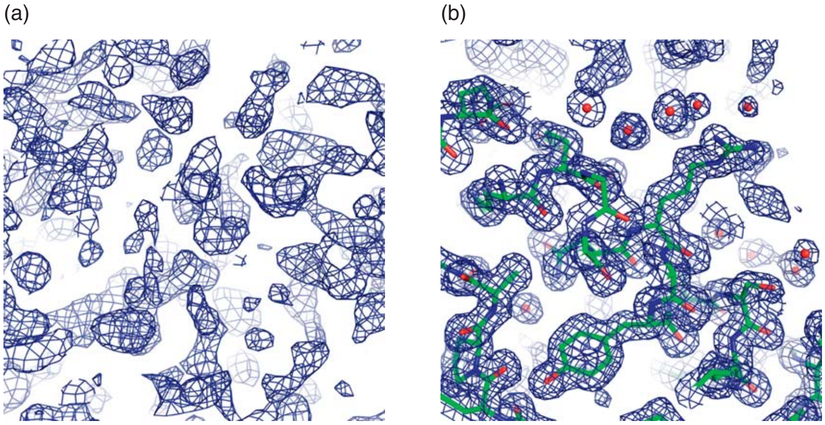


Figure 28 (a) An electron density map calculated from the most likely phases (also known as centroid phases) of selenium SAD-phasing of conserved dopamine neurotrophic factor (CDNF) (V.-M. Leppänen, unpublished). (b) The final map with the refined model.

We can then calculate new phases (α_{calc}) based on this improved map as in Equation (15). These new phases will be closer to the correct phases than the original ones, and therefore a map calculated using the new phases and new amplitudes (see below) will be better than the initial one. In the case of unresolved phase ambiguity, such as in SAD, density modification can be used to find the right phase angle. A map calculated from the statistically most likely phases – halfway between the right and wrong ones (**Figure 28**) is very rarely interpretable. The phases are too far from correct. Iterative density modification with or without symmetry averaging can be used to converge to the correct phases, since only those phases will lead to a map with a clear contrast between the molecule and the solvent (**Figure 28**).

9.03.10.2 Interpreting the Electron Density Map

From the above, it should be clear that electron density maps are the actual *result* of a crystallographic structure determination and, like most raw results, provide little insight on their own. What we want is an interpretation in terms of atoms – and yet in macromolecular crystallography, the electron density maps are almost never good enough to position all the atoms unambiguously. The model is only one of many possible interpretations.

Even before there is a model, assumptions are made about the properties of the electron density. Indeed, that is precisely what the density modification described above is: modifying the density based on what we *expect* it should look like. Regions that are the same should look the same; regions outside the macromolecule should be zero; regions inside the macromolecule should have positive electron density. The next step is to interpret the electron density map in terms of a molecular model, based on the chemical composition of the biological macromolecules in the unit cell, and on the geometrical properties (such as bond lengths and angles) of organic molecules with similar structures, such as the C═ONH linkages of peptide bonds.

While this makes perfect chemical sense, it causes a problem. Whatever the source of the phases used to calculate a new electron density map, some features of the model (or old electron density map) will show up in the new map, because the phases dominate its appearance (**Figure 15**), as mentioned in Section 9.03.9.3. If our model (and hence the phases calculated from it) is correct, this is not a problem, but since the process of structure refinement that we discuss below is iterative, the correctness of the model must be assessed carefully. In crystallography, what you see is what you put in – also known as model bias.

During the process of completing a structure determination the refinement of the structure and calculating and inspecting electron density maps proceed hand in hand. We will first discuss the maps, and then the refinement of the structure.

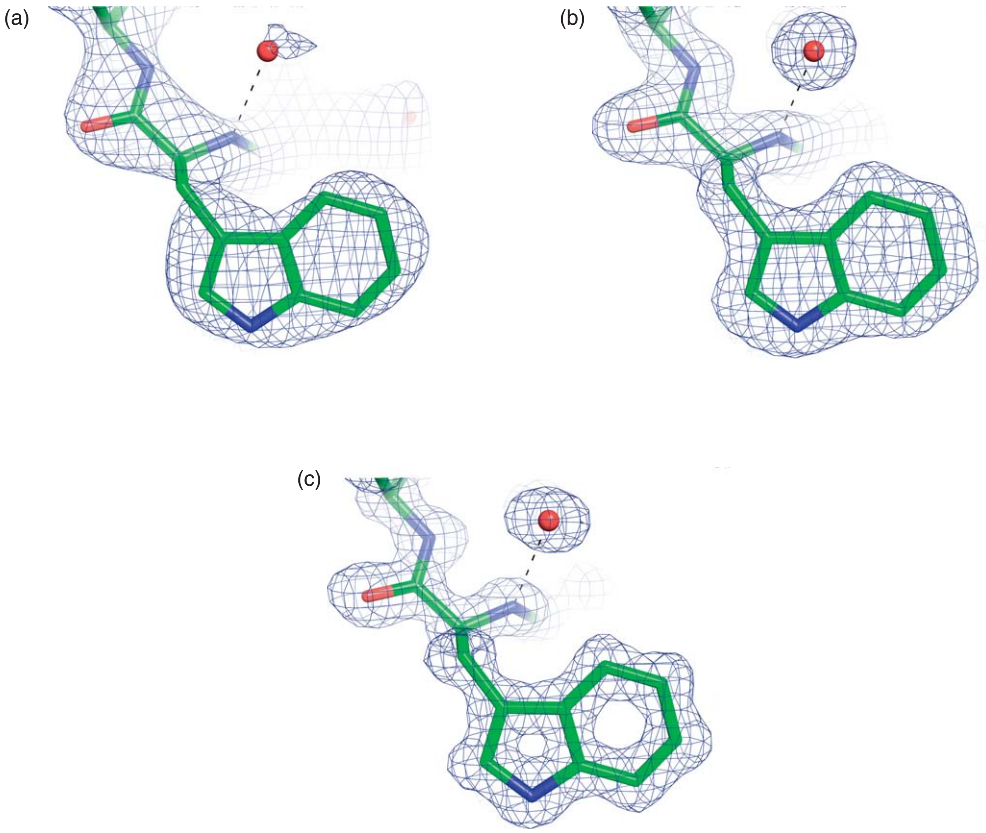


Figure 29 A $2F_{\mathrm{obs}}$–F_{calc} electron density map around a tryptophan residue in *Aspergillus flavus* urate oxidase contoured at 1.5σ level calculated at (a) 3.0 Å (b) 2.0 Å, and (c) 1.1 Å resolution (M. Spano, unpublished).

The resolution of the data available affects the appearance of the maps dramatically – and thus what can be understood from them. Figure 29 shows the same map calculated at various resolutions; at 3 Å the side chain positions can be distinguished, but not their conformation except for very large residues, like the tryptophan in the figure. At 2.0 Å, on the other hand, the side chain conformations are clearly identifiable and water molecules are seen, while at 1.1 Å, both the benzene and indole rings have holes in them, and the positions of the protons on the side chains become visible. In addition, as was discussed above (Figure 15), the phases dominate the appearance of the electron density map, so even a high-resolution map with poor phases can be difficult to interpret.

To avoid the model bias problem, various kinds of electron density maps are calculated as the structure is solved, and are often presented in publications involving macromolecular structures. To evaluate a structure or structural paper critically, one has to inspect the maps. A simple Fourier map with experimental $|\mathbf{F}_{\mathrm{obs}}|$s as amplitudes and model α_{calc}s is never used in practice. It has terrible model bias because, as mentioned above, the phases dominate the appearance of the maps. To reduce this, maps these days are also weighted by a factor σ_{A} related to the coordinate errors in the model. Often amplitudes such as $2|\mathbf{F}_{\mathrm{obs}}|-|\mathbf{F}_{\mathrm{calc}}|$ are used instead of $|\mathbf{F}_{\mathrm{obs}}|$, because this increases the size of electron density peaks due to differences between the $|\mathbf{F}_{\mathrm{obs}}|$s and $|\mathbf{F}_{\mathrm{calc}}|$s.[15] Those differences are due to errors in the model; in other words, a map calculated with $2|\mathbf{F}_{\mathrm{obs}}|-|\mathbf{F}_{\mathrm{calc}}|$ for amplitudes and α_{calc} for phases will have less model bias than one calculated with $|\mathbf{F}_{\mathrm{obs}}|$ for amplitudes and

[15] The formula is this, where $|\mathbf{F}_{\mathrm{obs}}|$ and $|\mathbf{F}_{\mathrm{calc}}|$ are observed and calculated amplitudes, and the phase of the wave is given by α_{calc}. $\rho(x,y,z)=\sum_h\sum_k\sum_l\left(2\left|\mathbf{F}^{\mathrm{obs}}_{h,k,l}\right|-\left|\mathbf{F}^{\mathrm{calc}}_{h,k,l}\right|\right)\mathrm{e}^{\mathrm{i}\alpha_{\mathrm{calc}}}\mathrm{e}^{2\pi\mathrm{i}(hx+ky+lz)}$.

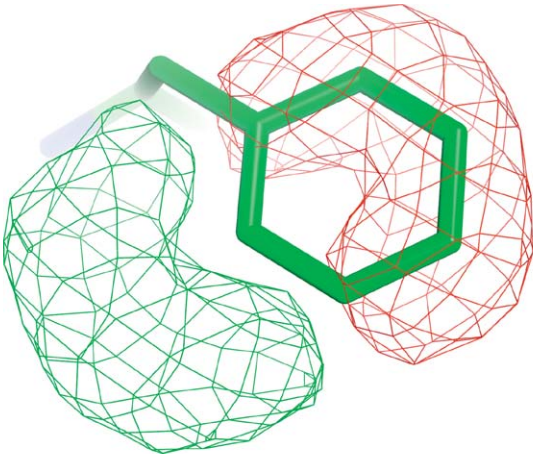


Figure 30 The $m|\mathbf{F}_{\text{obs}}|–D|\mathbf{F}_{\text{calc}}|$ map around Phe331 of the second PDZ domain of SAP97 contoured at $–3\sigma$ (red) and 3σ (green) after misplacing the side chain. The positive density (green) shows where the side chain should be and the negative density (red) shows where atoms should be removed.

α_{calc} for phases. The actual form of the (complex) structure factor is $\mathbf{F} = (2m|\mathbf{F}_{\text{obs}}| - D|\mathbf{F}_{\text{calc}}|)e^{i\alpha_{\text{calc}}}$, where, based on the expected error in the model, m and D differ from one. This minimizes the effect of model bias.[18]

A difference map with $|\mathbf{F}_{\text{obs}}| - |\mathbf{F}_{\text{calc}}|$ as amplitudes helps to identify discrepancies between the observed and calculated data. A negative peak (a hole) in the difference electron density map indicates something in the model that is not supported by the experimental data (**Figure 30**), while a positive peak indicates some feature in the data that is not in the model (**Figure 30**). As difference maps by definition subtract out all the real features currently accounted in our model, they are noisy, with many peaks and holes at the level of one to two standard deviations (σ). We thus interpret only peaks that are above 3σ. Difference maps are often presented as evidence of the presence of atoms or molecules not covalently bound to the macromolecule, such as a bound ligand.

Another method of reducing model bias is the omit map. A part of the model of which we are uncertain, such as a ligand or a loop region, is omitted from the model. The structure is then refined with this part left out and the phases calculated. The omitted part should nonetheless appear in the map, provided it is a real feature of the molecule.

9.03.11 Model Building and Refinement

Let us turn to the other part of the process: what do we do with our electron density map once we are certain that the initial solution to the phase problem is as correct as possible – that the map cannot be improved? We interpret the electron density map in terms of a molecular model: of atoms at given positions x, y, z. For a protein, this means following the path of the polypeptide backbone through the map; this is called a main chain trace. Following this, we add all the other chemical components we know about: side chains, water molecules[16] and, if present, prosthetic groups, small molecule ligands, metal ions, and so on. If data are available to sufficiently high resolution and the starting phases are good, the building of the protein can be automated, but usually it has to be done manually. If the phases originate from molecular replacement, an initial model is already available and only needs to be modified – with the model bias *caveat* mentioned above.

[16] Not all solvent molecules are disordered; some occupy the same position in all unit cells and hence show up in the map. These ordered waters can be structurally or functionally important.

The initial model is never the best obtainable but, in order to improve it, we need to know how to assess and compare its correctness. By far the best-known and most widely used measure is the crystallographic residual or *R*-factor:

$$R = \frac{\sum_{h,k,l}[|\mathbf{F}_{\text{obs}}| - |\mathbf{F}_{\text{calc}}|]}{\sum_{h,k,l}[|\mathbf{F}_{\text{obs}}|]} \tag{16}$$

where $|\mathbf{F}_{\text{obs}}|$ are the observed amplitudes and $|\mathbf{F}_{\text{calc}}|$ the amplitudes calculated from the model. The *R*-factor for a model that exactly matches the calculated structure factors would be 0, and for a completely random model it would be 0.59. However, a static atomic model is not a very accurate description of the actual contents of the unit cell, and so the *R*-factor never reaches zero even for good models; it is only a rough guide to the correctness of the structure. It does, however, allow us to compare the fit of two models to the experimental diffraction data and see which is better.

Thus, how can the initial model be improved? When a model is fitted to the observed data, the errors are often assumed to be normally distributed, and the statistically best fit is then obtained by minimizing the sum of the squared differences between the data points ($|\mathbf{F}_{\text{obs}}|$) and the values predicted by the model ($|\mathbf{F}_{\text{calc}}|$). This is called *crystallographic refinement*. (This is exactly the mathematical argument behind the linear regression formula for finding the best line through a set of points: we write a function:

$$y = mx + c + \varepsilon \tag{17}$$

where ε is a normally distributed error function.) The method of least-squares refinement is still used in small molecule crystallography, but in macromolecular crystallography, maximum likelihood refinement is almost exclusively used nowadays. The approximation that the errors are normally distributed is actually not very good, because the phases are not measured. Maximum likelihood refinement gives, as a final model, the one that is most likely to have produced the data that was actually measured.[17] This reduces model bias and the model produced is *not* the same as the model that minimizes the $(F_{\text{obs}} - F_{\text{calc}})^2$ differences.

Whether we use maximum likelihood or not, we need to optimize the three positional coordinates *x*, *y*, *z* and a temperature factor *B* (discussed below) for each atom. This is analogous to solving a system of linear equations and, as in that or any optimization problem, the number of observations must at least equal the number of variables or parameters. Usually the observations are not without error so, in order to arrive at a reliable result, the problem has to be overdetermined. There must be more observations than parameters. For macromolecular structure refinements, the number of observations (the observations are the individual reflection, the number of which is determined by the resolution of the data) is very rarely sufficient. Again, we take prior chemical knowledge from studies of small organic molecules about what the bond lengths, angles, and planar fragments (such as aromatic rings) should be. We then keep the model close to these values during refinement. The process is called restrained refinement, and the properties being used (lengths, angles, and so forth) are called restraints. This effectively reduces the number of parameters because the atoms are no longer free to move independently during refinement.[18] The problem is thus more overdetermined. The restraints may be formulated either as target values of the geometrical parameters or pseudo-energies. The progress and convergence of the refinement can be monitored by the change in the *R*-factor.

The problem with the *R*-factor is that it almost always decreases as refinement progresses, even if the change introduced is incorrect. This is expected; we are, after all, minimizing the difference between F_{obs} and F_{calc} – precisely what the *R*-value measures. A way around this is the R_{free}-value, which is an independent indicator of structure correctness. Before refinement starts, the observed reflections are divided into a working set used for refinement and a test set (typically 5–10% of the data), unused during refinement. R_{free} is calculated with only the reflections of the test set. It is therefore independent of the bias inherent in the normal *R*-value, called in this case R_{work}.

[17] If the description of maximum likelihood refinement sounded like a statement in Bayesian statistics, it was supposed to. You can find more out about Bayesian statistics in crystallographic refinement from McCoy[19] or Tronrud[20].

[18] If this is not clear, imagine the benzene ring in phenylalanine. To a first approximation, all the atoms in the ring move together, so instead of six atoms times four parameters (*x*, *y*, *z*, *B*), or 24 parameters, restrained refinement reduces the number of parameters to about seven: *x*,*y*,*z* for the center of mass of the benzene ring, α,β,γ to specify its orientation, and a temperature factor.

The problem in crystallographic refinement is nonlinear (x in Equation (17) has an exponent other than one), so unlike in linear least squares (Equation (17)), there are local minima. The refinement 'stalls' and the R-factor no longer decreases, but there are still errors in the structure, and some other – usually only slightly different – structure would fit the data even better but still be consistent with the restraints. The problem is that the intermediate geometries have higher R-factors, and by definition our minimization algorithm will not cross such barriers. Think of a roller coaster. Without the energy imparted by being dragged up the first hill to the highest point, all of the coasters would remain at the nearest low-point – the local minimum. This limitation can be overcome by using molecular dynamics; the fit to the data is expressed as a pseudo-energy and the molecule 'heated' in the computer to an artificially high temperature to move the atoms around during the molecular dynamics simulation, after which it is annealed – slowly 'cooled' and refined at the same time. This method is often called simulated annealing.[21]

In practice, structural refinement is an iterative process of alternating steps of model building and refinement. When the model improves, the phases also improve. New features appear in the electron density maps. The existing model from the previous refinement step needs to be manually modified by adding new atoms to the new electron density and removing them where the electron density has disappeared. In high-resolution structures alternate conformations of side chains are modeled (with occupancies below one; see below), and at the very highest resolutions, even hydrogen atoms can be visible. When further refinement or manual model building lowers the R-factors, the refinement has converged, and we have a final model to start interpreting.

9.03.11.1 Modelling Disorder: Temperature Factor and Occupancy

We mentioned earlier that a static atomic model does not describe a macromolecular crystal particularly well. This is because of disorder, which can be roughly divided into two kinds: static and dynamic. The former is due to differences between one unit cell and the next, while examples of the latter are molecular vibrations in the crystal and free rotations of side chains like valine. In static disorder, the average position is different in different unit cells, while in dynamic disorder it is the same. How can we model it? The usual way is to assume that each atom j is in isotropic (or spherically symmetrical) harmonic motion about its equilibrium position x_j, y_j, z_j, defined by its temperature factor, B_j.[19] This is convenient, because the distribution of a harmonic oscillator is Gaussian, simplifying the mathematics immensely. The B-factor does *not* mean that atoms actually oscillate harmonically; it merely gives the width of the Gaussian probability distribution of the atom around its mean position. If we assume this model, then U, the mean square deviation of the atom from its mean position, is given by $B = 8\pi^2 U$. The units of both are thus Å^2. Atoms with low B-factors will thus have a better-defined position than atoms with high B-factors. For example, an atom with a B-factor of 20 Å^2 will have a root mean square deviation $\sqrt{U}$ of 0.5 Å from the equilibrium position (see **Table 2**).

Table 2 Some values of the B-factor in the range typically observed in protein structures and the corresponding root mean square deviations $\sqrt{U}$

B (Å^2)	*$\sqrt{U}$ (Å)*
10	0.36
20	0.50
30	0.62
40	0.71
50	0.80

[19] In the very highest-resolution structures, we can allow deviations from this approximation, using something called anisotropic temperature factors. This describes the motions of the atoms by tensors, which are outside the scope of this chapter. For a description, see Schneider[22].

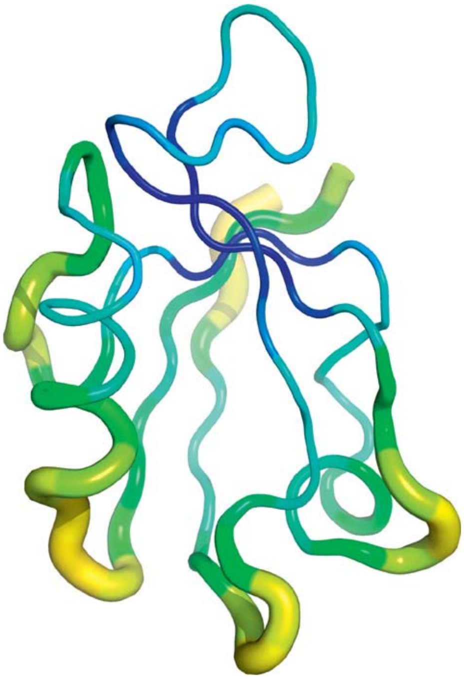


Figure 31 A 'B-factor sausage' representation of the PDZ2 domain of SAP97. The width of the sausage and the color encode the B-factor. Blue corresponds to lower and yellow to higher B-factor. The B-factors are highest for loop regions and the termini.

B-factors thus contain essential information about the local reliability of atomic coordinates; higher *B*-factors occur in mobile loops, termini, and on long flexible side chains like arginine. A 'sausage' representation of a typical protein structure (**Figure 31**) shows the C_α *B*-factors as the width of the sausage.

A special case of static disorder occurs when an atom is present in only a fraction of the unit cells in the crystals. For instance, a ligand might be present in only half of the unit cells, or a side chain might have two or more possible conformations. This type of disorder can be modeled with an occupancy parameter *O*, ranging from one (fully present) to zero (absent). It is simply the fraction of unit cells where the atom in question is present. Low occupancy is, however, difficult to distinguish from high *B*-factor. Indeed *B*-factors and occupancies are statistically correlated. Therefore, occupancies are normally used only when there are sufficient data (i.e., high resolution) to justify a more complicated disorder model.

9.03.12 Model Validation

How do we validate a final model? How can we identify potential problem areas during refinement? How do we estimate the general reliability of the structure? Because we have restrained the bond lengths and angles in the refinement, they are likely to be close to small molecule values in any case. The peptide torsion angles ϕ and ψ (**Figure 32**) of proteins, on the other hand, are not specifically restrained, but adopt only certain pairs of values due to steric hindrance,[20] as first pointed out by Ramachandran in the eponymous plot. The plot identifies suspicious parts of the model at a glance. In a well-refined structure, there is usually a convincing chemical explanation when the ϕ and ψ torsion angles adopt noncanonical values, as in for instance the structure of reindeer β-lactoglobulin[23] (**Figure 32**). Even if the Ramachandran plot is not explicitly shown in a structural

[20] Glycine and proline are exceptions and follow different distributions of torsion angles. These residue types are commonly excluded from Ramachandran plots.

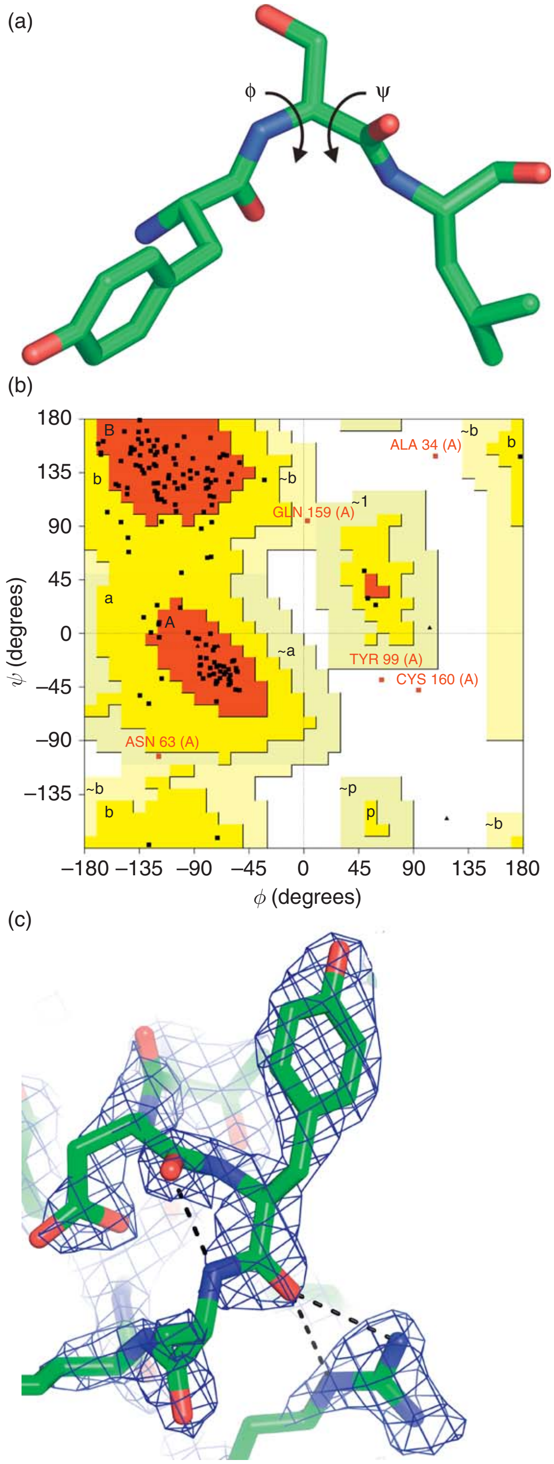


Figure 32 The Ramachandran plot (a) The mainchain dihedral angles; ϕ and ψ form the Ramachandran plot. (b) The Ramachandran plot (produced by the program PROCHECK[28] of reindeer β-lactoglobulin. (c) Tyr99 in reindeer β-lactoglobulin is part of a tight turn, where the hydrogen bonding interactions force it to adopt a disallowed conformation as shown by the ($2F_{\mathrm{obs}}$–F_{calc}) electron density map contoured at 1.5σ.

paper, the numbers of residues outside the allowed region are often reported. Ramachandran plots can also be calculated from the coordinates with online tools such as Molprobity (http://molprobity.biochem.duke.edu/).

Another way of evaluating the model geometry is to look at the deviations of geometrical parameters (bond lengths, angles, ring planarities) from small molecule values. These are usually expressed as root mean square (r.m.s.) deviations from 'ideal' values, usually those defined by Engh and Huber.[24] However, these ideal values, too, are used to restrain the model during refinement. If the restraints are too tight, the r.m.s. deviations will be small, even if the model fits the data poorly. In such a case of over-restraining, the *R*-values (which depend on the fit of the model to the data) will be large. If things do not turn out the way they should, it is thus possible to get a very good fit (low *R*) with a chemically unreasonable model (high r.m.s. deviations) or produce a chemically 'perfect' model (low r.m.s. deviations) that does not fit the data (high *R*). In practice, a compromise between the two has to be found.

Both the *R*-value and the average r.m.s. deviations are overall indicators of structure correctness. However, even structures of high overall quality may have disorder or model building errors at some specific location. These issues are usually clear upon inspecting the electron density map, but quantitative descriptors of the local fit to the diffraction data exist, one such being the real-space correlation coefficient (RSCC).[25] It measures the similarity of the map calculated with the experimental structure factor amplitudes $|\mathbf{F}_{\mathrm{obs}}|$ and that calculated from the model amplitudes $|\mathbf{F}_{\mathrm{calc}}|$ and its value ranges from −1 (perfect anticorrelation) to 1 (perfect correlation). It can be calculated for anything from an individual residue to an entire unit cell. A very useful Internet resource is the Uppsala Electron Density Server (http://fsrv1.bmc.uu.se/eds/), which, in addition to calculating and displaying electron density maps in a web-based interface, also calculates a number of useful validation statistics.

9.03.13 An Example of a Crystal Structure Determination

The determination of a macromolecular crystal structure is a complicated and often laborious process, but it is often rather briefly described in the materials and methods section of a journal article. We will use the paper by Leppänen *et al.*[26] as an example of a typical crystallographic structure determination and analyze the information presented.

> Crystals of the GFR1 domain 3 were grown at +4°C in sitting drops over a reservoir solution of 50 mM MES, pH 6.5, 0.2 M $MgCl_2$ and 10% (v/v) 2-methyl-2,4-pentanediol (MPD). The drops were prepared by mixing 2 µl of the reservoir solution and 2 µl of the protein solution at 3 mg/ml. The crystals belong to spacegroup $P6_1$ (a, $b = 61.3$ Å, $c = 65.2$ Å) with one molecule per asymmetric unit and solvent content of 51%. For data collection at −180°C, crystals were frozen in liquid nitrogen with the well solution containing MPD at 20% (v/v).[26]
> (Reprinted with permission from *EMBO Journal*)

The vapor diffusion method in a sitting drop (**Figure 5**) was used for reaching supersaturation. Crystallization in a cold room is often slower and produces better quality crystals. The combination of a buffer, an inorganic salt, and an organic precipitant is a common combination for protein crystallization. Very often several conditions are tested in smaller volumes, requiring sometimes months of work before well-diffracting crystals are found. When the crystals were cooled to the temperature of liquid nitrogen, more MDP was necessary to prevent the formation of crystalline ice. This is known as *cryoprotection.* The space group was determined to be $P6_1$ (a sixfold screw axis), which means there are six molecules in the unit cell related by a rotation of 60 and a translation of 1/6 along the axis *c* (10.9 Å), forming a sixfold helical 'spiral staircase' (**Figure 33**).

> MAD data on a selenomethionine derivative were collected to 1.8 Å using the BW7A beamline at EMBL Hamburg Outstation at three wavelengths (Table I). The remote wavelength data set was used for the final refinement. The data sets were processed with the programs DENZO and SCALEPACK (Otwinowski and Minor, 1997). CNS (Brünger *et al.*, 1998) was used to find the single selenium site and to estimate experimental phases at 2.0 Å (Table I). The spacegroup was shown to be $P6_1$ by calculating electron density maps both in $P6_1$ and $P6_5$, and choosing the one that gave clear protein–solvent boundaries. The electron density map obtained upon solvent flipping with CNS was used for initial model building.[26]
> (Reprinted with permission from *EMBO Journal*)

The data (**Table 3**) were collected at the DESY (Deutches Electronen Synchrotron) synchrotron source in Hamburg, Germany. The synchrotron facilities have experimental stations, or beam lines, designed for specific

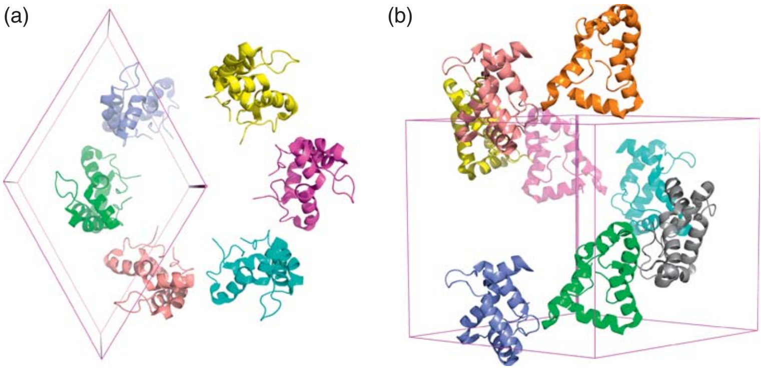


Figure 33 The P6$_1$ crystal of GFRα1. (a) A view along one sixfold screw axis of a GFRα1 crystal. Actually, the crystal consists of many such spirals ‘interwoven’ with each other, but only one is shown for clarity. (b) A view perpendicular to the screw axis showing the 1/6 unit cell translation.

Table 3 Data collection and refinement statistics from Leppänen *et al*.[26]

Data collection	λ_{peak}	λ_{remote}	$\lambda_{inflection\ point}$
Wavelength (Å)	0.9787	0.9635	0.9792
Resolution range (Å)[a]	20–1.8 (1.86–1.80)	20–1.8 (1.86–1.80)	20–1.8 (1.86–1.80)
Number of reflections			
Total	146 091	130 949	96 541
Unique	12 957	12 986	12 936
Completeness (%)[a]	100.0 (99.8)	100.0 (99.8)	100.0 (99.8)
I/σ[a]	43.7 (9.9)	41.2 (9.4)	34.6 (6.6)
R_{sym} (%)[a]	5.3 (30.5)	5.1 (25.5)	5.4 (35.1)
Number of Se-sites	1	1	1
Overall figure of merit for MAD phasing at 2.0 Å resolution			
Acentric		0.43	
Centric		0.51	
Refinement			
Resolution range (Å)		20–1.8	
Reflections		12 755	
R_{work} (%)		19.3	
R_{free} (%)[b]		20.8	
Average *B*-factor (Å^2)			
Protein		19.6	
Solvent		19.7	
r.m.s. deviation from ideal values			
Bond lengths (Å)		0.004	
Angles (°)		1.0	

[a] Values within parentheses correspond to the highest resolution shell.
[b] The R_{free} was calculated with 5% of the data omitted from structure refinement.
Reprinted from V. M. Leppänen; M. M. Bespalov; P. Runeberg-Roos; U. Puurand; A. Merits; M. Saarma; A. Goldman, *EMBO J.* **2004**, *23*, 1452–1462 with permission from *EMBO Journal* (Table I).

purposes. This particular station is operated by the European Molecular Biology Laboratory (EMBL) and is designed for multiple wavelength experiments. Crystals were grown from protein material where methionine was replaced by its selenium analogue. An X-ray absorption curve was recorded to determine the precise position of the absorption peak. Three datasets were collected: at wavelengths corresponding to the absorption peak, the inflection point of the absorption curve and a shorter (higher energy) wavelength sufficiently far from the peak as in **Figure 25**. These data were required to make use of the MAD phasing method (see Section 9.03.9.2). The anomalous differences are very small, so high redundancy data sets (7.5–11-fold, calculated as total reflections/unique reflections above) are needed to provide accurate enough data. The signal-to-noise ratios are also very good and R_{sym}'s low, also indications of accurate data. First, the anomalous scatterer substructure is solved. This consisted of one single selenium, which was nonetheless sufficient for the determination of phases.

The final choice of the space group could be made only based on the interpretability of electron density, because the space groups $P6_1$ and $P6_5$ differ only in the sense of the screw rotation (whether it is a left-handed or a right-handed helix). The electron density maps, however, are chemically reasonable only when the correct symmetry is used in the phasing procedure. In practice, the criterion for selecting the solution in $P6_1$ was clear boundaries between the ordered protein and disordered solvent, as is expected if the structure is correct. This illustrates the importance of prior information in macromolecular crystallography. A density modification method known as solvent flipping was used to improve the maps. The electron density values in the disordered solvent regions were inverted, thus increasing the contrast between the protein and the solvent. The resulting map (**Figure 34**) allowed a molecular model to be built into the density.

> Using the automated model-building tools in O (Jones *et al.*, 1991), the sequence was built for residues 239–300 and 309–346. This model was subjected to iterative rounds of building and refinement in CNS (Brünger *et al.*, 1998). Initial refinement was carried out using bulk solvent correction, torsion angle-simulated annealing, and *B*-factor refinement. Water molecules were added to peaks above 3.7σ in the (F_o–F_c) difference map if they had suitable hydrogen bonding geometry. The final model, with good stereochemistry (Table I), consists of 100 amino acids, one MPD molecule, and 97 water molecules. The N-terminal FLAG and 6His tags, residues 301–308 as well as the side chain of the first residue (239) are not seen in the electron density. PROCHECK (Laskowski *et al.*, 1993) was used to assign secondary structure elements and calculate the Ramachandran plot. Of all the non-Gly/non-Pro residues, 96.5% have main-chain torsion angles in the most favored regions, and there are no residues in the disallowed regions.[26]
>
> (Reprinted with permission from *EMBO Journal*)

The model was built in a partly automated manner, which was possible because of the good quality phases and high resolution. Automated model building usually does not perform as well as an experienced human crystallographer, but automation saves much time, so it is more often used in initial stages whenever possible. In order to reduce the number of parameters the molecular dynamics refinement was performed by varying just the torsion angles in the protein instead of the Cartesian coordinates of all atoms. As the model improved, ordered water molecules could be identified in the difference electron density maps. Not all of the protein is ordered in the crystal and hence is not visible even in the final maps. The model was validated based on both the main chain torsion angles (Ramachandran plot) and r.m.s. deviations from ideal values of bond lengths and angles.

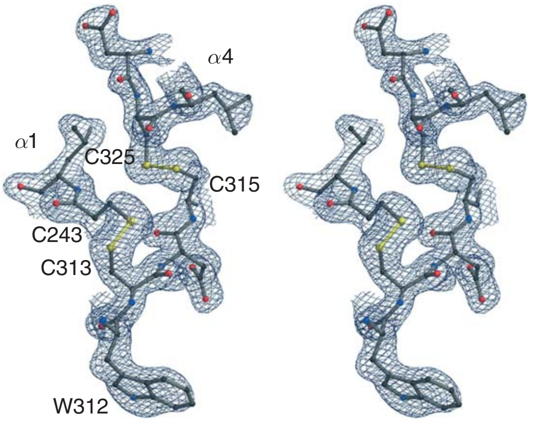


Figure 34 A stereo figure of the electron density around two of the S–S bridges in GFRα1 (Reprinted from V.-M. Leppänen; M. M. Bespalov; P. Runeberg-Roos; U. Puurand; A. Merits; M. Saarma; A. Goldman, *EMBO J.* **2004**, *23*, 1452–1462 with permission from *EMBO Journal*.) that join helices one and four, marked as α1 and α4. The electron density is sufficient to show that there are two disulfides even without supporting the chemical data.[27]

Acknowledgments

We thank Ilmari Karonen for writing the duck FT program at short notice, Monika Spano, Veli-Matti Leppänen, and Heidi Tuominen for giving us permission to use their unpublished results, and Heidi Repo for patiently reading earlier drafts of this chapter.

Nomenclature

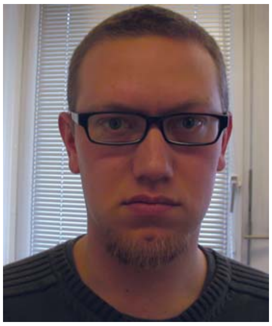
θ	angle of incidence on a Bragg plane
λ	wavelength (of X-rays)
$\rho(x, y, z)$	electron density at a point with coordinates *x*, *y*, and *z*
α, β, γ	euler angles for describing a general rotation
ϕ, ψ, κ	spherical polar angles for describing a general rotation
α_h	phase of a Fourier coefficient at harmonic number *h*
$\sigma_{h,k,l}$	standard deviation of the intensity of the reflection *h*, *k*, *l*
$\lvert\lvert\mathbf{F}_{h,k,l}\rvert\rvert$	amplitude of a structure factor of Miller indices *h*, *k*, and *l*
$<I_{h,k,l}>$	mean intensity of multiple observations of the reflection *h*, *k*, *l*
2θ	scattering angle (of X-rays)
$\mathbf{a}^*, \mathbf{b}^*, \mathbf{c}^*$	basis vectors of the reciprocal lattice
$\mathbf{a}, \mathbf{b}, \mathbf{c}$	basis vectors of a crystal coordinate system
B_j	temperature factor for atom *j*
$\mathbf{d}$	a vector normal to a Bragg plane
d	perpendicular distance of two Bragg planes
f'	dispersive difference
f''	anomalous difference
F_h	amplitude of a Fourier coefficient at harmonic number *h*
$\mathbf{F}_\mathrm{H}$	heavy atom contribution to the derivative structure factor
$\mathbf{F}_{h,k,l}$	structure factor of Miller indices *h*, *k*, and *l*
f_j	atomic scattering factor of atom *j*
$\mathbf{F}_\mathrm{P}$	native structure factor (only protein)
$\mathbf{F}_\mathrm{PH}$	derivative structure factor (protein and heavy atom)
h	harmonic number of a Fourier coefficient
h, k, l	Miller indices of a Bragg plane
I/σ	intensity over standard deviation, that is, signal-to-noise ratio
$I_{h,k,l}$	intensity of reflection of Miller indices *h*, *k*, and *l*
m	figure of merit for experimental phase
R_free	crystallographic residual for reflections not used in refinement but for cross-validation
R_meas	redundancy independent R_sym
R_sym	residual of the differences between observed intensities and the mean intensity
R-value	crystallographic residual
R_work	crystallographic residual for reflections used in refinement
$\mathbf{s}$	wavevector of the scattered X-ray beam
$\mathbf{s}_0$	wavevector of the incident X-ray beam
U_j	mean square deviation of atom *j* from its equilibrium position

References

1. V. Cherezov; D. M. Rosenbaum; M. A. Hanson; S. G. Rasmussen; F. S. Thian; T. S. Kobilka; H. J. Choi; P. Kuhn; W. I. Weis; B. K. Kobilka; R. C. Stevens, *Science* **2007**, *318*, 1258–1265.
2. T. A. Steitz; P. B. Moore, *Trends Biochem. Sci.* **2003**, *28*, 411–418.

3. I. G. Denisov; T. M. Makris; S. G. Sligar; I. Schlichting, *Chem. Rev.* **2005**, *105*, 2253–2277.
4. I. Schlichting; J. Berendzen; K. Chu; A. M. Stock; S. A. Maves; D. E. Benson; R. M. Sweet; D. Ringe; G. A. Petsko; S. G. Sligar, *Science* **2000**, *287*, 1615–1622.
5. C. Zubieta; X. Z. He; R. A. Dixon; J. P. Noel, *Nat. Struct. Biol.* **2001**, *8*, 271–279.
6. C. Giacovazzo; H. L. Monaco; G. Artioli; D. Viterbo; G. Ferraris; G. Gilli; G. Zanotti; M. Catti, *Fundamentals of Crystallography*, 2nd ed.; Oxford University Press: Oxford, 2002.
7. J. Kankare; T. Salminen; R. Lahti; B. S. Cooperman; A. A. Baykov; A. Goldman, *Acta Crystallogr. D* **1996**, *52*, 551–563.
8. M. Weselak; M. G. Patch; T. L. Selby; G. Knebel; R. C. Stevens, *Methods Enzymol.* **2003**, *368*, 45–77.
9. G. Rhodes, *Crystallography Made Crystal Clear*, 3rd ed.; Academic Press: Burlington, MA, 2006.
10. L. J. Smith; C. Redfield; R. A. G. Smith; C. M. Dobson; G. M. Clore; A. M. Gronenborn; M. R. Walter; T. L. Naganbushan; A. Wlodawer, *Nat. Struct. Biol.* **1994**, *1*, 301–310.
11. D. Blow, *Outline of Crystallography for Biologists*; Oxford University Press: Oxford, 2002.
12. R. Bracewell, *The Fourier Transform and Its Applications*; McGraw-Hill Book Company: New York, 1965; p 381.
13. K. Diederichs; P. A. Karplus, *Nat. Struct. Biol.* **1997**, *4*, 269–275.
14. M. G. Rossmann, Historical Background. In *Crystallography of Biological Macromolecules*; M. G. Rossmann, E. Arnold, Eds.; Kluwer Academic Publishers: Dordrecht, 2001; Vol. F, pp 4–9.
15. M. G. Rossmann, *Methods Enzymol.* **2003**, *368*, 11–21.
16. D. M. Blow; F. H. C. Crick, *Acta Crystallogr.* **1959**, *12*, 794–802.
17. W. A. Hendrickson; C. M. Ogata, *Methods Enzymol.* **1997**, *276*, 494–523.
18. R. J. Read, *Acta Crystallogr. A* **1986**, *42*, 140–149.
19. A. J. McCoy, *Acta Crystallogr. D* **2004**, *60*, 2169–2183.
20. D. E. Tronrud, *Acta Crystallogr. D* **2004**, *60*, 2156–2168.
21. A. T. Brunger; L. M. Rice, *Methods Enzymol.* **1997**, *277*, 243–269.
22. T. R. Schneider, What can we Learn from Anisotropic Temperature Factors? In *Proceedings of the CCP4 Study Weekend*, 1996; pp 133–144.
23. E. Oksanen; V. P. Jaakola; T. Tolonen; K. Valkonen; B. Akerstrom; N. Kalkkinen; V. Virtanen; A. Goldman, *Acta Crystallogr. D* **2006**, *62*, 1369–1374.
24. R. A. Engh; R. Huber, *Acta Crystallogr. A* **1991**, *47*, 392–400.
25. T. A. Jones; J. Y. Zou; S. W. Cowan; M. Kjeldgaard, *Acta Crystallogr. A* **1991**, *47*, 110–119.
26. V.-M. Leppänen; M. M. Bespalov; P. Runeberg-Roos; U. Puurand; A. Merits; M. Saarma; A. Goldman, *EMBO J.* **2004**, *23*, 1452–1462.
27. G. Evans; R. F. Pettifer, *J. Appl. Crystallogr.* **2001**, *34*, 82–86.
28. R. A. Laskowski; M. W. MacArthur; D. S. Moss; J. M. Thornton, *J. Appl. Crystallogr.* **1993**, *26*, 283–291.

Biographical Sketches



Esko Oksanen was born in Kauniainen, Finland in 1980. He obtained his M.Sc. in organic chemistry from the University of Helsinki in 2005 and is currently working on his Ph.D. in the group of Professor Goldman.

Professor Adrian Goldman was born in Durban, South Africa in 1958. He obtained his B.A. in Natural Sciences from Queens' College, Cambridge in 1980 and his Ph.D. in Thomas Steitz's laboratory at Yale University in 1985. After a short postdoctoral stay at Yale as a Markey Fellow, he became an assistant professor at Rutgers University (1987–1992), followed by appointments at the University of Turku (1992–1999) and at the University of Helsinki (1999–present). He is currently a research director at the Institute of Biotechnology and an adjunct professor in the Neuroscience Center. Adrian Goldman's interests focus on using X-ray crystallography to understand the chemistry behind simple enzyme systems and the biology behind more complex membrane-associated proteins and protein complexes.

9.04 Characterization by Circular Dichroism Spectroscopy

Nina Berova, George A. Ellestad, and Nobuyuki Harada, Columbia University, New York, NY, USA

9.04.1 Introduction

Chiral molecules are characterized by three-dimensional handedness and can exist in two enantiomeric forms of opposite absolute configuration (AC). Most natural products and biologically active compounds are chiral and their biological and molecular functions are closely related to their chirality, that is, AC and conformation. Furthermore, many drugs derived from natural products or of purely synthetic origin are currently used in enantiopure form. Therefore, the unambiguous determination of the AC of chiral compounds is critical for the studies of natural products and biomolecular systems.[1]

In this chapter, the characterization of natural products by circular dichroism (CD) spectroscopy in the UV/Vis region, called electronic CD (ECD), as well as a brief explanation of optical rotation (OR) $[\alpha]$ values, which are fundamental constants for chiral compounds, is reviewed. ECD is a very sensitive diagnostic tool for determining not only AC and conformation, but also for monitoring intermolecular interactions where chiral systems are involved.[2,3] There will be no discussion of vibrational CD (VCD) even though this method has made significant advances in the last decade in combination with theoretical *ab initio* calculations of VCD spectra.[4–6]

9.04.2 OR of Chiral Compounds

When linearly polarized light passes through a medium of chiral material, for example, a solution of chiral compound, the polarization plane of the transmitted light is rotated by an angle α compared to that of the incident light.[7,8] This phenomenon is called OR, and the angle α is named the rotation angle

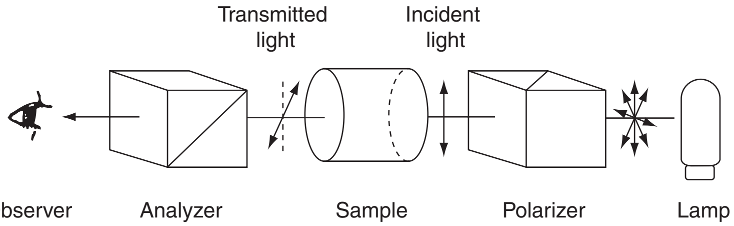


Figure 1 Concept of optical rotation. Redrawn from N. Harada, Optical Rotation, Optical Rotatory Dispersion, and Circular Dichroism. In *Handbook of Instrumental Analysis*, Part 2, 2nd ed.; Y. Izumi, M. Ogawa, S. Kato, J. Shiokawa, T. Shiba, Eds.; Kagakudojin: Kyoto, Japan, 2005; pp 119–133.

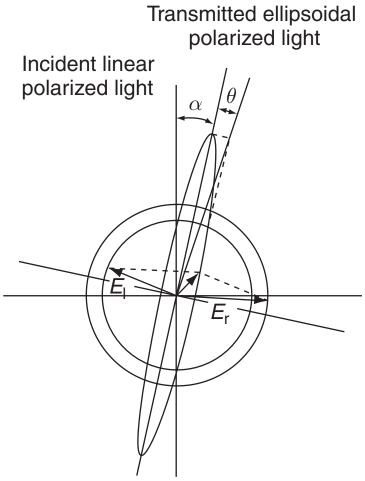

Transmitted ellipsoidal polarized light
Incident linear polarized light
α
θ
E_l
E_r

Figure 2 Definition of optical rotation and circular dichroism. When linearly polarized light passes through the medium of a chiral material, the transmitted light appears as an ellipsoidally polarized light. This figure shows this light as seen from the observer shown in **Figure 1**: α, rotation angle; θ, ellipticity angle; E_l and E_r, electric field vectors of left- and right-circularly polarized lights, respectively. Redrawn from N. Harada; K. Nakanishi, *Circular Dichroic Spectroscopy – Exciton Coupling in Organic Stereochemistry*; University Science Books: Mill Valley, CA, and Oxford University Press: Oxford, 1983.

(**Figures 1 and 2**). If the polarization plane is rotated clockwise as seen by an observer in **Figure 1**, the chiral material is defined as dextrorotatory and is given a plus sign, while if rotated counterclockwise, it is defined as levorotatory and given a minus sign. The instrument for OR measurements is called a polarimeter.

Linearly polarized light is composed of left- and right-circularly polarized light (l-CPL and r-CPL) with equal intensity, and the electromagnetic fields of these circularly polarized lights are mirror images of each other. When these l-CPL and r-CPL pass through a medium of chiral material the interaction of the chiral material with l-CPL is not equal to that with r-CPL, thus generating a difference in light velocity and refractive index, which leads to OR. Another phenomenon is that the absorption intensity of l-CPL with a chiral material is unequal to that of r-CPL, and this phenomenon generates CD.

As a parameter showing OR of a chiral compound, specific rotation $[\alpha]_\lambda{}^t$ is used and defined as follows:

$$[\alpha]_\lambda^t = \frac{100\alpha}{lc}(\text{solution}) \tag{1}$$

$$[\alpha]_\lambda^t = \frac{\alpha}{l\rho}(\text{liquid}) \tag{2}$$

where t is temperature (°C), λ wavelength of light (nm), α observed rotation angle (degree), l cell length (dm = 10 cm), c concentration (sample weight, g/100 cm^3 solution), and ρ the density of liquid (g cm^{-3}).

9.04.2.1 Measurement of $[\alpha]_D{}^t$

Wavelength. OR measured as a function of the wavelength of light is known as optical rotatory dispersion (ORD). However, it is general to use the D-line (589.3 nm) of the sodium lamp, and therefore, specific rotation is expressed as $[\alpha]_D{}^t$. When light of shorter wavelength is used, $[\alpha]_\lambda{}^t$ generally becomes larger. Therefore, when $[\alpha]_D{}^t$ value is small, the use of a mercury lamp light (578.0, 546.1, 435.8 nm) may give larger values and be useful. However, when comparing $[\alpha]_\lambda{}^t$ values with those reported by other researchers, difference of wavelength is inconvenient. Therefore, the $[\alpha]_D{}^t$ value at the sodium D-line is used as a common standard.

Cell. For solution samples, glass or quartz strainless cylindrical cells made specially for OR measurements are used.

cell A: cell length 10 cm, inner diameter 1 cm, volume approximately 8 cm^3, luminous flux 0.8 cm
cell B: cell length 10 cm, inner diameter 0.35 cm, volume approximately 1 cm^3, luminous flux 0.3 cm

For liquid samples, rectangular strain-free quartz cells (1.0–0.1 cm) made for CD measurements are used. In this case, the density ρ of the liquid is required, and therefore ρ has to be measured or taken from the literature.

Solvent. The following solvents are usually used: water, MeOH, EtOH, 1,4-dioxane, $CHCl_3$, benzene, cyclohexane, hexane. Because most organic compounds are easily soluble in chloroform it is generally the solvent of choice. However, it should be noted that trace acidic impurities and/or ethanol added as a stabilizer may react with the solute and affect the $[\alpha]_\lambda{}^t$ value.

Description of Data. Using Equation (1) or (2), $[\alpha]_D{}^t$ value is calculated and expressed as follows. For example,

$$[\alpha]_D{}^{20} \ +54.2\,(c\,0.323,\mathrm{EtOH})$$

where it is a general rule that concentration c is expressed in the unit of (g solute/100 cm^3 solution).

Standard Sample for Polarimeter. To calibrate the polarimeter, the standard sample and data are reported as follows:

sucrose or saccharose: solution (1.00 g sucrose/100 cm^3 solution in distilled water), 10.0 cm cell, 20 °C, Na-D line, observed rotation angle $\alpha = +0.665°$.

It should be noted that solutions of sucrose do not undergo mutarotation unlike other sugars. It is thus suggested to check the polarimeter using the standard solutions of sucrose.

9.04.3 Circular Dichroism Spectra

As described above, when linearly polarized light passes through the medium of a chiral material, the polarization plane of the transmitted light is rotated by an angle α compared to that of the incident light. At the same time the transmitted light is no longer linearly polarized but is ellipsoidally polarized (**Figure 2**). This phenomenon is caused by the difference in the absorption intensity for left- and right-circularly polarized light and is called CD. The original linearly polarized light is composed of left- and right-circularly polarized beams of equal intensity. However, when these two oppositely polarized beams pass through a solution of a chiral compound their absorption coefficients change and become unequal. The angle θ of the ellipsoidally polarized transmitted light is defined as the ellipticity angle.[3,8]

$$\tan\theta = \frac{\text{short axis of ellipsoid}}{\text{long axis of ellipsoid}} \tag{3}$$

CD is expressed by molar ellipticity $[\theta]$ or molar CD $\Delta\varepsilon$.

$$[\theta] = \frac{\theta M}{lc} \tag{4}$$

where θ is the observed ellipticity angle (degree = °), M the molecular or formula weight, l the cell length (dm = 10 cm), and c the concentration (g solute/100 cm^3 solution). In the case of organic compounds, molar CD $\Delta\varepsilon$ ($dm^3\,mol^{-1}\,cm^{-1}$) is preferentially used and $\Delta\varepsilon$ is the difference between the absorption coefficient ε_l for l-CPL and coefficient ε_r for r-CPL.

$$\Delta\varepsilon = \varepsilon_l - \varepsilon_r \tag{5}$$

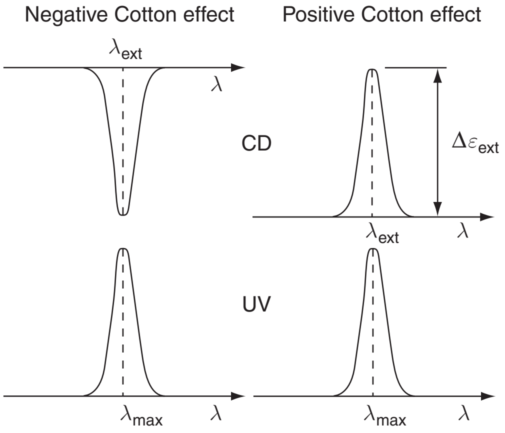


Figure 3 The general pattern of CD and UV spectra.

Molar ellipticity $[\theta]$ is correlated to molar CD $\Delta\varepsilon$ as follows:

$$[\theta] = 3300\Delta\varepsilon \tag{6}$$

In practice, the next equation is useful for obtaining molar CD $\Delta\varepsilon$.

$$\frac{\theta}{33} = \Delta A = \Delta\varepsilon \times c' \times l' \tag{7}$$

where ΔA is CD absorbance, c' molar concentration (mol solute/dm^3 solution), and l' cell length (cm).

The intensity scale (or sensitivity) of a CD spectropolarimeter is generally expressed in millidegrees (m°)/cm, and therefore molar CD $\Delta\varepsilon$ is calculated as follows:

$$\frac{\text{sensitivity}(\text{m}^\circ/\text{cm}) \times \text{CD signal}(\text{cm})}{33\,000} = \Delta A = \Delta\varepsilon \times c' \times l' \tag{8}$$

A curve plotting $\Delta\varepsilon$ or $[\theta]$ against wavelength λ (nm) is called the CD spectrum, which generally shows a positive or negative band at the region of the corresponding UV absorption band (**Figure 3**). These bands are called Cotton effects. A positive CD band is named a positive Cotton effect while a negative CD band is called a negative Cotton effect.

9.04.3.1 Measurement of CD Spectra

CD Spectropolarimeter. To protect the optical system from ozone, which may be generated by UV light from a xenon lamp, dry nitrogen gas (2–5 dm^{3}min^{-1}) is flushed through the CD spectropolarimeter. By flushing a large amount of nitrogen gas (50–70 dm^{3}min^{-1}), it may be possible to measure the CD spectrum down to approximately 170 nm (vacuum–UV region).

Cell. For solution samples, strain-free and high-quality quartz cells specially made for CD measurements are used. Cylindrical cells (cell length 100–0.1 mm), rectangular cells (cell length 20–1 mm), and U-shape cells (cell length 0.5–0.1 mm) are commercially available.

Sampling. In CD spectroscopy, $\Delta\varepsilon$ or $[\theta]$ value provides important information of chirality and, therefore, samples have to be weighed accurately using an electro-microbalance to the order of 10^{-6} g. Usually it is suggested to weigh a sample of more than 1.0 mg.

Solvent. The following solvents that are transparent in the UV–Vis region are usually used: water, MeOH, EtOH, acetonitrile, cyclohexane, and hexane. When a sample is weakly soluble in EtOH, a mixed solvent is used; the sample is dissolved in a small amount of 1,4-dioxane and then diluted with EtOH. Spectrograde solvents should be used.

Scanning of CD. In the CD measurement, the concentration of solution and cell length should be adjusted so that the absorbance A at the UV maximum region should be smaller than 1.5. Scanning starts from long wavelength toward short wavelength. If the voltage of the photomultiplier is scaled out or the CD curve becomes too noisy, measurement is repeated using a diluted solution or a thinner cell. For measurements around 230–200 nm, thinner cells (1.0–0.1 mm) are desirable since the absorption of the solvent in this spectral range interferes with the CD measurement. In most cases, measurement around 200 nm is possible with the proper selection of cell and solute concentration.

Blank Measurement. After the sample has been measured, a blank measurement must be carried out under the same conditions and the same solvent. The sample curve minus a blank curve gives the net CD curve.

Measurement of UV–Vis Spectra. For the interpretation of CD data, the information of UV–Vis spectra is necessary, and therefore it is strongly suggested to measure the UV–Vis spectra at the same time using the solutions employed for CD.

Description of CD Data. Using Equation (7) or (8), a $\Delta\varepsilon$ value is calculated and plotted against wavelength λ. For the CD Cotton effects, the wavelength at an extremum (λ_{ext}) and the intensity ($\Delta\varepsilon$) are described as follows. For example,

$$\text{CD (EtOH)}\ \ \lambda_{\text{ext}}\, 320.5\,\text{nm}(\Delta\varepsilon - 63.1), 308.8(0.0), 295.5(+39.7)$$

If necessary, the zero crossing point may also be included.

Standard Sample for CD Spectropolarimeter. To calibrate the CD spectropolarimeter, the following standard samples, conditions, and CD data are suggested:

androsterone: solution (c 0.0500 g/100 cm^3 solution in 1,4-dioxane), λ_{ext} 304 nm ($\Delta\varepsilon$ +3.39)

(−)-pantolactone: solution (c 0.0150 g/100 cm^3 solution in water), λ_{ext} 219 nm ($\Delta\varepsilon$ −5.00)
(+)-10-camphorsulfonic acid ammonium salt: solution (c 0.0600 g/100 cm^3 solution in water), λ_{ext} 290.5 nm ($\Delta\varepsilon$ +2.40).

9.04.3.2 CD Spectra and Rotational Strength

The rotational strength R, a parameter representing the sign and strength of CD Cotton effect, is formulated in Equation (9) and is experimentally obtainable from the observed CD spectra.[8–10]

$$R = 2.296 \times 10^{-39} \int \frac{\Delta\varepsilon(\sigma)}{\sigma}\, d\sigma \quad \text{cgs unit} \tag{9}$$

where σ is wavenumber (cm^{-1}).

The rotational strength is theoretically expressed by Equation (10), which was derived by Rosenfeld in the early days of the development of quantum mechanical theory.

$$R = \text{Im}\{<0|\boldsymbol{\mu}|a> \bullet <a|\boldsymbol{M}|0>\} \tag{10}$$

where Im denotes the imaginary part of the terms in brackets { }, < > denotes the integration over configurational space, and $\boldsymbol{\mu}$ and $\boldsymbol{M}$ are operators of electric and magnetic moment vectors, respectively. The dot • stands for the scalar product of two vectors, 0 and a are wavefunctions of ground and excited states, respectively. The rotational strength R is thus equal to the imaginary part of the scalar product of electric and magnetic transition moments. If the electric and magnetic moment vectors, $<0|\boldsymbol{\mu}|a>$ and $<a|\mathbf{M}|0>$, are parallel to each other, the rotational strength R is positive leading to a positive Cotton effect. On the other hand, if they are antiparallel, the rotational strength R is negative and a negative CD Cotton effect is observed.

The electric and magnetic moments, $<0|\boldsymbol{\mu}|a>$ and $<a|\mathbf{M}|0>$, are based on the linear and angular momentum of electrons involved in the transition. The angular momentum corresponds to the rotational motion of an electron, while the linear momentum corresponds to the linear motion of the electron. If the electric and magnetic moment vectors, $<0|\boldsymbol{\mu}|a>$ and $<a|\mathbf{M}|0>$, are parallel to each other in one enantiomer, they should be antiparallel in the other enantiomer. Therefore, the rotational strength R, OR $[\alpha]_{\lambda}{}^{t}$, and CD spectra of enantiomers are opposite in sign but of equal intensity. The problem of how to determine the ACs of

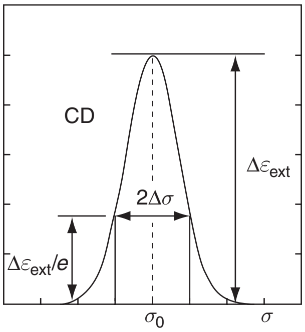


Figure 4 CD Cotton effect curve approximated by the Gaussian distribution, where $2\Delta\sigma$ is the bandwidth at $1/e$ peak height, $e = 2.71828$, and $1/e = 0.36788$.

chiral compounds in a theoretical manner is to calculate the rotational strength R by the use of Equation (10). For this purpose, several theoretical methods have been developed as described in Section 9.04.3.5.

If the CD Cotton effect is approximated by a Gaussian distribution, the CD curve is formulated as

$$\Delta\varepsilon(\sigma) = \Delta\varepsilon_{\text{ext}} \exp\left\{-\left(\frac{\sigma - \sigma_{\hat{0}}}{\Delta\sigma}\right)^2\right\} \tag{11}$$

where $\Delta\varepsilon_{\text{ext}}$ is the extremum value of the Cotton effect, $\sigma_{\hat{0}}$ the central wavenumber of the Cotton effect, and $\Delta\sigma$ half the bandwidth at $1/e$ peak height of the Gaussian curve (**Figure 4**).

Substitution of Equation (11) into Equation (9) and integration gives

$$R = 2.296 \times 10^{-39}\sqrt{\pi}\Delta\varepsilon_{\text{ext}}\frac{\Delta\sigma}{\sigma_{\hat{0}}} \tag{12}$$

From Equations (11) and (12), the calculated CD curve is formulated as

$$\Delta\varepsilon(\sigma) = \left(\frac{\sigma_{\hat{0}}}{2.296 \times 10^{-39}\sqrt{\pi}\Delta\sigma}\right) R \exp\left\{-\left(\frac{\sigma - \sigma_{\hat{0}}}{\Delta\sigma}\right)^2\right\} \tag{13}$$

where $\Delta\sigma$ can be evaluated from the corresponding observed UV–Vis spectra. Thus, if the rotational strength R is theoretically calculated by Equation (10), the CD spectral curve is reproducible by theoretical calculation.

9.04.3.3 UV–Vis Spectra and Dipole Strength

In analogy with the case of rotational strength, the dipole strength D representing the transition probability of the UV–Vis absorption band is estimated from the observed spectra, as follows:

$$D = 9.184 \times 10^{-39} \int \frac{\varepsilon(\sigma)}{\sigma} \mathrm{d}\sigma \quad \text{cgs unit} \tag{14}$$

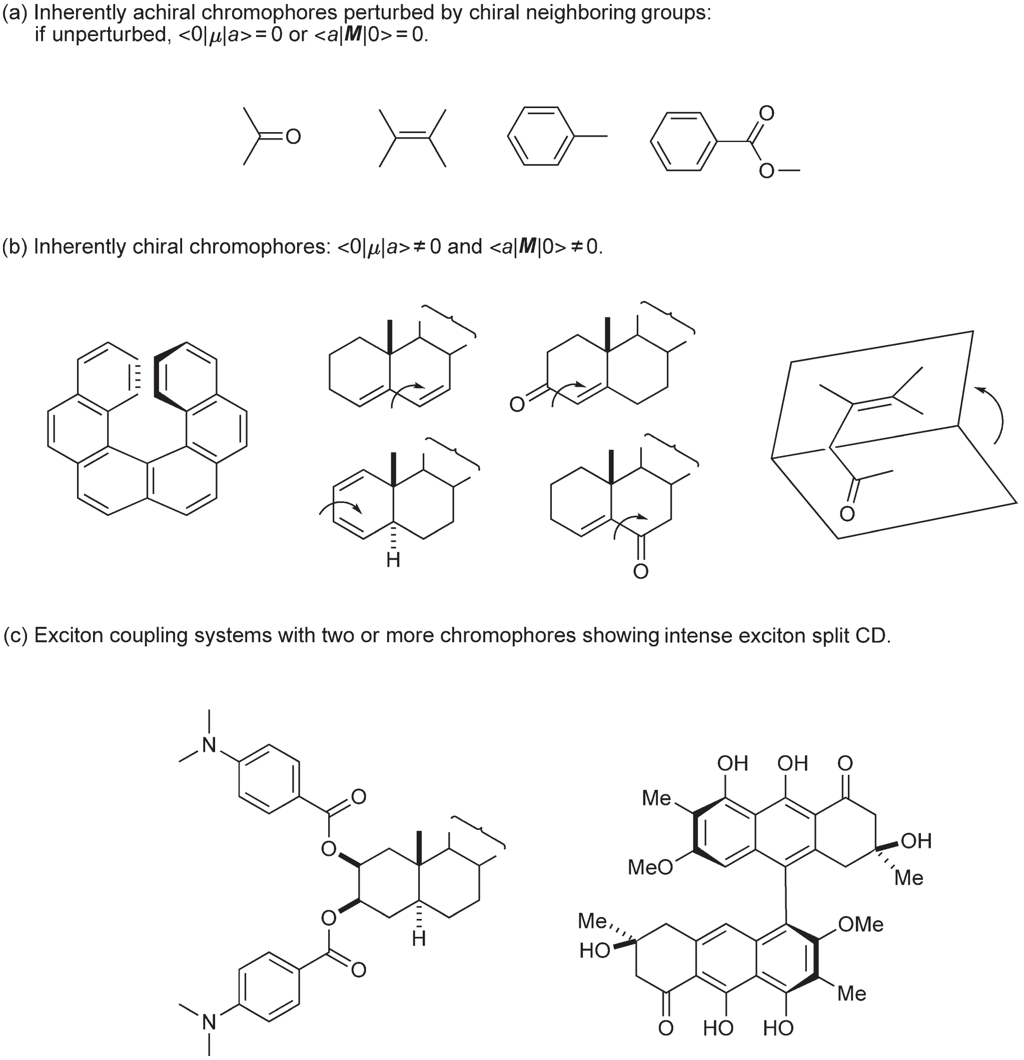
The dipole strength is theoretically expressed by Equation (15).

$$D = \{<0|\boldsymbol{\mu}|a>\}^2 \tag{15}$$

Thus dipole strength D is equal to the square of the electric transition moment $<0|\boldsymbol{\mu}|a>$.

9.04.3.4 CD Chromophores

The CD Cotton effects are classified into three groups depending on the chromophore and generation mechanism (**Figure 5**).



(a) Inherently achiral chromophores perturbed by chiral neighboring groups: if unperturbed, $<0|\mu|a> = 0$ or $<a|\boldsymbol{M}|0> = 0$.

(b) Inherently chiral chromophores: $<0|\mu|a> \neq 0$ and $<a|\boldsymbol{M}|0> \neq 0$.

(c) Exciton coupling systems with two or more chromophores showing intense exciton split CD.

Figure 5 Classification of chromophores and CD Cotton effects.

1. *Achiral chromophores*[3]: Chromophore itself is symmetrical and achiral and, therefore, the electric transition moment $<0|\boldsymbol{\mu}|a>$ or magnetic transition moment $<a|\mathbf{M}|0>$ is zero. For example, the $n-\pi^*$ transition of ketones is electrically forbidden but magnetically allowed. In chiral ketones, the carbonyl chromophore is perturbed by chiral neighboring groups, and this perturbation generates a small electric transition moment $<0|\boldsymbol{\mu}|a>$. Therefore, the $n-\pi^*$ transition becomes CD active, but in general, the CD Cotton effect at the $n-\pi^*$ transition is weak. On the other hand, the $\pi-\pi^*$ transition of olefins and aromatic compounds is electrically allowed but magnetically forbidden. In chiral compounds, chromophores are similarly perturbed by chiral neighboring groups. So, the $\pi-\pi^*$ CD Cotton effect of these compounds is again weak.
2. *Inherently chiral chromophores*[3]: Chromophore itself is twisted and chiral and, therefore, the $\pi-\pi^*$ and/or $n-\pi^*$ transitions become CD active, because both electric transition moment $<0|\boldsymbol{\mu}|a>$ and magnetic transition moment $<a|\mathbf{M}|0>$ take nonzero values. The CD intensity of this group is stronger than that of group (a).
3. *Exciton-coupled CD*[8,11]: The systems have two or more chromophores, which are placed in chiral positions to each other. Each chromophore exhibits intense $\pi-\pi^*$ transition generating a large electric transition moment $<0|\boldsymbol{\mu}|a>$, and at the same time, these moments make a large magnetic transition moment

$<a|\mathbf{M}|0>$, because of the mutual chiral displacement. Therefore, the CD Cotton effects of this group are much stronger than those of groups (a) and (b). So, when determining the ACs by CD spectroscopy, it is advisable to take advantage of the large Cotton effects of this group, even in the cases of theoretical calculations.

9.04.3.5 Theoretical Calculation of CD Spectra

As described above, if rotational strength R defined by Equation (10) can be calculated by quantum mechanical theory, CD spectra can be reproduced by Equation (13) leading to the determination of ACs. For the calculation of rotational strength R, there are several theoretical methods as follows:

1. *CD exciton chirality method*[8,11]: the most simple and reliable method applicable to a variety of natural products, because the exciton-coupled CD is based on the coupled oscillator theory and the mechanism of this method has already been established as will be briefly explained in the following sections. Therefore, numerical calculations using a computer are not necessary.
2. *De Voe calculation*[12,13]: a simple method based on the coupled oscillator theory, which is applicable to more complex chiral molecules composed of two or more groups. This method needs numerical calculations using a computer. Some examples are listed in the section of applications.
3. *π-Electron SCF-CI-DV MO (Self-Consistent Field-Configuration Interaction-Dipole Velocity-Molecular Orbital) method*[8,14]: a molecular orbital (MO) method with π-electron approximation, which is applicable to chiral molecules having twisted π-electron systems. As this method treats only π-electrons, computation time is shorter than the cases treating all electrons. Some examples are shown in the application sections.
4. *Ab initio MO calculations*[15]: Recent years have witnessed a great advancement in the first-principle calculations of chiroptical properties. The development of *ab initio* methodologies, which include Hartree–Fock, density functional theory (DFT), as well as high-level correlation methods, such as coupled cluster theory, have enabled theoretical simulations of CD, OR, and other chiroptical properties. Since these methods treat all electrons including σ-electrons, a large amount of computation is necessary, especially for calculations on conformationally flexible molecules. Some examples are shown in the application sections. **Figure 6** shows the general scheme for determining the AC of a chiral compound based on theoretical calculations and experimental CD measurements.

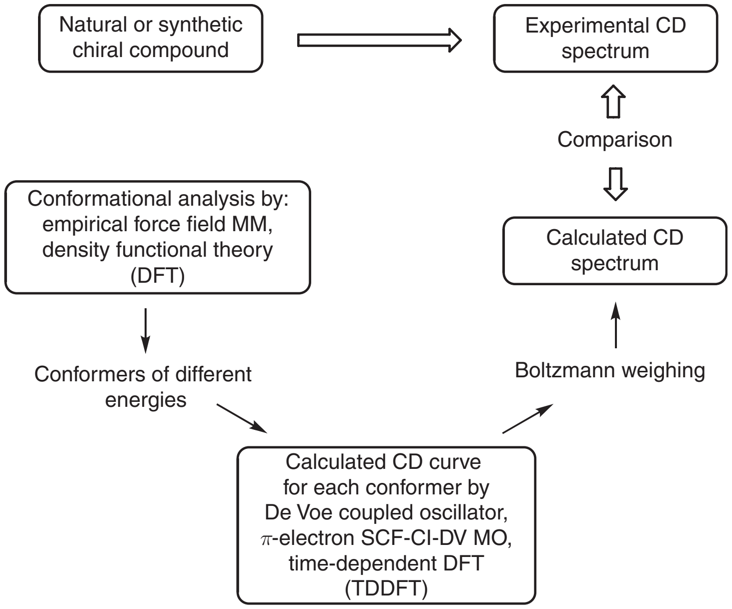

Natural or synthetic chiral compound
Experimental CD spectrum
Comparison
Conformational analysis by: empirical force field MM, density functional theory (DFT)
Calculated CD spectrum
Conformers of different energies
Boltzmann weighing
Calculated CD curve for each conformer by De Voe coupled oscillator, π-electron SCF-CI-DV MO, time-dependent DFT (TDDFT)

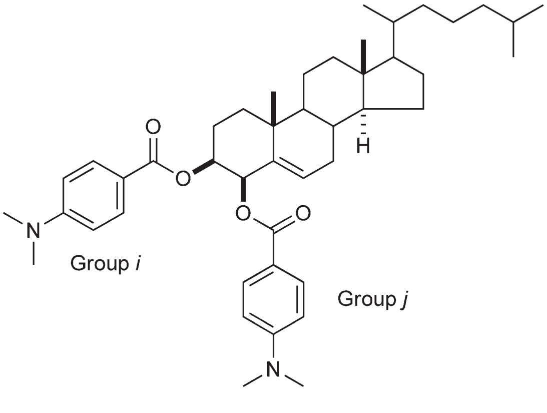
Figure 6 The scheme for determining the AC of a chiral compound by theoretical calculation of CD spectrum and comparison with the experimental CD spectrum.

9.04.4 CD Exciton Chirality Method

9.04.4.1 Basic Principles

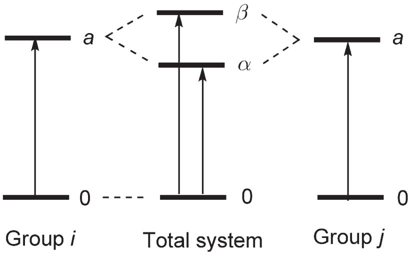
The CD exciton chirality method has been successfully applied to a variety of natural products to determine their ACs. This method enables one to deduce the AC of a chiral compound without any reference compound, and therefore, it is established as a nonempirical method. The principles of the CD exciton chirality method are explained using the steroidal bis(*p*-dimethylaminobenzoate) shown below as a model compound, where the nonempirical nature of this method is easily proved.[8,11,16]

As exemplified with cholest-5-ene-3β,4β-diol bis(*p*-dimethylaminobenzoate) **1** in **Figure 7**, when two identical chromophores (*i* and *j*), which exhibit intense UV absorption of their $\pi-\pi^*$ transition (ground state $0 \rightarrow$ excited state *a*), exist in a molecule, these two chromophores interact with each other and the excited state splits into two energy levels (α and β states).[8] The ground state (0) remains unsplit. This phenomenon is called exciton coupling or exciton interaction. Thus there are two electronic transitions, from ground 0 to excited states α and β, that is, transitions $0 \rightarrow \alpha$ and $0 \rightarrow \beta$. The wavefunction, energy, dipole strength, and rotational strength for the α-state and β-state are formulated as shown in **Figure 7**, where V_{ij} is defined as the interaction energy between two electric transition moments $\boldsymbol{\mu}_{i0a}$ and $\boldsymbol{\mu}_{j0a}$. If V_{ij} is positive, the α-state corresponds to the transition at longer wavelength, while the β-state corresponds to the transition at shorter wavelength. As shown in **Figure 7**, the rotational strength R^{α} of the α-state is opposite in sign to that of the β-state, R^{β}, but their absolute values are equal to each other. It should be noted that the sign and magnitude of R^{α} and R^{β} are governed by the triple product $\boldsymbol{R}_{ij} \bullet (\boldsymbol{\mu}_{j0a} \times \boldsymbol{\mu}_{j0a})$.[8]



Cholest-5-ene-3β,4β-diol bis(*p*-dimethylaminobenzoate) **1**

The mechanism of exciton CD is simple, and is easily solved by the exciton theory as formulated below:



By the exciton interaction between chromophores *i* and *j*, the excited state splits into two energy levels; the figure shows the case of $V_{ij} > 0$.

α-state:

Wave function, $\psi_a^{\alpha} = (1/\sqrt{2})(\phi_{ia}\phi_{j0} - \phi_{i0}\phi_{ja})$

Energy, $E^{\alpha} = E_a - V_{ij}$

Dipole strength, $D^{\alpha} = (1/2)(\mu_{i0a} - \mu_{j0a})^2$

Rotational strength, $R^{\alpha} = +(1/2)\pi\sigma_0 \boldsymbol{R}_{ij} \bullet (\mu_{i0a} \times \mu_{j0a})$

β-state:

Wave function, $\psi_a^{\beta} = (1/\sqrt{2})(\phi_{ia}\phi_{j0} + \phi_{i0}\phi_{ja})$

Energy, $E^{\beta} = E_a + V_{ij}$

Dipole strength, $D^{\beta} = (1/2)(\mu_{i0a} + \mu_{j0a})^2$

Rotational strength, $R^{\beta} = -(1/2)\pi\sigma_0 \boldsymbol{R}_{ij} \bullet (\mu_{i0a} \times \mu_{j0a})$

Interaction energy, $V_{ij} = R_{ij}^{-3}\{\mu_{i0a}\mu_{j0a} - 3R_{ij}^{-2}(\mu_{i0a}\boldsymbol{R}_{ij})(\mu_{j0a}\boldsymbol{R}_{ij})\}$

If $V_{ij} > 0$, the α-state is lower in energy than the β-state.

α-state, longer wavelength side => first Cotton effect.

β-state, shorter wavelength side => second Cotton effect.

Figure 7 Theoretical summary of the CD exciton chirality method.

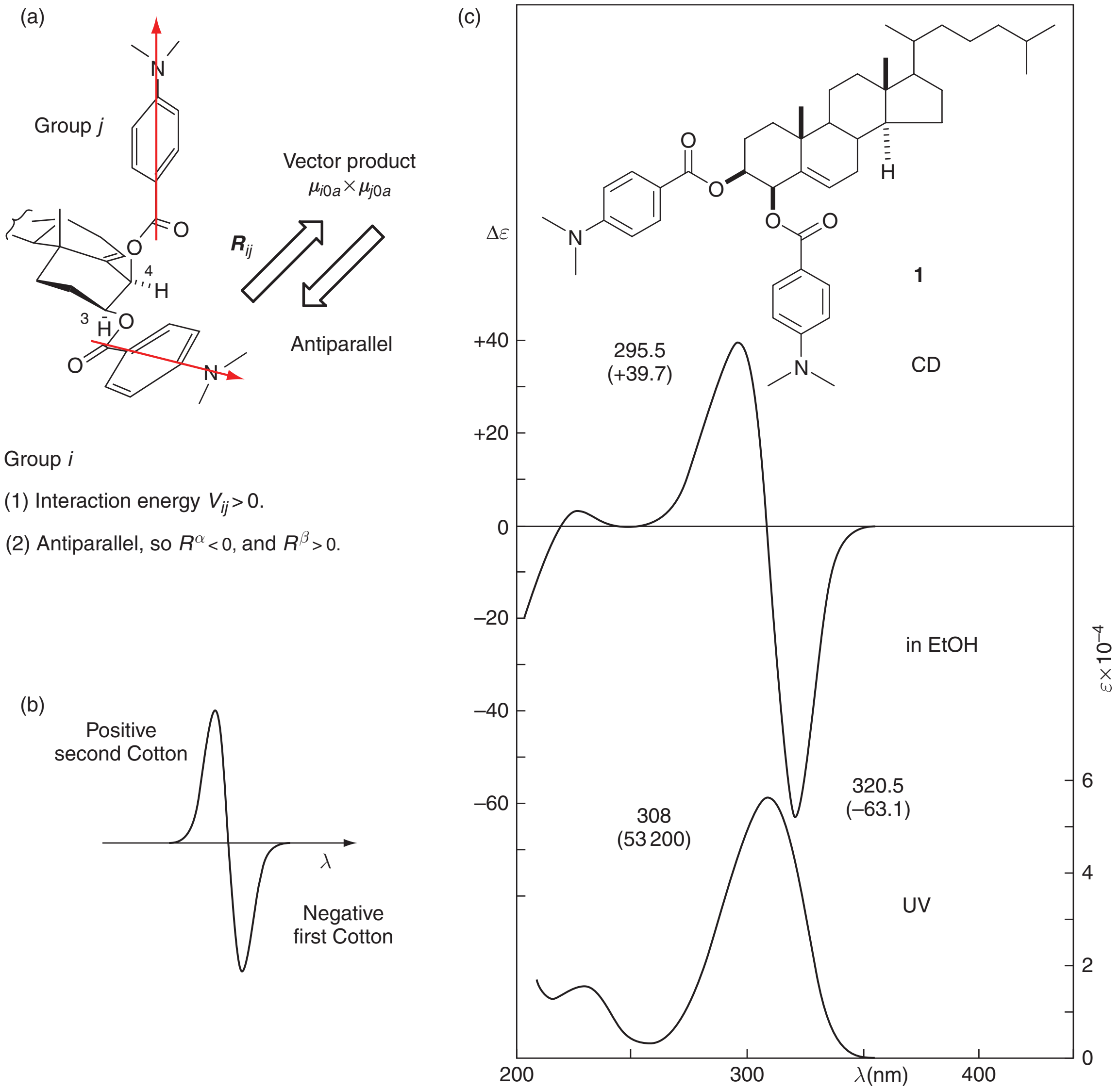


Figure 8 Application of the CD exciton chirality method to cholest-5-ene-3β,4β-diol bis(p-dimethylaminobenzoate) **1**: CD and UV spectra in EtOH. Redrawn from N. Harada; K. Nakanishi, *Circular Dichroic Spectroscopy – Exciton Coupling in Organic Stereochemistry*; University Science Books: Mill Valley, CA, and Oxford University Press: Oxford, 1983.

These equations are next applied to steroidal dibenzoate 1 in Figure 8.

For the two electric transition moments $\boldsymbol{\mu}_{i0a}$ and $\boldsymbol{\mu}_{j0a}$ in the benzoate chromophores (Figure 8), the interaction energy V_{ij} becomes positive, and therefore the α-state is lower in energy than the β-state. Two vectors $\boldsymbol{\mu}_{i0a}$ and $\boldsymbol{\mu}_{j0a}$ constitute a counterclockwise screw, and so the resultant vector $\boldsymbol{\mu}_{j0a} \times \boldsymbol{\mu}_{j0a}$ is antiparallel to the distance vector $\boldsymbol{R}_{ij}$. Therefore the triple product $\boldsymbol{R}_{ij}\bullet (\boldsymbol{\mu}_{j0a} \times \boldsymbol{\mu}_{j0a})$ becomes negative, and so R^{α} is negative while R^{β} is positive. This result leads to the CD spectral pattern as shown in Figure 8(b), where the Cotton effect at longer wavelength (named first Cotton effect) is negative and that at shorter wavelength (second Cotton effect) is positive. These exciton-coupled CD Cotton effects with opposite signs each other are called 'bisignate Cotton effects'. This is the theoretical deduction of exciton CD Cotton effects reflecting the AC of the two electric transition moments, that is, two chromophores.[8]

Figure 8(c) shows the UV and CD spectra of the actual compound, with cholest-5-ene-3β,4β-diol bis(p-dimethylaminobenzoate) 1, where the UV shows an intense π–π^* absorption band ($\lambda_{\max}$ 308 nm,

ε 53 200), which is polarized along the long axis of the chromophore. The CD spectrum shows negative first and positive second Cotton effects in agreement with the theoretical conclusion: first Cotton effect, λ_{ext} 320.5 nm, $\Delta\varepsilon$ −63.1 and second one λ_{ext} 295.5 nm, $\Delta\varepsilon$ +39.7. The amplitude of the exciton CD is defined as $A = \Delta\varepsilon_1 - \Delta\varepsilon_2$, where $\Delta\varepsilon_1$ and $\Delta\varepsilon_2$ are $\Delta\varepsilon$ values of first and second Cotton effects, respectively. In the case of dibenzoate **1**, $A = -102.8$. From these results, one can easily determine the AC of the original glycol.

In **Figure 9**, the UV spectrum of cholest-5-ene-3β,4β-diol bis(*p*-bromobenzoate) **2** shows the long-axis-polarized π–π^* transition at 244 nm, while the CD spectrum shows negative first and positive second Cotton effects ($A = -51.6$) in agreement with the negative screw sense between the two long axes. This counterclockwise screw sense is directly observed by the X-ray crystallographic stereoview shown in **Figure 9**, where 3β-equatorial benzoate chromophore is placed in front, while 4β-axial benzoate in the rear.

From the above-mentioned results, the exciton chirality governing the sign and intensity of Cotton effects is defined as shown in **Table 1**.[8]

The qualitative definition of exciton chirality is very simple: (1) if two transition moments constitute a clockwise screw sense, CD shows positive first and negative second Cotton effects. On the other hand, (2) if they describe a counterclockwise screw sense, negative first and positive second Cotton effects are observed. In most cases, intense exciton-coupled CD Cotton effects are observed at the long-axis-polarized transition, and therefore the above results are rephrased as follows:[8]

1. If the long axes of two interacting chromophores constitute a clockwise screw sense, the CD shows a positive first Cotton effect at a longer wavelength and a negative second Cotton effect at a shorter wavelength (**Table 1** and **Figure 10**).
2. If they make a counterclockwise screw sense, a negative first Cotton effect at a longer wavelength and a positive second Cotton effect at a shorter wavelength are observed.

In general, the CD zero crossing point corresponds to $\lambda_{\max}$ of UV band.

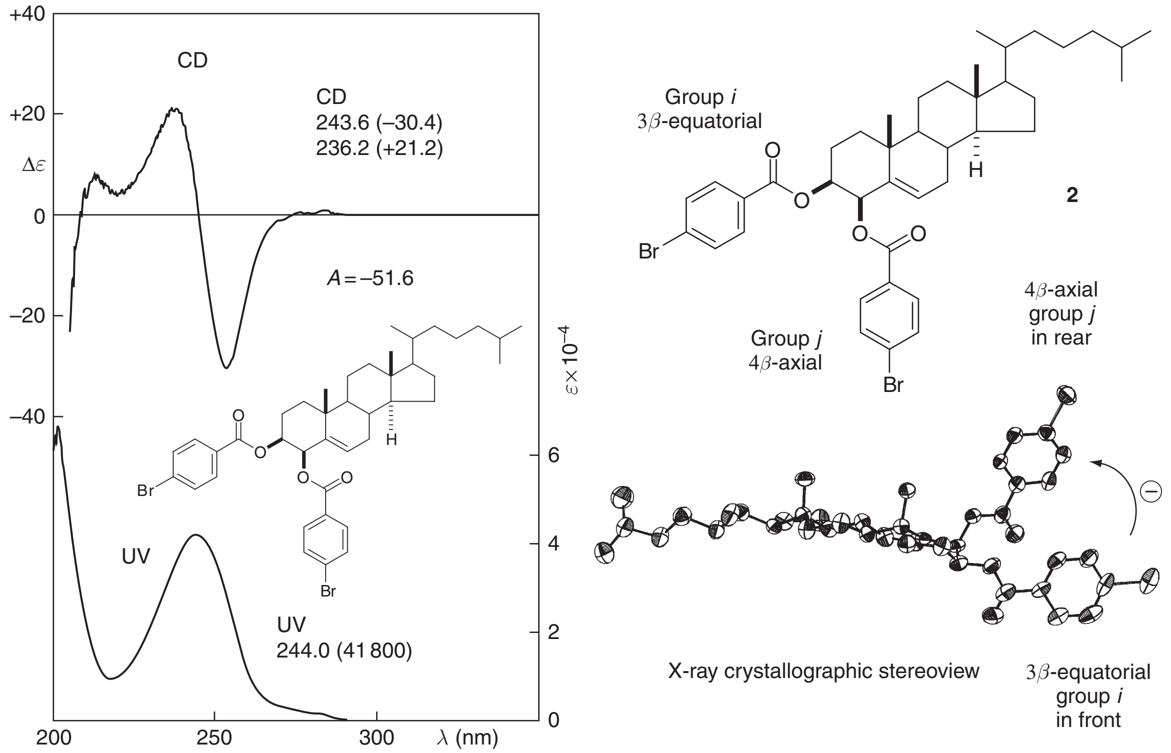
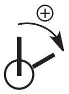
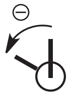


Figure 9 CD and UV spectra of cholest-5-ene-3β,4β-diol bis(*p*-bromobenzoate) **2** in 10% 1,4-dioxane/EtOH and X-ray crystallographic stereoview (X-ray, N. Harada, unpublished data).

Table 1 Definition of exciton chirality

Qualitative definition		*Quantitative definition*	*Cotton effects*
Positive exciton chirality		$\boldsymbol{R}_{ij} \bullet (\mu_{i0a} \times \mu_{j0a})\ V_{ij} > 0$	Positive first (at longer wavelength) and negative second (at shorter wavelength) Cotton effects
Negative exciton chirality		$\boldsymbol{R}_{ij} \bullet (\mu_{i0a} \times \mu_{j0a})\ V_{ij} < 0$	Negative first (at longer wavelength) and positive second (at shorter wavelength) Cotton effects

Redrawn from N. Harada; K. Nakanishi, *Circular Dichroic Spectroscopy – Exciton Coupling in Organic Stereochemistry*; University Science Books: Mill Valley, CA, and Oxford University Press: Oxford, 1983.

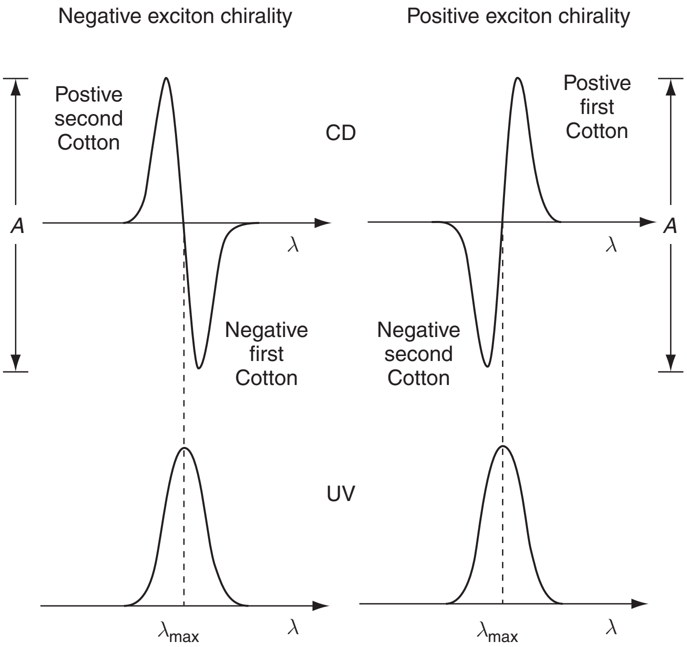


Figure 10 Typical pattern of exciton-coupled CD Cotton effects and UV absorption band.

From the quantitative definition of exciton chirality, some important features are derived.

1. The intensity of the exciton CD (A-value) is inversely proportional to the square of the interchromophoric distance R_{ij} provided the remaining angular part is the same.[8]

$$A \propto R_{ij}^{-2} \tag{16}$$

2. The A-value of exciton split CD is the function of the dihedral angle between two transition moments. In the case of vicinal glycol dibenzoates, the sign of the exciton split Cotton effects remains unchanged from 0 to 180°. Therefore, the qualitative definition shown in **Table 1** is applicable to a dibenzoate with the dihedral angle of more than 90°. The maximum A-value is around 70°.[8]
3. In the case of chiral 1,1′-binaphthyl and related compounds, it was theoretically calculated that the sign of the exciton CD Cotton effects changes from plus to minus, or vice versa, when the dihedral angle between the two naphthalene planes is changed from 0 to 180°; the zero point being around 110°.[17–19] Therefore, when the CD

exciton method is applied to these compounds, the information of the dihedral angle is necessary. However, the X-ray data of some compounds of this series revealed that the dihedral angle is distributed in the range of 68–92°.[20]

4. The A-value is proportional to the square of absorption coefficient ε of the chromophore. Therefore, it is advisable to use chromophores undergoing intense $\pi-\pi^*$ transition.
5. For the exciton coupling systems with three or more chromophores, it was found that the so-called additivity rule holds. For example, for a trimer,

$$A(\text{total}) = A(1,2) + A(1,3) + A(2,3) \tag{17}$$

where $A(1,2)$, $A(1,3)$, and $A(2,3)$ are the A-values of component pairs.

6. The Cotton effects of α- and β-states have identical rotational strength of opposite signs. Namely, the two split Cotton effects are conservative, and satisfy the sum rule.

$$\sum R^k = 0 \tag{18}$$

7. Since rotational strength R is a physically observable quantity, rotational strength should be origin independent. Equations shown in **Figure 7** satisfy the origin independence of rotational strength.

9.04.4.2 The Consistency between X-Ray Crystallographic Bijvoet and CD Exciton Chirality Methods

It is well known that the AC of chiral compounds was first determined by X-ray crystallography using the anomalous dispersion effect of heavy atoms by Bijvoet *et al.* in 1951.[21–23] As discussed above, the CD exciton chirality method enables one to determine ACs in a nonempirical manner without any reference compounds with known ACs. These methods are based on totally different physical phenomena, but it is natural that for a specific chiral compound, X-ray and CD exciton methods should come to the same AC. However, it was claimed in 1972 that the ACs determined by X-ray and CD exciton methods disagreed with each other, and the ACs determined by the X-ray Bijvoet method should be revised.[24–26] This conclusion was based on the X-ray and CD analyses of compounds (−)-**5** and (+)-**6** in **Figure 11**, where the CD of the weak ^1L_b transition (~290 nm) of aniline chromophore polarized along the short axis was analyzed as an exciton couplet. However, this claim was subsequently retracted as a wrong assignment. Thus the exciton chirality method should be applied to the intense UV transition as shown in **Figure 11**, but not to the weak UV transition.

In 1976, the synthesis and CD spectra of the most ideal chiral cage compound (+)-**3** with two anthracene chromophores were reported;[27–29] the results completely proved the consistency between X-ray Bijvoet and CD exciton methods (**Figure 11**). Compound (+)-**3** was synthesized starting from diester (+)-**4**, which was chemically correlated with compounds (−)-**5** and (+)-**6**. The ACs of (−)-**5** and (+)-**6** had previously been determined by X-ray Bijvoet method.[30,31] As expected, compound (+)-**3** shows extremely intense exciton-coupled CD Cotton effects at the strong ^1B_b transition of anthracene chromophore polarized along the long axis: λ_{ext} 268.0 nm ($\Delta\varepsilon + 931.3$), 249.7 (−720.8), $A = +1652.1$ (**Figure 11**). It was thus shown that the use of strong UV transition gives rise to intense exciton-coupled CD.

Since the UV $\pi-\pi^*$ transition at 267.2 nm is polarized along the long axis of the anthracene chromophore, exciton split Cotton effects at 268.0 and 249.7 nm are generated by the exciton coupling between these two transition moments of each anthracene group. The long axes of the anthracene moieties constitute a clockwise screw sense leading to positive first and negative second Cotton effects, and therefore the AC of compound (+)-**3** was determined as shown. This result agrees with that determined by the X-ray Bijvoet method (**Figure 11**). It is now established that both X-ray and CD exciton chirality methods give correct ACs.[27]

As discussed above, the CD exciton chirality method enables one to determine ACs in a nonempirical manner. A striking example proving the nonempirical nature and utility of the CD exciton chirality method is the reversal of the ACs of clerodin **7** and related diterpenes **8**, **9** as briefly explained below (**Figure 12**).

In 1962, the AC of clerodin **7**,[32] a key compound of the clerodane diterpenes, was determined as shown in **Figure 12** by the X-ray Bijvoet method.[33] Since this AC was believed to be correct, clerodin **7** was then treated as a reference compound for newly isolated members of this diterpene family. For example, in 1974, the ACs of caryoptin **8** and 3-epicaryoptin **9** were determined to be as shown by CD and/or chemical correlation with **7**.

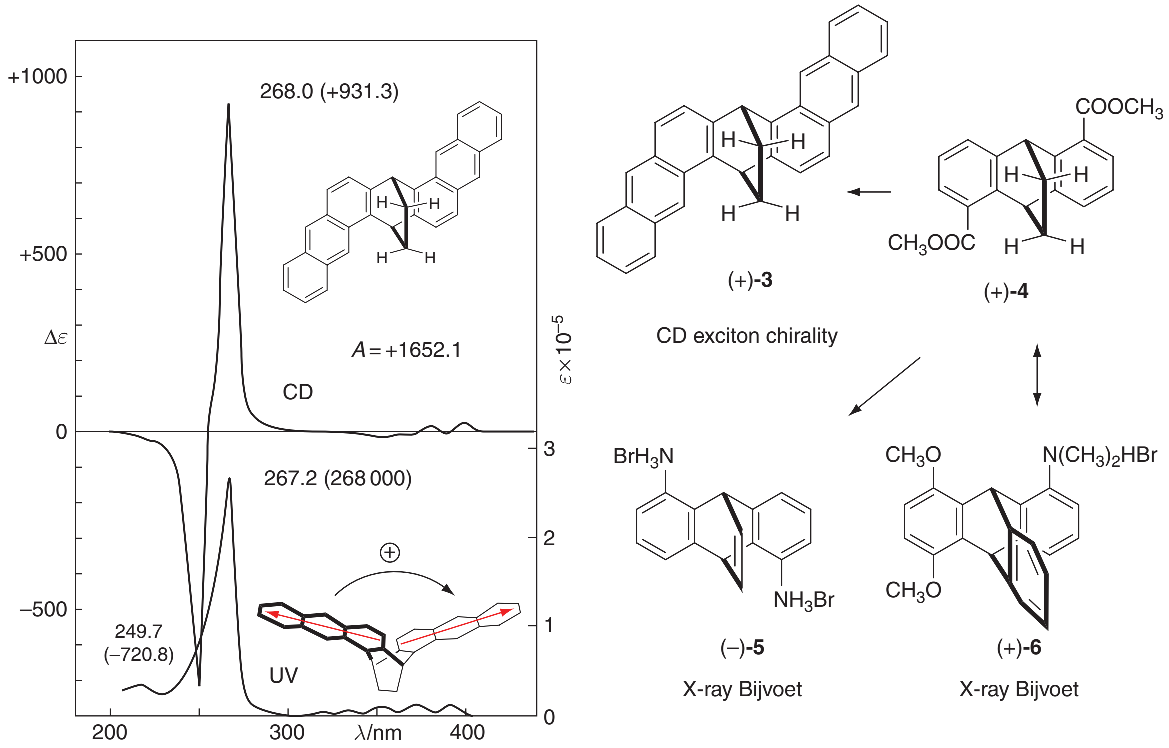


Figure 11 CD and UV spectra of (6*R*,15*R*)-(+)-6,15-dihydro-6,15-ethanonaphtho[2,3-*c*]pentaphene **3** in dioxane/EtOH and chemical correlation between compound **3** and related compounds, where the most ideal chiral cage compound **3** with two anthracene chromophores shows intense exciton-coupled CD Cotton effects, establishing the consistency between X-ray Bijvoet and CD exciton chirality methods. Redrawn from N. Harada; K. Nakanishi, *Circular Dichroic Spectroscopy – Exciton Coupling in Organic Stereochemistry*; University Science Books: Mill Valley, CA, and Oxford University Press: Oxford, 1983.

However, the observed positive exciton couplet of 3,6-bis(*p*-Cl-benzoate) **11** derived from 3-epicaryoptin **9** disagreed with the negative one expected from the ACs of **11** and **11a** (**Figure 12**). To explain the discrepancy between the CD and AC, the conformation **11b** was proposed in which one of the benzoate groups adopted a twisted conformation due to an intramolecular hydrogen bond (H-bond), thus generating a positive twist (**Figure 12**). This result was reported as an exception of the CD exciton chirality method.[34]

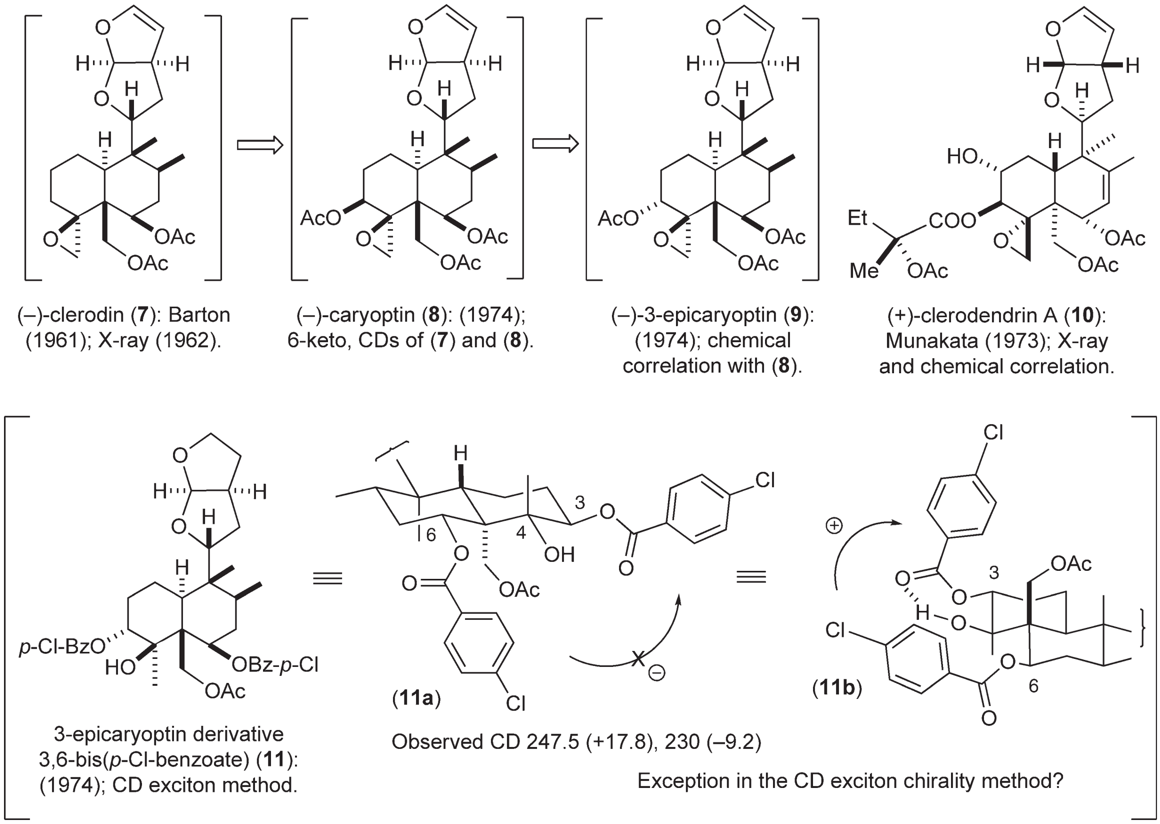
On the other hand, in 1973, the AC of newly isolated clerodendrin A **10** was independently determined by X-ray crystallography and chemical correlation and was shown to have the opposite AC to that of **7**.[35–37] Thus, it was classified as *ent*-clerodane (*enantiomeric* clerodane) (**Figure 12**).

As described above, 3,6-bis(*p*-Cl-benzoate) **11** was reported as the exception of the CD exciton chirality method. To resolve this problem, in 1978 the steroidal model compound **12** was synthesized.[38] Compounds **12** and **11** have the same relative configurations at key positions as shown in **12a** and **11a** (**Figure 13**). The CD spectrum of **12** showed a positive couplet in agreement with the positive twist of conformation **12a**, indicating that the conformation of the benzoate group is not twisted by an intramolecular H-bond. Since 3,6-bis(*p*-Cl-benzoate) **11** and **12** showed CD couplets of the same sign, **11a** should have the same AC as **12a** as shown in **Figure 13**. From these results, the ACs of **11**, **9**, and **7** were reversed.[38]

The above results prompted new X-ray analyses of **7** and **9**. The results showed that the original X-ray analysis used to determine the AC assignment for clerodin was incorrect and that the AC as originally assigned should be reversed.[39]

9.04.4.3 The Use of Preexisting Chromophores in Natural Products for Exciton Coupling

Some natural products already have one or two chromophores, which are useful for observing exciton CD to determine their ACs. The following chromophores are commonly found in natural products.



(−)-clerodin (**7**): Barton (1961); X-ray (1962).

(−)-caryoptin (**8**): (1974); 6-keto, CDs of (**7**) and (**8**).

(−)-3-epicaryoptin (**9**): (1974); chemical correlation with (**8**).

(+)-clerodendrin A (**10**): Munakata (1973); X-ray and chemical correlation.

3-epicaryoptin derivative 3,6-bis(*p*-Cl-benzoate) (**11**): (1974); CD exciton method.

Observed CD 247.5 (+17.8), 230 (−9.2)

Exception in the CD exciton chirality method?

Figure 12 The absolute configurations of clerodane diterpenes as determined by X-ray crystallography/CD/chemical correlation. However, the ACs of compounds in brackets were later reversed.

9.04.4.3.1 Substituted benzene and polyacene chromophores for the CD exciton chirality method

As demonstrated in Section 9.04.4.2, the ^1B_b transition of polyacene chromophores is ideally suitable for observing exciton-coupled CD. The UV data of some polyacenes with D_{2h}-symmetry are shown in Figure 14. In the polyacene systems, there is no ambiguity for determining the long and short axes, and therefore the CD exciton chirality method offers more reliable and definite conclusions of AC.

9.04.4.3.2 Conjugated dienes, enones, ene-esters, ene-lactones, and diene-esters as exciton CD chromophores

The conjugated dienes, enones, etc., shown in Figure 14 are useful chromophores for the CD exciton chirality method. The transition moment of their $\pi-\pi^*$ band is almost parallel to the long axis of the chromophores as depicted in the tables.

9.04.4.3.3 Natural products with two chromophores showing exciton CDs: Nondegenerate and degenerate cases

The exciton coupling CD mechanism is applicable also to compounds having two different chromophores, which exhibit long-axis-polarized $\pi-\pi^*$ transitions at different wavelengths. This case is called the nondegenerate system because of the different transition energies. On the other hand, if a compound has two identical chromophores, for example, steroidal bis(*p*-dimethylaminobenzoate) 1 in Figure 7, it is called a degenerate system, because of the same transition energies (degenerated excited state).

The ACs of some natural products, such as those in Figure 15, were established conveniently by direct analysis of their CD spectra without any additional chemical modification. In such cases the interaction of at

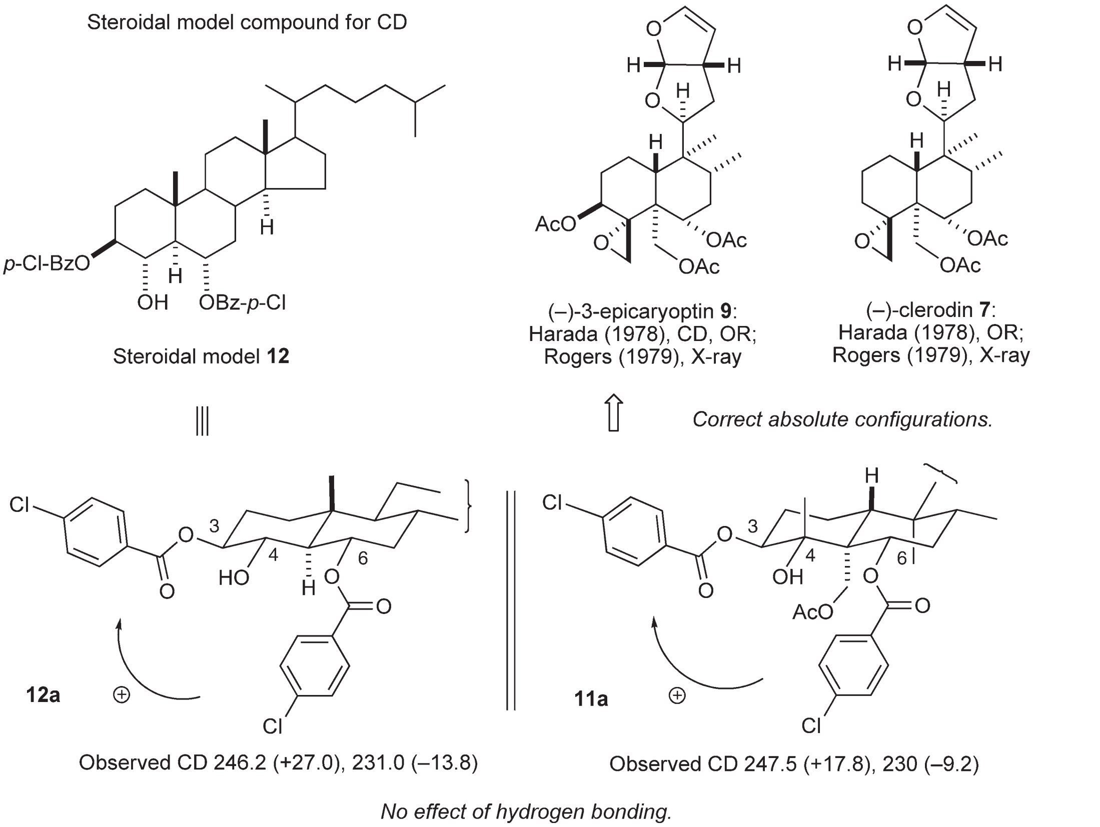


Figure 13 Comparison of exciton CD of model compound **12a** with that of **11a** led to the reversal of ACs of clerodane and related compounds. The same conclusion was obtained later by X-ray crystallography.

least two preexisting chromophores with suitable electronic and geometrical attributes leads to a very diagnostic exciton split CD band, and hence to assignment of the AC.

Dihydro-β-agarofuran **13** itself has cinnamate and benzoate chromophores, which exhibit exciton CDs at 270.7 and 227.8 nm (see Section 9.04.6.3.4).[40] In vinblastine **14**, indole and indoline chromophores interact with each other generating exciton CDs (see Section 9.04.6.3.13).[41] For abscisic acid **15**, the opposite AC was once assigned, but it was later revised as shown by several studies. One was the application of exciton CD to the interaction between the enone and the diene-carboxylic acid chromophores showing a positive couplet.[42] The AC of dendryphiellin F **16** was determined on the basis of exciton CDs generated by the interaction between diene and diene-carboxylate chromophores.[43] The case of quassin **17** is unique because of the exciton coupling between two identical chromophores, that is, α-methoxy-enone groups.[44] The exciton coupling between dehydro-tetralone and phthalide chromophores enabled the determination of AC of arnottin II **18**.[45]

9.04.4.4 Suitable Chromophores for the CD Exciton Chirality Method

In general, natural products contain either only one useful chromophore or none at all. For this reason the selection of suitable chromophore(s) to be introduced into the substrate by chemical derivatization is an important issue when it comes to the determination of AC by CD. With the so-called monochromophoric approach, applicable mainly to rigid substrates, two identical chromophores are introduced in a one-step reaction, usually by acylating the primary or secondary hydroxyl or amino groups. In such cases the exciton coupling provides an intense CD and allows for uncomplicated AC assignments.[9,11]

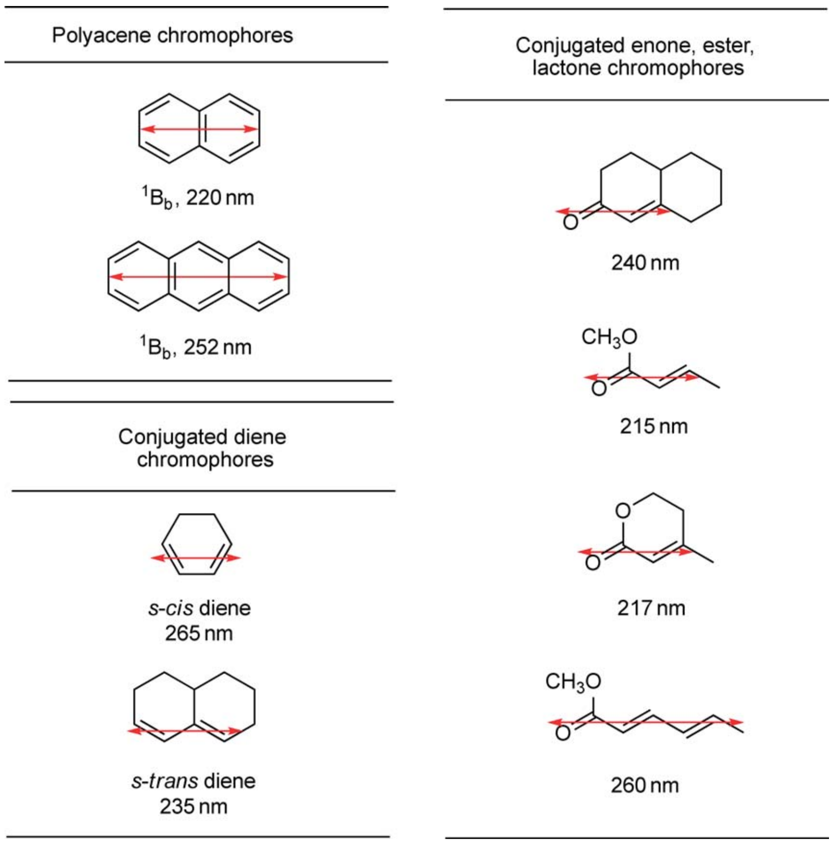


Figure 14 Some exciton CD chromophores found in natural products.

The chromophores used for the CD exciton chirality method have to satisfy the following requirements: (1) presence of an intense $\pi-\pi^*$ transition and (2) direction of the transition moment is clear in the geometry of the chromophore. Therefore, in general, chromophores of high symmetry are desirable.

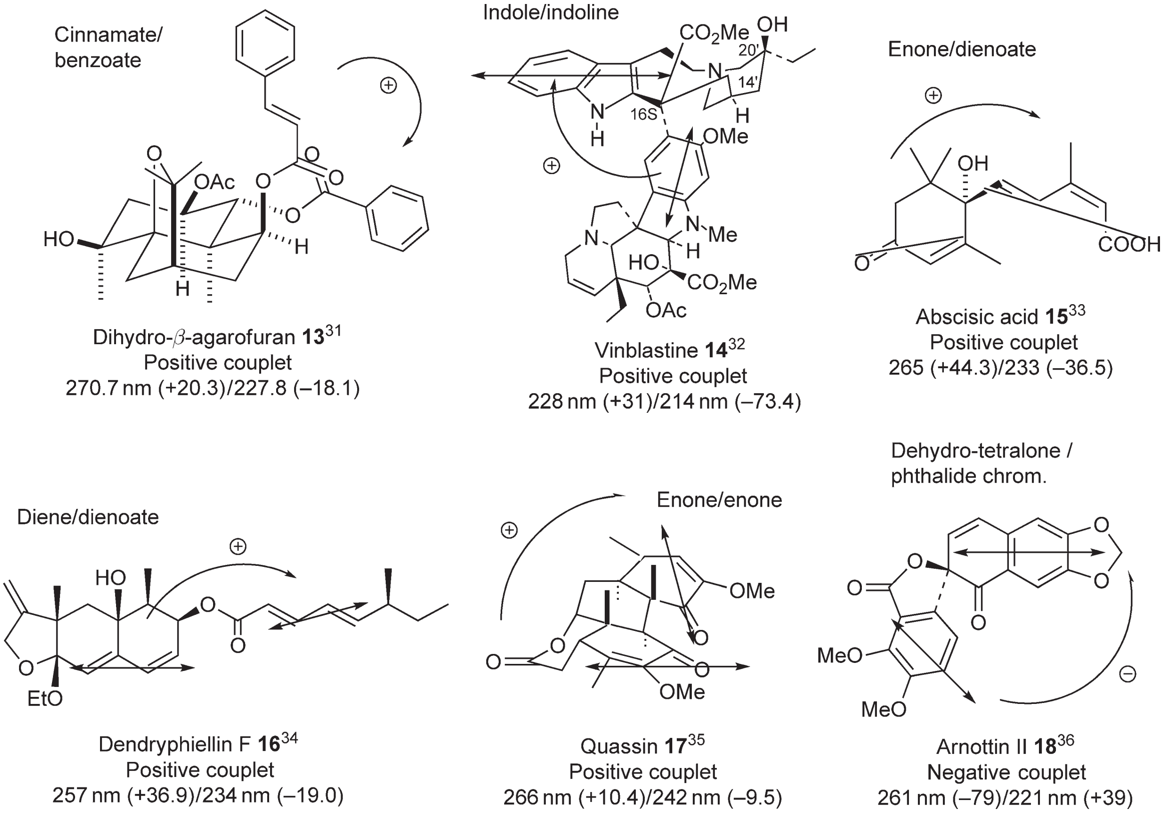
Figure 16 shows typical chromophores useful for the CD exciton chirality method, where arrows indicate the direction of the transition moment responsible for the exciton-coupled CD. In general, the long-axis-polarized transitions are suitable for exciton CD, because of the larger UV intensity; as discussed above, the exciton coupling between strong UV transition moments gives rise to strong CD Cotton effects. The newly introduced chromophores are selected either because of their suitability for exciton coupling with another preexisting chromophore in the substrate or to avoid interaction with them if the latter possess an electronically complicated structure.

9.04.4.4.1 Para-substituted benzoate chromophores for glycols

As discussed in the above examples of natural products, the intramolecular CT or ^1L_a transition (230–310 nm) of *para*-substituted benzoate chromophores is useful for determining the AC of glycols.[8] The intramolecular CT transition is polarized along the long axis of the benzoate chromophore, which is almost parallel to the alcoholic C–O bond. Therefore, the AC of the glycol part can be determined from the exciton CD data. On the other hand, *ortho*- and *meta*-substituted benzoate chromophores are not suitable for the CD exciton chirality method because their transition moments are not parallel to the alcoholic C–O bond.

9.04.4.4.2 Cinnamate, β-naphthoate, and other chromophores for glycols

These chromophores are also useful because of their absorption at longer wavelength and/or strong absorption intensity.



Dihydro-β-agarofuran **13**[31]
Positive couplet
270.7 nm (+20.3)/227.8 (–18.1)

Vinblastine **14**[32]
Positive couplet
228 nm (+31)/214 nm (–73.4)

Abscisic acid **15**[33]
Positive couplet
265 (+44.3)/233 (–36.5)

Dendryphiellin F **16**[34]
Positive couplet
257 nm (+36.9)/234 nm (–19.0)

Quassin **17**[35]
Positive couplet
266 nm (+10.4)/242 nm (–9.5)

Arnottin II **18**[36]
Negative couplet
261 nm (–79)/221 nm (+39)

Figure 15 Examples of natural products with two preexisting chromophores showing exciton CD.

9.04.4.4.3 Tetraphenyl-porphyrin-carboxylic acid

Among the chromophores shown in Figure 16 the tetraphenporphyrins and metalloporphyrins (see also Figure 17) deserve special attention. They possess a very intense sharp and narrow Soret band (ε 450 000–550 000), shifted to the red (at ~420 nm). They are also endowed with many other unique geometrical and electronic properties, such as fluorescence, facile modification, variable solubility, and approximately planar geometry. Therefore, the porphyrins and their Zn and Mg derivatives belong to the most powerful and versatile CD chromophores. A detailed discussion on the application of porphyrins as CD reporter groups as well as an account of the theoretical analysis of porphyrin–porphyrin exciton interactions is available.[46,47]

The Soret band originates from the two degenerate transitions B_x and B_y (Figure 17), which are perpendicular to each other; therefore, theoretically the porphyrin Soret band should be considered as a circular oscillator.[47] However, due to rotational flexibility around the meso porphyrin 5-C-phenyl junction (librational averaging), the transitions B_x and B_y can be represented by one effective transition moment along the 5–15 axis (Figure 17), and the exciton CD reflects the chirality between two effective transition moments. So, tetraphenyl-porphyrin-carboxylic acid (TPP-COOH) is very useful for observing exciton CD because of its large red shift and large ε value.

Most of the chromophores shown in Figure 16 are useful for exciton split CD analysis for short to medium interchromophoric distances of 13–15 Å. For distances up to 50 Å, only the porphyrins and metalloporphyrins can provide couplets sufficiently intense for configurational analysis. Therefore when the AC of remote stereogenic centers is sought, the tetraarylporphyrin (TPP) and metalloporphyrins make an excellent choice.

In cases where the configurational analysis involves remote stereogenic centers with C–C distances of approximately 8–9 Å and interchromophoric distances with R_{ij} of approximately 13–14 Å, the observed CD couplet becomes very weak or even undetectable with chromophores with weak or even moderate absorption

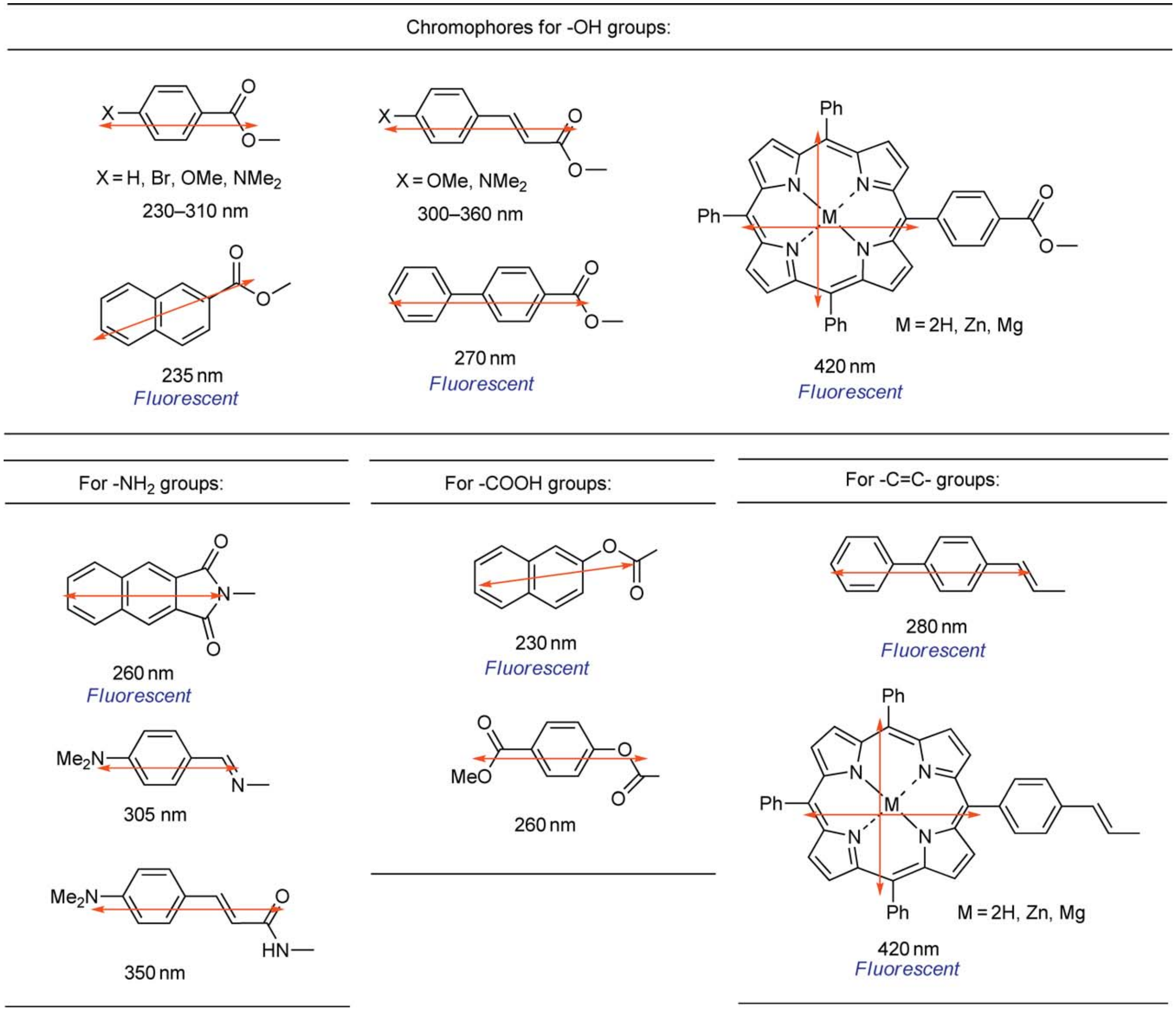


Figure 16 Typical chromophores useful for the CD exciton chirality method, where arrows show the direction of transition moment responsible for the exciton-coupled CD.

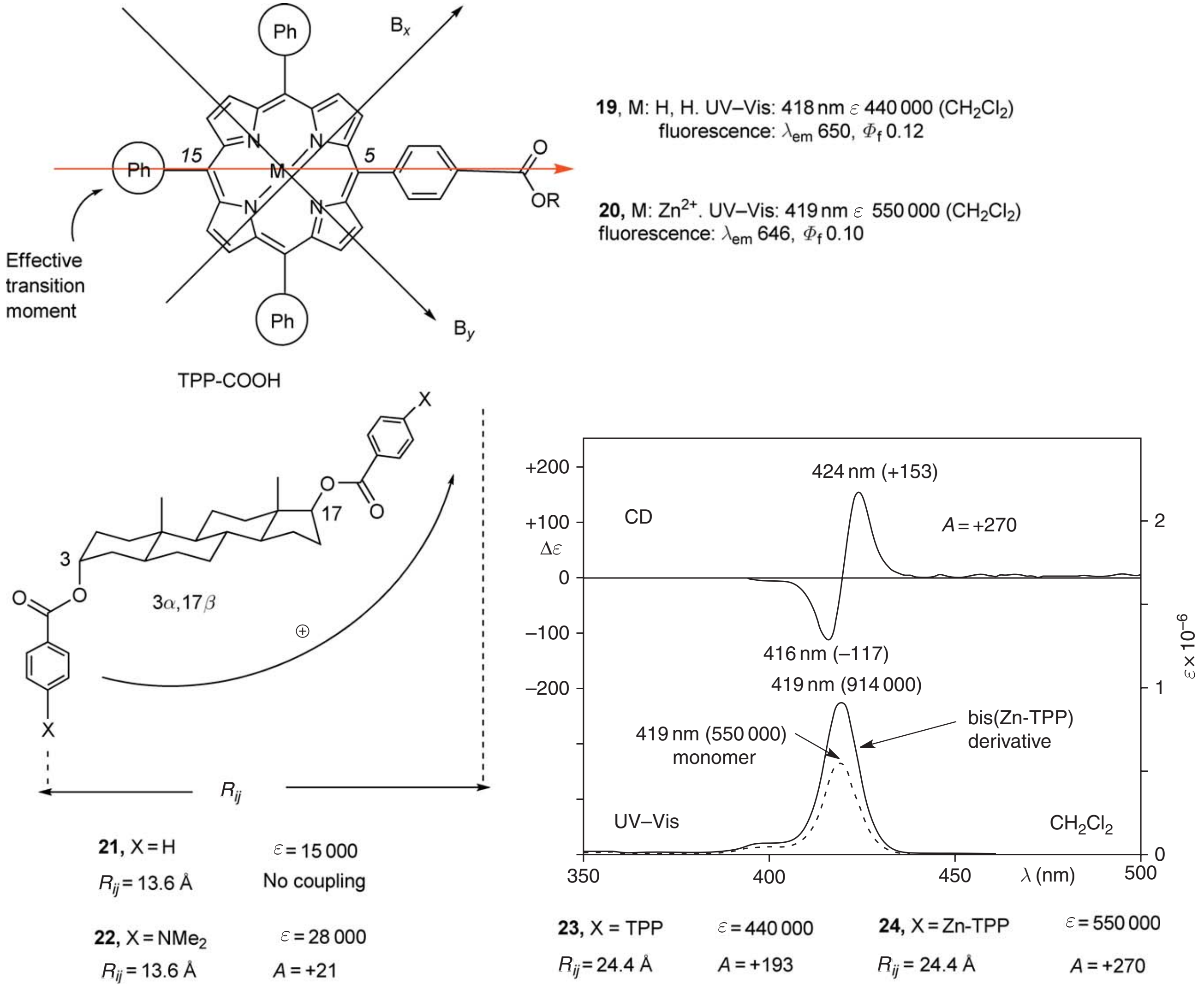


Figure 17 UV–Vis and fluorescence data for TPP-COOH **19** and Zn-TPP-COOH **20**. Bottom: bis(tetraarylporphyrin) derivative of 5α-cholestane-3α,17β-diol **23** and **24**: a positive helicity between the two effective transition moments defined in direction 5C/15C, the interchromophoric distance R_{ij}, UV–Vis and CD spectra of in CH_2Cl_2; UV ε, R_{ij} values, and CD amplitudes *A* of other bischromophoric derivatives of 5α-cholestane-3α,17β-diol **21** and **22**.

bands, such as benzoate or substituted benzoates. This is because the amplitude A is inversely proportional to the square of the interchromophoric distance R_{ij} (see Equation (16)). A striking increase in the A-value is seen in TPP and its Zn derivative (Zn-TPP). The Zn-TPP derivative exhibits an A-value more than 10-fold larger than *p*-dimethylaminobenzoate at an R_{ij} distance of 24.0 Å. Other examples for efficient porphyrin–porphyrin CD coupling over 40–50 Å can be found in Matile *et al.*[48]

9.04.4.4.4 Benzamido and C_{2v}-symmetrical 2,3-naphthalenedicarboximido chromophores for amino alcohols and diamines

The CD exciton chirality method is also applicable to the intramolecular CT band of benzamido groups. The transition is polarized along the long axis of the chromophore. However, in some cases, the benzamide moiety exists as a mixture of (E) and (Z) isomers, and therefore, the mutual orientation of the transition moments is uncertain. Thus, in these situations, one should be cautious in assigning AC by CD.

The chromophore of 2,3-naphthalenedicarboximide exhibits an intense ^1B_b transition around 260 nm, which is polarized along the long axis of the chromophore. This C_{2v}-symmetrical chromophore is ideally suitable for the CD exciton chirality method because the long-axis-polarized transition moment is exactly parallel to the C–N bond of amine moiety. This is an advantage of the 2,3-naphthalenedicarboximide group and hence the use of this chromophore is highly recommended for primary amines.

9.04.4.4.5 Bichromophoric methods and derivatization

For acyclic or conformationally flexible natural products, the bichromophoric approach is suitable, where chromophores with very different λ_{max} are introduced selectively by two-step protocols. When chromophores whose absorption maxima span 50–100 nm are introduced, the coupling leads to a CD curve with unique, fingerprint shapes, depending on the absolute twist between the interacting chromophores and the conformational population in the solvent employed. The comparison of such curves characteristic for each solvent with corresponding reference curves of known standards lead to a configurational assignment, although in a semiempirical manner, of several stereogenic centers at the same time. This approach was successfully applied to 1,2- and mixed 1,2-/1,3-polyols and amino alcohols.[49–51]

Figure 18 illustrates a submicroscale chemical protocol developed for the analysis of sphingosines and dihydrosphingosines isolated from new cell lines. First, the NH_2 group of D-*erythro*-sphingosine **25** was blocked as a naphthimido group yielding a derivative **26**. Then the OH groups were converted to 2-naphthoate groups affording a derivative **27** that can be sensitively detected by high-performance liquid chromatography (HPLC), mass spectrometry, CD, and fluorescence analysis. Upon comparison of the observed CD with the standard CD curves of *erythro*- and *threo*-sphingosines/dihydrosphingosines, the relative configuration and AC can be assigned.[52]

9.04.4.4.6 Chromophores for carboxylic acids and olefin compounds

The chromophores suitable for chiral carboxylic acids are listed in **Figure 16**. The application of the exciton chirality method to olefin compounds is unique and interesting. The isolated olefin group shows a $\pi-\pi^*$ transition below 200 nm, and therefore the exciton method is not applicable in a straightforward manner. However, by the use of olefin metathesis, the chromophores shown in **Figure 16** can be introduced and exciton CD can be used for determining ACs.

9.04.4.4.7 Natural products with one preexisting chromophore useful for exciton coupling

If a natural product contains one chromophore, which is useful as a partner of exciton coupling, the second chromophore can be introduced by chemical derivatization to determine the AC by exciton CD. The newly introduced chromophore is selected for optimal exciton coupling with the preexisitng chromophore. Thus the chromophore showing similar UV λ_{max} to that of the preexisitng chromophore is effective for exciton coupling CD.

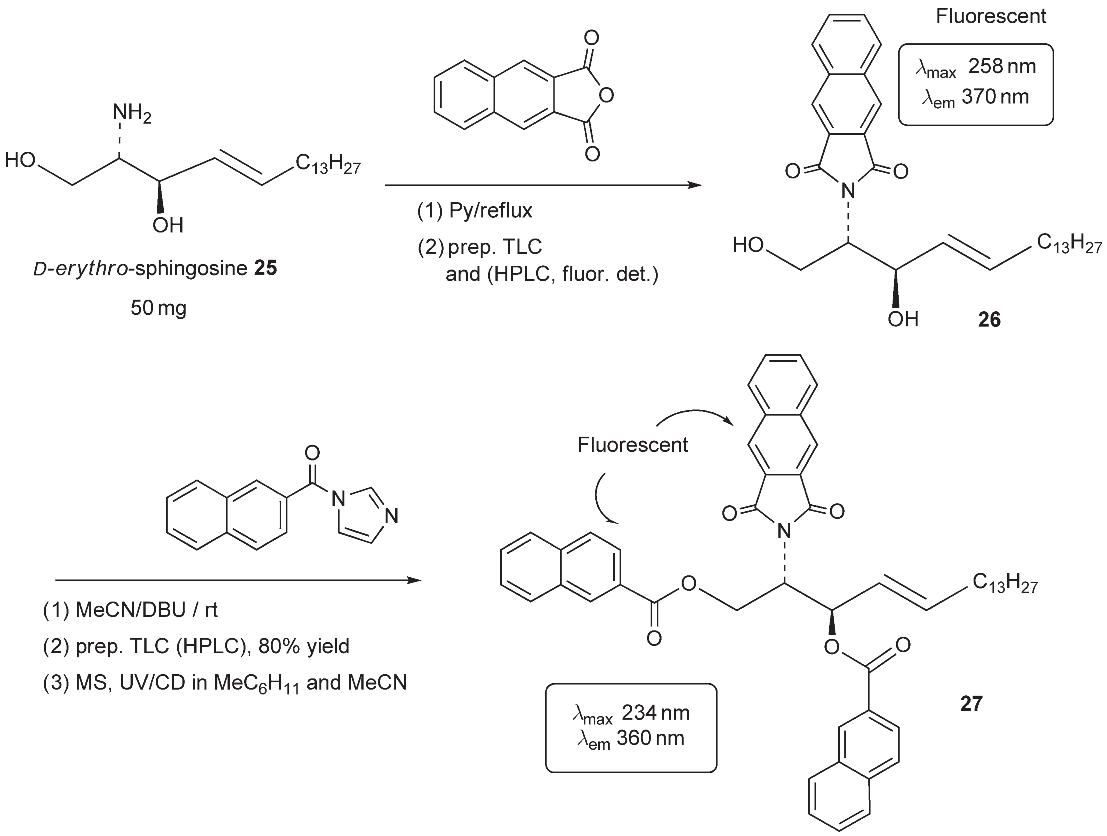


Figure 18 By a selective two-step microscale chemical derivatization procedure, two different types of chromophores are introduced in D-*erythro*-sphingosine **25**.

9.04.4.4.8 Natural products with preexisting chromophore not useful for exciton coupling: Use of red-shifted chromophores

If a natural product has a preexisting chromophore, which may disturb the observation of exciton CD, it is advisable to choose chromophores with longer wavelength UV $\lambda_{\max}$ than that of the preexisting chromophore to avoid the overlap of Cotton effects. Red-shifted chromophores shown in **Figure 19** are useful for this purpose.

As shown in **Figure 19**, taxinine derivative α-glycol **28a** shows a very intense positive CD Cotton effect due to the $\pi-\pi^*$ transition of the strongly strained enone group around 263 nm. In the previous application of the exciton chirality method, unsubstituted benzoate chromophores were used and a negative exciton couplet was clearly observed despite the overlap with the enone Cotton effect.[53] To avoid the overlap of exciton CDs with the enone Cotton effect, a red-shifted chromophore (**chrom-3**) was used for derivatization yielding ester **28b**. As expected, the CD of **28b** exhibited a clear negative exciton couplet indicating a counterclockwise screw sense between the two hydroxyl groups in full agreement with the previous report of the AC.[54]

9.04.4.5 Supramolecular Approach in Exciton Chirality Method – Application of Porphyrin Tweezers

Recently, the use of TPPs and their metal derivatives as useful CD chromophores was extended by the development of a totally new supramolecular approach for the determination of the AC of chiral compounds that contain a single stereogenic center and only one site for chromophoric derivatization. This group includes various natural products carrying only a single functionality, such as secondary hydroxyl, primary or secondary

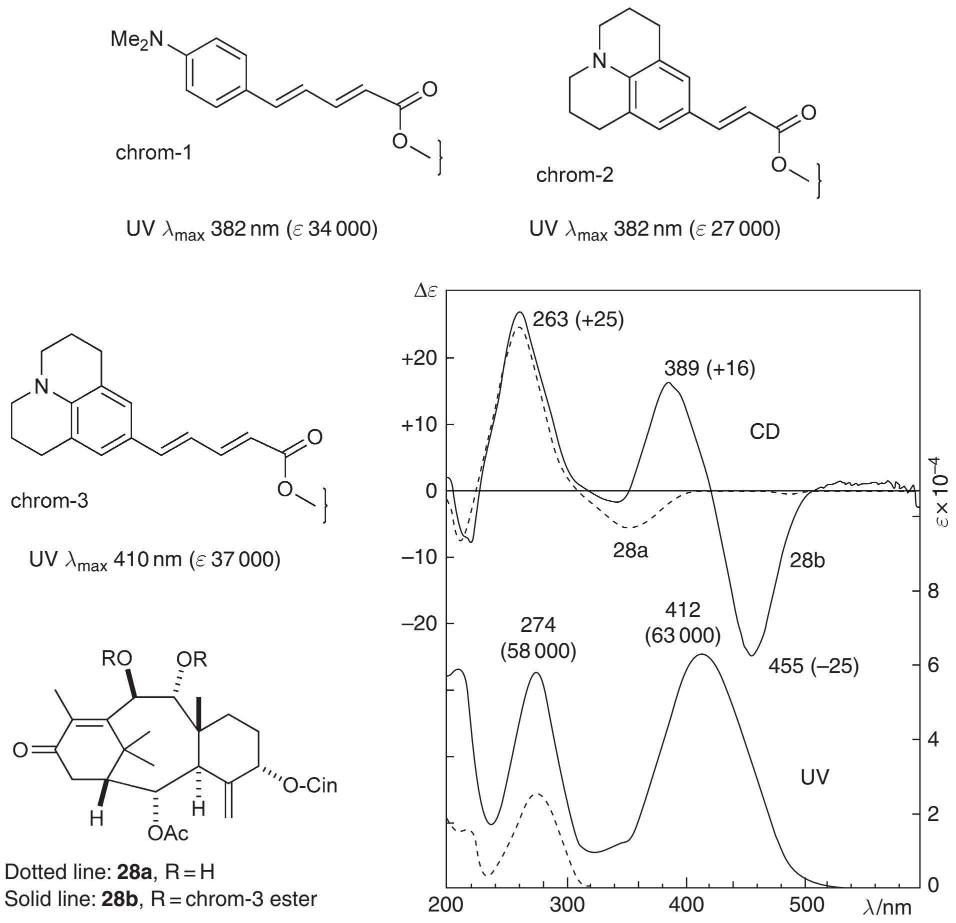


Figure 19 Red-shifted chromophores and application to taxinine system.

amino, and carboxyl groups. They are unsuitable for application of conventional exciton chirality approach where at least two intramolecularly interacting chromophores are necessary.

The supramolecular approach mentioned above employs a dimeric zinc porphyrin reagent, now available under the name 'Zn-tweezer'. The latter is capable of forming 1:1 host–guest complexes upon adding a solution of *N*,*N*-bidentate conjugate, prepared by reacting the chiral substrate with an achiral trifunctional bidentate carrier as shown in Figure 20.[55,56]

Interestingly, the observed facile N/Zn coordination to a Zn-porphyrin tweezer and formation of 1:1 sandwiched chiral host–guest complex proceeds under steric control and usually leads to a very intense exciton-coupled bisignate CD spectrum in the Soret region. The origin of such intense CD couplets lies in the predominant presence of conformers with a preferred interporphyrin helicity where the larger group L protrudes from the binding pockets in order to avoid unfavorable steric interactions. Therefore the chiral sense of twist between the two porphyrins in the complex is dictated by the steric orientation of L and M at the stereogenic center of the substrate. In case there is no ambiguity in the assignment of L and M groups, the sign of the couplet determines the AC at this center.

Over the past years the search for more reliable discrimination of L/M relative steric size and theoretical prediction of the preferred interporphyrin helicity of the host–guest complex has led to the development of molecular mechanics calculations protocol using the Merck Molecular Force Field (MMFF) approach coupled to Monte Carlo-based conformational analysis.[57]

The porphyrin tweezers method is now well established and has allowed a successful determination of AC of some natural products, such as isotomenoic acid **36**, an irregular diterpene,[58] and bovidic acid **37**, an 18-carbon

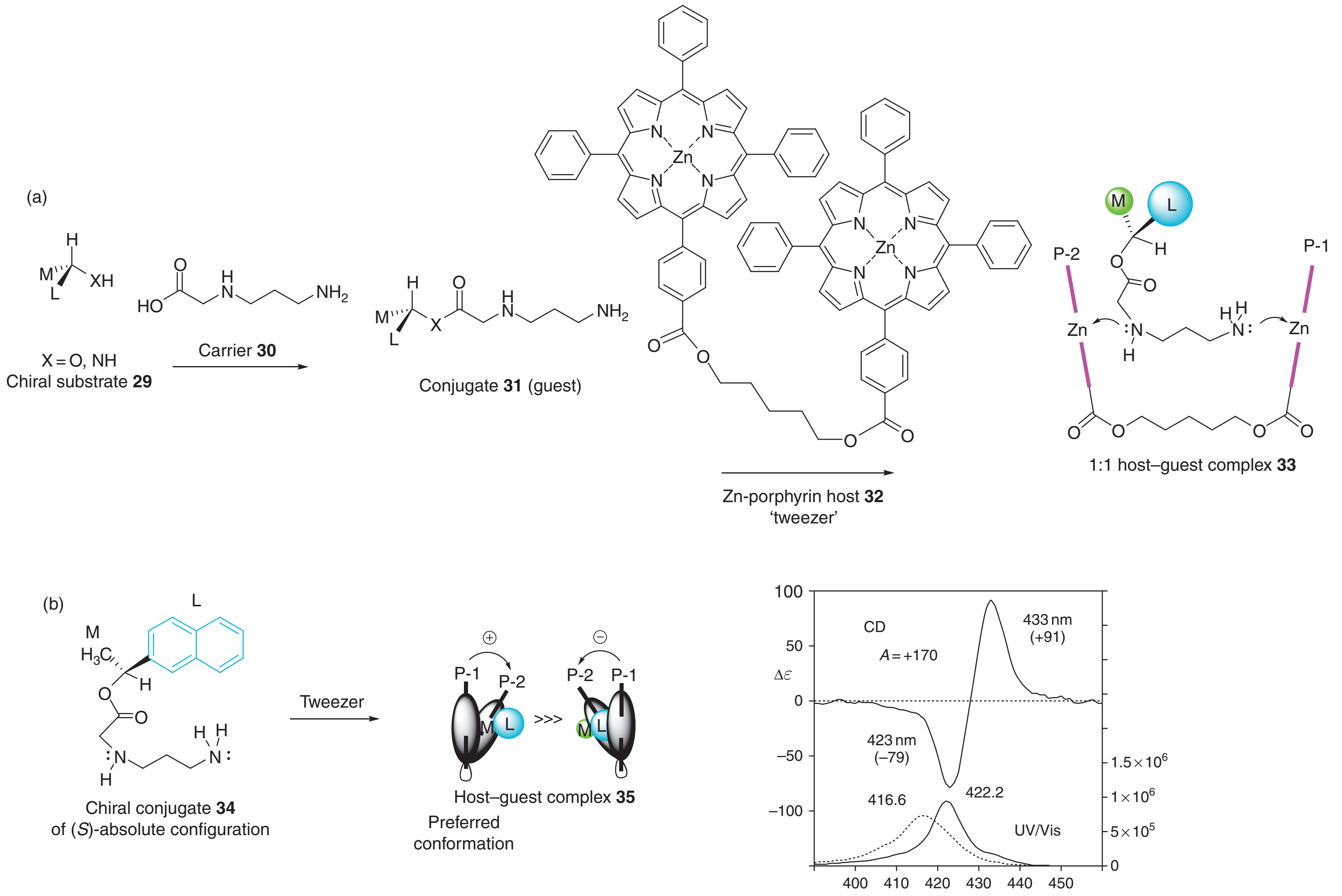


Figure 20 Formation and CD of 1:1 host–guest complex between achiral Zn-porphyrin tweezer and chiral substrate. (a) A reaction of the carrier molecule **30** with a starting substrate **29** (secondary alcohol or primary amine) leads to formation of bidentate chiral conjugate **31** (guest), which upon mixing with an achiral Zn-porphyrin tweezer **32** yields a 1:1 host–guest complex **33**. (b) Example for the formation with (*S*)-α-(2-naphthyl)ethanol **34** of a host–guest complex **35** in two conceivable conformations with opposite interporphyrin twist. The one where the L (larger) group protrudes away from P-1/P-2 binding pocket is preferred and has a positive twist between the two porphyrins. This gives rise to a characteristic exciton split CD with positive amplitude $A = +170$ (in methylcyclohexane) in agreement with the (*S*)-absolute configuration of starting substrate. Redrawn from N. Berova; L. Di Bari; G. Pescitelli, *Chem. Soc. Rev.* **2007**, *36*, 914–931.

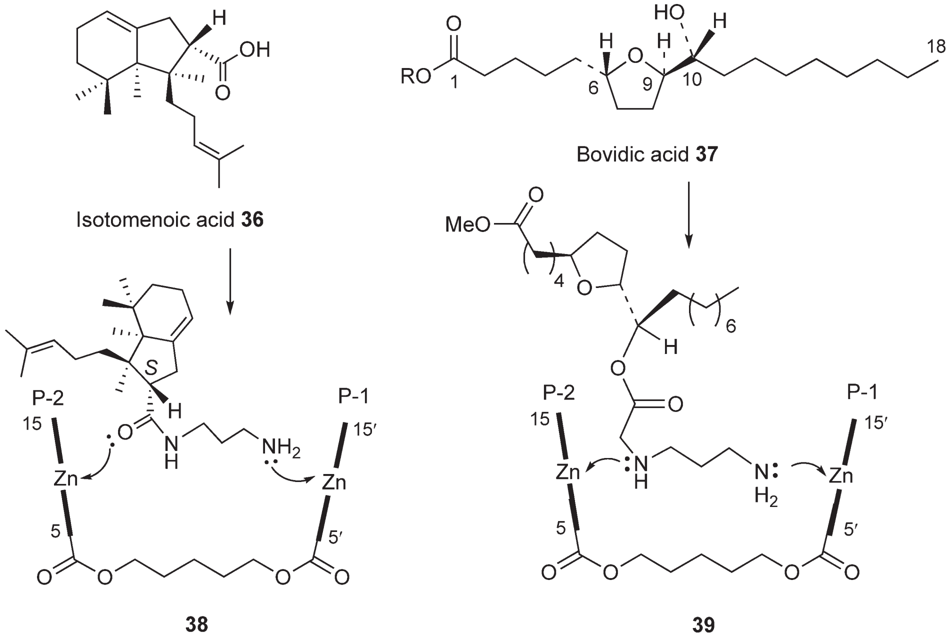


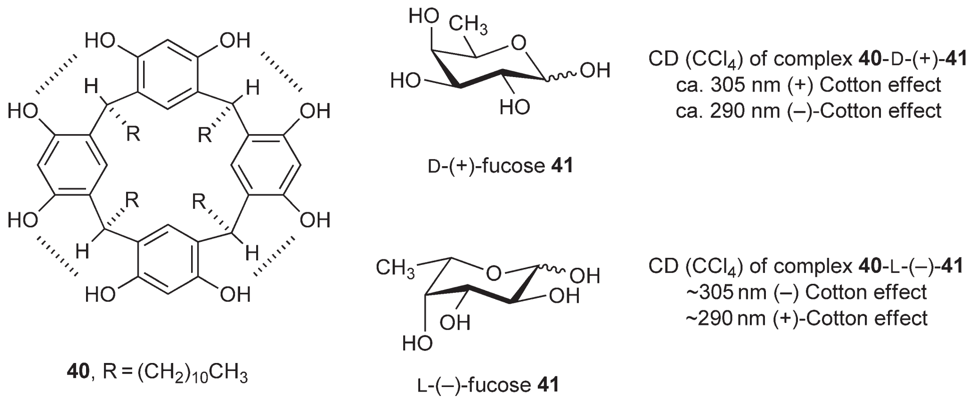
Figure 21 Applications of the porphyrin tweezers method to natural products.

hydroxyfuranoid acid[59] (**Figure 21**). More recently, other types of porphyrin-based tweezers have been developed. Structural changes in the tweezer, such as introduction of various substituents at the aryl groups and in the bridge between the two porphyrins allow for tuning the complexation ability of the tweezer and extension of its application to other types of chiral substrates.[60–62]

9.04.5 Induced CD

The enormous attention and advance in supramolecular chemistry in the past few decades has stimulated interest in CD arising from different types of *intermolecular* interactions. Four typical situations are encountered: (1) A chiral (nonracemic) 'guest' and an achiral chromophoric compound as 'host', for example, crown ethers, calixarenes, atropisomeric biaryls, and bis-porphyrin systems, can form a chiral host–guest complex, which exhibits an induced CD (ICD) within the absorption bands of the host.[63] (2) Inversely, a small guest molecule that is achiral and hence its chromophore is chiroptically inactive, upon binding to a biopolymer host, such as proteins,[64] polypeptides, oligonucleotides,[65] oligosaccharides (notably including cyclodextrins),[63] may produce an ICD due to the chiral perturbation by the biopolymer host. (3) A third case is when a coupling between several guest molecules bound to different sites of a macromolecular host results in a diagnostic CD spectrum.[66] (4) A chiral, nonchromophoric ligand binds to a metal ion with observable *d*- or *f*-type transitions in the UV–Vis spectrum, making them CD active. In several cases, CD lends itself not only to the detection of host–guest interactions, but also to the analysis of binding modes, such as association–dissociation kinetics and thermodynamics (see Section 9.04.6.5.2).

Figure 22 shows an interesting example of ICD of type (1), where achiral resorcinol-dodecanal cyclotetramer **40** interacts with D-(+)-fucose **41** to form a chiral host–guest complex, the CD spectrum of which shows positive and negative Cotton effects around 305 and 290 nm, respectively. Upon the host–guest interaction, host **40** takes chiral conformations, in which four resorcinol rings are chirally twisted to generate induced bisignate CD. When L-(−)-fucose **41** was used, opposite CDs were observed. Based on these results, the use of host **40** as a supramolecular probe for the assignment of ACs of chiral guests was reported.[67]



40, R = $(CH_2)_{10}CH_3$

D-(+)-fucose **41**

L-(−)-fucose **41**

CD (CCl_4) of complex **40**-D-(+)-**41**
ca. 305 nm (+) Cotton effect
ca. 290 nm (–)-Cotton effect

CD (CCl_4) of complex **40**-L-(–)-**41**
~305 nm (–) Cotton effect
~290 nm (+)-Cotton effect

Figure 22 Induced CD of complexes, achiral host **40** and chiral sugar guests **41**: CD λ_{ext} data were obtained from the published spectra.

9.04.6 Characterization of Natural Products by CD – Selected Examples

As discussed above, CD spectroscopy is useful for the characterization of natural products. In the following, the application of CD spectroscopy to the structural studies of natural products is exemplified and explained. The cases are (1) CD and solvent-dependent conformational change, (2) determination of AC by comparison of CD spectra, (3) application of CD exciton chirality method, (4) CD of atropisomers, (5) determination of ACs by theoretical calculation of CD spectra, and (6) supramolecular systems and CD spectra.

9.04.6.1 CD and Solvent-Dependent Atropisomerism of Antibiotic FD-594

Antibiotic FD-594 **42** exhibited almost opposite CD curves in $CHCl_3$ and MeOH due to the solvent-dependent atropisomerism, which was confirmed by ^1H nuclear magnetic resonance (NMR) coupling constants[68] (**Figure23**).

The AC of **42** was determined by X-ray crystallography as shown. A strong negative CD around 270 nm in $CHCl_3$ implies a negative exciton coupling between the two aromatic chromophores. In MeOH, the helicity is inverted to generate a strong positive CD around 270 nm. Similar behavior was observed with aglycon **43**.

9.04.6.2 Determination of Absolute Configuration by Comparison of CD Spectra

9.04.6.2.1 Absolute configuration of thysanone isolated from Thysanophora penicilloides

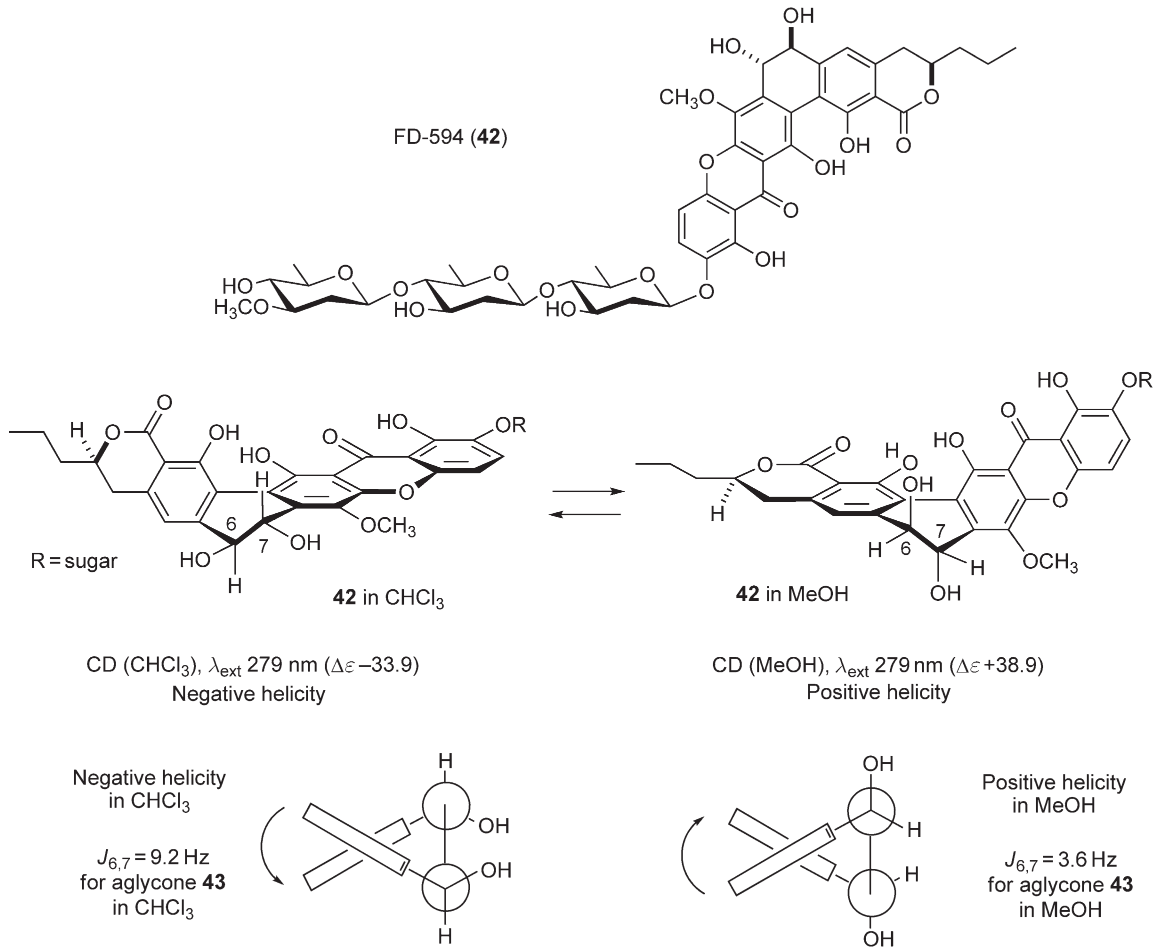
The (1*R*,3*S*) AC of thysanone **44**, a fungal benzoisochromanquinone with potent rhinovirus 3C-protease inhibitory activity was determined by comparison of the CD spectra of the authentic natural thysanone with that of a synthetic sample prepared by total synthesis from (*S*)-ethyl lactate[69] (**Figure 24**).

9.04.6.2.2 Absolute configurations of mutafurans A–G isolated from Bahamian sponge Xestospongia muta

The ACs of mutafurans A–G **45–51**, brominated ene-yne tetrahydrofurans (THFs), isolated from Bahamian sponge *X. muta* were determined by comparison of CD spectra as shown in **Figure 25**. The observed CD Cotton effects are very weak because of the weak perturbation of a conjugated ene-yne chromophore by the chirality in a THF ring. On the other hand, the terminal bromo-diene or bromo-ene chromophore does not contribute to the CD because of remote distance from the chiral THF ring. As reference compounds, two model compounds (−)-**d 52** and (+)-**e 53** with the ene-yne THF moiety were synthesized starting from (*R*)-(+)-epoxyhexane. Since the CD Cotton effects of (−)-**d 52** are the same in sign as those of natural products **45–51**, their ACs were determined as shown.[70]

9.04.6.2.3 Absolute configuration of ciguatoxin

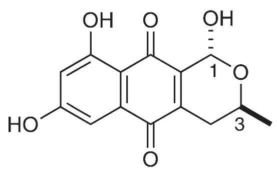
The 2*S* configuration of ciguatoxin (CTX, **54**)[71] was assigned on the basis of the CD exciton chirality data of tetrakis(*p*-Br-benzoate) of **54** and tris(*p*-Br-benzoate) of the AB fragment (**Figure 26**). This was later confirmed


FD-594 (42)
HO
OH
CH$_3$O
OH
OH
O
O
O
OH
HO
H$_3$CO
HO
HO
R = sugar
42 in CHCl$_3$
42 in MeOH
6
7
H
OCH$_3$
OR
Negative helicity in CHCl$_3$
Positive helicity in MeOH
$J_{6,7}$ = 9.2 Hz for aglycone 43 in CHCl$_3$
$J_{6,7}$ = 3.6 Hz for aglycone 43 in MeOH

CD (CHCl$_3$), λ_{ext} 279 nm ($\Delta\varepsilon$ −33.9)
Negative helicity

CD (MeOH), λ_{ext} 279 nm ($\Delta\varepsilon$ +38.9)
Positive helicity

Figure 23 Solvent-dependent atropisomerism of antibiotic FD-594. Redrawn from T. Eguchi; K. Kondo; K. Kakinuma; H. Uekusa; Y. Ohashi; K. Mizoue; Y.-F. Qiano, *J. Org. Chem.* **1999**, *64*, 5371–5376.



(1*R*,3*S*)-thysanone **44**

Natural thysanone

CD (MeOH), λ_{ext} 296 nm ($\Delta\varepsilon$ −3.7)
λ_{ext} 259 nm ($\Delta\varepsilon$ +3.8)

Synthetic sample from (*S*)-ethyl lactate

CD (MeOH), λ_{ext} 296 nm ($\Delta\varepsilon$ −3.1)
λ_{ext} 257 nm ($\Delta\varepsilon$ +2.8)

Figure 24 Thysanone and CD data.

by chemical degradation and comparison with an authentic sample. The AC of C5 in CTX4A 55 was determined by comparison of the CD spectrum of stereoselectively synthesized *p*-Br-benzoate 56a, containing the AB ring fragment of CTX4A, with that of tris(*p*-Br-benzoate) 55a of CTX4A. Both compounds show intense exciton CDs of positive chirality, which are caused by the interaction between 1,3-diene and *p*-Br-benzoate chromophores. Since the relative configurations of CTXs have been determined by intensive NMR spectral studies, the ACs of CTXs were determined as illustrated. It should be noted that because of the extremely limited availability of CTXs, these studies were carried out using 5–100 μg samples. These ACs of the CTXs were later confirmed by total synthesis.

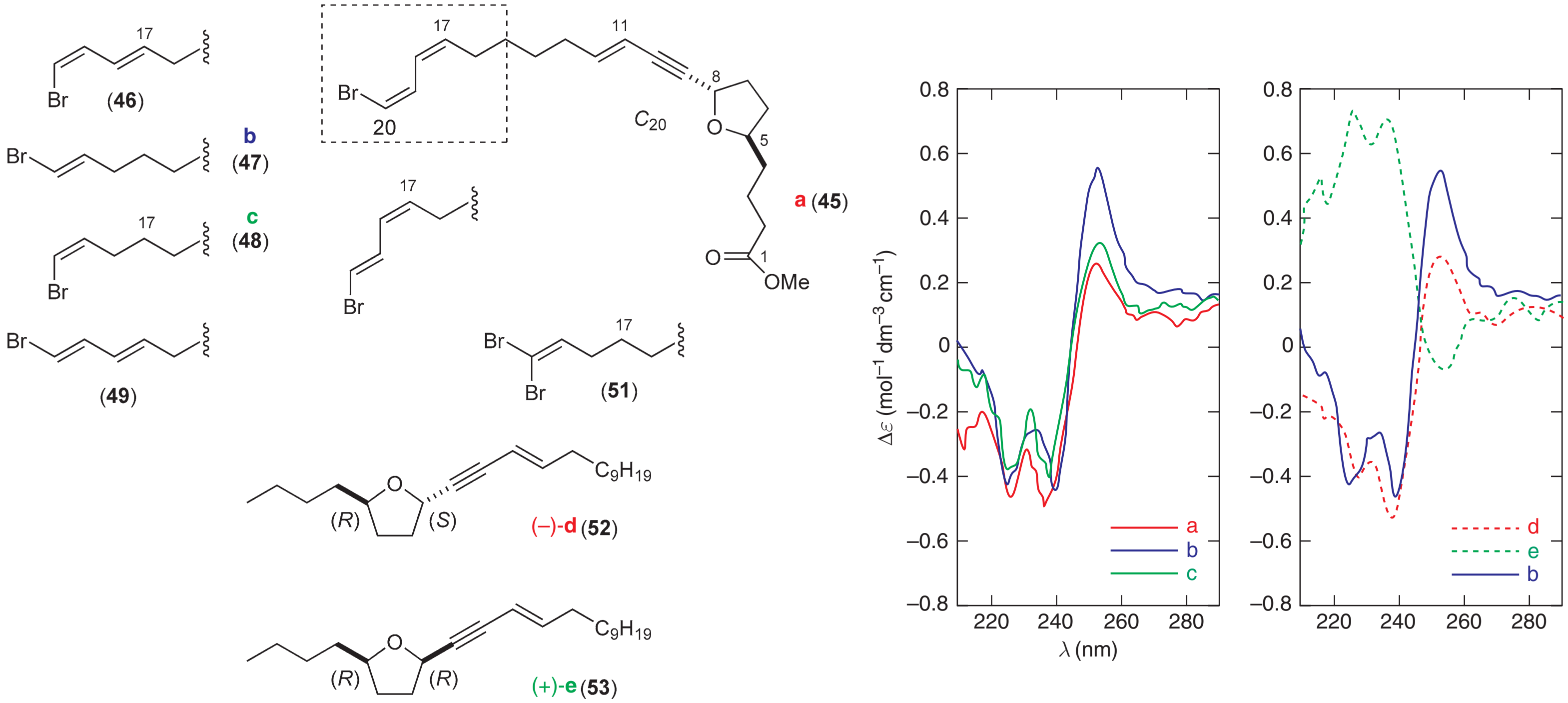
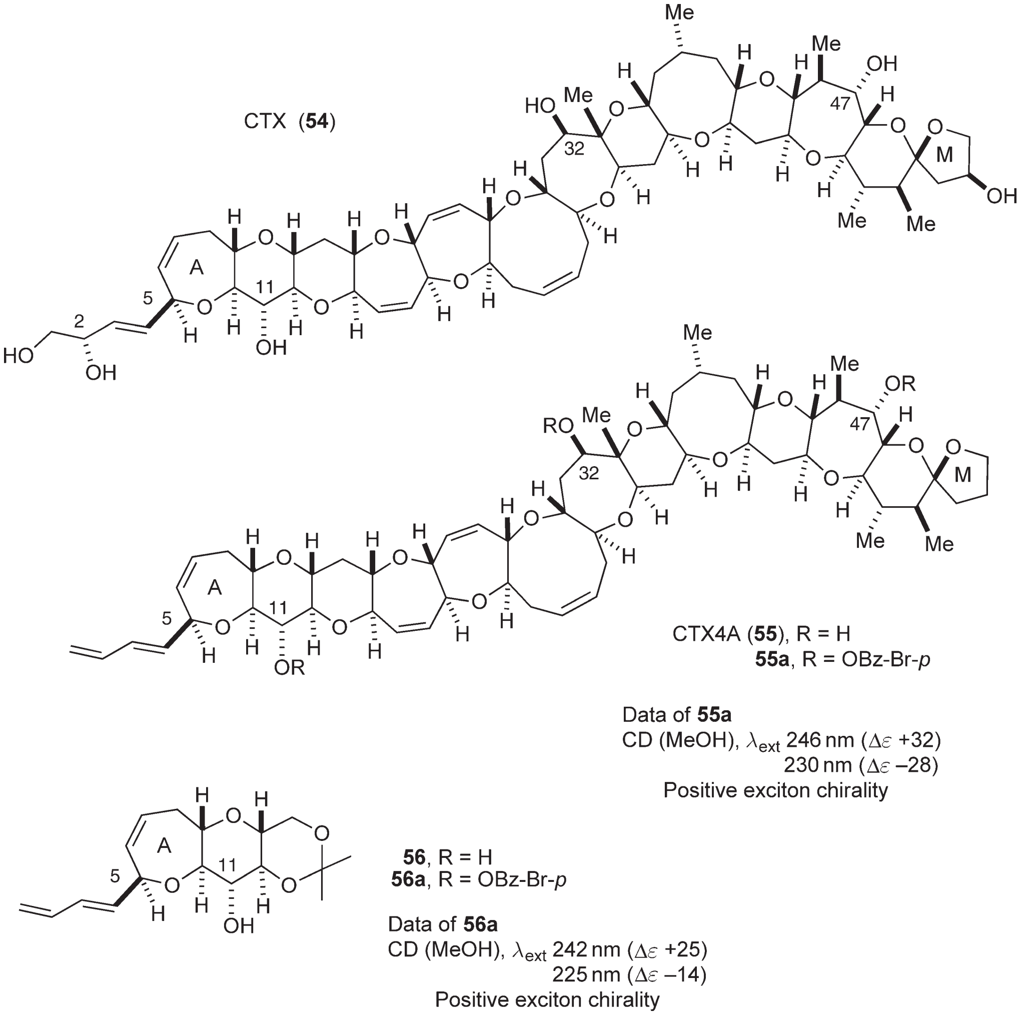


Figure 25 Mutafurans A–G **45–51** and CD spectra of compounds **a–e** at 25 °C in hexane. Redrawn from B. I. Morinaka; C. K. Skepper; T. F. Molinski, *Org. Lett.* **2007**, *9*, 1975–1978.



CTX (**54**)

CTX4A (**55**), R = H
55a, R = OBz-Br-*p*

Data of **55a**
CD (MeOH), λ_{ext} 246 nm ($\Delta\varepsilon$ +32)
230 nm ($\Delta\varepsilon$ –28)
Positive exciton chirality

56, R = H
56a, R = OBz-Br-*p*

Data of **56a**
CD (MeOH), λ_{ext} 242 nm ($\Delta\varepsilon$ +25)
225 nm ($\Delta\varepsilon$ –14)
Positive exciton chirality

Figure 26 Absolute configuration of ciguatoxin and CD data.

9.04.6.3 Determination of Absolute Configuration by the CD Exciton Chirality Method

9.04.6.3.1 Application of the CD exciton chirality method to acyclic 1,2-glycols

To determine the ACs of acyclic 1,2-glycols, the CD exciton chirality method has been applied to their dibenzoates or bis(2-anthroates), which show typical bisignate Cotton effects (see Section 9.04.4) as exemplified in Figures 27 and 28.[72,73] Acyclic dibenzoates or bis(2-anthroates) can rotate around the bond connecting two benzoate or 2-anthroate chromophores, and therefore the CD sign depends on the conformational equilibrium. From the data of many examples, general rules were derived as shown in Figures 27 and 28.

In the case of the diesters of a terminal 1,2-glycol, CD and AC are correlated as shown in Figure 27. For example, the diester 57 (bis(*p*-Br-benzoate)[72] or bis(2-anthroate)[73]) with the AC as shown adopts three rotational conformers 57A, 57B, and 57C, among which the conformer 57B is unstable because of two *gauche* relationships among three bulky groups. On the other hand, the conformers 57A and 57C have one *gauche* relationship between two bulky groups, respectively, and therefore, they are stable and dominant in the equilibrium. The stable conformer 57A has a positive exciton chirality between two chromophores, while in conformer 57C the two chromophores are in a *trans*-relationship, and therefore, no exciton chirality is generated. Thus the CD spectrum of diester 57 reflects a positive exciton chirality of conformer 57A. The ^{1}H NMR coupling constants (J(*trans*) = 6.8–8.4 Hz, J(*gauche*) = 3.6 Hz) support this conclusion. The CD

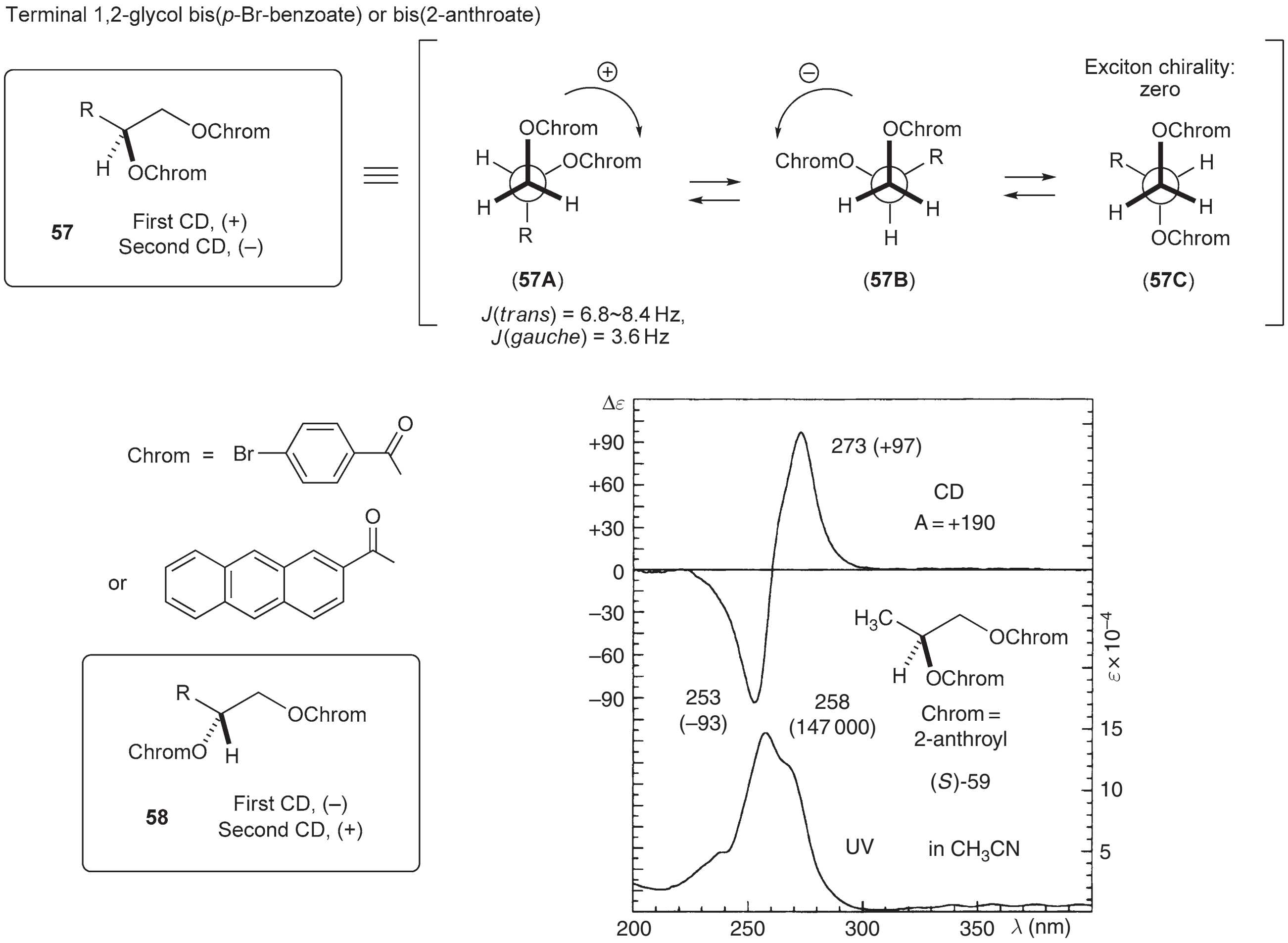


Figure 27 Applications of the CD exciton method to acyclic terminal 1,2-glycols. Redrawn from I. Akritopoulou-Zanze; K. Nakanishi; H. Stepowska; B. Grzeszczyk; A. Zamojski; N. Berova, *Chirality* **1998**, *9*, 699–712.

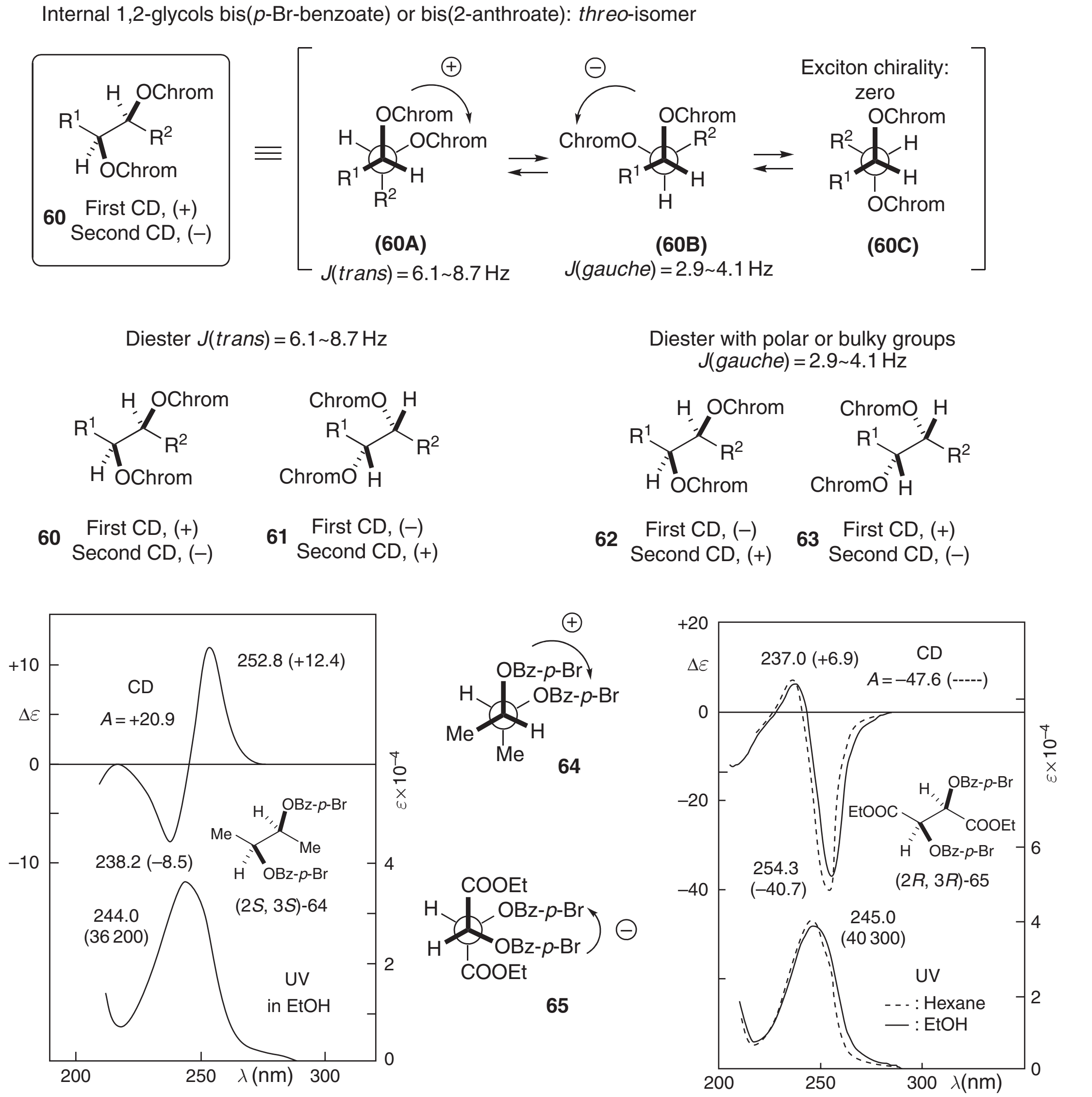


Figure 28 Applications of the CD exciton method to acyclic internal 1,2-glycols with *threo*-configuration. Redrawn from N. Harada; A. Saito; H. Ono; S. Murai; H.-Y Li; J. Gawronski; K. Gawronska; T. Sugioka; H. Uda, *Enantiomer* **1996**, *1*, 119–138.

spectrum of (*S*)-1,2-propanediol bis(2-anthroate) **59** shows very intense exciton Cotton effects, from which the AC of this compound could be assigned (**Figure 27**).[73] If a terminal 1,2-diester adopts the opposite AC as shown in **58**, the opposite CD is obtained. Thus the AC of terminal 1,2-glycols can be determined by the CD exciton chirality method.

In a similar manner, the CD exciton chirality method is applicable to internal 1,2-glycols with *threo*-configuration (**Figure 28**).[72,73] The exciton chirality between two chromophores depends on the rotational conformation. For example, the diester **60** (bis(*p*-Br-benzoate) or bis(2-anthroate)) with the AC as shown adopts three rotational conformers **60A, 60B, and 60C**, among which the conformers **60B and 60C** are unstable because of three *gauche* relationships among four bulky groups. On the other hand, the conformer **60A** has two *gauche* relationships between bulky groups, and hence it is stable and dominant in the equilibrium. The conformers **60A and 60B** have positive and negative twists between two chromophores, respectively, while in the conformer **60C** two chromophores are in the *trans*-relationship, and therefore, no exciton chirality is generated. The ^{1}H NMR coupling constant ($\mathcal{J}$(*trans*) = 6.1 – 8.7 Hz) supports the preference of the conformer

60A. After all, the CD spectrum of diester **60** reflects a positive chirality of conformer **60A**. The CD spectrum of (2*S*,3*S*)-2,3-butanediol bis(*p*-Br-benzoate) **64** shows a positive exciton couplet, from which the AC of this compound could be assigned (**Figure 28**).[72] If an internal 1,2-glycol has the opposite AC, the opposite CD Cotton effects are observed as shown in **61**.

The above relationship between the AC and the exciton CD Cotton effects holds for most internal 1,2-glycols. However, if a glycol has polar or extremely bulky groups (R^1 and R^2), the conformational equilibrium is changed. In such a case, the two polar or extremely bulky groups R^1 and R^2 adopt a *trans*-relationship to diminish the electric repulsive force or steric repulsion, and therefore the conformer **60B** becomes dominant. The preference of the conformer **60B** is supported by the ^{1}H NMR coupling constant ($\mathcal{J}$(*gauche*) $= 2.9-4.1$ Hz). The CD spectrum of (2*R*,3*R*)-diethyl tartrate bis(*p*-Br-benzoate) **65**, in which the two polar ethyl ester groups adopt a *trans*-relationship, shows a negative exciton couplet reflecting the preference of the conformer **60B**.[72] Thus the AC of terminal 1,2-glycols can be determined by the CD exciton chirality method in conjunction with ^{1}H NMR analysis.

If the groups R^1 and R^2 are identical, the ^{1}H NMR vicinal coupling constant between two methine protons cannot be obtained from the routine NMR spectrum because of the same chemical shift. In such a case, the ^{1}H NMR^{13}C satellite band method is useful to determine the $\mathcal{J}_{\text{vic}}$ value.[72,74]

In the case of *erythro*-1,2-glycols, the determination of AC is more difficult. If the two groups R^1 and R^2 are identical, the glycol is a *meso*-isomer and hence achiral. If they are different, the glycol is chiral. In general, the exciton CD Cotton effects of *erythro*-diester are weak and depend on the equilibrium of the rotational conformations. Therefore, the assignment of ACs needs the further conformational analysis by other methods, for example, nuclear overhauser effect (NOE).[73]

The AC of 1,3-glycols can also be assigned in a similar manner.[75–77]

9.04.6.3.2 Absolute configuration of urothion

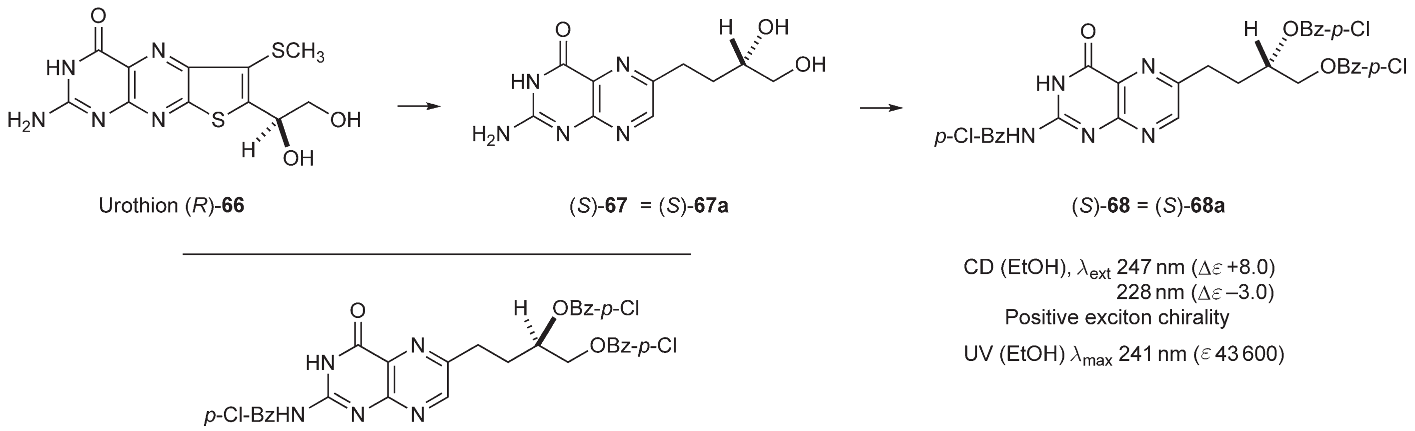
To determine the AC of urothion **66**, a yellowish pteridine pigment isolated from human urine, the compound was subjected to desulfurization with Raney-Ni yielding a product **67**, which was converted to tris(*p*-Cl-benzoate) **68**[78] (**Figure 29**). On the other hand, authentic samples of (*S*)-**67a** and (*R*)-**67b** were synthesized starting from D-glucose. Since the $[\alpha]_D$ values of **67, 67a, and 67b** were too small to assign their ACs by comparison, tris(*p*-Cl-benzoates) **68, 68a, and 68b** were prepared and their CD spectra compared. The CD spectrum of **68** agreed with that of (*S*)-**68a**, and therefore, the AC of urothion **66** was determined to be *R*. The bisignate Cotton effects at 247 and 228 nm originate mainly from the exciton coupling between the two benzoate groups in the side chain. According to the exciton chirality method applied to acyclic 1,2-glycols (Section 9.04.6.3.1), the positive sign of the first Cotton effect leads to the *S* configuration, which agrees with that obtained by comparison of CD spectra.

9.04.6.3.3 Absolute configuration of cephalocyclidin A, a five-membered ring cis-α-glycol

The unprecedented pentacyclic structure of cephalocylidin A **69**, a cephalotaxus alkaloid, was elucidated on the basis of X-ray crystallography, ^{1}H-NMR, and CD analysis[79] (**Figure 30**). The presence of a tetra-substituted benzenoid ring in the intact cephalocylidin does not allow assignment of AC from CD. However, upon *p*-methoxycinnamolylation of the secondary hydroxyl groups at 2-C, 3-C, the derivative **70** provided a useful, though rather weak bisignate CD band associated with a negative exciton coupling due to the small dihedral angle between the cinnamate chromophores. The CD couplet, sufficiently removed from other aromatic transitions, allowed for a straightforward assignment of the (2*R*,3*S*) AC, and eventually of the remaining four stereogenic centers, which were determined by taking into account the known relative configurations from NMR and X-ray analysis.

9.04.6.3.4 Absolute configuration of dihydro-β-agarofuran sesquiterpene

A screening program of South American medicinal plants for drugs resistant to parasites yielded a number of dihydro-β-agarofuran sesquiterpenes from the roots of *Maytensus magellanica*.[40] Because of their unique ability to block P-glycoprotein exporter activity, these compounds are considered to be privileged structures (**Figure 31**). Compound **13** (see **Figure 15**), the most active of this series, is representative of these new sesquiterpenes that were isolated based on their activity against a multidrug-resistant strain of *Leishmania tropica*



Urothion (*R*)-**66**

(*S*)-**67** = (*S*)-**67a**

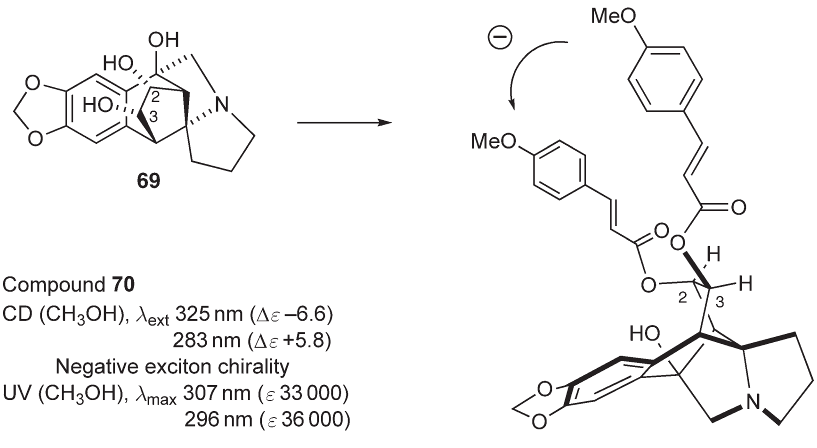
(*S*)-**68** = (*S*)-**68a**

CD (EtOH), λ_{ext} 247 nm ($\Delta\varepsilon$ +8.0)
228 nm ($\Delta\varepsilon$ –3.0)
Positive exciton chirality

UV (EtOH) λ_{max} 241 nm (ε 43 600)

(*R*)-**68b** CD of negative exciton chirality

Figure 29 Urothion and CD data.

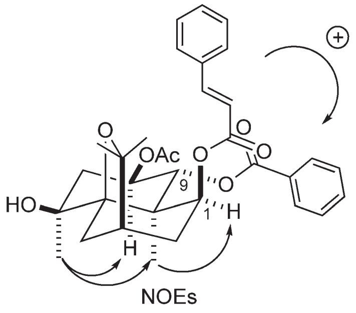


69

Compound **70**
CD (CH_3OH), λ_{ext} 325 nm ($\Delta\varepsilon$ −6.6)
283 nm ($\Delta\varepsilon$ +5.8)
Negative exciton chirality
UV (CH_3OH), λ_{max} 307 nm (ε 33 000)
296 nm (ε 36 000)

70

Figure 30 Absolute configuration of cephalocyclidin A 69 as determined by CD exciton chirality method.



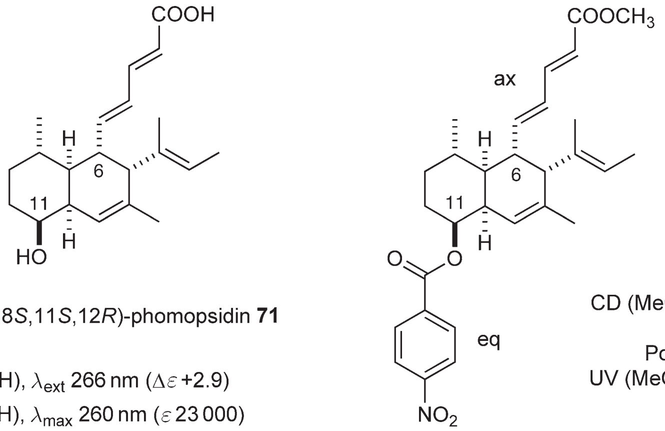
CD (CH_3CN), λ_{ext} 270.7 nm ($\Delta\varepsilon$ +20.3)
227.8 nm ($\Delta\varepsilon$ −18.1)
Positive exciton chirality

Figure 31 Dihydro-β-agarofuran sesquiterpene 13 and CD data.

overexpressing a P-glycoprotein-like exporter. As seen in the structure, 13 contains cinnamate and benzoate esters on carbons 1 and 9 in an ideal 1–3 relationship for determining their AC by taking advantage of the sign of the anticipated exciton couplet in its CD spectrum. In other words, the long axes of the transition dipole moments of these esters, depending on their absolute stereochemical relationship, will describe either a right- or a left-handed twist as evidenced by the sign of the exciton couplet in its CD spectrum. In the event, the CD shows a clear and positive exciton couplet at around 270 and 226 nm, respectively. Thus, this positive right-handed relationship defines the absolute stereochemistry of this family of sesquiterpenes.

9.04.6.3.5 Absolute configuration of phomopsidin

The CD spectrum of phomopsidin 71, a marine-derived fungal metabolite shows only one very weak Cotton effect at 266 nm associated with the diene-carboxylic acid chromophore at C-6 with a moderate UV absorption at 266 nm[80] (**Figure 32**). The Cotton effect due to the absorption of the two isolated double bonds below 200 nm were difficult to measure. Since the observed single Cotton effect was unsuitable for a determination of AC, the phomopsidin methyl ester was subjected to esterification with *p*-nitrobezoyl chloride. As expected the CD spectrum of the corresponding 11-*p*-nitrobezoate derivative 72 exhibited a clear-cut positive exciton couplet arising from a through-space interaction between the dienoate and *p*-nitrobezoate chromophores, whose electric transition moments and twisted axial/equatorial orientation, respectively, fit well the requirements for a nondegenerate exciton coupling.



(6*S*,7*S*,8*S*,11*S*,12*R*)-phomopsidin **71**

CD (MeOH), λ_{ext} 266 nm ($\Delta\varepsilon$ +2.9)
UV (MeOH), λ_{max} 260 nm (ε 23 000)

p-Nitrobenzoate **72**
CD (MeOH), λ_{ext} 271 nm ($\Delta\varepsilon$ +22.4)
λ_{ext} 244 nm ($\Delta\varepsilon$ −7.9)
Positive exciton chirality
UV (MeOH), λ_{max} 264 nm (ε 35 000)

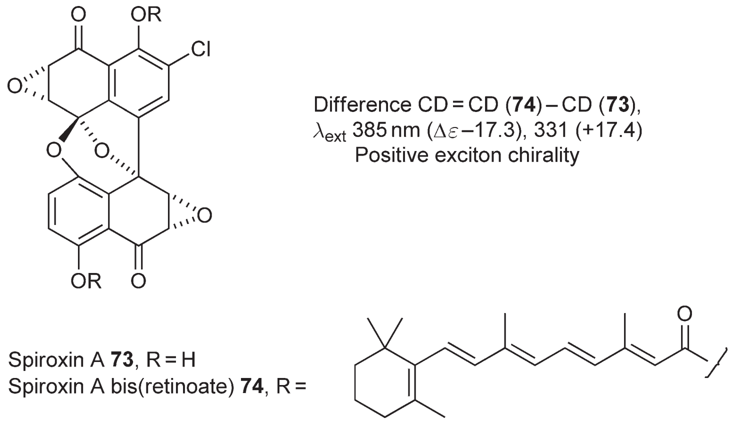
Figure 32 CD and UV data of phomopsidin 71 and phomopsidin methyl ester *p*-nitrobezoate 72.

9.04.6.3.6 Absolute configuration of spiroxin A, a bis-acetophenone fungal metabolite

Spiroxin A 73 is the major component of a group of metabolites isolated from fermentations of the fungus LL-37H248[81] (Figure 33). It is a bis-acetophenone with a spiroketal grouping at carbon 4 that locks the two conjugated chromophores in either a right- or a left-handed twist. The relative stereochemistry had previously been established by NMR. The CD spectrum of spiroxin A 73, however, exhibits a complex CD in the 200–280 nm region, which was difficult to interpret. Thus, the AC could not be assigned from the CD of the intact molecule. To resolve this issue, esterification of the two phenolic hydroxyls with retinoic acid was carried out because all-*trans* retinoic acid methyl ester has a λ_{max} 356 nm (ε 39 500) and is well red-shifted from the absorption of the existing chromophore. Furthermore, the transition dipole moment is aligned parallel to the all-*trans* polyene providing for potentially clear interpretation of the CD of the two interacting retinoate chromophores. Microscale derivatization gave access to spiroxin A bis(retinoate) 74 whose CD showed no clear exciton couplet in the retinoic acid region. Thus the CD of spiroxin A 73 itself was subtracted from that of the bis(retinoate) 74 to give the difference spectrum which then showed a clear, negative exciton couplet at λ_{ext} 385 nm ($\Delta\varepsilon$ −17.3) and 331 nm ($\Delta\varepsilon$ +17.4), which permits an unequivocal assignment of the twist as left-handed and thus the AC as shown.

9.04.6.3.7 Absolute configuration of pinellic acid

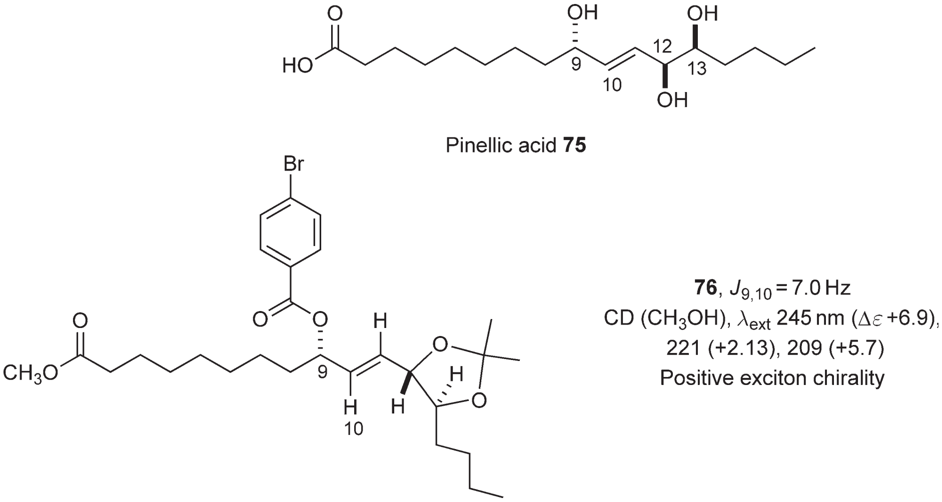
A useful application of the allylic benzoate CD method for determining the AC of allylic alcohols was that used for pinellic acid 75[82,83] (Figure 34). Pinellic acid was isolated from *Pinelliae tuber*, a component of Japanese



Difference CD = CD (**74**) − CD (**73**),
λ_{ext} 385 nm ($\Delta\varepsilon$ −17.3), 331 (+17.4)
Positive exciton chirality

Spiroxin A **73**, R = H
Spiroxin A bis(retinoate) **74**, R =

Figure 33 Spiroxin A 73 and CD data.



Pinellic acid **75**

76, $J_{9,10}$ = 7.0 Hz
CD (CH_3OH), λ_{ext} 245 nm ($\Delta\varepsilon$ +6.9),
221 (+2.13), 209 (+5.7)
Positive exciton chirality

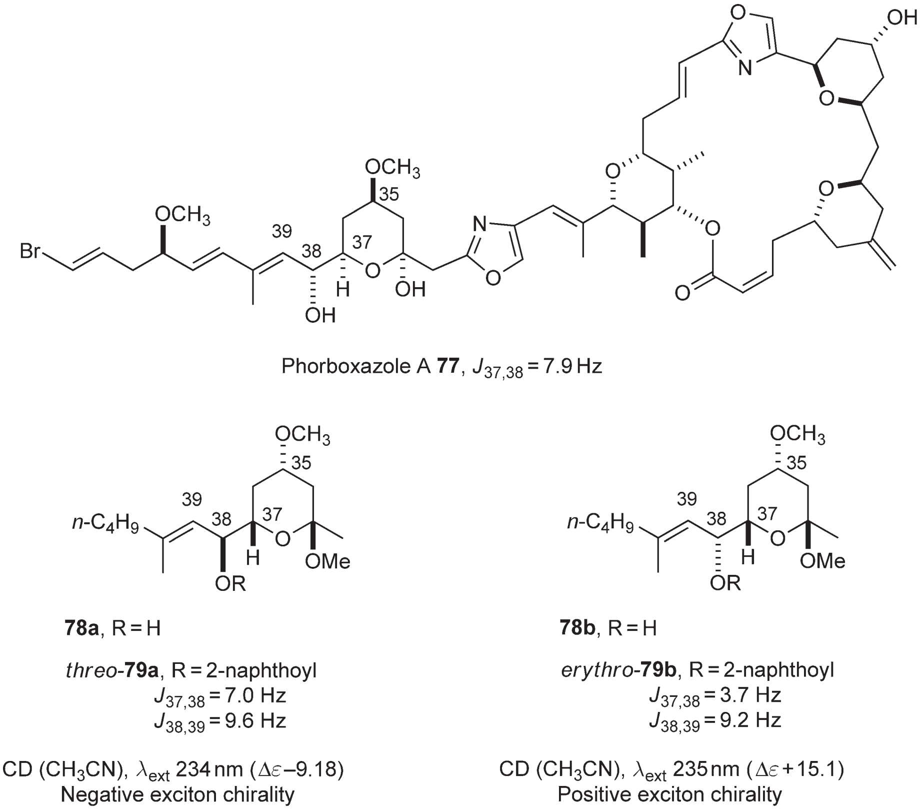
Figure 34 Pinellic acid **75** and exciton CD.

herbal medicine, and exhibits oral adjuvant activity for nasal influenza vaccine. The relative configuration of the three stereogenic centers at carbons 9, 12, and 13 were determined by NOE studies of the methyl ester of its acetonide. This established the *syn* configuration for the 12-C, 13-C vicinal diols. The acetonide was then converted to the *p*-bromo-benzoate **76**, the ^{1}H NMR spectrum of which indicated an antiperiplanar relationship between the 9 and 10 protons ($J_{9,10} = 7.0$ Hz). The CD of the bromo-benzoate allylic ester showed a positive Cotton effect at λ_{ext} 245 nm ($\Delta\varepsilon$ +6.97) indicative of the *S* configuration at 9-C. This predicted the AC of pinellic acid to be either (9*S*,12*S*,13*S*) or (9*S*,12*R*,13*R*). The question was resolved by a stereospecific synthesis of both isomers. Comparison of spectral data of the two synthetic preparations with those of the natural product indicated that the AC of the natural product was (9*S*,12*S*,13*S*) as shown.

9.04.6.3.8 Absolute configuration of phorboxazole

Phorboxazoles are marine natural products that exhibit strong cytostatic activity. The AC of phorboxazole A **77** was assigned as shown by total synthesis except for the configuration of 38-C[84] (**Figure 35**). The AC at the 38-C allylic alcohol had originally been assigned as *R* by application of the Mosher methoxy trifluoromethyl phenyl acetic acid (MTPA) method. However, there was an anomaly in the NMR $\Delta\delta$ data. To corroborate the assigned *R* configuration, the following CD studies were carried out. The *threo* and *erythro* model compounds **78a** and **78b** were synthesized by several steps from (*S*)-malic acid and the derived allylic alcohols converted to 2-naphthoate esters **79a** and **79b**. The NMR vicinal coupling constants of 37-H/38-H were observed to be $J_{37,38} = 7.0$ Hz for **79a** and $J_{37,38} = 3.7$ Hz for **79b**. The data for **79a** are similar to that of natural product **77**, $J_{37,38} = 7.9$ Hz, indicating that compound **77** has the same relative configurational relationship as **79a**.

With this relative configurational relationship established, examination of the CD spectrum of these two model compounds permitted the assignment of the 38-C ACs in **79a** and **79b**. The minor *threo* ester **79a** showed a strong negative Cotton effect at λ_{ext} 234 nm ($\Delta\varepsilon$ −9.2) indicating a negative twist between the esterified alcohol and the allylic double bond. In contrast, the major *erythro* product **79b** showed a similar CD to that of the *threo* compound **79a** except for the sign of the Cotton effect, λ_{ext} 234 nm ($\Delta\varepsilon$ +15.1) describing a positive helicity. These exciton couplets are ascribed to exciton interactions between the transition dipole moments of the ^1B_b band of the 2-naphthoate chromophore and the $\pi-\pi^*$ transition of the 39-C/40-C double bond. The NMR coupling constant between 38-H and 39-H was observed as $J_{38,39} = 9.6$ Hz for **79a** and $J_{38,39} = 9.2$ Hz for **79b** indicating that these two protons are in *trans*-relationship in their stable conformations. Based on these data, the AC for 38-C was determined to be *S* in *threo*-**79a** and *R* in *erythro*-**79b**. Because the two model compounds were derived from (*S*)-malic acid, they have opposite ACs to phorboxazole A between 33-C and 37-C. And because both phorboxazole A **77** and *threo*-**79a** have the same relative configurations at 33-C through 38-C, these analyses corroborated the originally assigned 38*R* configuration of the natural product.



Phorboxazole A **77**, $J_{37,38} = 7.9$ Hz

78a, R = H

threo-**79a**, R = 2-naphthoyl
$J_{37,38} = 7.0$ Hz
$J_{38,39} = 9.6$ Hz

CD (CH_3CN), λ_{ext} 234 nm ($\Delta\varepsilon$–9.18)
Negative exciton chirality

78b, R = H

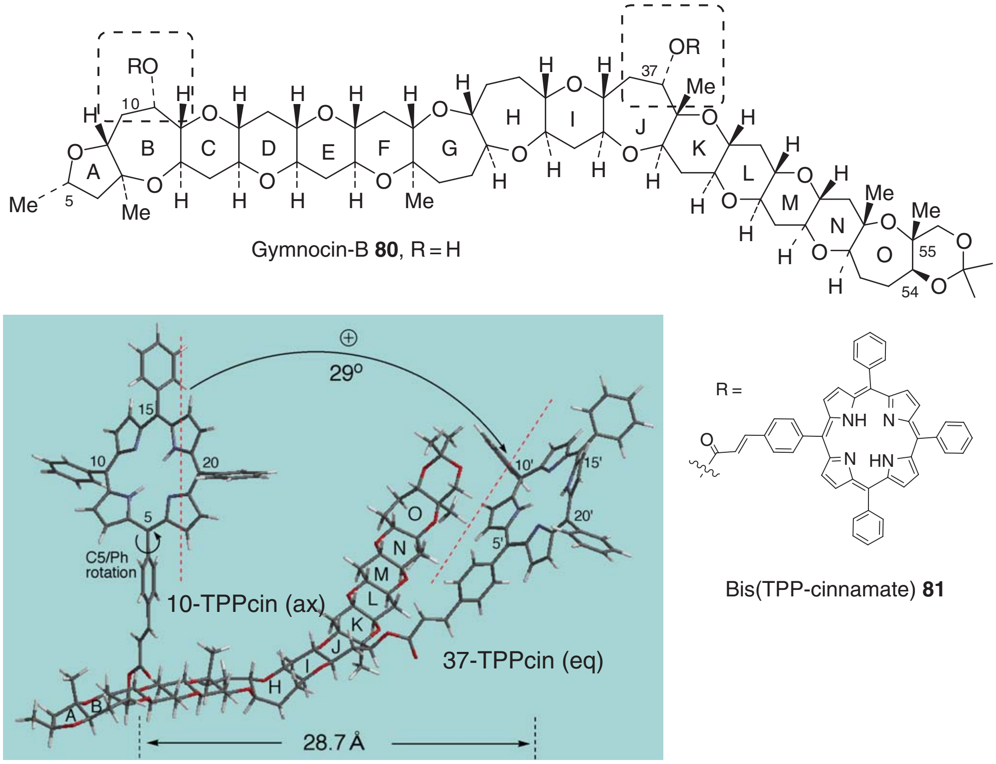
erythro-**79b**, R = 2-naphthoyl
$J_{37,38} = 3.7$ Hz
$J_{38,39} = 9.2$ Hz

CD (CH_3CN), λ_{ext} 235 nm ($\Delta\varepsilon$ + 15.1)
Positive exciton chirality

Figure 35 Phorboxazole A 77 and allylic benzoate method.

9.04.6.3.9 Absolute configuration of gymnocin-B

Several challenges were encountered during the course of the configurational assignments of gymnocin-B 80, a cytotoxic marine natural product containing the largest 15-polyether skeleton isolated so far[85] (**Figure 36**). Along with the conformational flexibility arising from the presence of five, seven-membered rings, investigation of the sterically hindered and remote critical hydroxyl groups at 10-C and 37-C in the B and J rings was especially difficult. In addition, the sample isolated from the red-tide dinoflagellate was available in extremely limited amount. On the basis of known relative configurations at all 31 stereogenic centers previously assigned by NMR, the formidable task of determining their ACs was achieved by direct CD analysis at the critical 10-C and 37-C secondary hydroxyl groups. The potent but bulky triphenylporphyrin-cinnamate chromophore was chosen and introduced into 10-OH and 37-OH by acryloylation/cross metathesis under microscale conditions. Owing to very intense UV–Vis porphyrin absorption, this gymnocin-B derivative 81 did show a clear-cut exciton split CD even though the two stereogenic centers bearing the porphyrins were approximately 30 Å apart. However, only after extensive conformational analysis of this derivative could the observed positive couplet be rationalized regarding the ACs at 10-C/37-C. For this purpose a conformational analysis by MMFF94s/Monte Carlo calculation was first carried out on a few truncated models and then finally on the entire gymnocin-B bis(triphenylporphyrin-cinnamate) derivative 81. This analysis permitted correlation of the positive inter-porphyrin twist in the preferred 10-axial and 37-equatorial TPP-cin conformations of the arbitrarily chosen (10*S*,37*S*)-configuration with the observed positive exciton-coupled CD. In addition, the Boltzman-weighted calculated CD by De Voe's coupled oscillator method was in full agreement with the experimental results.



Gymnocin-B **80**, R = H

Bis(TPP-cinnamate) **81**

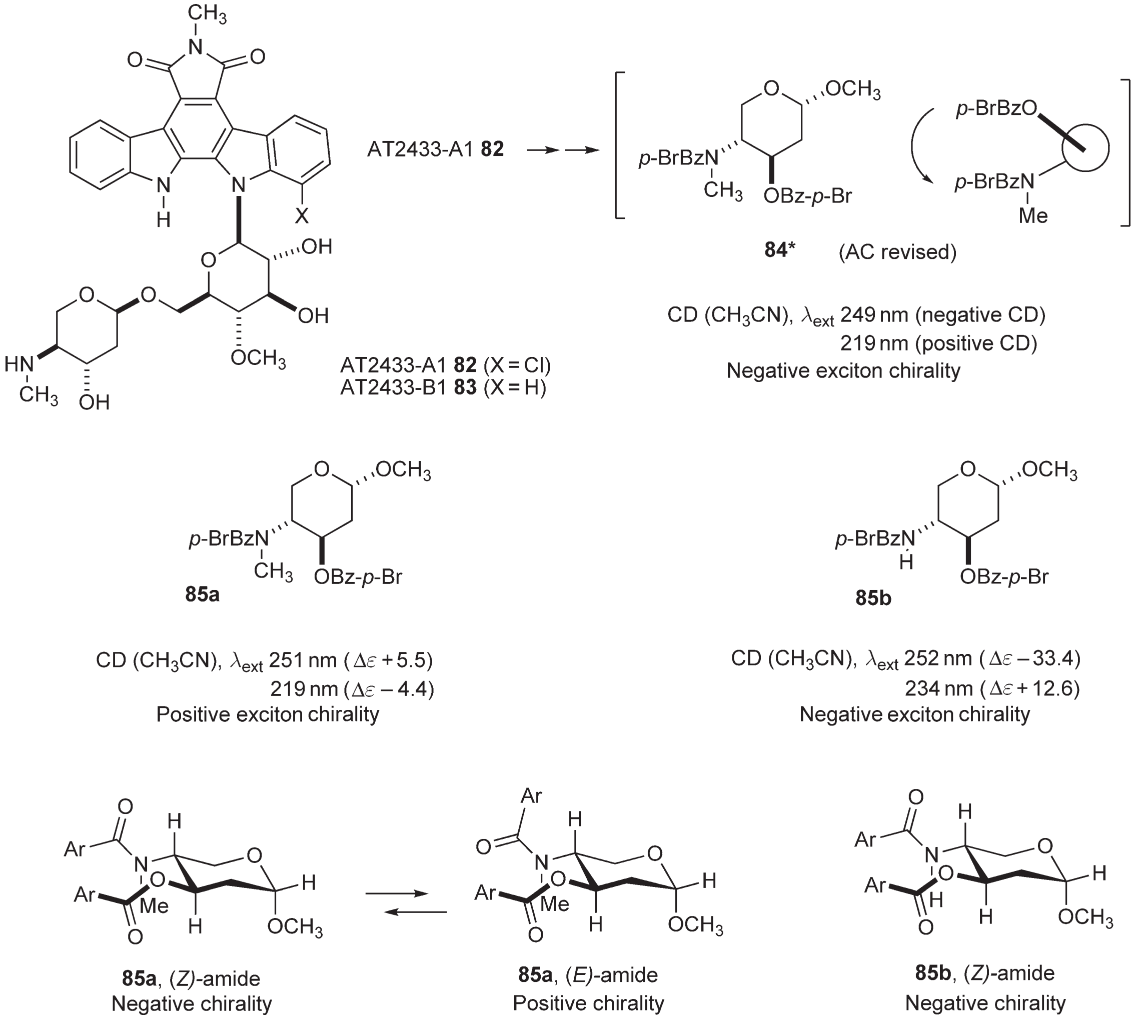
Figure 36 Gymnocin-B 80 and the lowest-energy conformation of its 10,37-bis(TPP-cinnamate) derivative 81 obtained by Monte Carlo/MMFF94s with Spartan 02. Experimental CD (in MeOH, $c = 3.0 \times 10^{-6}$): 419 nm ($\Delta\varepsilon$ +11), 414 nm ($\Delta\varepsilon$ −15); Boltzmann weighted (at 298 K) average CD calculated by De Voe's method: 420 nm ($\Delta\varepsilon$ +25), 414 nm ($\Delta\varepsilon$ −25). Redrawn from K. Tanaka; Y. Itagaki; M. Satake; H. Naoki; T. Yasumoto; K. Nakanishi; N. Berova, *J. Am. Chem. Soc*. **2005**, *127*, 9561–9570.

9.04.6.3.10 Absolute configuration of antitumor antibiotic AT2433-A1 containing a secondary amino group

To determine the AC of antitumor antibiotic AT2433-A1 82, amino sugar bis(*p*-Br-benzoyl) derivative **84*** was prepared from the natural product[86] (**Figure 37**). Its CD spectrum showed a negative couplet leading to the AC as shown. However, it was later found that this assignment was wrong as explained below. The authentic samples 85a and 85b were synthesized from a starting material with known AC. Surprisingly, the CD of 85a showed a weak positive exciton couplet, while that of 85b showed a strong negative one. The ^{1}H NMR of 85a, a benzamide derivative of the secondary amine, indicated the existence of (*Z*) and (*E*) amide isomers, which adopt negative and positive exciton chiralities, respectively. They cancel each other to some extent and the remaining CD is governed by the (*E*) amide. On the other hand, the CD of the primary amine derivative 85b reflects its AC in a straightforward way because of its (*Z*) conformation. Therefore, when the exciton chirality method is applied to secondary amines, the analysis of (*E*) and (*Z*) conformations is critical. The total synthesis of AT2433-B1 83 was carried out confirming the AC of the secondary amine.

9.04.6.3.11 Absolute configuration of chiral binaphthoquinones

The synthesis of (−)-8′-hydroxyisodiospyrin 86, a naturally occurring bi(naphthoquinone), was carried out as shown; the coupling of chiral compound 88 with bromide 89 yielded a product that was treated with MeI giving iodide 90 as crystals.[87] The AC of 90 was determined as shown by X-ray crystallography. Compound 90 was converted to binaphthalene 91, the CD of which shows typical and intense exciton-coupled Cotton effects as shown in **Figure 38**. From the positive sign of the first Cotton effect, an *S* configuration was assigned to 91. The oxidative demethylation and treatment with $AlCl_3$ furnished (*S*)-(+)-8′-hydroxyisodiospyrin 87, which was identified to be the enantiomer of natural product 86. The AC of the natural product was thus determined to be



AT2433-A1 **82** (X = Cl)
AT2433-B1 **83** (X = H)

84* (AC revised)

CD (CH_3CN), λ_{ext} 249 nm (negative CD)
219 nm (positive CD)
Negative exciton chirality

85a

CD (CH_3CN), λ_{ext} 251 nm ($\Delta\varepsilon$ +5.5)
219 nm ($\Delta\varepsilon$ − 4.4)
Positive exciton chirality

85b

CD (CH_3CN), λ_{ext} 252 nm ($\Delta\varepsilon$ −33.4)
234 nm ($\Delta\varepsilon$ + 12.6)
Negative exciton chirality

85a, (*Z*)-amide
Negative chirality

85a, (*E*)-amide
Positive chirality

85b, (*Z*)-amide
Negative chirality

Figure 37 Application of the exciton chirality method to secondary amine (numerical CD data were obtained from the spectrum reported in Chisholm *et al.*[86]).

(*R*)-(−)-86. The CD spectrum of (*S*)-(+)-87 shows two positive and one negative Cotton effects around 360–260 nm, but their $\Delta\varepsilon$ values are smaller than those of 91 and the CD curve deviated from the ideal pattern of the exciton coupling. Thus, to determine the ACs by the exciton method, it is important to select the most appropriate chromophores, that is, binaphthalene rather than binaphthoquinone as used in this case.

9.04.6.3.12 Absolute configuration of pre-anthraquinones

Atropisomeric pigments 92, 93, 94, and 95 were isolated from indigenous Australian toadstools belonging to the genus *Dermocybe* (**Figure 39**). The structures of these pigments were deduced by spectroscopic methods, and their ACs of atropisomerism were determined by CD spectra.[88] The CD spectrum of 93 shows intense negative first and positive second Cotton effects at 272 and 251 nm, respectively, and therefore the AC with negative helicity between two long axes of aromatic chromophores was assigned. The same helicity was assigned to 92 showing a similar CD curve. The AC at the 3′ position was determined by chemical correlation; the reductive cleavage of 93 yielded (*R*)-torosachrysone methyl ether with known AC. Pigments 94 and 95 are diastereomers of each other, but their CD curves are almost mirror images leading to opposite helicity between the two aromatic chromophores. The weak Cotton effects of 95 as compared to those of 93 reflect a smaller dihedral angle between the two aromatic chromophores in the rigid structure of 95.

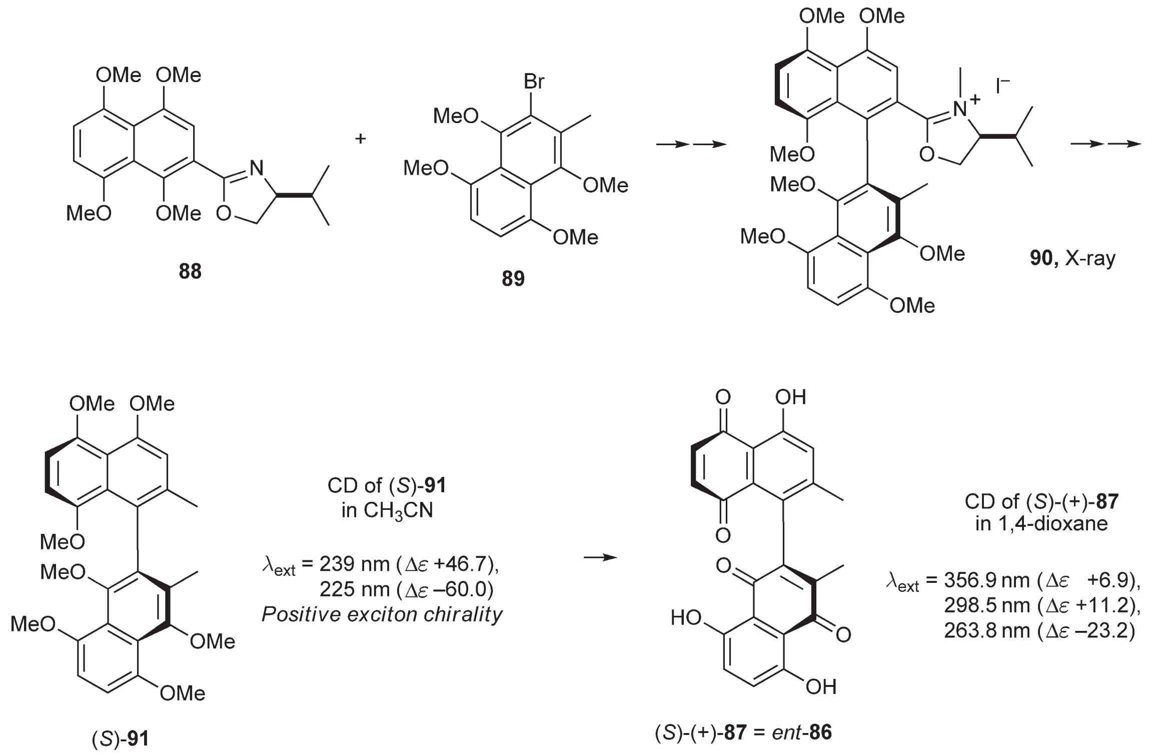


Figure 38 Synthesis and absolute configuration of axially chiral binaphthoquinones (numerical CD data were obtained from the spectra reported in Baker *et al.*[87]).

9.04.6.3.13 Absolute configuration of vinblastine

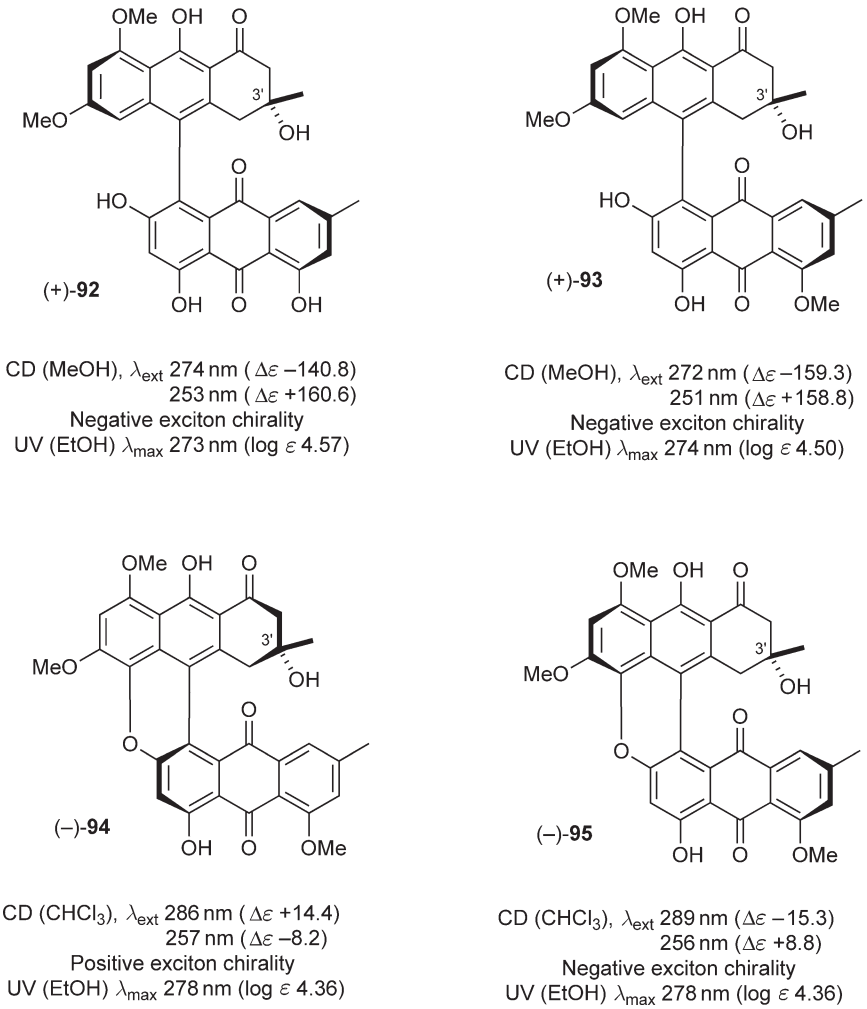
The AC of vinblastine **14**, one of the best-known *Vinca* alkaloids, was originally established by X-ray crystallography by Moncrief and Lipscomb.[89] CD experimental and theoretical studies have also been carried out for the purpose of determining the AC of vinblastine and its natural and synthetic analogues.[41,90]

The CD spectrum of vinblastine **14** consists of intrinsic CD bands associated with the isolated transitions in chiral cleavamine and indoline (‘half-molecules’) alone, together with the exciton CD due to the through-space interaction between the indole and vindoline chromophores (**Figure 40**). This exciton CD reflects the AC at the 16′-C stereogenic center of vinblastine **14**. Therefore, to obtain the net exciton CD, the intrinsic CD bands were subtracted from the CD of **14** giving ‘difference CD’, which showed an intense positive couplet around 220 nm (**Figure 40**(**b**)). This exciton CD is generated by the interaction between two ^1B_b transition moments of indole and indoline chromophores. The analysis of these CD spectra thus illustrates not only the validity of the general CD additivity rule but, importantly, it also reflects the positive exciton chirality and *S* AC at 16′-C.

9.04.6.4 Absolute Configurations by Theoretical Calculation of CD Spectra

9.04.6.4.1 Absolute configuration of a biflavone as determined by π-electron SCF-CI-DV MO

The AC of a natural biflavone atropisomer, (−)-4′,4‴,7′,7″-tetra-*O*-methylcupressuflavone **96**, has been determined to be *aR* (axial chirality, *R*) by theoretical calculation of its CD spectrum[91] (**Figure 41**). The π-electron system of biflavone **96** is strongly twisted to produce intense CD Cotton effects as shown. For the molecular structure of (*aR*)-**96** with a counterclockwise screw sense, UV and CD spectra were calculated by the π-electron SCF-CI-DV MO method. The calculated CD and UV curves are in excellent agreement with the observed spectra. Therefore, the AC of biflavone (−)-**96** was determined to be *aR*. The theoretically



(+)-**92**

CD (MeOH), λ_{ext} 274 nm ($\Delta\varepsilon$ −140.8)
253 nm ($\Delta\varepsilon$ +160.6)
Negative exciton chirality
UV (EtOH) λ_{max} 273 nm (log ε 4.57)

(+)-**93**

CD (MeOH), λ_{ext} 272 nm ($\Delta\varepsilon$ −159.3)
251 nm ($\Delta\varepsilon$ +158.8)
Negative exciton chirality
UV (EtOH) λ_{max} 274 nm (log ε 4.50)

(−)-**94**

CD ($CHCl_3$), λ_{ext} 286 nm ($\Delta\varepsilon$ +14.4)
257 nm ($\Delta\varepsilon$ −8.2)
Positive exciton chirality
UV (EtOH) λ_{max} 278 nm (log ε 4.36)

(−)-**95**

CD ($CHCl_3$), λ_{ext} 289 nm ($\Delta\varepsilon$ −15.3)
256 nm ($\Delta\varepsilon$ +8.8)
Negative exciton chirality
UV (EtOH) λ_{max} 278 nm (log ε 4.36)

Figure 39 Pre-anthraquinones and CD data.

determined AC of this biflavone was later confirmed by total synthesis of the natural product (−)-96. This theoretical approach should be a promising tool for determination of the AC of various natural products with a twisted π-electron system.

9.04.6.4.2 Absolute configurations of naturally occurring dihydroazulene and marine natural product halenaquinol as determined by SCF-CI-DV MO

The AC of a naturally occurring (+)-1,8a-dihydro-3,8-dimethylazulene 97 was similarly determined by π-electron SCF-CI-DV MO (Figure 42).[92] Dihydroazulene 97 shows intense CD Cotton effects reflecting its twisted π-electron system. The CD and UV spectra of a model compound (8a*S*)-98 were calculated giving CD data as shown in Figure 42, which were similar in position and sign to those of the natural product (+)-97. Therefore, the AC of (+)-97 was theoretically determined to be 8a*S*. To verify this theoretical determination in an experimental manner, a model compound (8a*S*)-(+)-99, which has a methyl group at the angular position and therefore is inert toward the oxidation to azulene, was synthesized. The observed CD data of (8a*S*)-(+)-99 were also similar to those of (+)-97. Therefore, the 8a*S* AC of (+)-97 was established and it was proved that the π-electron SCF-CI-DV MO method gives a correct AC.[92]

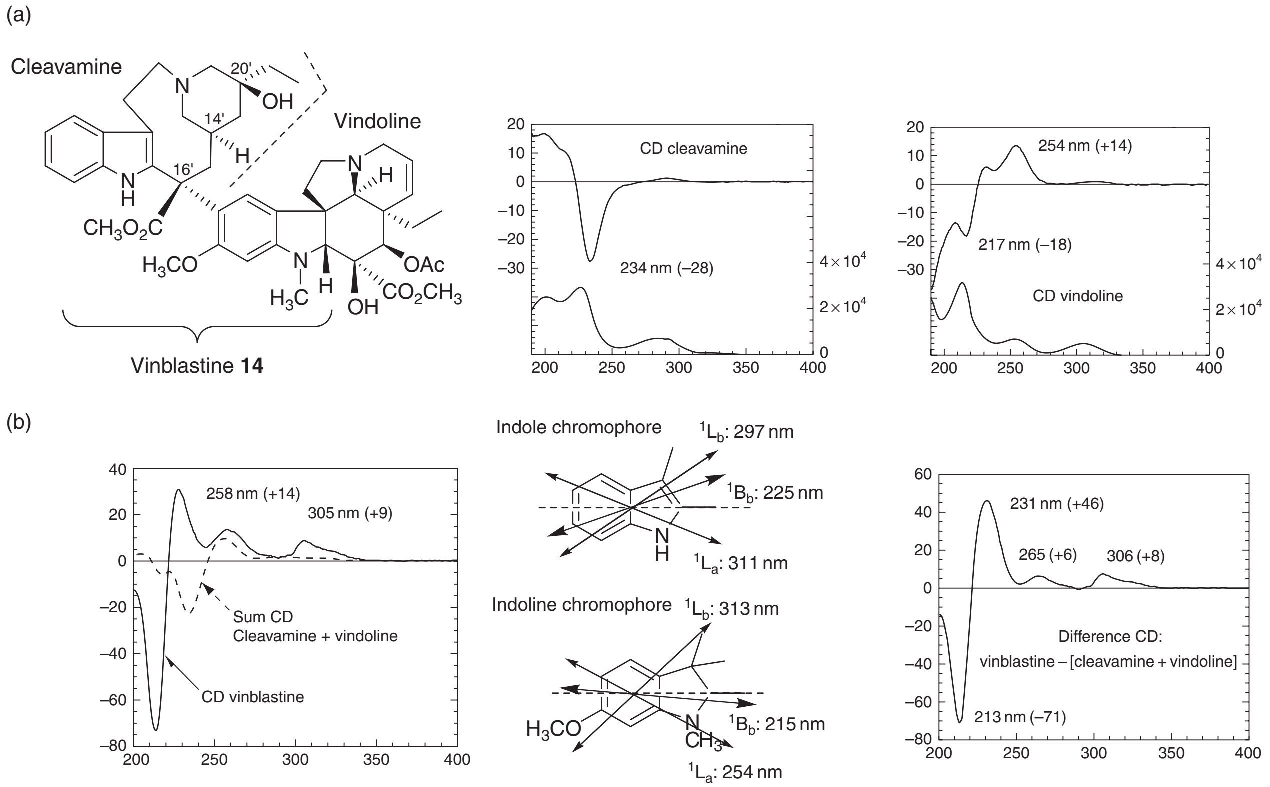
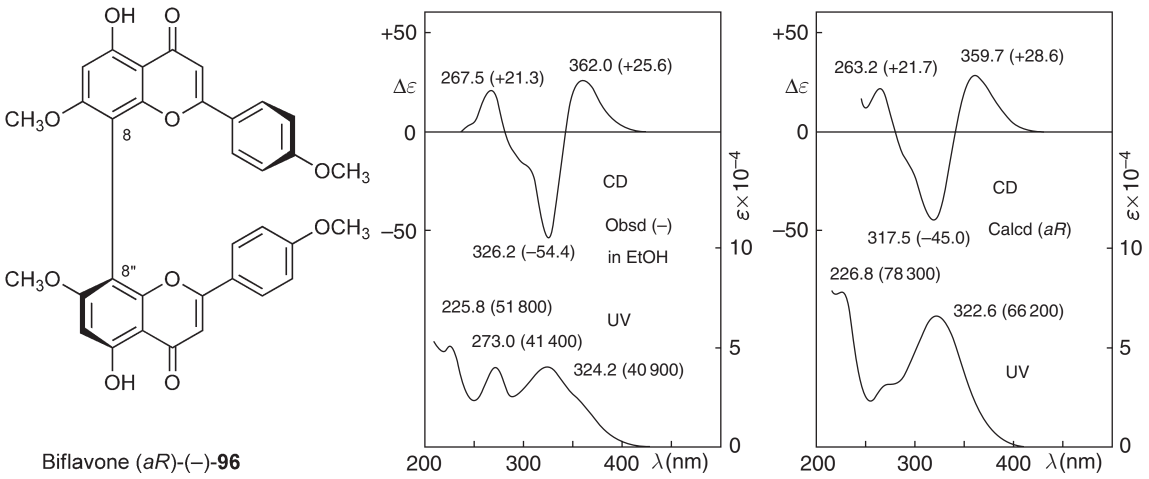
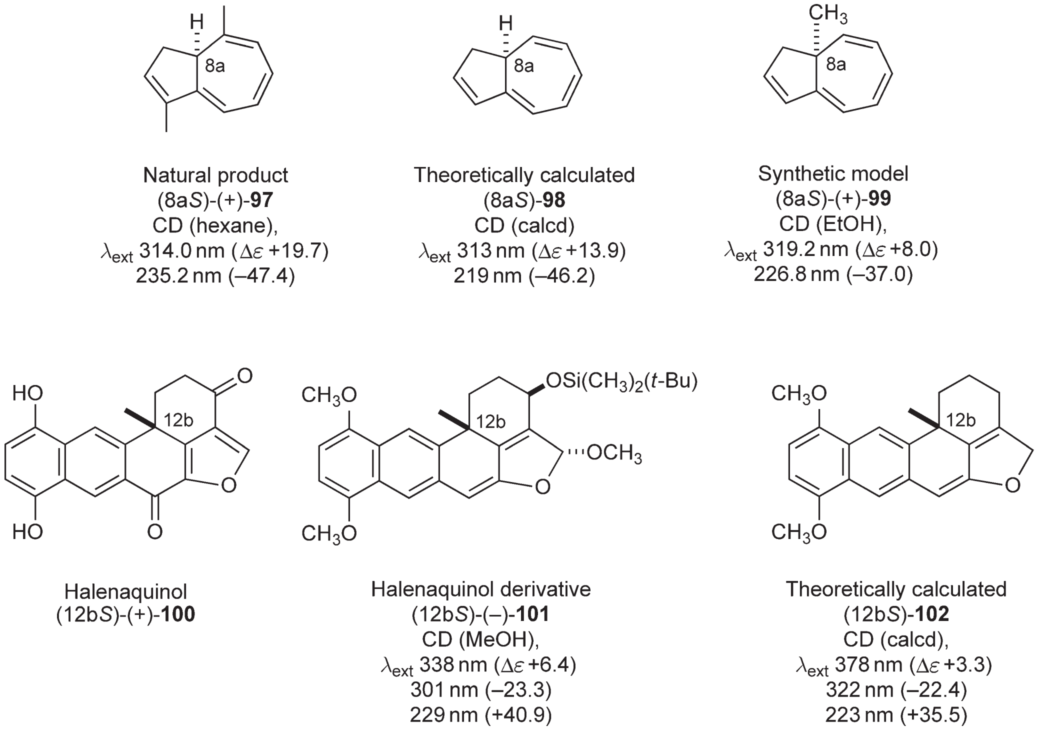


Figure 40 (a) Vinblastine 14 consists of two half-molecules, cleavamine and vindoline: the CD spectra of the component molecules. (b) Left: The sum CD spectrum (dotted line) = CD(cleavamine) + CD(vindoline) and CD spectrum of vinblastine: Center: Electric transition moments of indole and indoline chromophores; Right: Difference CD = CD(vinblastine) − [CD(cleavamine) + CD(vindoline)]. Redrawn from C. A. Parish; J.-G. Dong; W. G. Bornmann; J. Chang; K. Nakanishi; N. Berova, *Tetrahedron* **1998**, *54*, 15739–15758.



Biflavone (*a*R)-(–)-**96**

Figure 41 Absolute configuration of atropisomer, biflavone **96**, as determined by CD calculation. Redrawn from N. Harada; H. Ono; H. Uda; M. Parveen; N. U.-D. Khan; B. Achari; P. K. Dutta, *J. Am. Chem. Soc.* **1992**, *114*, 7687–7692.



Natural product
(8a*S*)-(+)-**97**
CD (hexane),
λ_{ext} 314.0 nm (Δε +19.7)
235.2 nm (–47.4)

Theoretically calculated
(8a*S*)-**98**
CD (calcd)
λ_{ext} 313 nm (Δε +13.9)
219 nm (–46.2)

Synthetic model
(8a*S*)-(+)-**99**
CD (EtOH),
λ_{ext} 319.2 nm (Δε +8.0)
226.8 nm (–37.0)

Halenaquinol
(12b*S*)-(+)-**100**

Halenaquinol derivative
(12b*S*)-(–)-**101**
CD (MeOH),
λ_{ext} 338 nm (Δε +6.4)
301 nm (–23.3)
229 nm (+40.9)

Theoretically calculated
(12b*S*)-**102**
CD (calcd),
λ_{ext} 378 nm (Δε +3.3)
322 nm (–22.4)
223 nm (+35.5)

Figure 42 Dihydroazulene and halenaquinol compounds with a twisted π-electron system and their CD data.

The ACs of marine natural products, halenaquinol 100 and related compounds were also determined by π-electron SCF-CI-DV MO (Figure 42).[93] Halenaquinol 100 has a twisted π-electron system, but its CD Cotton effects are weak. On the other hand, halenaquinol derivative (−)-101 showed intense CD Cotton effects as listed in Figure 42. Therefore, the π-electron system of this compound was selected for theoretical calculation. The CD spectrum of the model compound (12b*S*)-102 was calculated giving the data as shown in Figure 42. Although the wavelength position of the calculated Cotton effects deviated from those of the observed values, the basic pattern of CD spectrum was well reproduced by the calculation. Therefore, the ACs of (−)-101 and (+)-100 were determined to be 12b*S*. This AC was later confirmed by the total synthesis of halenaquinol (+)-100 and related compounds.[93]

9.04.6.4.3 TDDFT calculation of ECD of β-lactam antibiotics

Recently, J. Frelek and coworkers proposed an empirical helicity rule relating the configuration of the bridgehead carbon atom in clavams and oxacephams to the sign of observed Cotton effect at approximately 220–240 nm[94–97] (**Figure 43**). The rule was established empirically on the basis of X-ray data and a tentative assignment of the electronic transition at 220 (oxacephams) and 240 nm (clavams) to an n,π^*-amide transition in the azetidinone system.

According to this rule, which was found experimentally to be correct for a variety of oxacephams[94,95] and clavams,[96,97] a positive sign of the 220 nm Cotton effect corresponds to the (*R*)-AC at the bridgehead carbon atom whereas a negative sign indicates the (6*S*)-AC. The rationale for this rule relied on an assumption for conformational rigidity of the bicyclic system and localization of most of the molecular excitations within the amide chromophore. Having in mind the long-recognized utility of the β-lactam helicity rule, it is not surprising that this rule was one of the first to prompt quantum mechanical investigations on its validity. In a recent study, Frelek *et al.* confirmed by time-dependent density functional theory (TDDFT) calculations the validity of the β-lactam helicity rule to a series of clavams.[98,99] Furthermore, by using for the first time a combination of TDDFT calculations with full quantum mechanical Born–Oppenheimer molecular dynamics, Frelek *et al.* were able to show a surprisingly high sensitivity of CD to molecular conformations of cephams and their carba and oxa analogues.[98,99]

9.04.6.4.4 TDDFT calculation of ECD of quadron and related compounds

The study of four sesquiterpenes, quadron **110**, suberosenone **111**, suberosanone **112**, and suberosenol A acetate **113**, represents the first attempt to apply the *ab initio* DFT methodology for simulations of three different chiroptical properties, namely, OR and electronic and vibrational CD (ECD and VCD) for the purpose of determining the ACs of natural products[100] (**Figure 44**). For example, in the case of quadron **110**, the same AC is obtained from all three chiroptical properties, which leads to an AC of the highest reliability. It is to be noted that quadron belongs to a molecular type where the establishment of the AC is safely assigned using only the ECD

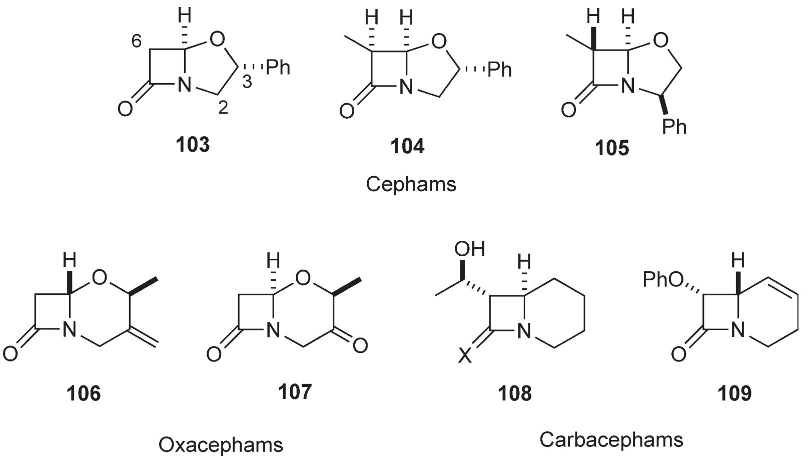
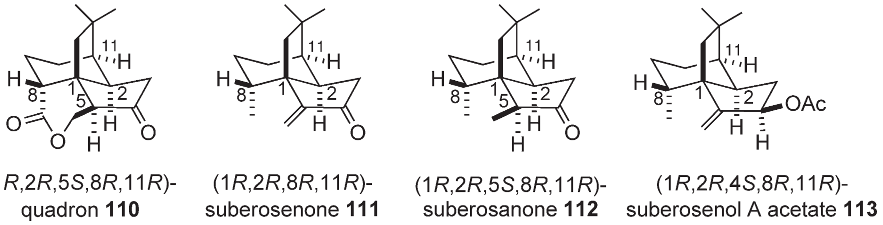


Figure 43 β-Lactam antibiotics.



(1*R*,2*R*,5*S*,8*R*,11*R*)-
quadron 110
(1*R*,2*R*,8*R*,11*R*)-
suberosenone 111
(1*R*,2*R*,5*S*,8*R*,11*R*)-
suberosanone 112
(1*R*,2*R*,4*S*,8*R*,11*R*)-
suberosenol A acetate 113
OAc

Figure 44 Quadron and related compounds.

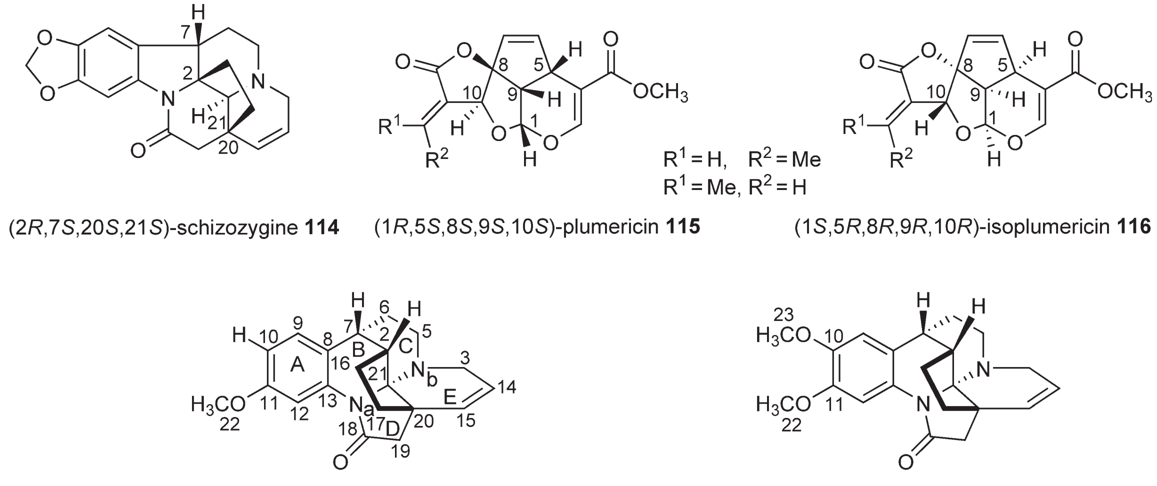
method. In such types the low-energy transitions, such as the carbonyl group $n-\pi^*$ transition, are limited in number and density, therefore they are well resolved experimentally. However, in general, when the assignment of AC is the main goal, the more experimental chiroptical data for the substrate that are available, the better the selection of suitable theoretical method(s). For example, the assignment of the AC of suberosenone **111**, suberosanone **112**, and suberosenol A acetate **113** was made on the basis of calculated OR values and comparison with the only currently available experimental ORs. It would have greatly benefited the AC assignments if ECD and VCD data had been provided.

9.04.6.4.5 TDDFT concerted calculation of CD, VCD, and OR of schizozygine, plumericin, and related compounds

The alkaloid schizozygine **114**, and the iridoids plumericin **115** and isoplumericin **116** are natural products where the concerted application of two or three chiroptical methods has led to more reliable assignment of ACs by TDDFT (**Figure 45**). Owing to the presence of multisignate CD bands in the CD spectrum of schizozygine **114** and an imperfection of the B3LYP functional, the calculated ECD alone did not permit a safe AC assignment.[101] Therefore it required additional support by OR and VCD data. Plumericin **115** and isoplumericin **116** also exhibited trisignate CD spectra that precluded the sole application of the ECD method. The unequivocal assignment of their ACs was made after the experimental VCD spectra were compared with the calculated data.[102] A recent study by Stephens *et al.* on iso-schizogaline **117** and iso-schizogamine **118** provides further insights on the difficulties encountered in some cases to assign the AC only on the basis of TDDFT simulations of ECD data.[103] According to the authors, when the chiral molecules contain a substantial density of low-energy electronic states, this may give rise to electronic excitations in the near-UV that in turn will prevent the resolution and assignment of individual transitions. In such cases the assignment of AC will require additional support by VCD and OR data.

9.04.6.4.6 TDDFT calculation of CD of alkaloid chimonantine

This study on chimonantine **119** provides an example of the enormous advance in the past decade in computational analysis, which makes feasible the correct assignment of the AC of large and conformationally flexible molecules by *ab initio* calculations of ECD and OR[104] (**Figure 46**). Such calculations are free of the limitations typical for coupled oscillator approaches, where the presence of chromophores with certain electronic and geometrical attributes is a prerequisite for a straightforward configurational analysis. The molecular flexibility of pyrrolo[2,3-*b*]indoline alkaloids, including chimonantine, led Mason and Vane in 1966 to conclude that it was impossible to deduce their AC from chiroptical data.[105] The recent study by



(2*R*,7*S*,20*S*,21*S*)-schizozygine **114**

(1*R*,5*S*,8*S*,9*S*,10*S*)-plumericin **115**

(1*S*,5*R*,8*R*,9*R*,10*R*)-isoplumericin **116**

(2*R*,7*R*,20*S*,21*S*)-iso-schzogaline **117**

(2*R*,7*R*,20*S*,21*S*)-iso-schizogamine **118**

Figure 45 Schizozygine, plumericin, and related compounds.

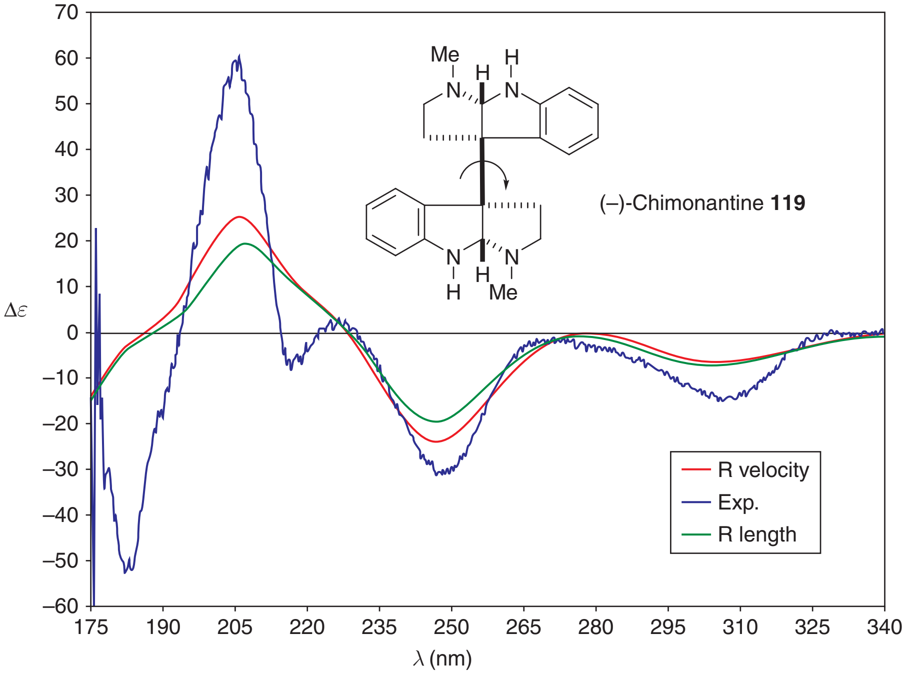
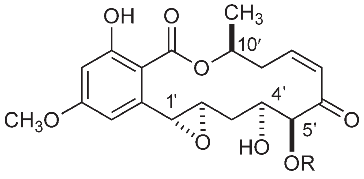


Figure 46 (−)-Chimonantine **119** experimental CD spectrum in cyclohexane and two calculated in the velocity and length formalism. The TDDFT/B3LYP/6-31G* calculated spectra were obtained as Boltzmann average upon the total six conformers taking into account the 40 lowest energy transitions and assuming a Gaussian distribution with $\sigma_i = 0.15$ eV. Redrawn from E. Giorgio; K. Tanaka; L. Verotta; K. Nakanishi; N. Berova; C. Rosini, *Chirality* **2007**, *19*, 434–445.

Giorgio *et al.* shows, indeed, that 40 years later the situation has changed dramatically in favor of theoretical predictions of AC when the natural products possess challenging molecular complexity.

9.04.6.4.7 Absolute configuration of hypothemycin by TDDFT calculation and solid-state CD

The theoretical calculations of the CD spectrum of the antitumor macrolide hypothemycin 120 were based on a geometry derived from X-ray analysis without further optimization and TDDFT methodology (B3LYP/ TZVP)[106] (Figure 47). The comparison of the calculated CD spectrum with experimental CDs, measured in the solid state (in KBr) and in solution, revealed some differences of approximately 270–300 nm. In other recent studies,[107] a good agreement of calculated CD with the spectra measured in the solid state as well as in solution was found. However, it seems the hypothemycin example illustrates a boarder line example for the application of this new solid-state CD/TDDFT approach. It does point out the need for caution in the interpretation of solid-state experimental/theoretical data. Most likely, intermolecular H-bonding of hypothemycin in the solid state and perhaps also in solution, which has not been taken into account by the calculation in vacuum, is responsible for the observed differences.



(1′*R*,2′*R*,4′*S*,5′*S*,10′*S*)-hypothemycin **120**

Figure 47 Absolute configuration of hypothemycin.

9.04.6.5 CD of Supramolecular Systems

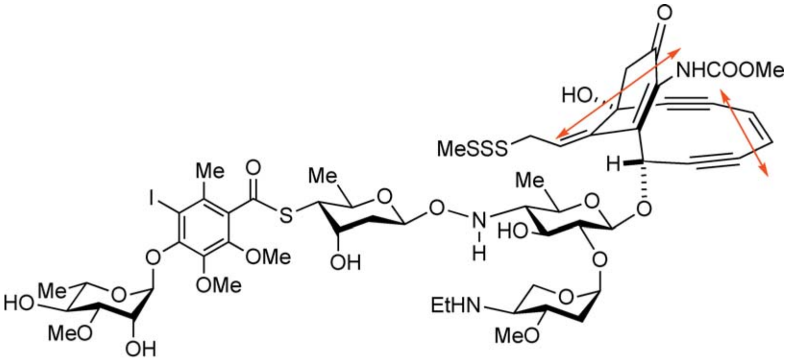
CD has been particularly useful in providing insight into the chirality of supramolecular assemblies. And in some cases monitoring the change in the CD signal on titration of the host molecule with a ligand can lead to information about the mode of binding and dissociation constants. Thermodynamic parameters may be obtained when monitoring the CD signal as a function of temperature. Examples of these cases are given below.

9.04.6.5.1 Theoretical simulation of CD spectrum of calicheamicin

The aglycon portion of calicheamicin **121** and other 10-membered ring enediyne antitumor antibiotics contain dienonecarbamate and enediyne chromophores in a unique bicyclic ring structure in which these two subunits are essentially orthogonal to each other. The CD spectrum of calicheamicin **121**, as well as the other members of this family, all of which contain the same bicyclic system, exhibits a characteristic and strongly negative exciton-coupled CD at 310 and 270 nm, which was used to assign the stereochemical relationship between these two chromophores at the time of the structure determination. This was later confirmed by stereospecific total synthesis of the (−)-aglycon, calicheamicinone **122**. Additional confirmation was then obtained by calculating the theoretical CD spectra of the whole calicheamicin aglycon A, and the dienone and enediyne chormophores individually, by using DFT and the De Voe's coupled oscillator method.[108] The necessary input geometry was obtained from the X-ray structure of the synthetic (−)-calicheamicinone **122**. In order to simplify the calculations, the allylic sulfur of the trisulfide moiety was replaced with a methyl group and the carbohydrate tail portion replaced with a hydrogen. The DFT calculations showed that the enediyne chromophore alone contributes very weakly to the exciton couplet whereas the twisted dienone chromophore makes a more significant contribution. However, only by taking into account both chromophores simultaneously could the shape of the experimental CD spectrum be adequately reproduced, but with only about a third of the experimental intensity. The De Voe calculations support those of the DFT method. The dipole transition moments used for the De Voe calculations are shown in the structure of calicheamicin (**Figure 48**).

9.04.6.5.2 Calicheamicin binding to an oligonucleotide

CD titration studies of the calicheamicin–DNA interaction provided a dissociation constant and evidence for a calicheamicin-induced DNA conformational change. As mentioned in Section 9.04.6.5.1, calicheamicin is a potent antitumor agent that binds to double-stranded DNA at specific sequences and subsequently cleaves both strands of the double helix. The carbohydrate tail portion of the molecule is the DNA recognition part of the molecule. Measurement of the change in the CD signal of a duplex DNA 12-mer (3′-GGGCCAGGATTC-5′ hybridized with its complementary sequence) on titration with calicheamicin at a wavelength not masked by the absorption of the antibiotic gave a binding isotherm with saturation as well as a prominent isobestic point.[109]



Calicheamicin $\gamma_1{}^I$ (**121**)
CD (CH_3CN) λ_{ext} 312 nm ($\Delta\varepsilon$ −44), 272 (+46)

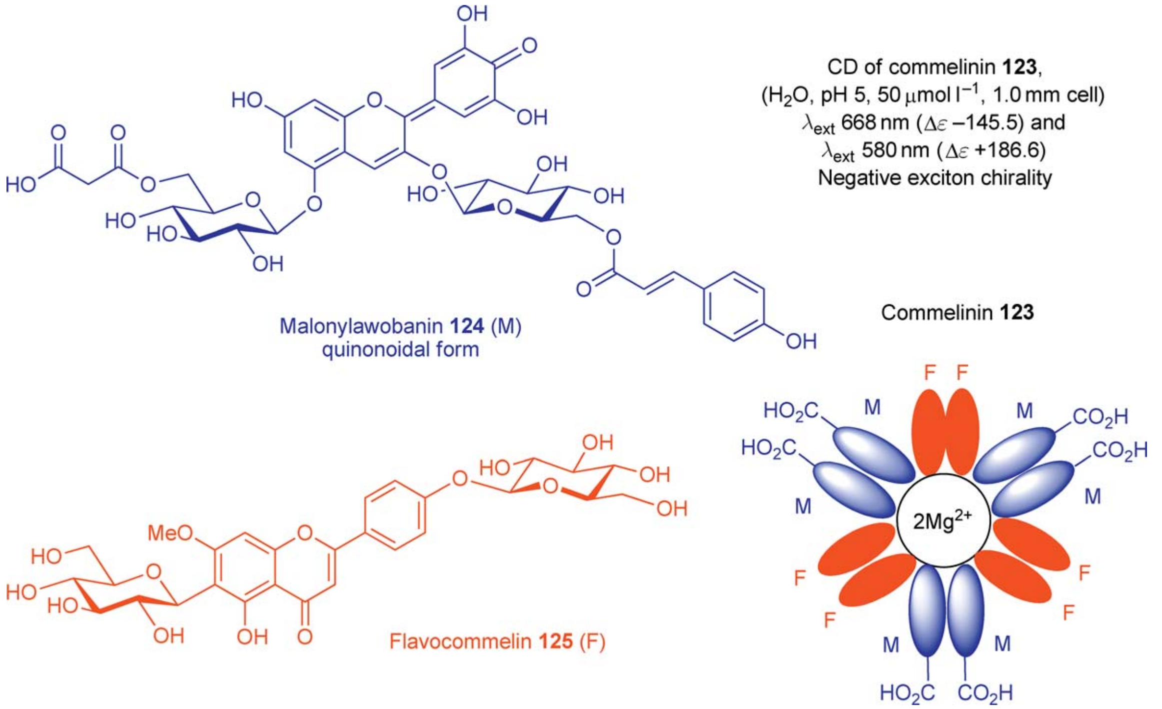
Figure 48 Calicheamicin **121** and CD data.

The calculated dissociation constant of a few micromolar agreed well with the value obtained from a direct measurement carried out using microcalorimetry.[110] Furthermore, the CD titration showed that the DNA conformation was condensed somewhat as evidenced by a decrease in the CD signal of the DNA due to the binding of the hydrophobic antibiotic in the DNA minor groove.

9.04.6.5.3 Chiral stacking of anthocyanin flower pigments as revealed by CD spectroscopy

The molecular basis of flower pigments has fascinated organic chemists for the past 100 years. However, it is only recently that this enigma has finally been resolved primarily by the seminal effort of Japanese researchers led by Goto and Kondo,[111–113] and more recently by Takeda and coworkers.[114] Goto and Kondo determined that the deep blue color of *Commelina communis* flower is due to a high molecular weight pigment, commelinin **123**.[111,112] This supramolecular, Mg-containing pigment consists of a flattened spherical cluster with six molecules each of an anthocyanin and a copigment flavone glycoside, which surrounds two Mg^{2+} ions located in the center of the complex (**Figure 49**) in a threefold axis of symmetry. The structure was subsequently confirmed by X-ray analysis. The structures of two other similar supramolecular pigments, protodelphin[115] and protocyanin,[113,114] were later determined to be responsible for the color of the blue flowers of *Salvia patens* and the cornflower, *Centaurea cyanus*, respectively.[116]

CD spectroscopy, in addition to X-ray analysis of commelinin **123** and protocyanin, showed that all three of these supramolecular pigments were stacked in a left-handed helical assembly with complex, large, negative exciton couplets in the visible region (e.g., commelinin **123**, λ_{ext} 668 nm ($\Delta\varepsilon$ −145.5) and 580 nm ($\Delta\varepsilon$ +186.4)) of the spectra matching the maxima in the UV–Vis spectra. The pendent glucose sugars are all of the *D* form and this homochirality drives the formation of the left-handed chiral supramolecular assembly by specific hydrogen bonding between the sugars of the anthocyanin **124** and copigment flavone **125**. Metal chelation and hydrophobic interactions between the aromatic chromophores also play an important role in the assembly. The observed exciton couplet in the CD of commelinin **123** clearly indicates that the homodimeric anthocyanins in the quinonoidol tautomeric form assemble in a left-handed helical twist. The dimeric flavone copigments are intercalated between the anthocyanins and are also assembled in a left-handed offset. Flavocommelin **125** itself



CD of commelinin **123**,
(H_2O, pH 5, 50 μmol l^{-1}, 1.0 mm cell)
λ_{ext} 668 nm ($\Delta\varepsilon$ −145.5) and
λ_{ext} 580 nm ($\Delta\varepsilon$ +186.6)
Negative exciton chirality

Figure 49 Flower color pigment, commelinin **123** and CD data. Redrawn from G. A. Ellestad, *Chirality* **2006**, *18*, 134–144.

stacks in a left-handed geometry with a clear and negative exciton couplet at approximately λ_{ext} 365 nm (negative) and 320 nm (positive), but this region is obscured in the CD of the supramolecule.

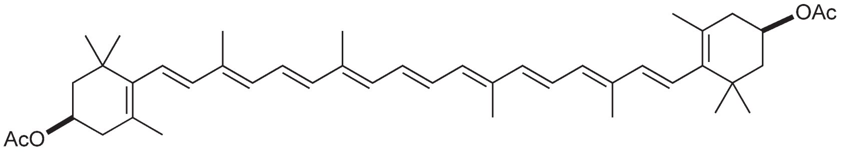
9.04.6.5.4 CD of chirally stacked carotenoid pigments

An unusual application of CD published by Zisla *et al.* has to do with the chirality of self-assemblies of carotenoids and their esters in intact orange and yellow flower petals[117] (**Figure 50**). Carotenoid esters themselves have been found to stack intermolecularly in either a right- or a left-handed twist as evidenced by strong, complex exciton couplets in the visible region of the CD spectrum. The handedness of the stacking relates to the absolute stereochemistry of the stereogenic centers at each end of the carotenoids. For example, the CD spectrum of lutein diacetate **126** recorded in aqueous ethanol shows a positive exciton couplet between 450 and 500 nm indicative of right-handed chirality of the stacked carotenoid esters. The aqueous solvent promotes the aggregation and apparently simulates the plant cell's aqueous environment. A remarkably clear CD spectrum of intact flower petals from *Chelidonium majus* – obtained by using freshly picked petals pressed between two quartz windows – matched closely the above-mentioned spectrum of lutein diacetate demonstrating the validity of using intact petals.

Based on their CD spectrum, each plant species appears to produce flowers with a distinctive CD that is characteristic of that species. Furthermore, there are usually two or more types of carotenoids in the petal and the CD is influenced by not only this chemical heterogeneity but by the presence of the proteins and lipids that co-occur with the pigments.

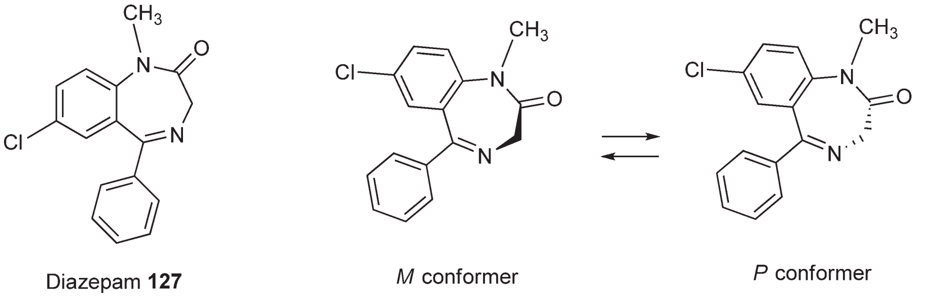
9.04.6.5.5 CD of diazepam–HSA and diazepam–AGP complexes

ICD was used to determine the bound conformations of the 1,4-benzodiazepine, anxiolytic drug, diazepam **127**, to the two main serum proteins, human serum albumin (HSA) and α_1-acid glycoprotein (AGP).[118,119] This is an interesting application of CD because diazepine lacks a stereogenic center, but due to the rapid inversion of the nonplanar seven-membered ring, the drug is in equimolar equilibrium between two chiral conformers, *P* (plus) and *M* (minus) (**Figure 51**). This study shows that the two serum proteins display different conformer priorities. The ICD of HSA-bound diazepam is strongly positive at λ_{ext} 260 nm ($\Delta\varepsilon$ +46.5) with a smaller negative band at λ_{ext} 321 nm ($\Delta\varepsilon$ −8.3), which indicates an *M* conformer preference. In contrast, the CD of AGP-bound diazepam shows a negative signal at λ_{ext} 261 nm ($\Delta\varepsilon$ −9.6) and a positive one at λ_{ext} 313 nm ($\Delta\varepsilon$ +1.6) indicating a *P* conformer preference. Diazepam has been shown to bind to the domain III region of HSA, as shown by photolabeling studies, although it binds to a low-affinity domain I site with an inverse ICD



Lutein diacetate **126**

Figure 50 Carotenoid pigment.



Diazepam **127**

M conformer

P conformer

Figure 51 Equilibrium of diazepam **127** between two chiral conformers.

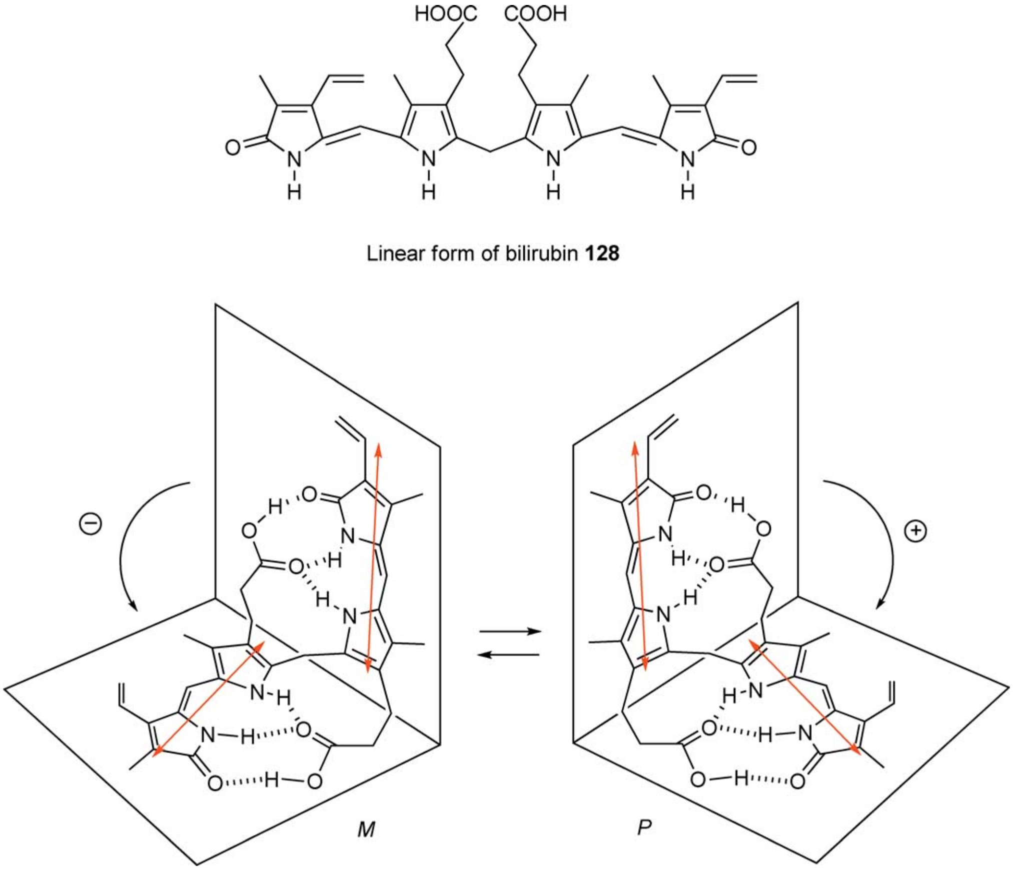
spectrum. This is presumably due to a preference for the *P* conformation. Thus AGP/diazepam binding seems to mimic this minor binding site of HSA. The binding affinities to the two serum proteins are similar as shown by ultrafiltration experiments.

9.04.6.5.6 CD of bilirubin bound to human and bovine serum albumins

Bilirubin **128**, the cytotoxic yellow pigment of jaundice, is an achiral tetrapyrrole with no stereogenic centers.[120] It consists of an equilibrium mixture of two equimolar conformers, *P* and *M* (**Figure 52**). Although it does not show any CD signal in aqueous solution, upon binding to HSA that serves as a chiral selector, the *P* conformer is preferentially bound as evidenced by the appearance of a positive exciton split CD at 457 nm ($\Delta\varepsilon$ +49.5) and 407 nm ($\Delta\varepsilon$ −29.5). Interestingly, a negative exciton couplet at 457 nm ($\Delta\varepsilon$ −62.5) and 407 nm ($\Delta\varepsilon$ +23.7) is observed with bilirubin bound to bovine serum albumin (BSA), which correlates to a preference for the *M* conformer in this binding site. Thus, this methodology provides important insight into the drug-binding properties of serum proteins. Another publication described the interesting preference for bilirubin enantiomeric conformations in biomembrane models composed of chiral micellar aggregates formed from enantiomeric *N*-alkyl-*N*,*N*-dimethyl-*N*-(1-phenyl)ethylammonium bromides, as determined by CD. This study points to a possible correlation between conformer-specific bilirubin neuromembrane alterations and bilirubin neurotoxicity.[121]

9.04.7 Concluding Remarks and Outlook

As can be seen from the great variety of examples set forth in this chapter ranging from more traditional small molecule natural products and drugs to supramolecular flower pigments and the yellow jaundice pigment



Linear form of bilirubin **128**

Figure 52 Enantiomeric conformations *P* and *M* of bilirubin **128** with electric transition moments. Redrawn from R. V. Person; B. R. Peterson; D. A. Lightner, *J. Am. Chem. Soc*. **1994**, *116*, 42–59.

bilirubin, the interest in applying CD method for the determination of AC in natural products has increased dramatically in recent years. The reasons for this are twofold. First, the importance of stereochemical relevance to biological activity has made the determination of molecular and supramolecular chirality extremely critical for understanding the behavior and interaction of molecules. This is especially true for the binding of natural products to the cellular receptors that mediate their biological activity. Second, there has been a tremendous advance during the past 10 years in the theoretical treatment of optical activity and the computational methods that support the *ab initio* calculations of CD spectra. Furthermore, these calculations in conjunction with experimental results obtained with the more sophisticated instrumentation presently available make the assignment of the ACs, and in certain cases even bioactive conformations, a straightforward and highly efficient process. It is these tremendous advances in the past decade in the development of *ab initio* calculations of CD spectra that has put the interpretation of CD spectra on more solid ground.

There are good reasons to remain optimistic that recent momentum in technological progress in chiroptical instrumentation and in development of new more sophisticated *ab initio* computational methodologies will continue in the future at faster pace. This will make the calculations of optical activity properties, including CD, a truly indispensable and widely affordable approach in the stereochemical analysis of natural products and their interactions on molecular and supramolecular levels.

References

1. S. F. Mason, *Molecular Optical Activity and the Chiral Discrimination*; Cambridge University Press: Cambridge, 1982.
2. N. Berova; K. Nakanishi; R. W. Woody, Eds., *Circular Dichroism: Principles and Applications,* 2nd ed.; Wiley-VCH: New York, 2000.
3. E. L. Eliel; S. H. Wilen; L. N. Mander, *Stereochemistry of Organic Compounds*; John Wiley & Sons Inc.: New York, 1994; Chapter 13, pp 991–1118.
4. J. R. Cheeseman; M. J. Frisch; F. J. Devlin; P. J. Stephens, *Chem. Phys. Lett.* **1996**, *252*, 211–220.
5. P. L. Polavarapu, *Vibrational Spectra: Principles and Applications with Emphasis on Optical Activity*; Elsevier: Amsterdam-Lausanne-New York-Shannon-Tokyo, 1998.
6. T. B. Freedman; X. Cao; R. K. Dukor; L. A. Nafie, *Chirality* **2003**, *15*, 743–758.
7. N. Harada, Optical Rotation, Optical Rotatory Dispersion, and Circular Dichroism. In *Handbook of Instrumental Analysis*, Part 2, 2nd ed.; Y. Izumi, M. Ogawa, S. Kato, J. Shiokawa, T. Shiba, Eds.; Kagakudojin: Kyoto, Japan, 2005; pp 119–133.
8. N. Harada; K. Nakanishi, *Circular Dichroic Spectroscopy – Exciton Coupling in Organic Stereochemistry*; University Science Books: Mill Valley, CA, and Oxford University Press: Oxford, 1983.
9. N. Berova; L. Di Bari; G. Pescitelli, *Chem. Soc. Rev.* **2007**, *36*, 914–931.
10. D. A. Lightner; J. E. Gurst, *Organic Conformational Analysis and Stereochemistry from Circular Dichroism Spectroscopy*; Wiley-VCH: New York, 2000.
11. N. Berova; K. Nakanishi, Exciton Chirality Method: Principles and Application. In *Circular Dichroism: Principles and Applications*, 2nd ed.; N. Berova, K. Nakanishi, R. W. Woody, Eds.; Wiley-VCH: New York, 2000; Chapter 12, pp 337–382.
12. H. De Voe, *J. Chem. Phys.* **1964**, *41*, 393–400; **1965**, *43*, 3199–3208.
13. S. Superchi; E. Giorgio; C. Rosini, *Chirality* **2005**, *16*, 422–451.
14. C. M. Kemp; S. F. Mason, *Tetrahedron* **1966**, *22*, 629–635.
15. T. D. Crawford, *Theor. Chem. Acc.* **2006**, *115*, 227–245.
16. N. Berova; N. Harada; K. Nakanishi, Electronic Spectroscopy: Exciton Coupling, Theory and Applications. In *Encyclopedia of Spectroscopy and Spectrometry*; J. Lindon, G. Tranter, J. Holmes, Eds.; Academic Press: London, 2000; pp 470–488.
17. S. F. Mason; R. H. Seal; D. R. Roberts, *Tetrahedron* **1974**, *30*, 1671–1682.
18. I. Hanazaki; H. Akimoto, *J. Am. Chem. Soc.* **1972**, *94*, 4102–4106.
19. L. Di Bari; G. Pescitelli; P. Salvadori, *J. Am. Chem. Soc.* **1999**, *121*, 7998–8004.
20. K. Harata; J. Tanaka, *Bull. Chem. Soc. Jpn.* **1973**, *46*, 2747–2751.
21. J. M. Bijvoet; A. F. Peerdeman; A. J. Van Bommel, *Nature* **1951**, *168*, 271–272.
22. J. Trommel; J. M. Bijvoet, *Acta Crystallogr.* **1954**, *7*, 703–709.
23. J. M. Bijvoet; A. F. Peerdeman, *Acta Crystallogr.* **1956**, *9*, 1012–1015.
24. J. Tanaka; F. Ogura; H. Kuritani; M. Nakagawa, *Chimia* **1972**, *26*, 471–473.
25. J. Tanaka; K. Ozeki-Minakata; F. Ogura; M. Nakagawa, *Nature (London) Phys. Sci.* **1973**, *241*, 22–23.
26. J. Tanaka; K. Ozeki-Minakata; F. Ogura; M. Nakagawa, *Spectrochim. Acta A* **1973**, *29*, 897–924.
27. N. Harada; Y. Takuma; H. Uda, *J. Am. Chem. Soc.* **1976**, *98*, 5408–5409.
28. N. Harada; Y. Takuma; H. Uda, *Bull. Chem. Soc. Jpn.* **1977**, *50*, 2033–2038.
29. N. Harada; Y. Takuma; H. Uda, *J. Am. Chem. Soc.* **1978**, *100*, 4029–4036.
30. N. Sakabe; K. Sakabe; K. Ozeki-Minakata; J. Tanaka, *Acta Crystallogr. B* **1972**, *28*, 3441–3446.
31. J. Tanaka; C. Katayama; F. Ogura; H. Tatemitsu; M. Nakagawa, *Chem. Commun.* **1973**, 21–22.
32. D. H. R. Barton; H. T. Cheung; A. D. Cross; L. M. Jackman; M. Martin-Smith, *J. Chem. Soc.* **1961**, 5061–5073.
33. I. C. Paul; G. A. Sim; T. A. Hamor; J. M. Robertson, *J. Chem. Soc.* **1962**, 4133–4145.
34. S. Hosozawa; N. Kato; K. Munakata, *Tetrahedron Lett.* **1974**, *15*, 3753–3756.

35. N. Kato; S. Shibayama; K. Munakata; C. Katayama, *Chem. Commun.* **1971**, 1632–1633.
36. N. Kato; K. Munakata; C. Katayama, *J. Chem. Soc., Perkin Trans. 2* **1973**, 69–73.
37. N. Kato; M. Shibayama; K. Munakata, *J. Chem. Soc., Perkin Trans. 1* **1973**, 712–719.
38. N. Harada; H. Uda, *J. Am. Chem. Soc.* **1978**, *100*, 8022–8024.
39. D. Rogers; G. G. Unal; D. J. Williams; S. V. Ley; G. A. Sim; B. S. Joshi; K. R. Ravindranath, *J. Chem. Soc., Chem. Commun.* **1979**, 97–99.
40. M. L. Kenney; F. Cortés-Selva; J. M. Perez-Victoria; I. A. Jiménez; A. G. Gonzalez; O. M. Muñoz; F. Gamarro; S. Castanys; A. G. Ravelo, *J. Med. Chem.* **2001**, *44*, 4668–4676.
41. C. A. Parish; J.-G. Dong; W. G. Bornmann; J. Chang; K. Nakanishi; N. Berova, *Tetrahedron* **1998**, *54*, 15739–15758.
42. N. Harada, *J. Am. Chem. Soc.* **1973**, *95*, 240–242.
43. A. Guerriro; M. D'Ambrosio; V. Cuomo; F. Vanzanella; F. Pietra, *Helv. Chim. Acta* **1989**, *72*, 438–446.
44. M. Koreeda; N. Harada; K. Nakanishi, *J. Am. Chem. Soc.* **1974**, *96*, 266–268.
45. T. Ishikawa; M. Murota; T. Watanabe; T. Harayam; H. Ishii, *Tetrahedron Lett.* **1995**, *36*, 4269–4272.
46. X. Huang; K. Nakanishi; N. Berova, *Chirality* **2000**, *12*, 237–255.
47. G. Pescitelli; S. Gabriel; Y. Wang; J. Fleischhauer; R. W. Woody; N. Berova, *J. Am. Chem. Soc.* **2003**, *125*, 7613–7628.
48. S. Matile; N. Berova; K. Nakanishi, *Chem. Biol.* **1996**, *3*, 379–392.
49. N. Zhao; P. Zhou; N. Berova; K. Nakanishi, *Chirality* **1995**, *7*, 636–651.
50. D. Rele; N. Zhao; K. Nakanishi; N. Berova, *Tetrahedron* **1996**, *52*, 2759–2776.
51. N. Zhao; N. Berova; K. Nakanishi; M. Rohmer; P. Mougenot; U. J. Jurgens, *Tetrahedron* **1996**, *52*, 2777–2788.
52. A. Kawamura; N. Berova; V. Dirsch; A. Mangoni; K. Nakanishi; G. Schwartz; A. Bielawska; Y. Hannun; I. Kitagawa, *Bioorg. Med. Chem.* **1996**, *4*, 1035–1043.
53. N. Harada; K. Nakanishi, *J. Am. Chem. Soc.* **1969**, *91*, 3989–3991.
54. G. Cai; N. Bozhkova; J. Odingo; N. Berova; K. Nakanishi, *J. Am. Chem. Soc.* **1993**, *115*, 7192–7198.
55. T. Kurtán; N. Nesnas; Y.-Q. Li; X. Huang; K. Nakanishi; N. Berova, *J. Am. Chem. Soc.* **2001**, *123*, 5962–5973.
56. T. Kurtán; N. Nesnas; F. E. Koehn; Y.-Q. Li; K. Nakanishi; N. Berova, *J. Am. Chem.* **2001**, *123*, 5974–5982.
57. X. Huang; N. Fujioka; G. Pescitelli; F. E. Koehn; T. R. Williamson; K. Nakanishi; N. Berova, *J. Am. Chem. Soc.* **2002**, *124*, 10320–10335.
58. J. W. Van Klink; S.-H. Baek; A. J. Barlow; H. Ishii; K. Nakanishi; N. Berova; N. B. Perry; R. T. Weavers, *Chirality* **2004**, *16*, 549–558.
59. H. Ishii; S. Krane; Y. Itagaki; N. Berova; K. Nakanishi; P. J. Weldon, *J. Nat. Prod.* **2004**, *67*, 1426–1430.
60. Q. Yang; C. Olmsted; B. Borhan, *Org. Lett.* **2002**, *4*, 3423–3426.
61. X. Li; M. Tanasova; C. Vasileiou; B. Borhan, *J. Am. Chem. Soc.* **2008**, *130*, 1885–1893.
62. V. V. Borovkov; J. M. Lintuluoto; Y. Inoue, *J. Am. Chem. Soc.* **2001**, *123*, 2979–2989.
63. S. Allenmark, *Chirality* **2003**, *15*, 409–422.
64. G. A. Ascoli; E. Domenici; C. Bertucci, *Chirality* **2006**, *18*, 667–690.
65. B. Nordén; T. Kurucsev, *J. Mol. Recognit.* **1994**, *7*, 141–156.
66. M. Simonyi; Z. Bikadi; F. Zsila; J. Deli, *Chirality* **2003**, *15*, 680–698.
67. Y. Kikuchi; K. Kobayashi; Y. Aoyama, *J. Am. Chem. Soc.* **1992**, *114*, 1351–1358.
68. T. Eguchi; K. Kondo; K. Kakinuma; H. Uekusa; Y. Ohashi; K. Mizoue; Y.-F. Qiano, *J. Org. Chem.* **1999**, *64*, 5371–5376.
69. C. D. Donner; M. Gill, *J. Chem. Soc., Perkin Trans.1* **2002**, 938–948.
70. B. I. Morinaka; C. K. Skepper; T. F. Molinski, *Org. Lett.* **2007**, *9*, 1975–1978.
71. M. Satake; A. Morohashi; H. Oguri; T. Oishi; M. Hirama; N. Harada; T. Yasumoto, *J. Am. Chem. Soc.* **1997**, *119*, 11325–11326.
72. N. Harada; A. Saito; H. Ono; S. Murai; H.-Y. Li; J. Gawronski; K. Gawronska; T. Sugioka; H. Uda, *Enantiomer* **1996**, *1*, 119–138.
73. I. Akritopoulou-Zanze; K. Nakanishi; H. Stepowska; B. Grzeszczyk; A. Zamojski; N. Berova, *Chirality* **1998**, *9*, 699–712.
74. N. Harada; H.-Y. Li; N. Koumura; T. Abe; M. Watanabe; M. Hagiwara, *Enantiomer* **1997**, *2*, 349–352.
75. N. Harada; A. Saito; H. Ono; J. Gawronski; K. Gawronska; T. Sugioka; H. Uda; T. Kuriki, *J. Am. Chem. Soc.* **1991**, *113*, 3842–3850.
76. D. Rele; N. Zhao; K. Nakanishi; N. Berova, *Tetrahedron* **1996**, *52*, 2759–2776.
77. N. Zhao; N. Berova; K. Nakanishi; M. Rohmer; P. Mougenot; U. J. Jürgens, *Tetrahedron* **1996**, *52*, 2777–2788.
78. A. Sakurai; H. Horibe; N. Kuboyama; Y. Hashimoto; Y. Okumura, *J. Biochem.* **1995**, *118*, 552–554.
79. J. Kobayashi; M. Yoshinaga; N. Yoshida; M. Shiro; H. Morita, *J. Org. Chem.* **2002**, *67*, 2283–2286.
80. H. Kobayashi; S. Meguro; T. Yoshimoto; M. Namikoshi, *Tetrahedron* **2003**, *59*, 455–459.
81. T. Wang; O. Shirota; K. Nakanishi; N. Berova; L. A. McDonald; L. R. Barbieri; G. Carter, *Can. J. Chem.* **2001**, *79*, 1786–1791.
82. T. Sunazuka; T. Shirahata; K. Yoshida; D. Yamamoto; Y. Harigaya; T. Nagai; H. Kiyohara; H. Yamada; I. Kuwajima; S. Omura, *Tetrahedron Lett.* **2002**, *43*, 1265–1268.
83. K. Kouda; T. Ooi; K. Kaya; T. Kusumi, *Tetrahedron Lett.* **1996**, *37*, 6347–6350.
84. T. F. Molinski; L. J. Brzezinski; J. W. Leahy, *Tetrahedron: Asymmetry* **2002**, *13*, 1013–1016.
85. K. Tanaka; Y. Itagaki; M. Satake; H. Naoki; T. Yasumoto; K. Nakanishi; N. Berova, *J. Am. Chem. Soc.* **2005**, *127*, 9561–9570.
86. J. D. Chisholm; J. Golik; B. Krishnan; J. A. Matson; D. L. Van Vranken, *J. Am. Chem. Soc.* **1999**, *121*, 3801–3802.
87. R. W. Baker; S. Liu; M. V. Sargent, *Aust. J. Chem.* **1998**, *51*, 255–266.
88. M. S. Buchanan; M. Gill; P. Millar; S. Phonh-Axa; E. Raudies; J. Yu, *J. Chem. Soc. Perkin Trans. 1* **1999**, 795–801.
89. J. W. Moncrief; W. N. Lipscomb, *Acta Cryst.* **1966**, *21*, 322–331.
90. J. P. Kutney; D. E. Gregonis; R. Imhof; I. Itoh; E. Jahngen; A. I. Scott; W. K. Chan, *J. Am. Chem. Soc.* **1975**, *97*, 5013–5015.
91. N. Harada; H. Ono; H. Uda; M. Parveen; N. U.-D. Khan; B. Achari; P. K. Dutta, *J. Am. Chem. Soc.* **1992**, *114*, 7687–7692.
92. N. Harada; J. Kohori; H. Uda; K. Nakanishi; R. Takeda, *J. Am. Chem. Soc.* **1985**, *107*, 423–428.
93. N. Harada; H. Uda; M. Kobayashi; N. Shimizu; I. Kitagawa, *J. Am. Chem. Soc.* **1989**, *111*, 5668–5674.
94. R. Lysek; K. Borsuk; M. Chmielewski; Z. Kaluza; Z. Urbanczyk-Lipkowska; A. Klimek; J. Frelek, *J. Org. Chem.* **2002**, *67*, 1472–1479.
95. T. T. Danh; W. Bocian; L. Kozerski; P. Szczukiewicz; J. Frelek; M. Chmielewski, *Eur. J. Org. Chem.* **2005**, *67*, 429–440.

96. J. Frelek; R. Lysek; K. Borsuk; J. Jagodzinski; B. Furman; A. Klimek; M. Chmielewski, *Enantiomer* **2002**, *7*, 107–114.
97. M. Cierpucha; J. Solecka; J. Frelek; P. Szczukiewicz; M. Chmielewski, *Biorg. Med. Chem.* **2004**, *12*, 405–416.
98. M. Chmielewski; M. Cierpucha; P. Kowalska; M. Kwit; J. Frelek, *Chirality* **2008**, *20*, 621–627.
99. J. Frelek; P. Kowalska; M. Masnyk; A. Kazimierski; A. Korda; M. Woznica; M. Chmielewski; F. Furche, *Chem. Eur. J.* **2007**, *13*, 6732–6744.
100. P. J. Stephens; D. M. McCann; F. J. Devlin; A. B. Smith III, *J. Nat. Prod.* **2006**, *69*, 1055–1064.
101. P. J. Stephens; J.-J. Pan; F. J. Devlin; M. Urbanova; J. Hajicek, *J. Org. Chem.* **2007**, *72*, 2508–2524.
102. P. J. Stephens; J.-J. Pan; F. J. Devlin; K. Kron; T. Kurtan, *J. Org. Chem.* **2007**, *72*, 3521–3536.
103. P. J. Stephens; J.-J. Pan; F. J. Devlin; M. Urbanova; O. Julinek; J. Hajicek, *Chirality* **2008**, *20*, 454–470.
104. E. Giorgio; K. Tanaka; L. Verotta; K. Nakanishi; N. Berova; C. Rosini, *Chirality* **2007**, *19*, 434–445.
105. S. F. Mason; G. W. Vane, *J. Chem. Soc. B* **1966**, 370–374.
106. H. Hussain; K. Krohn; U. Florke; B. Schulz; S. Draeger; G. Pesitelli; P. Salvadori; S. Antus; T. Kurtan, *Tetrahedron Asymmetry* **2007**, *18*, 925–930.
107. H. Hussain; K. Krohn; U. Floerke; B. Schulz; S. Draeger; G. Pescitelli; S. Antus; T. Kurtan, *Eur. J. Org. Chem.* **2007**, 292–295.
108. E. Giorgio; K. Tanaka; W. Ding; G. Krishnamurthy; K. Pitts; G. Ellestad; C. Rosini; N. Berova, *Bioorg. Med. Chem.* **2005**, *13*, 5072–5079.
109. G. Krishnamurthy; W.-D. Ding; G. A. Ellestad, *Tetrahedron* **1994**, *50*, 1341–1349.
110. M. Chatterjee; P. J. Smith; C. A. Townsend, *J. Am. Chem. Soc.* **1996**, *118*, 1938–1948.
111. T. Goto; T. Kondo, *Angew. Chem. Int. Ed. Engl.* **1991**, *30*, 17–33.
112. T. Kondo; K. Yoshida; A. Nakagawa; T. Kawai; H. Tamura; T. Goto, *Nature* **1992**, *358*, 515–518.
113. T. Kondo; M. Ueda; H. Tamura; K. Yoshida; M. Isobe; T. Goto, *Angew. Chem. Int. Ed. Engl.* **1994**, *33*, 978–979.
114. M. Shiono; N. Matsugaki; K. Takeda, *Nature* **2005**, *436*, 791.
115. T. Kondo; K. Oyama; K. Yoshida, *Angew. Chem. Int. Ed. Engl.* **2001**, *40*, 894–897.
116. G. A. Ellestad, *Chirality* **2006**, *18*, 134–144.
117. F. Zisla; J. Deli; M. Simonyi, *Planta* **2001**, *213*, 937–942.
118. I. Fitos; J. Visy; F. Zsila; G. Mády; M. Simonyi, *Bioorg. Med. Chem.* **2007**, *15*, 4857–4862.
119. M. Pistolozzi; C. Bertucci, *Chirality* **2008**, *20*, 552–558.
120. R. V. Person; B. R. Peterson; D. A. Lightner, *J. Am. Chem. Soc.* **1994**, *116*, 42–59.
121. C. Bombelli; C. Bernadini; G. Elemento; G. Manacini; A. Sorrenti; C. Villani, *J. Am. Chem. Soc.* **2008**, *130*, 2732–2733.

Biographical Sketches

Professor Nina Berova received her Ph.D. in chemistry in 1971 from the University of Sofia, Bulgaria. In 1982 she became an associate professor at the University of Sofia and at the Institute of Organic Chemistry, Bulgarian Academy of Sciences. In 1988 she joined the Department of Chemistry of Columbia University, New York, first as a visiting professor, and later she accepted her current position of research professor at the same department. She has been a recipient of many scholarships, among them, in 1989–92 a research fellowship at the University of Bochum, Germany, a visiting professorship in 1994 at Ecole Normal Superieure de Lyon, a lecturership in 1996 by the Japan Society for Promotion of Science (JSPS), and more recently visiting professorships at the University of Naples, the University of Santiago de Compostela, Spain, Tokyo Institute of Technology, and University 'Louis Pasteur', Strasbourg. Her research is focused on organic stereochemistry and chiroptical spectroscopy, in particular, on the electronic circular dichroism and its application in structural analysis. In 2000 she was the coeditor and coauthor of a comprehensive monograph *Circular Dichroism: Principles and Applications* (first edition 1994), coedited with K. Nakanishi and R. W. Woody, published by Wiley-VCH. She has received various awards including the Gold Medal 'Piero Pino' (2003), ACS/CA Editor Award (2005), and 'Chirality' Gold Medal (2007). Since 1998 she is the editor of the Wiley-Liss Journal *Chirality*.

Dr. George A. Ellestad obtained a B.S. and M.S. in chemistry from Oregon State University in 1957 and 1958, respectively, and a Ph.D. in organic chemistry from University of California, Los Angeles in 1962. After postdoctoral studies at the University of London he joined Lederle Laboratories in Pearl River, New York in 1964. His almost 40-year career at Lederle/Wyeth included structural and bioorganic studies on pharmacologically active mold metabolites; spermidine, glycopeptide, and tetracycline antibiotics; enediyne antitumor agents; and finally biophysical chemistry and enzymology for hit and lead characterization and assay development. His group's efforts on the structure and DNA cleavage chemistry of the enediyne antitumor agent calicheamicin helped lead to Mylotarg, an antibody conjugate for use in the treatment of acute myologenous leukemia. He also contributed to the development of Tygacil, a new semisynthetic tetracycline active against resistant bacterial infections that are no longer susceptible to previously useful antimicrobial agents. George retired from Wyeth in 2004 and in 2005 became an adjunct senior research chemist at Columbia University working in the laboratory of Professors Koji Nakanishi and Nina Berova studying the circular dichroism properties of porphyrin–DNA conjugates. Dr. Ellestad was the recipient of the 2006 ACS Medicinal Chemistry award.

Professor Nobuyuki Harada obtained his B.Sc., M.Sc., and Ph.D. degrees from Tohoku University in 1965, 1967, and 1970, respectively. In 1970 he joined the Chemical Research Institute of Nonaqueous Solutions, Tohoku University as research associate. After postdoctoral studies at the Department of Chemistry, Columbia University, U.S.A. (1973–75) he was promoted as associate professor at the Chemical Research Institute of Nonaqueous Solutions, Tohoku University. In addition, he was appointed as the adjunct associate professor, Institute for Molecular Science, Okazaki National Research Institutes, Japan (1980–82) and a visiting research scientist, R&D Department, Experimental Station, Du Pont de Nemours & Company, U.S.A. (1987). In 1992 he was promoted as professor, Institute for Chemical Reaction Science, Tohoku University. In 2006 he retired from Tohoku University and was appointed professor emeritus, Tohoku University. Since then he has been a visiting researcher and scholar, Department of Chemistry, Columbia University, U.S.A. Professor Harada's research field covers (a) natural products chemistry and structural organic chemistry; (b) theory and development of the CD exciton chirality method; (c) enantioresolution, absolute configurational and conformational studies of chiral compounds by NMR and X-ray methods using novel reagents; and (d) molecular machine, light-powered chiral molecular motors. In 1983 he published with Professor K. Nakanishi the monograph *Circular Dichroic Spectroscopy – Exciton Coupling in Organic Stereochemistry*, published by University Science Books, Mill Valley, California and Oxford University Press, Oxford. He has also contributed to the chemical community as an active editor of the journal *Enantiomer* (1996–2002), an associate editor of *Chirality* (2003–05), and was the organizer of CD Conference 2001, Sendai. He received the Academic Prize from the Chemical Society of Japan in 1984 and Molecular Chirality Award from the Molecular Chirality Research Organization, Japan in 2000.

9.05 Determination of Structure including Absolute Configuration of Bioactive Natural Products

Kenji Mori, The University of Tokyo, Tokyo, Japan

9.05.1 Introduction

Since the advent of modern physical tools such as UV, IR, ^{1}H-NMR, ^{13}C-NMR, mass spectrometry (MS), circular dichroism (CD), and X-ray analysis, the structure determination of bioactive small molecules is often regarded as a routine operation for natural products chemists.[1] Such a view by biologists and many chemists is contestable, and two reviews have appeared recently, both treating incorrectly assigned structures of many natural products.[2,3] Even X-ray analysis can be erroneous.[3] There are some cases in which the correctly proposed structures of the presumably bioactive molecules do not represent the structures of genuinely bioactive molecules, as shown by the bioassay of synthetic compounds with the proposed structures.[2] This type of error usually stems from the incorrect and nonreproducible bioassay methods employed for the biological phenomena in discussion.

In the case of the complex marine polyether brevenal (**1**), the structure as shown in the upper part of **Figure 1** was proposed by Bourdelais *et al.*[4] through extensive spectroscopic studies. After completing the synthesis of the proposed structure, Fuwa *et al.*[5] revised the structure of brevenal as **1**, because there were subtly distinct discrepancies of the chemical shifts in the ^{1}H- and ^{13}C-NMR spectra of the left-hand region of the synthetic material compared with those of the natural brevenal. It may therefore be important to evaluate the proposed structure by means of its synthesis.

9.05.2 Absolute Configuration and Sign of Optical Rotation

Semiochemicals are usually compounds with structures much simpler than that of brevenal. But even with simple compounds, there are possibilities of misassigning their absolute configuration. Male-produced pheromone components of the flea beetle *Aphthona flava* were isolated and identified in 2001 by Bartelt *et al.*[6] They proposed himachalene-type sesquiterpene structures **2–5** (**Figure 2**) to the components. In 2004, Mori and coworkers synthesized **2–5** and their enantiomers *ent*-**2**–*ent*-**5** from enantiomers of citronellal, and the

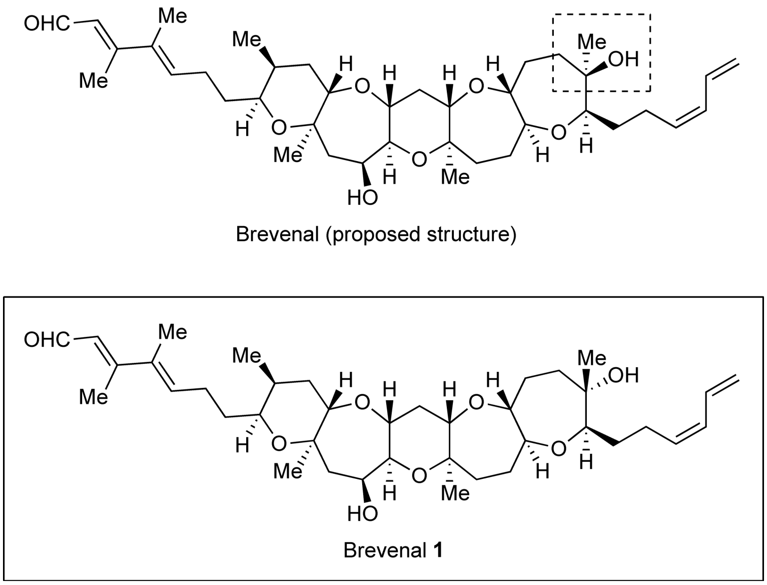


Figure 1 Structure of brevenal (**1**).

pheromone components were found to possess the absolute configuration as depicted in *ent*-**2**–*ent*-**5**.[7] Mori's assignments were opposite to those proposed by Bartelt *et al.*, and indeed *ent*-**2**–*ent*-**4** were pheromonally active against the Hungarian flea beetle *Phyllotreta cruciferae*, while **2**–**4** were inactive.[8]

Bartelt *et al.*[6] proposed the absolute configuration **5** for their pheromone component on the basis of its positive rotation (in hexane), because Pandey and Dev[9] reported positive rotation (in chloroform) of their synthetic **5**. Mori[10] synthesized *ent*-**5** by employing (*R*)-*ar*-turmerone (**6**) as the key intermediate, and found it to be dextrorotatory in hexane while levorotatory in chloroform. A simple mistake of using hexane as the solvent, instead of the reported chloroform, for measuring the optical rotation resulted in stereochemical misassignment of the absolute configuration of their pheromone components.

A similar example had been reported in 1976.[11] (1*S*,4*S*,5*S*)-*cis*-Verbenol (**7**) is a pheromone component of *Ips* bark beetles. Prior to Mori's work,[11] some researchers had called **7** (+)-*cis*-verbenol, while others referred to it as (−)-*cis*-verbenol. After synthesis of **7** and measurements of its optical rotations in different solvents, it became clear that **7** was dextrorotatory in acetone or methanol but levorotatory in chloroform. It is therefore of utmost importance to use the same solvent as reported by others, when one compares the sign of the optical rotation of a new sample with the previous data.

9.05.3 Elucidating the Structure of Pheromones of Stink Bugs

A simple example of the examination of a proposed structure through synthesis is provided in this section. In 2005, Takita[12] proposed the structure of the male-produced aggregation pheromone of the stink bug *Eysarcoris lewisi* as the sesquisabinene alcohol, (*E*)-2-methyl-6-(4′-methylenebicyclo[3.1.0]hexyl)hept-2-en-1-ol (**8**) (**Scheme 1**). Mori[13] synthesized (6*R*)-**8** and (6*S*)-**8** from the enantiomers of citronellal (**10**). The key steps were the intramolecular addition of an α-keto carbene to the alkene bond (**11** → **12**) and (*E*)-selective olefination of **13** to give **14**. The ^{1}H- and ^{13}C-NMR spectra of **8** around the trisubstituted double bond at C-2 were different from those of the natural pheromone.

(*Z*)-Alcohol **9** was therefore synthesized by (*Z*)-selective olefination of **13** with Ando's reagent **15**, giving **16**. ^{1}H- and ^{13}C-NMR spectra of synthetic (*R*,*Z*)-**9** and (*S*,*Z*)-**9** (both mixtures of diastereomers of C-1′ and C-5′) were very similar to those of the natural pheromone, and (*R*,*Z*)-**9** was pheromonally active against *E. lewisi*.[13] Mori's synthesis, however, could not determine the relative configuration at C-1′ and C-5′ of the pheromone.

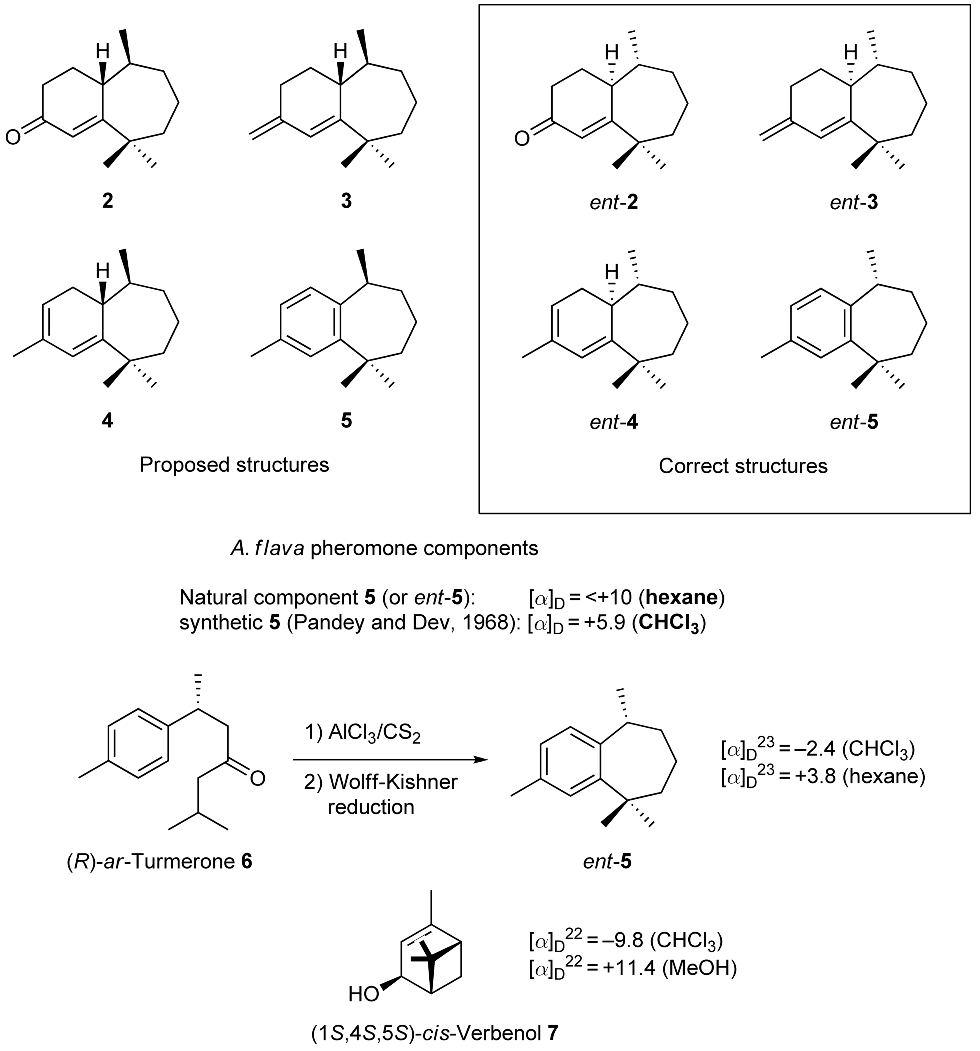

2
3
ent-2
ent-3
4
5
ent-4
ent-5
Proposed structures
Correct structures
A. flava pheromone components
Natural component 5 (or ent-5): [α]D = <+10 (hexane)
synthetic 5 (Pandey and Dev, 1968): [α]D = +5.9 (CHCl3)
1) AlCl3/CS2
2) Wolff-Kishner reduction
(R)-ar-Turmerone 6
ent-5
[α]D23 = –2.4 (CHCl3)
[α]D23 = +3.8 (hexane)
HO
[α]D22 = –9.8 (CHCl3)
[α]D22 = +11.4 (MeOH)
(1S,4S,5S)-cis-Verbenol 7

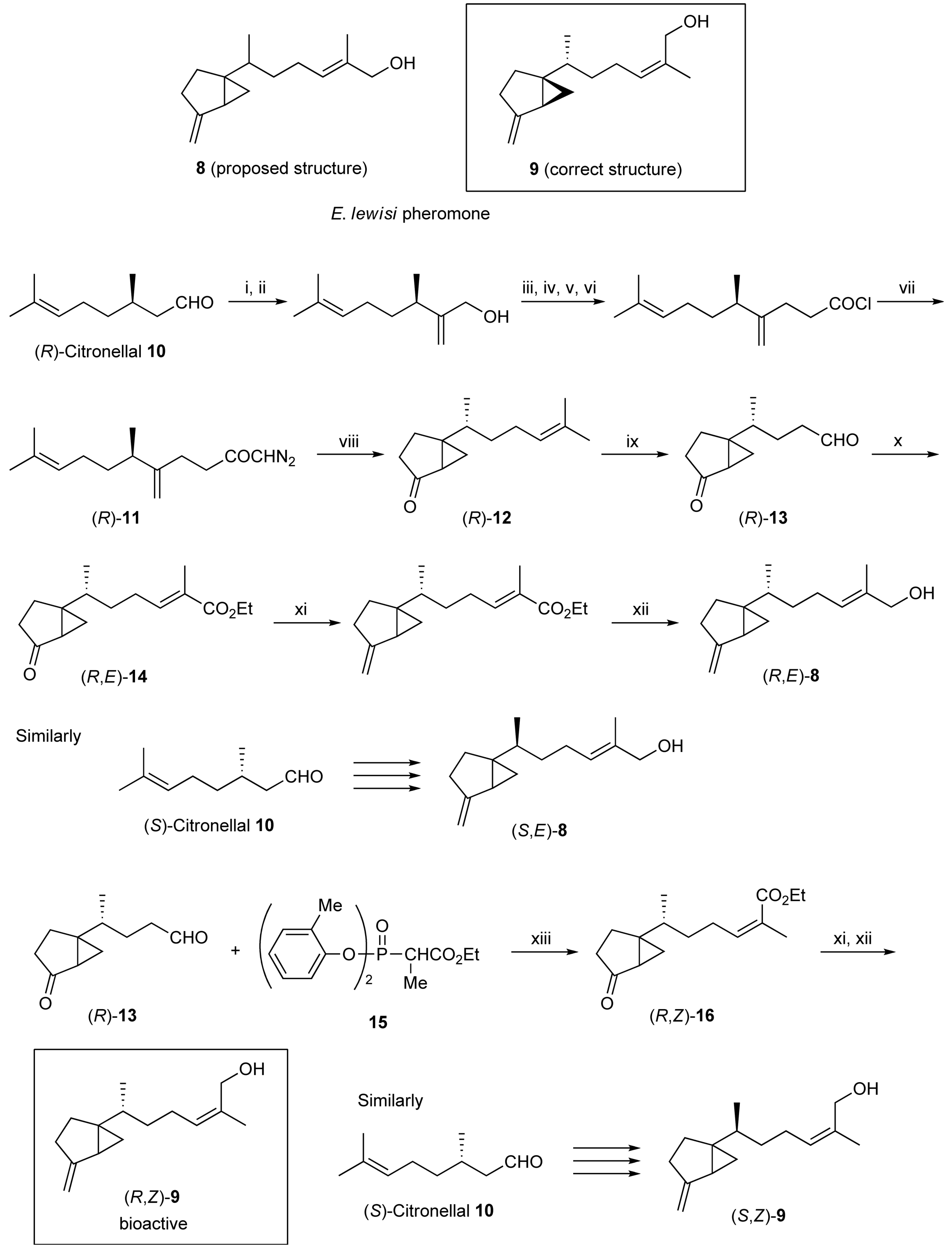
Figure 2 Absolute configuration of *Aphthona flava* pheromone components.

Mori *et al.*[14] finally determined the absolute configuration of the pheromone as (2*Z*,6*R*,1′*S*,5′*S*)-**9** by employing lipase-catalyzed asymmetric acetylation of **17′** as the key step (**Scheme 2**). Reduction of (6*R*)-**12** with L-selectride® afforded a mixture of **17** and **17′**, the latter of which could be acetylated with vinyl acetate in the presence of lipase PS-D (Amano) to give **18**. The remaining **17** was oxidized to give (6*R*,1′*S*,5′*R*)-**12**. Its absolute configuration was determined as depicted by CD comparison with (−)-sabina ketone **19** with a known absolute configuration. (2*Z*,6*R*,1′*S*,5′*S*)-**9** was synthesized from (6*R*,1′*S*,5′*R*)-**12**, while (6*R*,1′*R*,5′*S*)-**12** yielded (2*Z*,6*R*,1′*R*,5′*R*)-**9**. NMR and GC comparisons of these two products with the natural pheromone revealed (2*Z*,6*R*,1′*S*,5′*S*)-**9** to be the correct structure of the pheromone. Synthetic (2*Z*,6*R*,1′*S*,5′*S*)-**9** was biologically active, and none of its stereoisomers was either active or inhibitory.

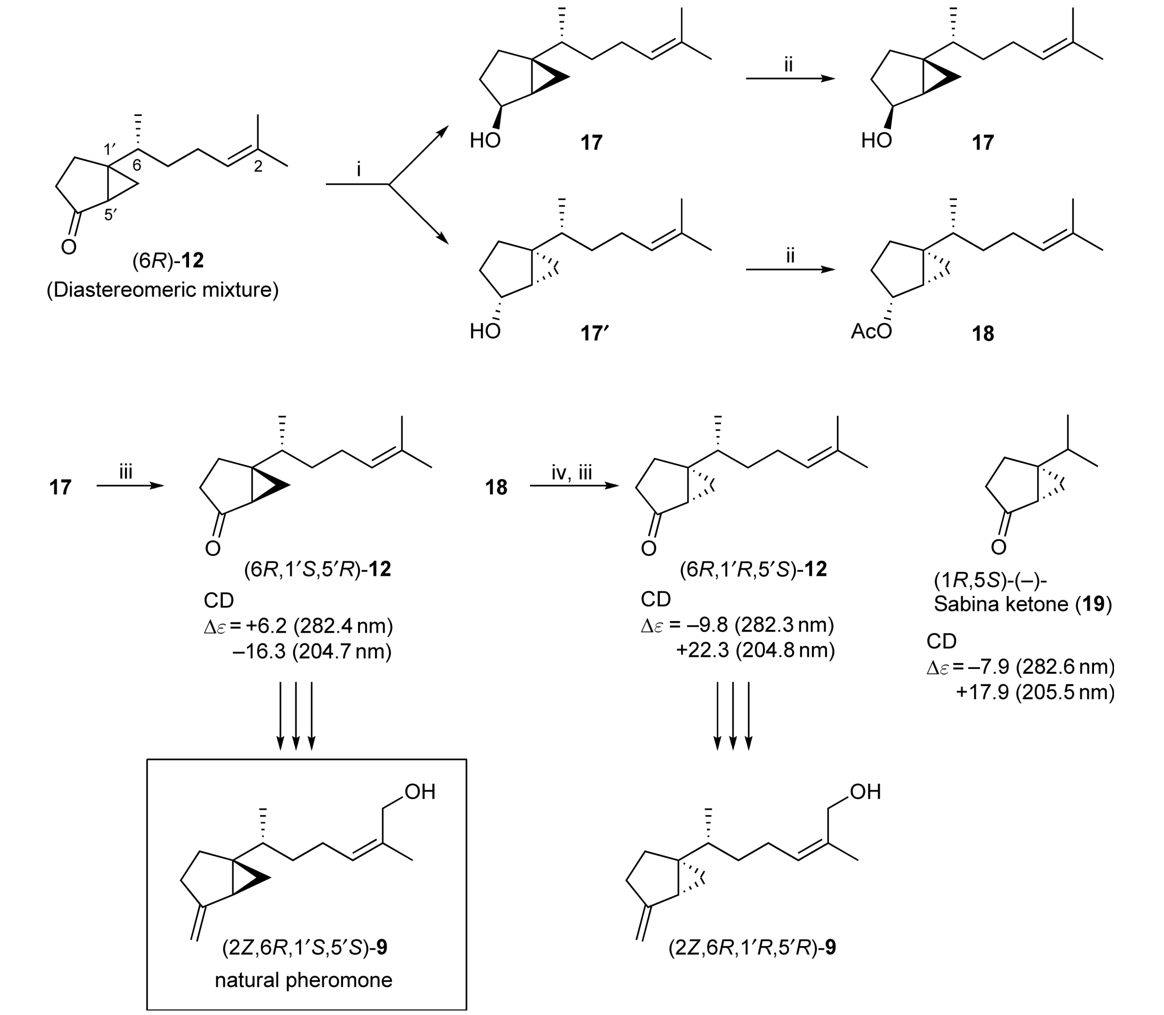
9.05.4 Absolute Configuration Involving Remote Stereocenters

9.05.4.1 German Cockroach Pheromone

In 1974, Nishida *et al.*[15–17] isolated and identified the components of the contact sex pheromone of the German cockroach *Blattella germanica*. They proposed the structures of the three components as **20**, **21**, and **22** (**Figure 3**). Their isolated amounts are shown in parentheses. As to the absolute configuration of **20** and **21**,



Scheme 1 Synthesis of the possible structures of the male-produced aggregation pheromone of the stink bug *Eysarcoris lewisi*. Reagents: (i) 37% CH_2O, $EtCO_2H$, pyrrolidine, Pr^iOH (90%); (ii) $LiAlH_4$, Et_2O (91%); (iii) $MeC(OEt)_3$, $EtCO_2H$, heat (95%); (iv) KOH, aq. EtOH (83%); (v) NaOEt, EtOH; (vi) $(COCl)_2$, C_5H_5N, hexane (quant., 2 steps); (vii) CH_2N_2, Et_2O (quant.); (viii) Cu, $CuSO_4$, cyclohexane, heat (58%); (ix) OsO_4, $NaIO_4$, THF, Bu^tOH, H_2O (quant.); (x) $Ph_3P{=}C(Me)CO_2Et$, THF, CH_2Cl_2 (57%); (xi) $Ph_3P(Me)Br$, Bu^nLi, THF (96%); (xii) Bu^i_2AlH, toluene (55%).



Scheme 2 Synthesis of the male-produced aggregation pheromone of the stink bug *Eysarcoris lewisi*. Reagents: (i) (a) LiBBu$^s{}_3$H, THF; (b) 30% H_2O_2, dil. NaOH (94%); (ii) (a) lipase PS-D (Amano), $CH_2{=}CHOAc$, Et_2O, room temperature, 10–13 h, repeat three times; (b) SiO_2 chromatography; (iii) $Pr^n{}_4NRuO_4$, NMO, MS 4A CH_2Cl_2, room temperature, 5 h (quant.); (iv) K_2CO_3, MeOH (quant.).

Nishida *et al.*[18] proposed the 3*S*-configuration on the basis of their optical rotatory dispersion (ORD) spectra coupled with NMR studies employing a chiral shift reagent. No information was available to assign the absolute configuration at C-11, because the stereocenter at C-11 was separated from the C-3 stereocenter by seven methylene groups.

Mori *et al.*[19] established the absolute configuration of 20 and 21 as 3*S*,11*S* by synthesizing all four stereoisomers of 20 and 21 and comparing their physical properties with those of the natural products. As shown in Figure 3, the stereoisomers of 20 and 21 were synthesized from (*R*)-isopulegol (23) via (*R*)-citronellic acid (24) of 92% ee (enantiomeric excess). Because the two stereocenters of 20 and 21 were separated, their stereoisomers showed identical ^{1}H- and ^{13}C-NMR spectra. However, their IR spectra as nujol mulls (i.e., as solid state and not as solutions) showed differences. Their optical rotations and melting points (mp's) were also very important in assigning the absolute configuration of natural 20 as shown in Table 1.

The natural ketone 20 was dextrorotatory in hexane, and (3*S*,11*S*)-20 as well as (3*S*,11*R*)-20 showed positive rotations, while (3*R*,11*R*)- and (3*R*,11*S*)-20 were levorotatory. The natural 20 must therefore be either (3*S*,11*S*)- or (3*S*,11*R*)-20. As chloroform solutions, all the stereoisomers of 20 showed IR spectra that were identical to

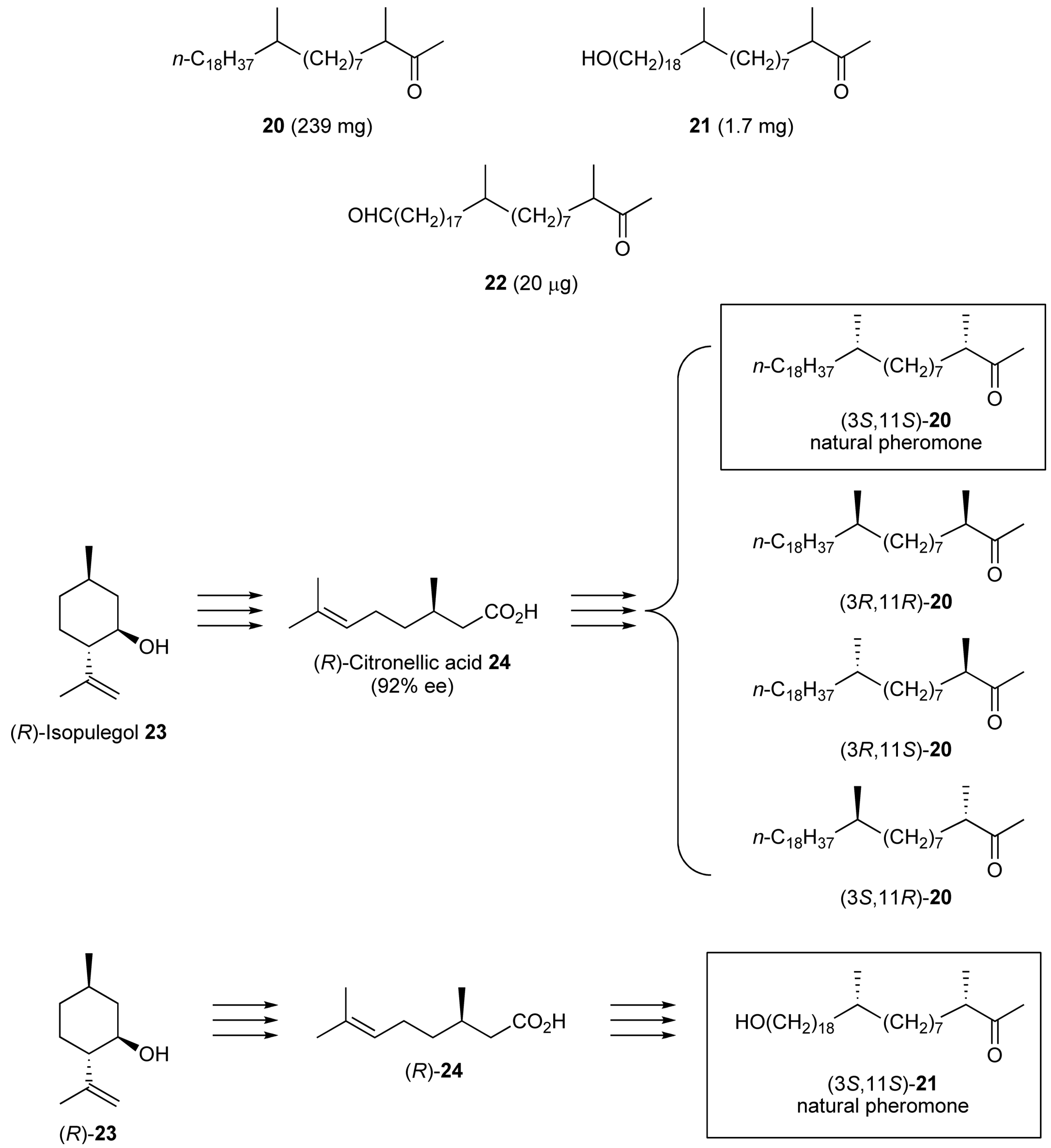


Figure 3 Structures of the sex pheromone components of the German cockroach *Blattella germanica* and related compounds.

each other. Their ^{1}H- and ^{13}C-NMR spectra were also indistinguishable. However, when their IR spectra were measured as nujol mulls, the stereoisomeric and crystalline ketones **20** showed subtle differences in the spectra due to the difference in their crystalline lattice structures. Thus, the IR spectrum of natural **20** was identical to those of (3*S*,11*S*)- and (3*R*,11*R*)-**20**. The natural **20** seemed to be (3*S*,11*S*)-**20** at this stage. To confirm this conclusion, the mp's of the four stereomers of **20** were measured, and the mixture mp determinations of the four isomers with the natural **20** were carried out. As can be seen from **Table 1**, (3*S*,11*S*)- and (3*R*,11*R*)-**20** showed the same mp as the natural **20**. Mixture mp determinations revealed (3*S*,11*S*)-**20** to be the natural **20**, because it showed no depression.[19] The classical method of mixture mp test is still useful in establishing the identity of two like samples. Similarly, the absolute configuration of the natural **21** could be established as 3*S*,11*S*.[19]

Later in 1990, highly pure (>99% ee) stereoisomers of **20** were synthesized from (*R*)-citronellal and ethyl (*R*)-3-hydroxybutanoate.[20] Bioassay of the four pure isomers of **20** by Schal and coworkers[21] showed that the natural pheromone (3*S*,11*S*)-**20** was the least effective of the four stereoisomers at eliciting courtship

Table 1 Specific rotations, IR spectra, and mp's of the natural and synthetic stereoisomers of **20** and their mixture mp's with natural **20**

Sample	$[\alpha]_D$ *(in hexane)*	*IR (nujol)*	*mp (°C)*	*Mixture mp with the natural* **20**
Natural **20**	+5.1 (c = 3.54)		45–46	–
(3*S*,11*S*)-**20**	+5.98 (c = 0.9)	same	44–44.5	44–45
(3*R*,11*R*)-**20**	–5.63 (c = 4.1)		44.5–45	35–37.5
(3*R*,11*S*)-**20**	–5.68 (c = 4.0)	different	39–39.5	34.5–35.5
(3*S*,11*R*)-**20**	+5.73 (c = 2.04)	same	38–38.5	33.5–35

responses in males. The German cockroach produces the least active (3*S*,11*S*)-**20** due to the stereochemical restriction in the course of its biosynthesis.

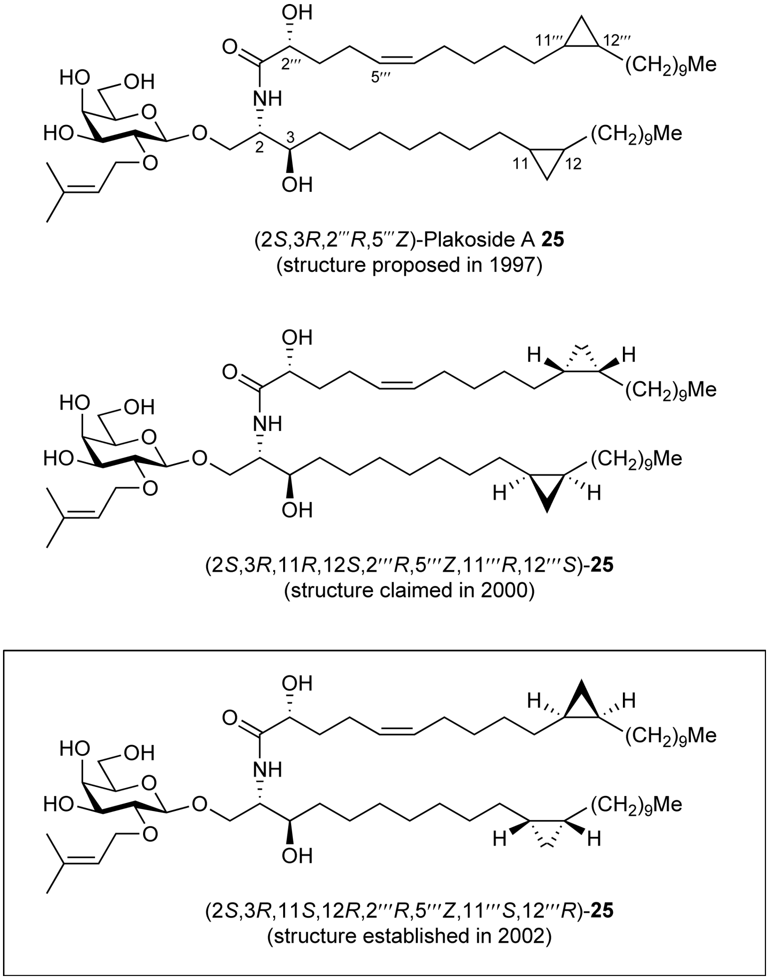
9.05.4.2 Plakoside A

In 1997, plakoside A (**25**) (**Figure 4**) was isolated by Fattorusso and coworkers[22] as an immunosuppressive metabolite of the Carribean sponge *Plakortis simplex*. It is a structurally unique glycosphingolipid with a prenylated D-galactose moiety and cyclopropane-containing alkyl chains. Its 2*S*,3*R*,2‴*R* stereochemistry was proposed on the basis of the CD measurements of its degradation products.[22] The absolute configuration at the two cyclopropane moieties of **25**, however, remained unknown, although the *cis*-stereochemistry was suggested by detailed ^{1}H-NMR analysis of **25**.[22]

In 2000, Nicolaou *et al.*[23] accomplished the synthesis of (2*S*,3*R*,11*R*,12*S*,2‴*R*,5‴*Z*,11‴*R*,12‴*S*)-**25**, and found its ^{1}H- and ^{13}C-NMR data to be identical to those reported for the natural **25**. They therefore claimed their synthetic product to be identical to the natural product. However, in 2001, Seki and Mori[24] synthesized both (2*S*,3*R*,11*R*,12*S*,2‴*R*,5‴*Z*,11‴*R*,12‴*S*)- and (2*S*,3*R*,11*S*,12*R*,2‴*R*,5‴*Z*,11‴*S*,12‴*R*)-**25**, both of which were spectroscopically indistinguishable from natural **25**. Then, which stereoisomer of **25** is plakoside A? In order to solve this problem, degradation studies of natural plakoside A (**25**) were executed as shown in **Figure 5**, and the degradation products were compared with the synthetic samples of known absolute configuration.[25]

Lipase-catalyzed asymmetric acetylation of *meso*-diol **26** gave enantiomerically pure **27**, which was converted to the enantiomers of the reference acids **28** and **29**. These acids were derivatized and analyzed by high-performance liquid chromatography (HPLC) according to Ohrui and coworkers.[26–28] Esterification of acids **28** and **29** with Ohrui's chiral and fluorescent reagent R*OH **30** yielded esters **31** and **32**. All of these derivatives were separable by reversed-phase HPLC at a column temperature of −50 °C. Owing to the presence of the anthracene system in **31** and **32**, their picogram quantities were detectable by fluorescence, and therefore minute amounts of degradation products could be analyzed.

Degradation of plakoside A pentaacetate (**33**) was executed by first treating it with nitrous acid in acetic anhydride through N-nitrosation at the amide nitrogen of **33** to give **34** and **35**, which were further cleaved to give **28** and **29**, respectively. A mixture of **28** and **29** was derivatized with **30**, and the products were subjected to HPLC analysis to show them to be (6*S*,7*R*)-**31** and (9*S*,10*R*)-**32**. Accordingly, the absolute configuration of plakoside A must be (2*S*,3*R*,11*S*,12*R*,2‴*R*,5‴*Z*,11‴*S*,12‴*R*)-**25**. The synthetic product (2*S*,3*R*,11*R*,12*S*,2‴*R*,5‴*Z*,11‴*R*,12‴*S*)-**25** of Nicolaou *et al.* turned out to be a diastereomer of plakoside A.[25] A combination of enantioselective synthesis and HPLC analysis is a powerful method for the determination of the absolute configuration of a compound with stereogenic centers remote from other functionalities and stereogenic centers.



(2*S*,3*R*,2‴*R*,5‴*Z*)-Plakoside A **25**
(structure proposed in 1997)

(2*S*,3*R*,11*R*,12*S*,2‴*R*,5‴*Z*,11‴*R*,12‴*S*)-**25**
(structure claimed in 2000)

(2*S*,3*R*,11*S*,12*R*,2‴*R*,5‴*Z*,11‴*S*,12‴*R*)-**25**
(structure established in 2002)

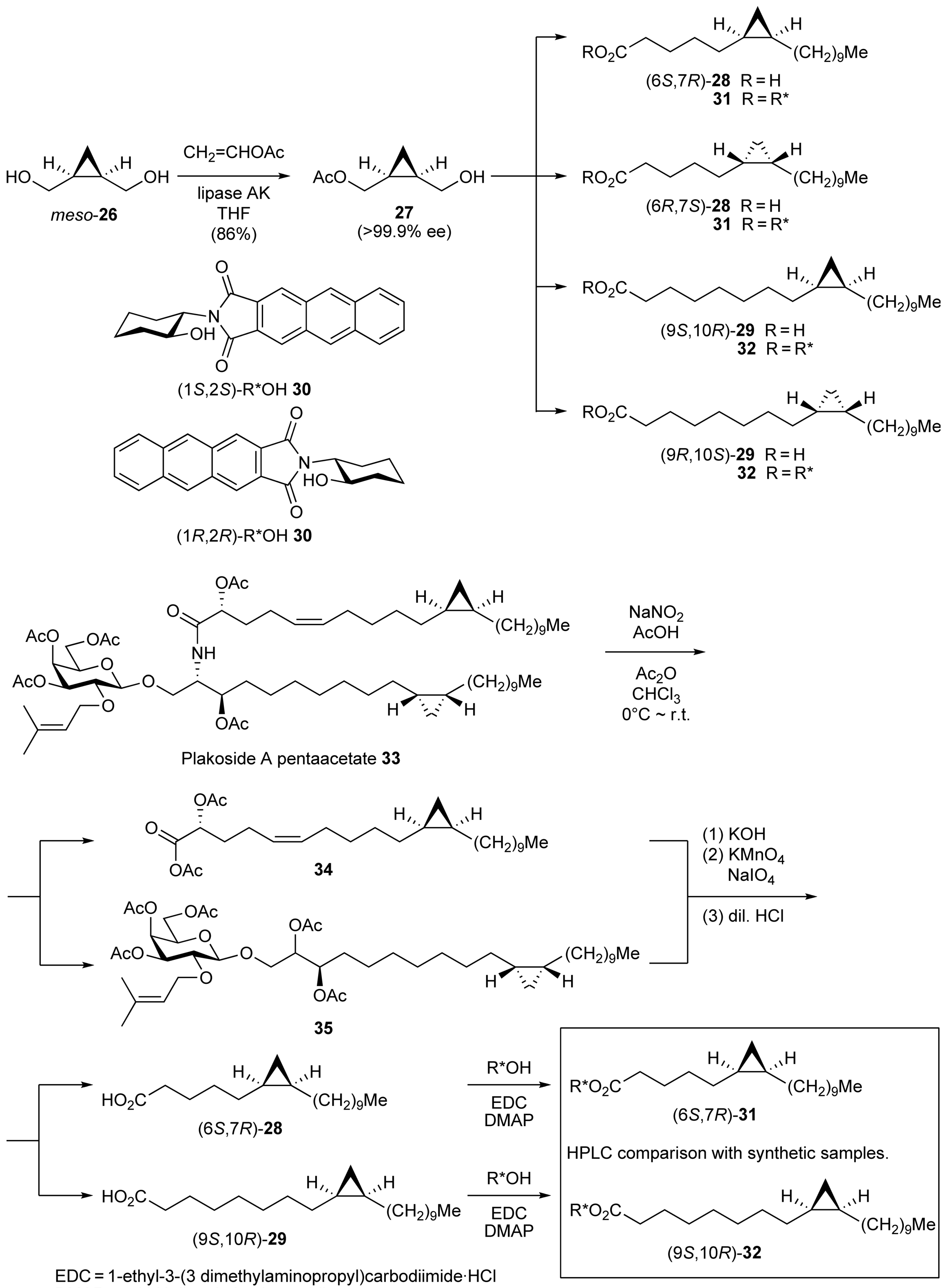
Figure 4 Structure of plakoside A (**25**).

9.05.5 Absolute Configuration Involving Stereocenters Separated by a Polymethylene Spacer

9.05.5.1 *cis*-Solamin

Annonaceus acetogenins, isolated from the plant species belonging to Annonaceae (custard apple family), are waxy solids with cytotoxic and antitumor activity. They are characterized by the presence of one or more 2,5-disubstituted tetrahydrofuran rings connected to a butenolide through a polymethylene spacer. As exemplified by *cis*-solamin A (**36**) and *cis*-solamin B (**37**) (**Figure 6**), metabolites of tropic fruit tree *Annona muricata*, two stereogenic moieties are separated by a polymethylene spacer. Within the tetrahydrofuran moiety, its relative configuration could be determined by NMR analysis as depicted, but its absolute configuration was difficult to determine. Brown and coworkers[29] synthesized four possible stereoisomers (**36**, *ent*-**36**, **37**, and *ent*-**37**) of *cis*-solamin by the route summarized in **Figure 6**.

The four synthetic isomers of *cis*-solamin were indistinguishable from each other and from natural *cis*-solamin on the basis of their IR, MS, ^{1}H-NMR, and ^{13}C-NMR spectra, owing to the length and flexibility of the spacer connecting the tetrahydrofurandiol and butenolide moieties. Optical rotation values obtained for each of the pairs of diastereomers were also very similar, and consistent with the known fact that the contribution to optical rotation from the butenolide moiety dominates that from a pseudosymmetrical



EDC = 1-ethyl-3-(3 dimethylaminopropyl)carbodiimide·HCl
DMAP = 4-*N*,*N*-dimethylaminopyridine

Figure 5 Determination of the absolute configuration of plakoside A.

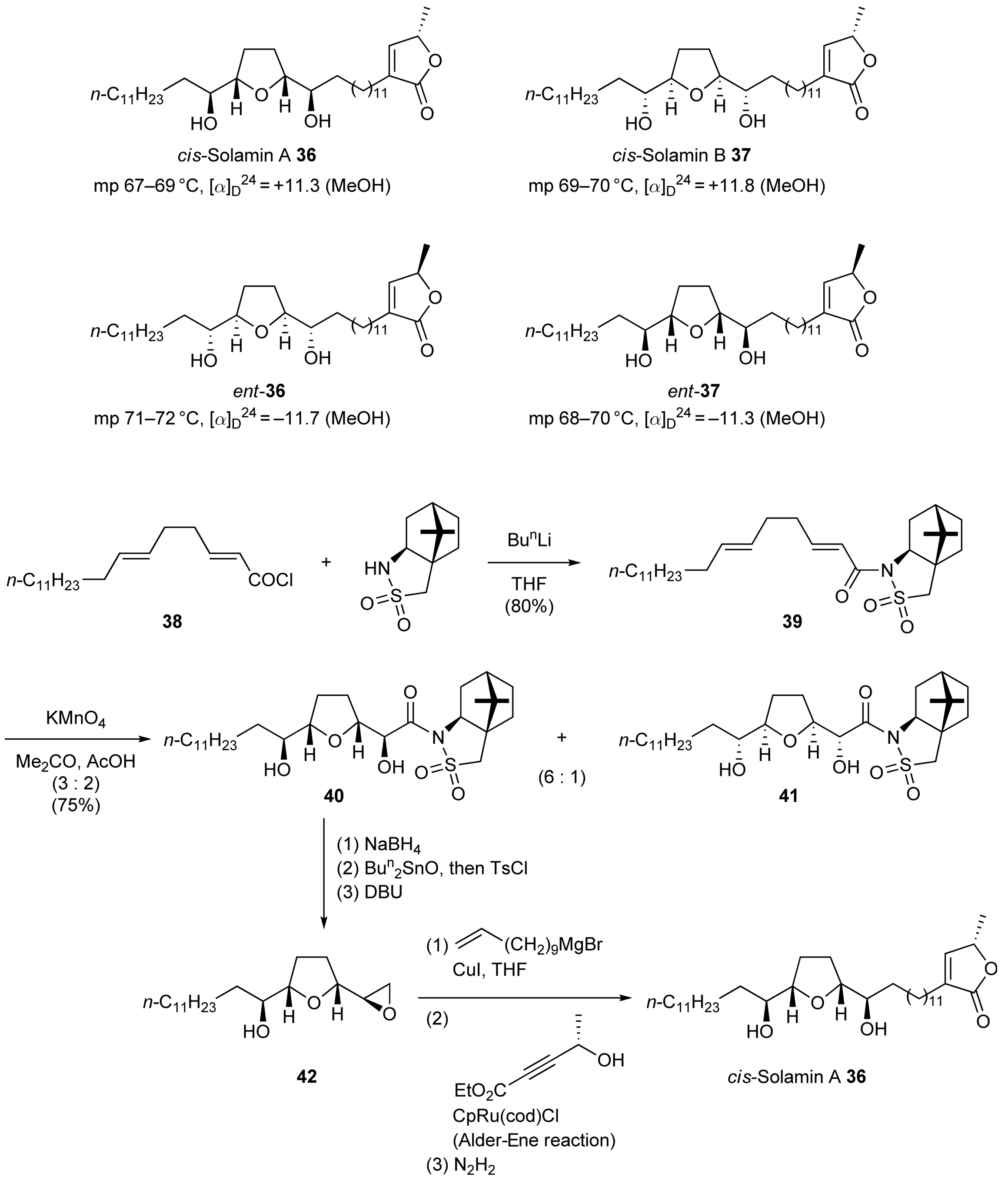


Figure 6 Structure and synthesis of *cis*-solamin A (36).

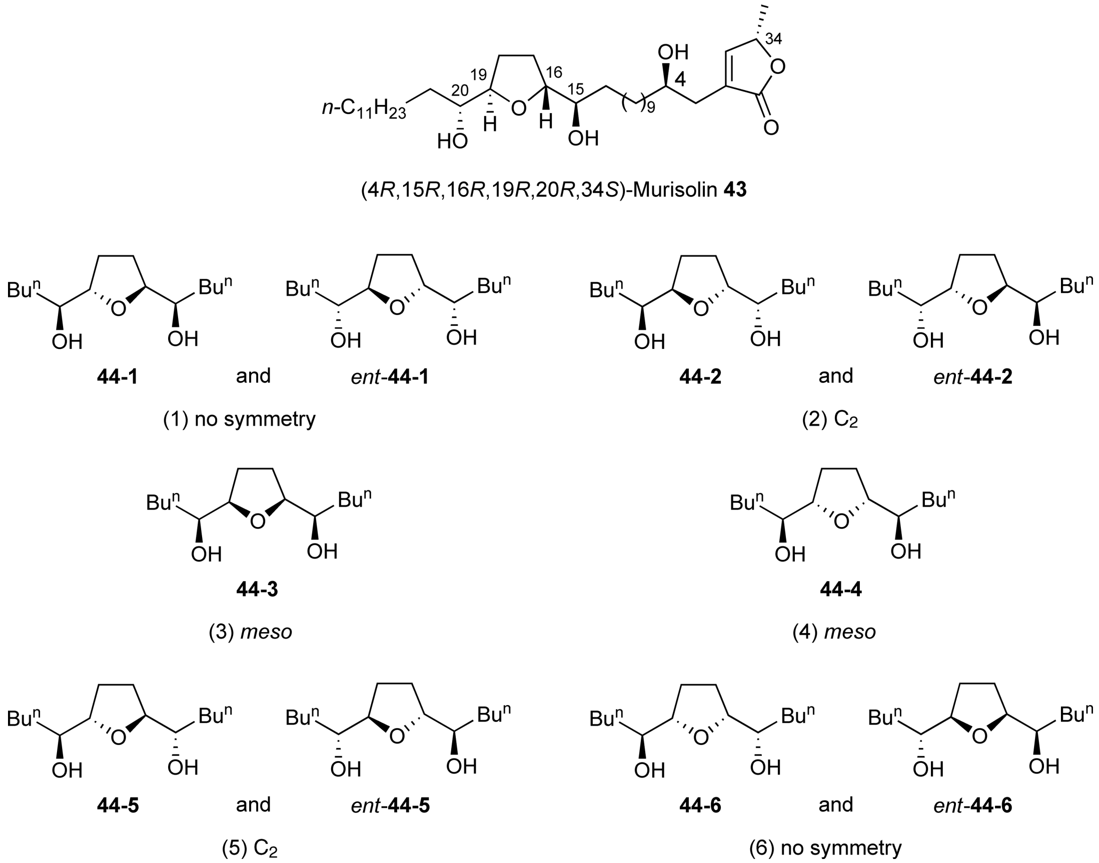
tetrahydrofurandiol region in acetogenins. In the course of this study, Brown and coworkers[29] found that all four isomers (36, *ent*-36, 37, and *ent*-37) were separable by enantioselective HPLC employing a cyclodextrin-based stationary phase.

Subsequently, Figadére and coworkers[30] analyzed natural *cis*-solamin by enantioselective HPLC, and found it to be a 9:8 mixture of *cis*-solamin A (36) and *cis*-solamin B (37). Thus, natural *cis*-solamin was stereochemically heterogeneous demonstrating that enantioselective chromatography is indeed a powerful technique in stereochemical studies of natural products. Although NMR and X-ray analysis are regarded as the most powerful techniques for structure elucidation, a chromatographic method gave the decisive evidence to show the heterogeneity of *cis*-solamin.

9.05.5.2 Murisolin

In 2006, Curran *et al.*[31] published an important paper on murisolin (**43**; **Figure 7**), another acetogenin, entitled 'On the proof and disproof of natural product stereostructures'. They synthesized two 16-member stereoisomer libraries of murisolin isomers that provided 24 of the 32 possible diastereomers of murisolin (**43**). Each member of the 16-member sublibrary of murisolins was subjected to NMR analysis at 600 MHz (^{1}H) and 150 MHz (^{13}C). The library members have 4*R*,34*S* configurations in the butenolide moiety with all possible configurations at the remaining stereogenic centers in the tetrahydrofurandiol fragment. Every NMR spectrum belongs to one of only six groups, and the spectra within each group are substantially identical. Symmetry considerations of simple model compounds **44** as shown in **Figure 7** help us to understand why there are only six groups. The butenolide group in **43** is substantially separated from the tetrahydrofurandiol moiety, and therefore ^{1}H-NMR spectrum of the former cannot be affected sufficiently to show differences due to the stereochemistry of the latter.

The six groups of the ^{1}H-NMR spectra were organized according to the local symmetry of the tetrahydrofurandiol moiety. On the basis of this NMR information, inspection of the NMR spectrum of a murisolin stereoisomer enables users to assign the relative configuration to its tetrahydrofurandiol region. Very small (≤0.1 ppm) differences were observed in the hydroxybutenolide region of the 150 MHz ^{13}C-NMR spectra of



(4*R*,15*R*,16*R*,19*R*,20*R*,34*S*)-Murisolin **43**

44-1 and *ent*-**44-1**

(1) no symmetry

44-2 and *ent*-**44-2**

(2) C2

44-3

(3) *meso*

44-4

(4) *meso*

44-5 and *ent*-**44-5**

(5) C2

44-6 and *ent*-**44-6**

(6) no symmetry

Figure 7 Structure of murisolin (**43**) and group classifications of its stereoisomers on the basis of simple model compounds **44**.

murisolin stereoisomers, based on the *syn/anti* relative configuration at C-4 and C-34. Derivatization of **43** and its stereoisomers to tris-(*S*)-Mosher esters followed by NMR measurements revealed that murisolin-Mosher ester stereoisomers exhibited one of only 10 sets of ^{1}H-NMR spectra. Through these observations it was possible to assign 4*R*,15*R*,16*R*,19*R*,20*R*,34*S* configuration to murisolin (**43**), which was in accord with the previous proposals. Curran *et al.* also comment that enantioselective HPLC is superior to either optical rotation or melting point comparisons to prove or disprove structures, if all the candidate isomers are available.

Curran *et al.*'s work informs us that construction of stereoisomer libraries followed by thorough studies on their NMR and enantioselective HPLC behaviors is an especially reliable way of elucidating the stereostructure of natural products. This type of thorough stereochemical analysis is likely to become more popular in the future in connection with the advances in parallel synthesis.

9.05.5.3 New World Screwworm Fly

Female-produced sex pheromones of the New World screwworm fly *Cochliomyia hominivorax* were first studied by Pomonis *et al.*[32] in 1993. They isolated 16 pheromone candidates from the female flies, but they were unable to identify the pheromonally active compounds. In 2002, Mori and coworkers[33] synthesized stereoisomeric mixtures of **45** and **46** (**Figure 8**), and they were found to be pheromonally active. Subsequently, all four stereoisomers of the more potent acetate **45** were synthesized as shown in **Figure 8**.[34] In the course of the synthesis, the parent alcohol **50** was esterified with the anthracene-containing acid (1*S*,2*S*)-**51**, and the derived ester **52** was analyzed by HPLC at −25 °C by the method of Ohrui and coworkers.[26–28] All four diastereomers of **52** were separable, and therefore the stereochemical purities of the four isomers of **45** could be estimated as depicted.[34]

The four isomers of **45** showed identical IR, ^{1}H-NMR, and ^{13}C-NMR spectra. In addition, all of them were equally bioactive as the sex pheromone. Accordingly, their derivatization to **52** followed by HPLC analysis was the only way to distinguish the stereoisomers. Finally, the natural pheromone component was shown to be (6*R*,19*R*)-**45** by its derivatization to **52** followed by HPLC analysis.[35] Enantioselective HPLC or gas chromatography (GC) and chromatographic analysis after derivatization with Ohrui's reagent seems to be the most sensitive method for discrimination of stereoisomers. Thus, it may be concluded that structural analysis must, from time to time, be carried out by employing various kinds of different analytical methods. Otherwise, mistakes are likely to occur.

9.05.6 Origin of Biological Homochirality

A characteristic hallmark of life is believed to be its 'homochirality'.[36] In general, it is true, although natural products are not always enantiomerically pure.[37] The origin of biomolecular homochirality is discussed in depth by MacDermott.[36] Those who are interested to see whether the parity-violating weak force is the cosmic dissymmetry that Pasteur was looking for should read her chapter in the book entitled '*Chirality in Natural and Applied Science*'.

Soai *et al.*[38] discovered and developed asymmetric autocatalysis (**Figure 9**), in which the structures of the chiral catalyst (*S*)-**54** and the chiral product (*S*)-**54** are the same after the addition of diisopropylzinc to aldehyde **53**. Consecutive asymmetric autocatalysis starting with (*S*)-**54** of 0.6% ee amplifies its ee, and yields itself as the product with >99.5% ee. Even chiral inorganic crystals, such as quartz or sodium chlorate, act as chiral inducers in this reaction. Soai *et al.*'s asymmetric autocatalysis gives us an insight to speculate on the early asymmetric reactions on this planet Earth. However, it can be argued whether such strictly anhydrous organometallic reactions are possible under the nonartificial conditions or not.

A phenomenon that may be related to the origin of biological homochirality was recently reported by Cooks and coworkers:[39] Serine sublimes with spontaneous chiral amplification. Sublimation of near racemic sample of serine **55** (**Figure 9**) yields a sublimate that is enriched in the major enantiomer. The chiral purity maximizes at 190–210 °C, and then falls as thermolysis becomes favorable. This simple one-step sublimation may represent a possible mechanism for the chiral amplification step to explain the origin of biological homochirality.

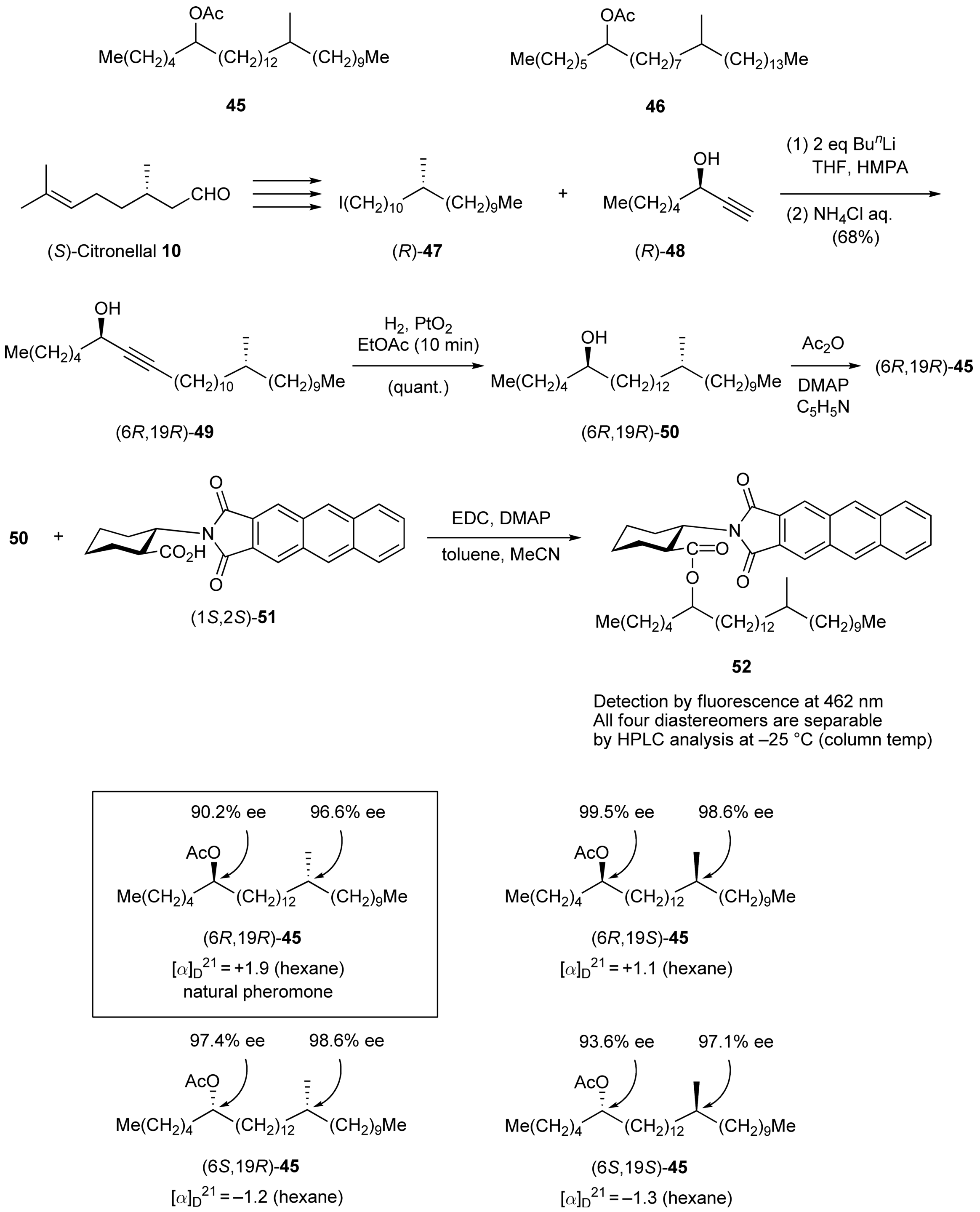


Figure 8 Synthesis of four stereoisomers of the most potent component **45** of the female sex pheromone of the New World screwworm fly *Cochliomyia hominivorax*.

9.05.7 Exceptions to Biological Homochirality

Until recently, we believed our human bodies to be constituted from L-amino acids only. Advances in analytical methods now indicate that there are a number of D-amino acids in human bodies as detailed in the review by Fujii and Saito.[40] Free D-serine was observed predominantly in mammalian brain, and free D-aspartic acid

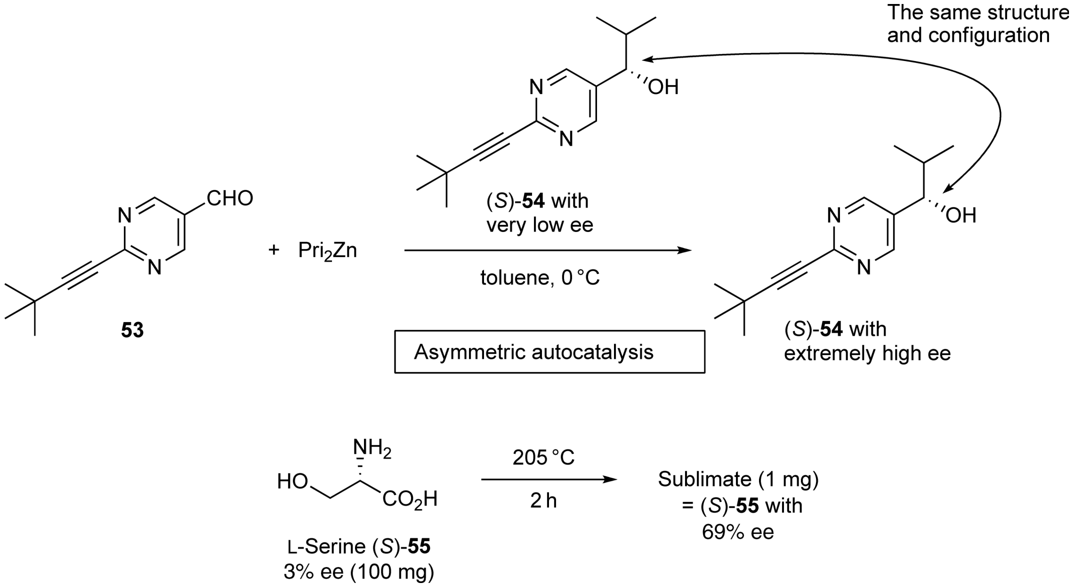


Figure 9 Structures of compounds in connection with biological homochirality (1).

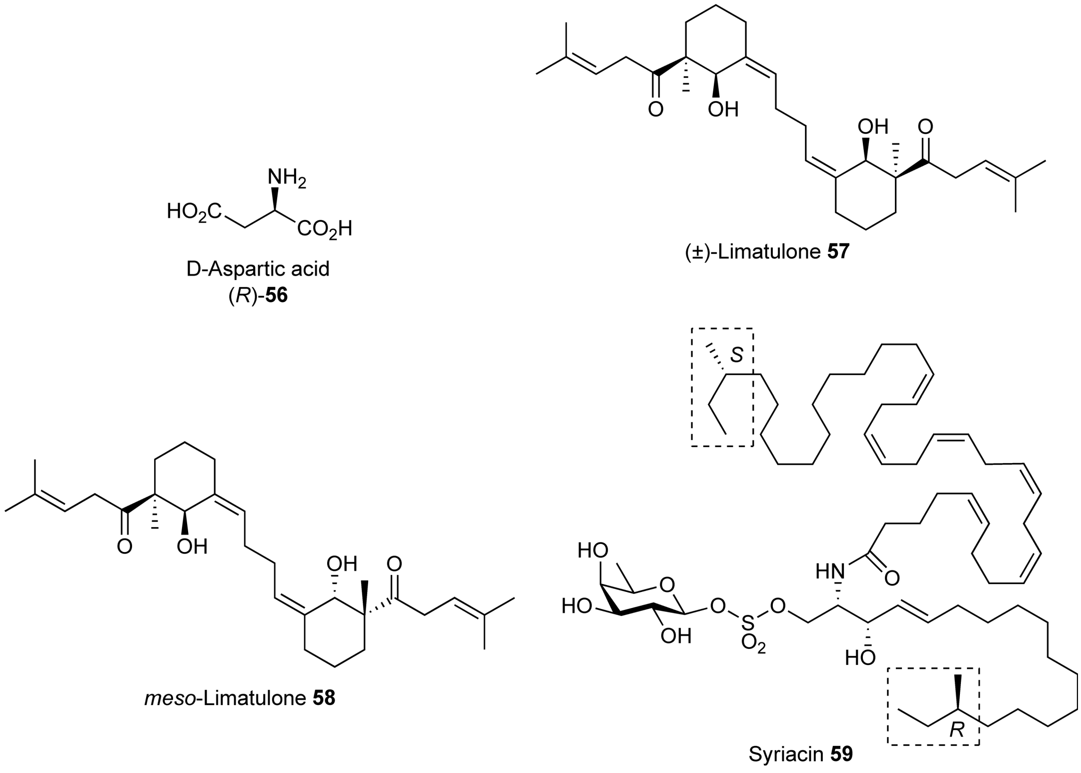


Figure 10 Structures of compounds in connection with biological homochirality (2).

(56; Figure 10) exists in various mammalian tissues. For example, in the prefrontal cortex of human brain, as much as 60% of the total aspartic acid was in the D-form at week 14 of gestation, but rapidly decreased to trace levels by the time of birth.

D-Amino acids were detected in various aged human tissues such as tooth, bone, aorta, brain, erythrocytes, eye lens, skin, ligament, and lung. Especially D-serine was found in the β-amyloid protein of Alzheimer's disease. Stereoinversion of L-aspartic acid to D-aspartic acid takes place in alpha A and alpha B crystallins of human lens. Thus, aged persons possess more D-aspartic acid in the lens. This phenomenon seems to be related to cataract.[40]

The fluctuation of the amount of free D-amino acids in living bodies suggests that D-amino acids might be one of the factors controlling the generation and differentiation of cells or tissues. D-Amino acids in proteins can be interpreted as molecular markers of aging.[40] Another review is available concerning the occurrence and functions of free D-aspartic acid and its metabolizing enzymes.[41]

As already described in the first edition of this Comprehensive Series,[42] the limpet *Achmeia (Collisella) limatula* produces a defensive metabolite, limatulone, both as the racemate (**57** and its enantiomer) and as the *meso*-isomer **58**.[43] This example demonstrates that nature does not always produce enantiomerically pure compounds.

In 2006, Rezanka *et al.*[44] isolated an antifeedant, syriacin (**59**), from the freshwater sponge *Ephydatia syriaca* in the Jordan river. It is an unusual sulfated ceramide glycoside with branched-chain sphingosine and also a branched-chain fatty acid. From the viewpoint of absolute configuration, **59** is very unusual, because it contains both (*R*)- and (*S*)-configured *sec*-butyl terminals in its alkyl chains. It seems to be biosynthesized from precursors with opposite absolute configuration.

9.05.8 Mimics of Bioactive Natural Products and Bioisosterism

There are practical demands for the invention of pheromone mimics, because pheromones are often too labile to be used in pest control. Various mimics have been prepared to date, several of which will be described in this section.

Tacke *et al.*[45] synthesized the enantiomers of sila-linalool (**61**) as shown in **Figure 11**. The starting material **60** was converted into (±)-**61**, which was resolved by GC to give both (+)-**61** and (−)-**61**. Both enantiomers were bioactive as tested by electroantennographic detection (EAD) on the males of the vernal solitary bee *Colletes cunicularius*. There was no major difference between the bioactivity of the sila-pheromone **61** and the natural linalool. The substitution of a carbon atom by silicon provides a good example of bioisosterism.

(1*S*,5*R*)-Frontalin (**62**) is the aggregation pheromone of bark beetles such as *Dendroctonus brevicomis* and *D. frontalis*. Strunz *et al.*[46] synthesized its isomer **63**, which was shown to be pheromonally active. Bravo *et al.*[47] synthesized the trifluoro analogue **64** of frontalin. Its bioactivity, however, was not reported.

(4*S*,5*R*)-Eldanolide (**65**) is the male-produced sex pheromone of the African sugarcane borer *Eldana saccharina*. Itoh *et al.*[48,49] reported the synthesis and pheromonal activity of its fluorinated analogues **66–68**. Two analogues, **66** and *ent*-**66**, were bioactive, while the remaining four analogues showed no activity as revealed by EAD.

(7*R*,8*S*)-Disparlure (**69**) is the female-produced sex pheromone of the gypsy moth *Lymantria dispar*. Plettner and coworkers[50] synthesized and bioassayed its 5-oxa analogues **70** and *ent*-**70**. GC-EAD bioassay revealed both **70** and *ent*-**70** to be bioactive. The dose–response curve for **70** and that for *ent*-**70** were similar. Interestingly, pheromone-binding protein 1 (PBP1), which binds (7*S*,8*R*)-*ent*-**69** strongly, binds **70** and *ent*-**70** with nearly the same affinity as *ent*-**69**. The affinity of PBP1 for naturally occurring (7*R*,8*S*)-**69** is known to be much weaker than for *ent*-**69**. Neither **70** nor *ent*-**70** functioned as a pheromone inhibitor. The concept of bioisosterism works in this case, too, although with a subtle difference.

(4*R*,8*R*)-4,8-Dimethyldecanal (**71**; tribolure) is the aggregation pheromone of the flour beetle *Tribolium castaneum* and *T. confusum*. Due to the air sensitivity of **71** as an aldehyde, the more stable formate ester **72** was synthesized. This was found to be bioactive and was used in commercial pheromone traps.[51] This is an example of bioisosterism by which a carbon atom is replaced by an oxygen atom.

(2*S*,3*R*,1′*R*)-Stegobinone (**73**) is the female sex pheromone of the drugstore beetle *Stegobium paniceum*. Its (2*S*,3*R*,1′*S*) isomer is a strong inhibitor of pheromone action. The methyl group at C-1′ of **73** is so readily epimerizable that the natural **73** soon becomes biologically inactive, and **73** cannot be used practically.

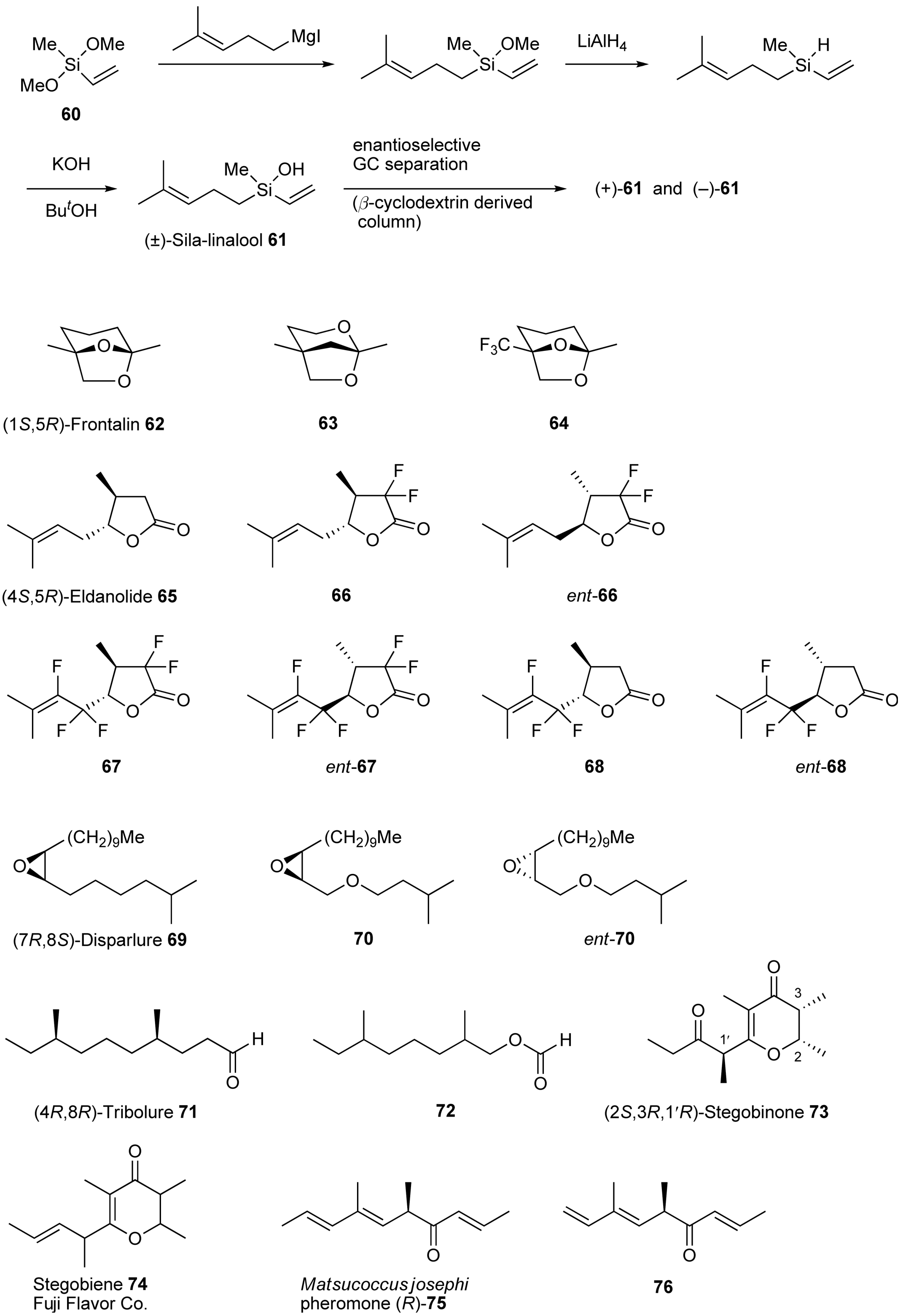


Figure 11 Structures of pheromones and their mimics.

Scientists at Fuji Flavor Co. synthesized stegobiene (**74**), which showed pheromone activity and could be used commercially to monitor the population of the drugstore beetle.

The female sex pheromone (*R*)-**75** of the Israeli pine bast scale *Matsucoccus josephi* is also a potent kairomone that attracts the scale insect's predator *Elatophilus hebraicus*. A mimic **76** of the pheromone **75** shows only the pheromone activity with no kairomone activity.[52,53] Accordingly, **76** is a more useful population-monitoring agent for *M. josephi* than the pheromone itself, which also catches the beneficial predator *E. hebraicus*.

9.05.9 Inventions of Pesticides and Medicinals

Natural products continue to be prototypes of pesticides and medicinals. Chemists' creativity and efforts have brought about many new mimics that are more potent, more economical, and more stable or safer than the original natural products.

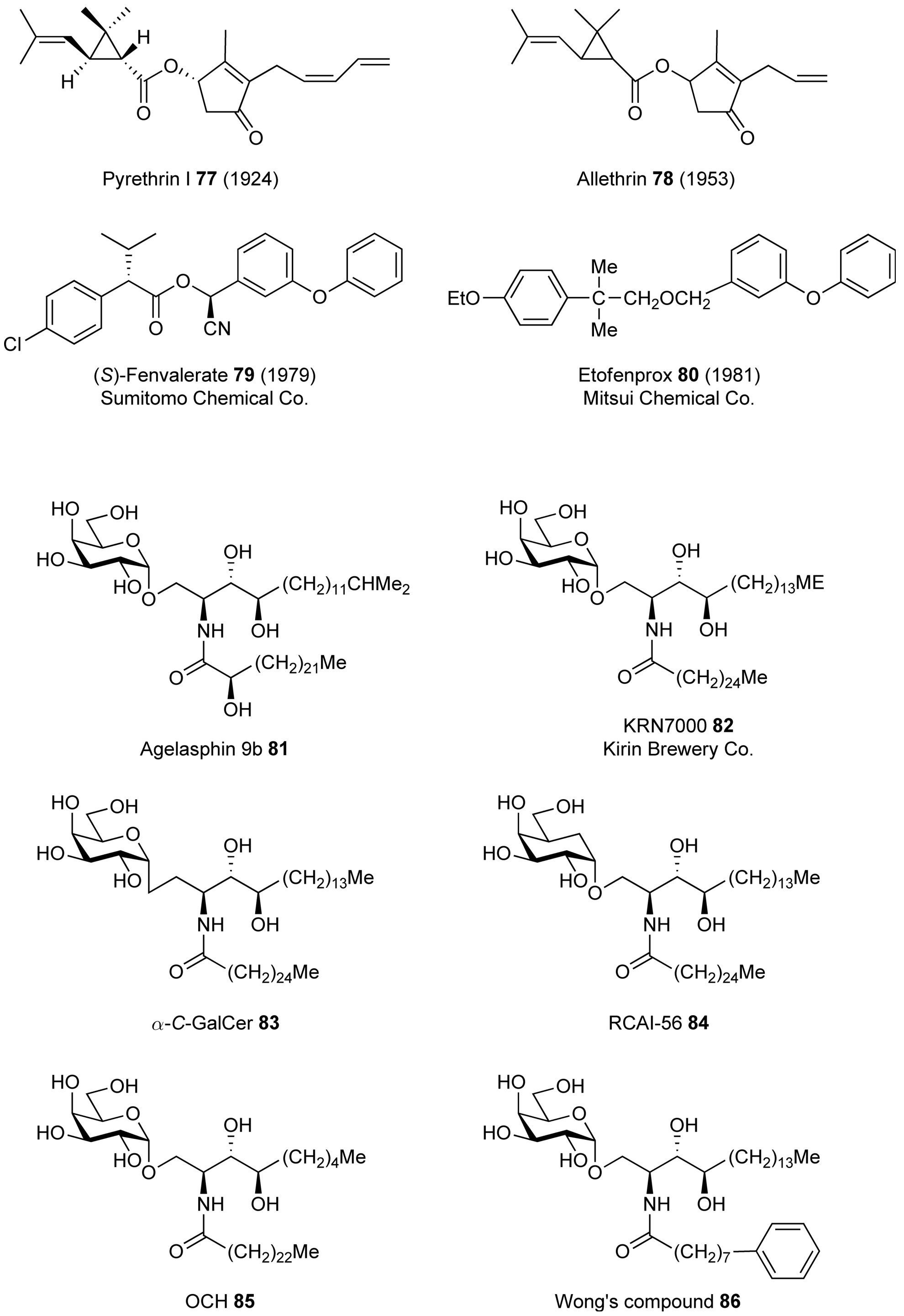
Pyrethrum powder is the dried flowerheads of *Chrysanthemum cinerariaefolium* and has been used widely as an insecticide. Its active principle was studied by L. Ruzicka, H. Staudinger, R. Yamamoto, and others, and pyrethrin I (**77**) (**Figure 12**) was identified as the major component.[54] Even now, after 80 years following the elucidation of the structure, modification of **77** continues to generate a group of insecticides called pyrethroids. Allethrin (**78**) was the first mimic to be manufactured in a large scale by Sumitomo Chemical Co. in 1953. Subsequently, in 1979 Sumitomo developed (*S*)-fenvalerate (**79**), while in 1981 etofenprox (**80**) was commercialized by Mitsui Chemical Co. These two compounds are stable in field conditions and are widely used as agricultural insecticides.[55]

Agelasphin 9b (**81**) and its relatives were isolated from the Okinawan marine sponge *Agelas mauritianus* as glycosphingolipids and they exhibited anticancer activity *in vivo* in mice and humans.[56] By simplifying the structure of **81**, researchers at Kirin Brewery Co. developed KRN7000 (**82**) as an anticancer drug candidate.[57]

It has been shown that KRN7000 (**82**) is a ligand that forms a complex with CD1d protein, a glycolipid presentation protein on the surface of the antigen-presenting cells of the immune system. Lipid alkyl chains of **82** are bound in the grooves in the interior of the CD1d protein and the galactose head group of **82** is presented to the invariant Vα14 antigen receptors of natural killer (NK)T cells. After activation by recognition of the CD1d–**82** complex, NKT cells release both helper T1 (Th1) and Th2 types of cytokines simultaneously in large quantities. Th1-type cytokines such as interferon (IFN)-γ mediate protective immune functions like tumor rejection, whereas Th2-type cytokines such as interleukin (IL)-4 mediate regulatory immune functions to ameliorate autoimmune diseases. Th1- and Th2-type cytokines can antagonize each other's biological actions. Because of this antagonism, the use of **82** for clinical therapy has not been successful yet. To circumvent this problem, many research groups modified the structure of KRN7000 (**82**) to develop new analogues of **82** that induce NKT cells to produce either Th1- or Th2-type cytokines.

Modification of the α-galactosyl part of **82** afforded α-*C*-GalCer (**83**)[58] and RCA1-56 (**84**).[59] These two compounds showed an enhanced Th1-type response *in vivo* to generate IFN-γ. Modification of the phytosphingosine part of **82** by shortening the alkyl chain to give OCH (**85**) resulted in an enhanced Th2-type response *in vivo* to produce IL-4.[60] Introduction of an aromatic ring at the end of the fatty acid chain to give **86** caused enhanced IFN-γ production.[61] Chemical modification of the parent KRN7000 (**82**) turned out to be a promising way to invent a more specific anticancer drug candidate.

It is true that we can see computer-generated docking models of the bioactive prototype compounds and their receptors. Even so, however, invention of mimics is restricted by the limited human capacity to imagine only conventional changes in functional groups and skeletons of the parent compounds. Natural products certainly will give us new and vast opportunities to find out unusual structures beyond our imagination. There are many mimics of natural products in the areas of taste, flavor, and fragrance. These interesting topics have been treated in Volume 4.



Pyrethrin I **77** (1924)

Allethrin **78** (1953)

(*S*)-Fenvalerate **79** (1979)
Sumitomo Chemical Co.

Etofenprox **80** (1981)
Mitsui Chemical Co.

Agelasphin 9b **81**

KRN7000 **82**
Kirin Brewery Co.

α-*C*-GalCer **83**

RCAI-56 **84**

OCH **85**

Wong's compound **86**

Figure 12 Structures of natural product-inspired pesticides and medicinals.

9.05.10 Conclusion

This chapter has treated the three important points in the studies of bioactive natural products.

First, recent examples of the determination of structure including the absolute configuration of bioactive natural products have been discussed, emphasizing the techniques to solve stereochemical problems among compounds with remote stereogenic moieties separated by a polymethylene spacer. Case studies with pheromones and acetogenins have been given to illustrate the problems and solutions.

Second, problems related to biological homochirality have been discussed to contemplate its origin and also to see exceptions to the homochirality principle. The presence and roles of D-amino acids in organisms as well as the presence of stereochemically heterogeneous compounds have been illustrated with examples.

Third, invention of mimics of bioactive natural products has been briefly discussed to show the importance of natural products as prototypes of pesticides and medicinals.

There are so many new discoveries in natural products research that no one can be an expert in all the areas unless they were content to be superficial. Thus, we have to remember the following words of Apostle Paul, "The person who thinks he knows something really does not know as he ought to know." (I Corinthians 8:2).

Abbreviations

EAD	electroantennographic detection
HPLC	high-performance liquid chromatography
IFN	interferon
IL	interleukin
NK	natural killer
mp	melting point
ORD	optical rotatory dispersion
PBP1	pheromone-binding protein 1

References

1. K. Mori, Ed., *Comprehensive Natural Products Chemistry, Vol. 8: Miscellaneous Natural Products including Marine Natural Products, Pheromones, Plant Hormones, and Aspect of Ecology*; Elsevier: Oxford, 1999; Chapter 1, pp 2–6.
2. K. Mori, *Chem. Rec.* **2005**, *5*, 1–6.
3. K. C. Nicolaou; S. A. Snyder, *Angew. Chem. Int. Ed. Engl.* **2005**, *44*, 1012–1044.
4. A. J. Bourdelais; H. M. Jacocks; J. L. C. Wright; P. M. Bigwarfe, Jr.; D. G. Baden, *J. Nat. Prod.* **2005**, *68*, 2–6.
5. H. Fuwa; M. Ebine; A. J. Bourdelais; D. G. Baden; M. Sasaki, *J. Am. Chem. Soc.* **2006**, *128*, 16989–16999.
6. R. J. Bartelt; A. A. Cossé; B. W. Zilkowski; D. Weisleder; F. A. Momany, *J. Chem. Ecol.* **2001**, *27*, 2397–2423.
7. S. Muto; M. Bando; K. Mori, *Eur. J. Org. Chem.* **2004**, 1946–1952.
8. M. Tóth; E. Csonka; R. J. Bartelt; A. A. Cossé; B. W. Zilkowski; S. Muto; K. Mori, *J. Chem. Ecol.* **2005**, *31*, 2705–2720.
9. R. C. Pandey; S. Dev, *Tetrahedron* **1968**, *24*, 3829–3839.
10. K. Mori, *Tetrahedron: Asymmetry* **2005**, *16*, 685–692.
11. K. Mori; N. Mizumachi; M. Matsui, *Agric. Biol. Chem.* **1976**, *40*, 1611–1615.
12. M. Takita, *Tohoku Nogyo Kenkyu Seika Joho* **2005**, *19*, 50–51.
13. K. Mori, *Tetrahedron Asymmetry* **2007**, *18*, 838–846.
14. K. Mori; T. Tashiro; T. Yoshimura; M. Takita; J. Tabata; S. Hiradate; H. Sugie, *Tetrahedron Lett.* **2008**, *49*, 354–357.
15. R. Nishida; H. Fukami; S. Ishii, *Experientia* **1974**, *30*, 978–979.
16. R. Nishida; H. Fukami; S. Ishii, *Appl. Entomol. Zool.* **1975**, *10*, 10–18.
17. R. Nishida; T. Sato; Y. Kuwahara; H. Fukami; S. Ishii, *J. Chem. Ecol.* **1976**, *2*, 449–455.
18. R. Nishida; Y. Kuwahara; H. Fukami; S. Ishii, *J. Chem. Ecol.* **1979**, *5*, 289–297.
19. K. Mori; S. Masuda; T. Suguro, *Tetrahedron* **1981**, *37*, 1329–1340.
20. K. Mori; H. Takikawa, *Tetrahedron* **1990**, *46*, 4473–4486.
21. D. Eliyahu; K. Mori; H. Takikawa; W. S. Leal; C. Schal, *J. Chem. Ecol.* **2004**, *30*, 1839–1848.
22. V. Costantino; E. Fattorusso; A. Mangoni; M. Di Rosa; A. Ianaro, *J. Am. Chem. Soc.* **1997**, *119*, 12465–12470.
23. K. C. Nicolaou; J. Li; G. Zanke, *Helv. Chim. Acta* **2000**, *83*, 1977–2006.
24. M. Seki; K. Mori, *Eur. J. Org. Chem.* **2001**, 3797–3809.
25. T. Tashiro; K. Akasaka; H. Ohrui; E. Fattorusso; K. Mori, *Eur. J. Org. Chem.* **2002**, 3659–3665.
26. H. Ohrui; H. Terashima; K. Imaizumi; K. Akasaka, *Proc. Jpn. Acad. Ser. B Phys. Biol. Sci.* **2002**, *78*, 69–72.
27. K. Imaizumi; H. Terashima; K. Akasaka; H. Ohrui, *Anal. Sci.* **2003**, *19*, 1243–1249.
28. H. Ohrui, *Proc. Jpn. Acad. Ser. B Phys. Biol. Sci.* **2007**, *83*, 127–135.
29. A. R. L. Cecil; Y. Hu; M. J. Vincent; R. Duncan; R. C. D. Brown, *J. Org. Chem.* **2004**, *69*, 3368–3374.
30. Y. Hu; A. R. L. Cecil; X. Frank; C. Gleye; B. Figadére; R. C. D. Brown, *Org. Biomol. Chem.* **2006**, *4*, 1217–1219.

31. D. P. Curran; Q. Zhang; H. Lu; V. Gudipati, *J. Am. Chem. Soc.* **2006**, *128*, 9943–9956.
32. J. G. Pomonis; L. Hammack; H. Hakk, *J. Chem. Ecol.* **1993**, *19*, 985–1007.
33. A. Furukawa; C. Shibata; K. Mori, *Biosci. Biotechnol. Biochem.* **2002**, *66*, 1164–1169.
34. K. Mori; T. Ohtaki; H. Ohrui; D. R. Berkebile; D. A. Carlson, *Eur. J. Org. Chem.* **2004**, 1089–1096.
35. K. Akasaka; D. A. Carlson; T. Ohtaki; H. Ohrui; K. Mori; D. R. Berkebile, to be submitted.
36. A. J. MacDermott, The Origin of Biomolecular Chirality. In *Chirality in Natural and Applied Science*; W. J. Lough,I. W. Wainer, Eds.; CRC Press: Boca Raton, FL, 2002; Chapter 2, pp 23–52.
37. K. Mori, *Acc. Chem. Res.* **2000**, *33*, 102–110.
38. K. Soai; I. Sato; T. Shibata, *Chem. Rec.* **2001**, *1*, 321–332.
39. R. H. Perry; C. Wu; M. Nefliu; R. G. Cooks, *Chem. Commun.* **2007**, 1071–1073.
40. M. Fujii; T. Saito, *Chem. Rec.* **2004**, *4*, 267–278.
41. R. Yamada; Y. Kera; S. Takahashi, *Chem. Rec.* **2006**, *6*, 259–266.
42. K. Mori, Overview. In *Comprehensive Natural Products Chemistry, Vol. 8: Miscellaneous Natural Products including Marine Natural Products, Pheromones, Plant Hormones, and Aspect of Ecology*; K. Mori, Ed.; Elsevier: Oxford, 1999; Chapter 1, pp 7–10.
43. K. Mori; H. Takikawa; M. Kido, *J. Chem. Soc. Perkin Trans. 1* **1993**, 169–179.
44. T. Rezanka; K. Sigler; V. M. Dembitsky, *Tetrahedron* **2006**, *62*, 5937–5943.
45. R. Tacke; T. Schmidt; M. Hofmann; T. Tolasch; W. Francke, *Organometallics* **2003**, *22*, 370–372.
46. G. M. Strunz; C.-M. Yu; L. Ya; P. S. White; E. A. Dixon, *Can. J. Chem.* **1990**, *68*, 782–786.
47. P. Bravo; E. Corradi; M. Frigerio; S. V. Meille; W. Panzeri; C. Pesenti; F. Viani, *Tetrahedron Lett.* **1999**, *40*, 6317–6320.
48. T. Itoh; K. Sakabe; K. Kudo; P. Zagatti; M. Renou, *Tetrahedron Lett.* **1998**, *39*, 4071–4074.
49. T. Itoh; K. Sudo; K. Yokota; N. Tanaka; S. Hayase; M. Renou, *Eur. J. Org. Chem.* **2004**, 406–412.
50. J. A. H. Inkster; I. Ling; N. S. Honson; L. Jacquet; R. Gries; E. Plettner, *Tetrahedron Asymmetry* **2005**, *16*, 3773–3784.
51. K. Mori; S. Kuwahara; M. Fujiwhara, *Proc. Indian Acad. Sci. (Chem. Sci.)* **1988**, *100*, 113–117.
52. S. Kurosawa; M. Takenaka; E. Dunkelblum; Z. Mendel; K. Mori, *ChemBioChem* **2000**, *1*, 56–66.
53. E. Dunkelblum; M. Harel; F. Assael; K. Mori; Z. Mendel, *J. Chem. Ecol.* **2000**, *26*, 1649–1657.
54. E. D. Morgan; I. D. Wilson, Insect Hormones and Insect Chemical Ecology. In *Comprehensive Natural Products Chemistry, Vol. 8: Miscellaneous Natural Products including Marine Natural Products, Pheromones, Plant Hormones, and Aspect of Ecology*; K. Mori, Ed.; Elsevier: Oxford, 1999; Chapter 5, pp 333–336.
55. K. Mori, Searching Environmentally Benign Methods for Pest Control: Reflection of a Synthetic Chemist. In *Pesticide Chemistry. Crop Protection, Public Health, Environmental Safety*; H. Ohkawa, H. Miyagawa, P. W. Lee, Eds.; Wiley-VCH Verlag: Weinheim, 2007; Chapter 2, pp 13–22.
56. T. Natori; M. Morita; K. Akimoto; Y. Koezuka, *Tetrahedron* **1994**, *50*, 2771–2776.
57. M. Morita; K. Motoki; K. Akimoto; T. Natori; T. Sakai; E. Sawa; K. Yamaji; Y. Koezuka; E. Kobayashi; H. Fukushima, *J. Med. Chem.* **1995**, *38*, 2176–2187.
58. G. Yang; J. Schmieg; M. Tsuji; R. W. Franck, *Angew. Chem. Int. Ed. Engl.* **2004**, *43*, 3818–3822.
59. T. Tashiro; R. Nakagawa; T. Hirokawa; S. Inoue; H. Watarai; M. Taniguchi; K. Mori, *Tetrahedron Lett.* **2007**, *48*, 3343–3347.
60. K. Murata; T. Toba; K. Nakanishi; B. Takahashi; T. Yamamura; S. Miyake; H. Annoura, *J. Org. Chem.* **2005**, *70*, 2398–2401.
61. M. Fujino; D. Wu; R. Garcia-Navarro; D. D. Ho; M. Tsuji; C.-H. Wong, *J. Am. Chem. Soc.* **2006**, *128*, 9022–9023.

Biographical Sketch

Kenji Mori was born in 1935. In all, he spent 42 years at the University of Tokyo. He holds B.Sc. (agricultural chemistry, 1957), M.Sc. (biochemistry, 1959), and Ph.D. (organic chemistry, 1962) degrees. He was appointed as assistant professor in the Department of Agricultural Chemistry at the University of Tokyo (1962), and was promoted to associate professor (1968) and professor (1978–95). Currently, he is Professor Emeritus. Dr. Mori worked for 7 years (1995–2001) as a professor at the Science University of Tokyo. At present, he is a research consultant at RIKEN (Institute of Physical and Chemical Research) and at

Toyo Gosei Co., Ltd. He was awarded the Japan Academy Prize (1981), the Silver Medal of the International Society of Chemical Ecology (1996), the American Chemical Society's Ernest Guenther Award in the Chemistry of Natural Products (1999), the Special Prize of the Society of Synthetic Organic Chemistry, Japan (2003), and the Frantisek Sorm Memorial Medal of the Academy of the Czech Republic (2003).

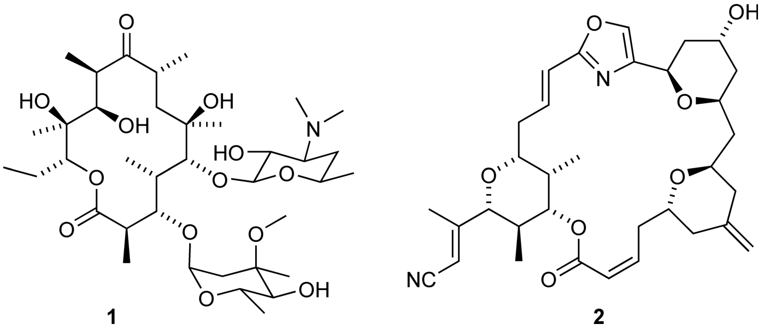
9.06 NMR – Small Molecules and Analysis of Complex Mixtures

Arthur S. Edison, University of Florida, Gainesville, FL, USA

Frank C. Schroeder, Cornell University, Ithaca, NY, USA

9.06.1 Introduction

Modern spectroscopic techniques have revolutionized compound identification and quantification. Only a few decades ago, identification of a structurally complex natural product would require multigram quantities of isolated material, which would then be subjected to series of derivatization and degradation experiments, aiming to deduce the unknown's structure from that of resulting derivatives or fragments that may represent known compounds. As a result of the tremendous advances in sensitivity and resolution of NMR spectroscopy over the past 30 years, identification of microgram quantities of new compounds has now become routine. For example, the structure of the polyketide antibiotic, erythromycin (**1**), was identified in 1957 only after extensive chemical and spectroscopic studies based on multigram amounts of isolated compound.[1–3] By the time its



1 2

Figure 1 Structures of the polyketide antibiotic erythromycin (**1**) and the marine polyketide hemi-phorboxazole A (**2**), which was recently identified based on a 16.5 μg (28 nmol) sample isolated from *Phorbas* sp.[7]

chemical structure was finally identified, erythromycin had already found extensive use in human medicine. Today, natural products of similar complexity, for example, hemi-phorboxazol (**2**), are routinely identified based on samples of 100 μg or less[4–7] (**Figure 1**). Of course, factors such as the structural complexity and novelty of the discovered compounds must be considered when making such comparisons. Whereas it may not be particularly challenging to design an analytical method that can reliably detect 10^{-9} mol of a known compound (e.g., a pesticide residue), determining the structure of an unknown natural product based on 10^{-9} mol of sample will likely present great difficulty.[8]

9.06.1.1 Sensitivity

A recent example illustrates how increases in sensitivity and the advent of multidimensional NMR spectroscopy have truly revolutionized organic structure determination. Identification of the first cardiotonic steroids from an invertebrate source in the late 1970s required the extraction of 28 000 *Photinus pyralis* fireflies. The crude extract was fractionated into five pure fractions, representing amounts from over 1 g down to 70 mg, which were then characterized by a combination of chemical and spectroscopic methods. Key structural information was afforded by one-dimensional ^{1}H- and ^{13}C-NMR spectroscopic analyses using a modest 250 MHz NMR spectrometer, resulting in identification of the bufadienolide (**3**).[9] Just over 25 years later, a similar analysis was carried out using a partially purified extract obtained from only 50 fireflies of the rare species, *Lucidota atra.* A 600 MHz spectrometer equipped with a microcoil probe was used,[10] allowing the characterization of 13 new bufadienolides (e.g., **4**) present in amounts ranging from 20 to 75 μg, corresponding to a decrease in sample requirement of roughly four orders of magnitude.[4] In another example, the disulfated steroid (**5**) was identified based on a sample of only 4 μg (6 nmol) and a 1.7 mm microprobe at 500 MHz. The steroid (**5**) functions as a 'sperm attracting and activating factor' (SAAF) in chemical signaling systems of the ascidian (sea squirt) *Ciona intestinalis* (**Figure 2**).[11]

Improvements in NMR spectroscopic sensitivity also benefit studies aimed at elucidation of the biological context of natural products. For example, NMR spectroscopic analysis of insect metabolites traditionally necessitated the pooling of material collected from multiple individuals, effectively eliminating NMR spectroscopy as a technique that could be used for the detailed analysis of metabolite dynamics in ecological studies. However, sensitivity increases derived from microsample NMR technology enabled Dossey *et al.*[12] to analyze metabolite mixtures within individual walking sticks, *Anisomorpha buprestoides*, permitting complete characterization of the iridioid anisomorphal (**6**) from a single insect specimen. Using the *A. buprestoides* secretion as a model system, subsequent work by Zhang *et al.*[13] demonstrated the application of covariance-based mixture analysis to automatically identify individual components in an unpurified sample. The ability to analyze individual specimens by both NMR spectroscopy and MS holds considerable promise for future biological studies.

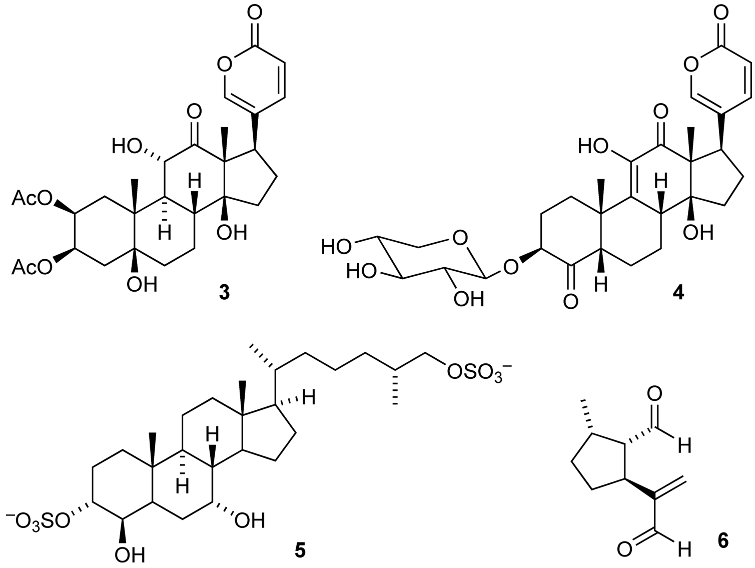


Figure 2 Bufadienolides (3 and 4) from fireflies, SAAF (5) from the ascidian *Ciona intestinalis*, and anisomorphal (6) from walking sticks.

9.06.1.2 Mixtures

Traditionally, detailed NMR spectroscopic characterization of natural product samples was not initiated until largely pure samples of individual compounds had been obtained, usually through extensive chromatographic fractionation. However, the potential advantages of structure identification of individual components in mixtures have been widely recognized since at least the mid-1990s. Several techniques based on diffusion-ordered spectroscopy were developed to aid in this process including DOSY (diffusion-ordered spectroscopy)[14] and DECODES (diffusion-encoded spectroscopy).[15] Unfortunately, these methods often fail to resolve multiply overlapping signals and suffer from low dynamic range, which reduces their utility for structure determination in complex mixtures of organic small molecules. As a result, DOSY and related methods were primarily used to analyze mixtures of synthetic products[16–19] and never found widespread use in natural products research.

Using 2D NMR spectroscopy for the analysis of complex natural products mixtures recently regained momentum, as several studies demonstrated that using simple dqfCOSY (double-quantum-filtered correlation spectroscopy), TOCSY, HSQC, or HMBC spectra for complex mixtures offers exciting new perspectives for natural products research and chemical biology. Compared to mass spectrometric (MS) analyses of small molecule mixtures, such 2D NMR spectroscopic investigations offer the benefit of more detailed structural information, which is of particular relevance for the detection of unanticipated chemotypes. Recent examples include the identification of sulfated nucleosides, such as 7, from spider venom,[20,21] the detection of ascarosides (e.g., 8) as part of the mating signal in the nematode *Caenorhabditis elegans*,[22,23] and the identification of the highly unstable polyketide bacillaene (9) from *Bacillus subtilis*.[24] These studies show that using state-of-the-art NMR spectroscopy even minor components of complex small molecule mixtures can be characterized. Such NMR spectroscopic analyses of complex mixtures may not always permit complete structural assignments; however, additional results from mass spectroscopic analyses frequently allow proposing complete structures. As a result, the need for chromatographic separations is greatly reduced, which not only accelerates compound discovery, but also offers distinct advantages for the discovery of chemically unstable metabolites. It seems likely that the pervasive use of chromatography in natural products chemistry has skewed our knowledge of secondary metabolism, because sensitive compounds often do not survive extended exposure to solvents or chromatography media. In fact, the original motivation to explore the utility of high-resolution 2D NMR spectroscopy for the characterization of small molecule mixtures arose because alkaloids present in the poison gland secretion of *Myrmicaria* ants were found to be highly unstable for chromatographic isolation.[25,26] Myrmicarin

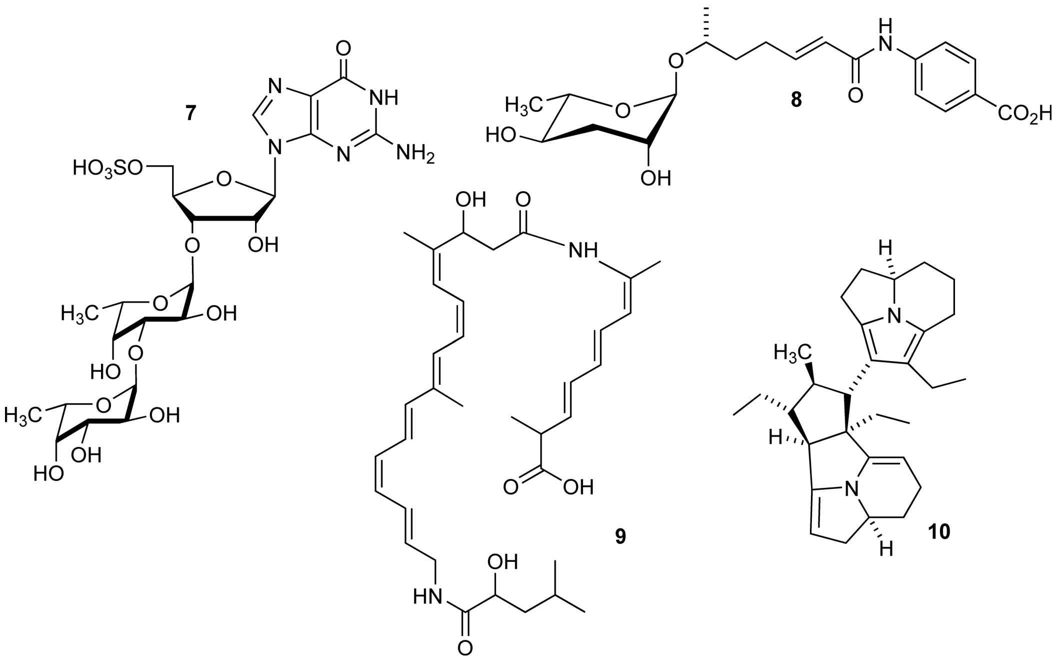


Figure 3 Natural products identified from complex metabolite mixtures.

430A (**10**), the most unstable of the *Myrmicaria* alkaloids identified so far, thus represents one of the first members of a growing class of natural products that have never been isolated in pure form (**Figure 3**).

Advanced processing of spectroscopic data, taking advantage of statistical tools originally developed for metabolomics studies such as STOCSY,[27] SHY,[28] and others[29–32] could further enhance the utility of NMR spectroscopy and MS for analyzing natural products mixtures and correlating chemical information with biological data. However, to date there have been few reports on the application of metabolomics techniques to natural products research.[33]

In this chapter, we start with a brief overview of the standard methods currently used for the NMR-spectroscopic identification of natural products and other types of organic small molecules, which is followed by a section dedicated to NMR spectroscopic characterization of small molecule mixtures and a discussion of approaches to increase NMR spectroscopic sensitivity.

9.06.2 Routine NMR Spectroscopy for Natural Products Structure Elucidation

Strategies for NMR spectroscopic structure elucidation of organic compounds have been reviewed extensively.[34–37] In this section, we briefly describe a set of the most commonly useful 2D NMR spectra that is sufficient for most (though certainly not all) organic structure determination problems, and we comment on specific modifications of acquisition parameters that facilitate the analysis of mixtures.

Any NMR spectroscopic analysis of an organic sample will normally begin with the examination of a simple ^{1}H-NMR spectrum, which serves to assess purity, concentration of minor components (if any), and overall complexity of the structures in the sample. Furthermore, the ^{1}H spectrum provides an opportunity to examine line shape characteristics of the sample's components, and, if necessary, reevaluate solvent choice, sample concentration, or acquisition temperature. If large quantities of pure compound are available, 1D ^{13}C-NMR spectra may also be useful. However, in most cases acquisition of a pair of (^{1}H,^{13}C)-HSQC and (^{1}H,^{13}C)-HMBC spectra will be a better use of spectrometer time, unless structural features are suspected that preclude full characterization by HSQC and HMBC. For example, compounds that feature quaternary carbon atoms that cannot be detected by HMBC will often require acquisition of a 1D ^{13}C spectrum and, if possible, a 2D ^{13}C-INADEQUATE (incredible natural abundance double quantum transfer experiment) spectrum.

Experienced natural products chemists may be able to recognize certain compound classes or characteristic structural features at this early stage of the analysis, and for known compounds such tentative structural assignments can often be confirmed through comparison with literature NMR data and additional mass spectrometric analyses. For unknown compounds, the next step in the structure elucidation process usually consists of acquisition of variants of four different types of 2D NMR spectra:

9.06.2.1 COSY and TOCSY

(1) (^{1}H,^{1}H)-COSY or TOCSY is used for characterization of the proton spin systems. A 'spin system' is represented by any group of protons that interact through scalar couplings, for example ethyl butanoate features two spin systems, one consisting of the five protons of the ethoxy group, and one consisting of the seven protons of the butanoyl group. There are many different COSY and TOCSY variants, differing in time requirement and type of (^{1}H,^{1}H)-coupling information provided. COSY spectra show crosspeaks only for directly coupled protons, whereas TOCSY spectra may show crosspeaks not only for protons directly coupled with each other, but also with other protons in the same spin system. For example, in COSY spectra, the alpha proton in the amino acid leucine will show crosspeaks to the adjacent beta-methylene protons, whereas in a TOCSY spectrum, depending on the experimental parameters, the alpha proton may have additional crosspeaks with protons of the gamma methine and the two methyl groups. TOCSY spectra are useful in situations where part of a spin system is obscured in the corresponding COSY spectrum, for example, due to an extensive overlap in the aliphatic region. TOCSY crosspeaks at the chemical shift of a nonobscured proton can often be used to reveal the obscured parts of the spin system. For example, TOCSY spectra are used extensively in the NMR-spectroscopic characterization of proteins where it is used to map complete amino acid spin systems onto the corresponding amide protons. TOCSY can also be useful for the analysis of natural products, especially for complex mixtures where overlap is often a problem. As described below, Bruschweiler's group has developed mixture analysis methods that are based on TOCSY spectra.[13,38,39] Acquisition parameters for TOCSY spectra can be roughly tuned to emphasize either COSY-type interactions with short mixing times or more complete spin system correlations with longer mixing times. Several mixing sequences are available on modern spectrometers, and some of the more popular are DIPSI-2[40] and MLEV-17.[41] Bax[42] has provided an excellent overview of TOCSY (also known as homonuclear Hartmann–Hahn or HOHAHA) that summarizes the basic principles and demonstrates applications.

For identifying new natural products, TOCSY spectra are often less straightforward to analyze than COSY spectra, because a TOCSY crosspeak does not necessarily indicate that two protons are coupled with each other – the presence of a crosspeak only shows that two protons are part of the same spin system. Furthermore, TOCSY spectra may be significantly more crowded than COSY spectra, and TOCSY crosspeak intensity is often difficult to correlate with structural properties. Finally, the fine structure of TOCSY crosspeaks is much less amenable to detailed analysis than dqfCOSY crosspeaks; as described below, the antiphase dqfCOSY crosspeaks contain information on proton multiplicities and scalar coupling constants, which TOCSY crosspeaks cannot provide.

Among variants of COSY, simple gradient COSY (gCOSY) spectra and phase-sensitive double-quantum-filtered COSY (dqfCOSY) spectra are the most useful for natural product analysis. gCOSY spectra can be acquired extremely fast – using a small molecule sample of >1 mg a decent spectrum can usually be acquired within 5–10 min. However, gCOSY spectra provide only very limited information. The fine structure of gCOSY crosspeaks is often poorly defined, which poses problems for differentiating signals of overlapping peaks and does not allow distinguishing crosspeaks that are due to large coupling constants from crosspeaks that are due to smaller couplings. This is of particular relevance for the analysis of complex spin systems where distinguishing between long-range couplings and stronger geminal or vicinal couplings is important, and where coupling constants may carry important information about relative configuration. In addition, artifacts are sometimes difficult to distinguish from 'real' crosspeaks in gCOSY spectra.

For these reasons, dqfCOSY spectra are usually a better choice for any compound or mixture sample that includes complex proton spin systems. If acquired using sufficiently long acquisition times (600 ms or more),

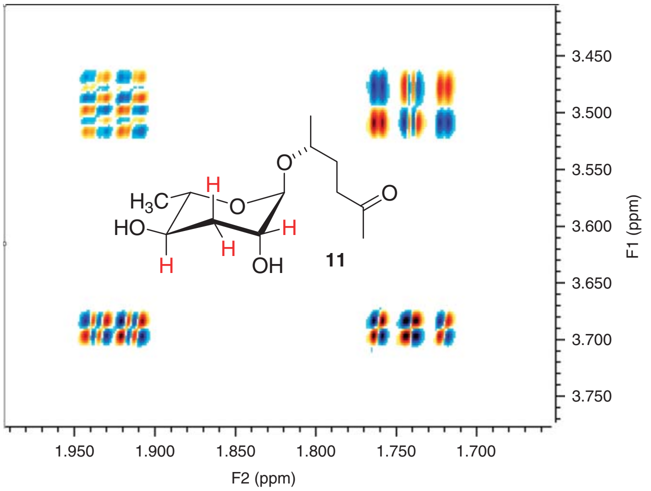
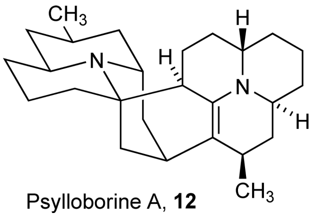


Figure 4 Part of the dqfCOSY spectrum of the ascaroside (**11**), a component of the *Caenorhabditis elegans* dauer pheromone. The fine structure of the four shown crosspeaks permits accurate determination and assignments of the geminal and all vicinal coupling constants of the two methylene protons (red).[43]

dqfCOSY crosspeaks closely reflect the splitting patterns of corresponding multipletts in one-dimensional ^{1}H spectra. Based on their splitting patterns, dqfCOSY crosspeaks belonging to a specific proton can be easily recognized and grouped together, and as a result, overlapping signals can be clearly distinguished. Furthermore, the characteristic antisymmetric fine structure of each crosspeak not only allows for fairly accurate determination of coupling constant values, but also permits determining the coupling partner responsible for the coupling constant (**Figure 4**). Therefore, crosspeaks due to small coupling constants can be easily distinguished from crosspeaks due to large coupling constants. Another advantage resulting from the highly characteristic appearance of dqfCOSY crosspeaks is that artifacts can be recognized very easily. dqfCOSY spectra should always be acquired using pulse sequences that employ phase-cycling for coherence selection. Although gradient-selected versions of dqfCOSY are available, line shapes in these gradient versions are usually extremely poor.

DqfCOSY spectra usually provide sufficiently accurate values for coupling constants that are larger than twice the line width of the corresponding proton signals, for example, coupling constants larger than 2–4 Hz. However, for signals of protons that have several similar though not identical coupling constants, the interpretation of the dqfCOSY crosspeaks may present considerable difficulty. For analysis of such highly complex spin systems, or in situations where precise knowledge of small coupling constants is required, E.COSY ('exclusive' COSY) spectra are better suited.[44,45] E.COSY crosspeaks are less complex than dqfCOSY crosspeaks and can be used to obtain highly accurate values even for very small long-range coupling constants. For example, E.COSY was used to determine coupling constants in the ladybird beetle alkaloid psylloborine A (**12**) (**Figure 5**), which features mostly an aliphatic heptacyclic ring system. As evident from the dqfCOSY spectrum shown in **Figure 6**, this compound's spin systems are extremely complex and some dqfCOSY crosspeaks are



Psylloborine A, **12**

Figure 5 Structure of the dimeric polyacetate alkaloid psylloborine from the ladybird beetle *Psyllobora vigintiduopunctata* (**12**).[46]

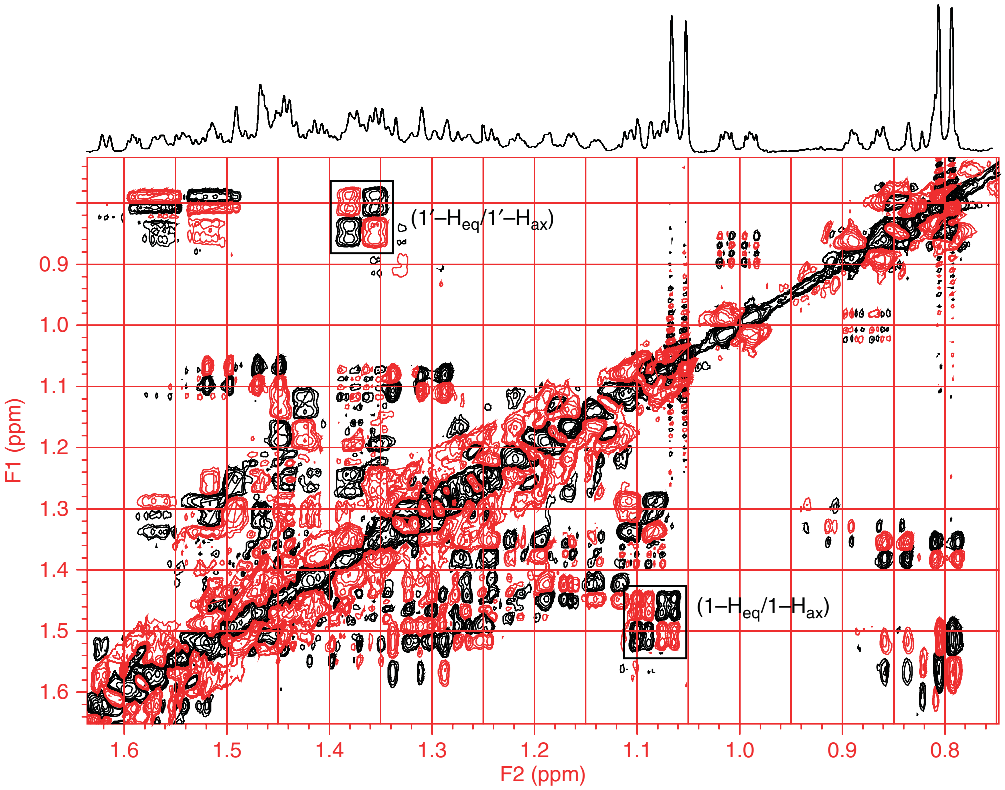


Figure 6 The 75–1.64 ppm region of the ^{1}H-NMR and the dqfCOSY spectrum of psylloborine A (**12**) (C_6D_6, 500 MHz). The vicinal coupling constants of the proton $1' - H_{eq}$ (1.37 ppm) cannot be directly extracted, due to poor resolution in F2 and overlap with other crosspeaks, for example, of the proton $8 - H_{eq}$ at 1.36 ppm. The E.COSY signals corresponding to the crosspeaks ($1 - H_{eq}/1 - H_{ax}$) and ($1' - H_{eq}/1' - H_{ax}$) are shown in **Figure 7**.

difficult to interpret. However, the interpretation of corresponding E.COSY spectrum was straightforward (**Figure** 7).[46] It should be noted, however, signal to noise (S/N) of E.COSY is considerably lower than that of dqfCOSY, and that E.COSY requires careful calibration of pulse width in order to minimize artifacts.

9.06.2.2 HSQC and HMQC

(^{1}H,^{13}C)-HSQC or HMQC serves to identify proton-bearing carbons and to associate these carbons with their attached protons. HSQC spectra feature generally better line shapes than HMQC and the commonly used multiplicity-edited HSQC versions offer the added benefit of distinguishing CH_3, CH_2, and CH groups. However, the HMQC pulse sequence is significantly shorter than the HSQC sequence, and therefore magnitude-mode HMQC spectra may have better S/N than magnitude-mode HSQC spectra. Use of adiabatic ^{13}C- pulses in the HSQC sequence can significantly reduce this sensitivity disadvantage. For the analysis of complex mixtures with many overlapping proton and/or carbon signals, HSQC is much better suited than HMQC because of better line shapes. Both HMQC and HSQC are usually acquired with ^{1}H-decoupling during acquisition and as a result feature one crosspeak per ^{1}H-resonance that is located at a chemical shift value close to that of the ^{12}C-attached protons. It should be noted, however, that the proton chemical shifts of HSQC or HMQC crosspeaks are not exactly identical to that of the main signals in the ^{1}H [or (^{1}H,^{1}H)-COSY and (^{1}H,^{13}C)-HMBC] spectra. The latter represent ^{12}C-bound protons (unless the sample is isotopically labeled), whereas the signals in HSQC/HMQC obviously represent ^{13}C-bound protons, whose chemical shift values may differ slightly, and to a varying extent. Such small differences in chemical shift can make precise calibration of HMQC/HSQC spectra difficult and may present a problem for samples that feature

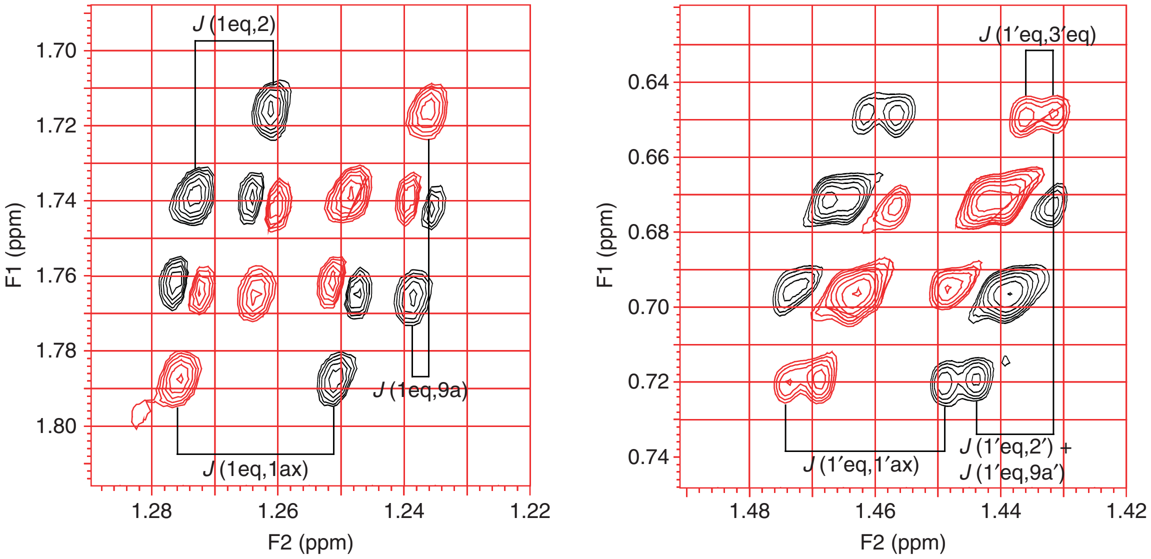


Figure 7 Left: Crosspeak of the geminal pair $1-H_{eq}/1-H_{ax}$ in the E.COSY spectrum of psylloborine A (**12**) (CD_2Cl_2, 500 MHz). The passive vicinal couplings J(1eq,2) and J(1eq,9a) and the active coupling $J_{1eq,1ax}$ can be determined without interference from components of opposite phase. Right: E.COSY crosspeak of the geminal pair $1'-H_{eq}/1'-H_{ax}$. Although, in comparison to the corresponding dqfCOSY spectrum (**Figure 2**), the crosspeak is greatly simplified, the passive vicinal coupling constants J(1′eq,9a′) and J(1′eq,2′) cannot be directly extracted. Owing to partial overlap, only the sum J(1′eq,9a′) + J(1′eq,2′) can be determined. Since J(1′eq,9a′) is accessible from the E.COSY crosspeak of $9a'-H/1'-H_{ax}$ (not shown), J(1′eq,2′) can be calculated. The E.COSY crosspeak $1'-H_{eq}/1'-H_{ax}$ also allows one to determine the active geminal coupling J(1′eq,1′ax). In addition, a small four-bond coupling J(1′eq,3′eq) is revealed.[46]

many overlapping proton signals of very similar chemical shift, for example, complex mixtures. Acquisition of coupled ('nondecoupled') HSQC spectra is sometimes advantageous, for example, in cases where one-bond ^{1}H–^{13}C coupling constants are of interest, or in cases where decoupling would result in low-quality spectra due to sample heating (this can be an issue when using long acquisition times, or when using polar solvents, especially in the presence of salts).

9.06.2.3 HMBC

(^{1}H,^{13}C)-HMBC provides correlations between protons and carbons that are two or three bonds apart from each other (though occasionally four-bond or even five-bond correlations may be observed). HMBC spectra are important for the detection of quaternary carbons and serve to link separate structural fragments obtained from analysis of COSY/TOCSY and HSQC/HMQC. However, interpretation of HMBC spectra often represents the most challenging step in the structure elucidation process, for several reasons. The intensity of crosspeaks in HMBC spectra is notoriously difficult to predict. Some two- or three-bond correlations may be extremely weak or not appear at all, and therefore the absence of an HMBC crosspeak cannot, *a priori*, be taken as evidence against a specific structural connection. Furthermore, there is no simple method for distinguishing two- and three-bond correlations. Additional difficulties may arise in situations where weak HMBC crosspeaks could represent either a three-bond or a four-bond (or rarely five-bond) correlation. Finally, because standard HMBC experiments are usually optimized for (^{1}H,^{13}C)-long range coupling constants of intermediate size, both very strong and very weak (^{1}H,^{13}C)-long-range couplings may give rise to weak crosspeaks in routine HMBC spectra, which can be source of considerable confusion. The latter concern can be addressed by acquiring two separate HMBC spectra using two different mixing delays, for example, 50 and 100 ms. It should be noted that when using long mixing delays, the acquisition time should be increased to at least twice the mixing delay, for example, for a delay of 100 ms, the acquisition time should be at least 200 ms, which is considerably above the default values in common HMBC parameter sets. Even when using standard parameters, HMBC signal intensity is often somewhat lower than that of HSQC or HMQC spectra, primarily because of the relatively long mixing delay (40–120 ms) in the HMBC pulse sequence. Often ^{1}H line shape is a good predictor of S/N in HMBC spectra: generally, narrow ^{1}H line widths correlate with good S/N of corresponding signals in the HMBC spectra.

For most organic small molecules, acquisition of one-dimensional ^{13}C spectra is not required when well-resolved HSQC and HMBC spectra are available. Exceptions include compounds that have quaternary carbons that simply do not show any HMBC correlations, for example, because there are no protons within two or three bonds of some carbons (see also Section 9.06.2.5). Another limitation of routine HMBC is that spectral resolution in the ^{13}C-chemical shift dimension is limited. As is the case for HMQC (but not HSQC), HMBC crosspeaks are broadened in the ^{13}C-chemical shift dimension by the (^{1}H,^{1}H)-coupling constants of the proton whose long-range (^{1}H,^{13}C)-coupling is observed. As a result, crosspeaks belonging to carbons with very similar chemical shifts can sometimes not be unambiguously assigned. However, the interfering (^{1}H,^{1}H)-couplings can be removed by using a constant-time variant of the standard HMBC experiment. Using band-selective, constant-time HMBC variants, spectra with extremely high resolution in the ^{13}C-dimension can be easily obtained.[47]

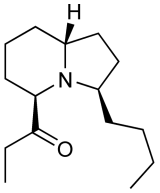
9.06.2.4 NOESY and ROESY

(^{1}H,^{1}H)-NOESY and ROESY provide information about spatial proximity of protons that are separated by up to about 5 Å, which can be used to determine relative configuration and, in some cases, conformation of organic small molecules. Other applications include the study of chemical exchange or investigations of the interaction of natural products with their protein targets. NOESY and ROESY spectra are similar, but the choice of which to use depends on the rate of tumbling of the molecule, which is roughly proportional to the molecular weight, but also depends on its polarity and on polarity and viscosity of the solvent. NOESY crosspeaks are opposite in sign for small and large molecules. For small molecules, the sign of NOESY crosspeaks is opposite to the sign of the diagonal, whereas for large molecules NOESY cross- and diagonal peaks have the same sign. Correspondingly, there is a range of molecules for which NOESY crosspeaks are close to zero,[48] and thus NOESY is not suitable. NOESY can generally be used for organic small molecules of molecular weights below 800 Da, unless the compound under investigation is very polar, or requires the use of a highly polar and/or viscous solvent, such as DMSO.

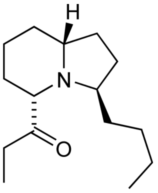
In ROESY spectra, the sign of the crosspeaks are always opposite to that of the diagonal peaks, and therefore ROESY is suitable for intermediate-sized molecules around 1 kDa and smaller molecules in viscous solvents or at low temperatures. Both NOESY and ROESY can also be used to investigate chemical exchange, and they show crosspeaks for nuclei that are in slow (relative to the chemical shift differences) exchange. Importantly, the sign of chemical exchange crosspeaks is the same as that of the diagonal, and so for small molecules both NOESY and ROESY can be used to distinguish chemical exchange crosspeaks from crosspeaks due to spatial proximity. In larger molecules, but sometimes also in small molecules, additional crosspeaks can occur through 'spin diffusion', a relay of magnetization along a chain of ^{1}H's that are close together that leads to TOCSY-type crosspeaks. Mixing times for acquiring natural product NOESY or ROESY spectra should be set to 500–800 ms (NOESY) and 200–400 ms (ROESY).

9.06.2.5 Other Techniques and Current Limitations

Not all types of natural products can be sufficiently characterized using routine 2D NMR spectroscopy as described in the preceding section. For example, in highly unsaturated compounds some carbons may not be detected by HMBC simply because there are no protons within three bonds of these carbons. There are few ways to address this problem with NMR-spectroscopic means, and in some cases, chemical modification (e.g., hydrogenation) or degradation may be necessary in order to complete structural assignments. In a rare event that large amounts of the compound in question are available, (^{13}C,^{13}C)-correlations that provide direct evidence for carbon–carbon bonds such as (^{13}C,^{13}C)-INADEQUATE (or its ^{1}H-detected cousin, ADEQUATE) can be useful.[49] However, due to the low natural abundance of ^{13}C, sensitivity of INADEQUATE is extremely low, as only pairs of adjacent ^{13}C atoms contribute to the signal. One of the



13



14

Figure 8 Myrmicarin 237A (13) and 237B (14), which were identified using (^{13}C,^{13}C)-2D-INADEQUATE.[50]

very few examples for the use of INADEQUATE for natural product samples is presented by the identification of the indolizidine alkaloids myrmicarin 237A and 237B (13 and 14) (Figure 8).[50] These compounds equilibrate through keto–enol tautomerism and therefore had to be characterized as a 1:1 mixture of diastereomers, which resulted in extremely crowded COSY and HMBC spectra. A (^{13}C,^{13}C)-COSY-type INADEQUATE was then used to unambiguously distinguish between the ^{13}C-resonances of the two diastereomers (Figure 9).

A different problem is posed by structures that include large numbers of NMR-inactive heteroatoms. In such cases, it may be impossible to assemble based on NMR spectroscopic data simply because there are too many possibilities for arranging the heteroatoms around the identified carbon- and hydrogen-based partial structures. For compounds that include nitrogen or phosphorus, ^{15}N- and ^{31}P-NMR spectroscopy can often supply important additional information.

As natural products chemists detect and investigate more and more complex structures, additional limitations of current NMR spectroscopic approaches have become apparent. Examples for compound classes that pose great difficulty for NMR spectroscopists include compounds with ill-defined conformations, natural products that occur as large families of structurally similar compounds, or oligomeric compounds whose

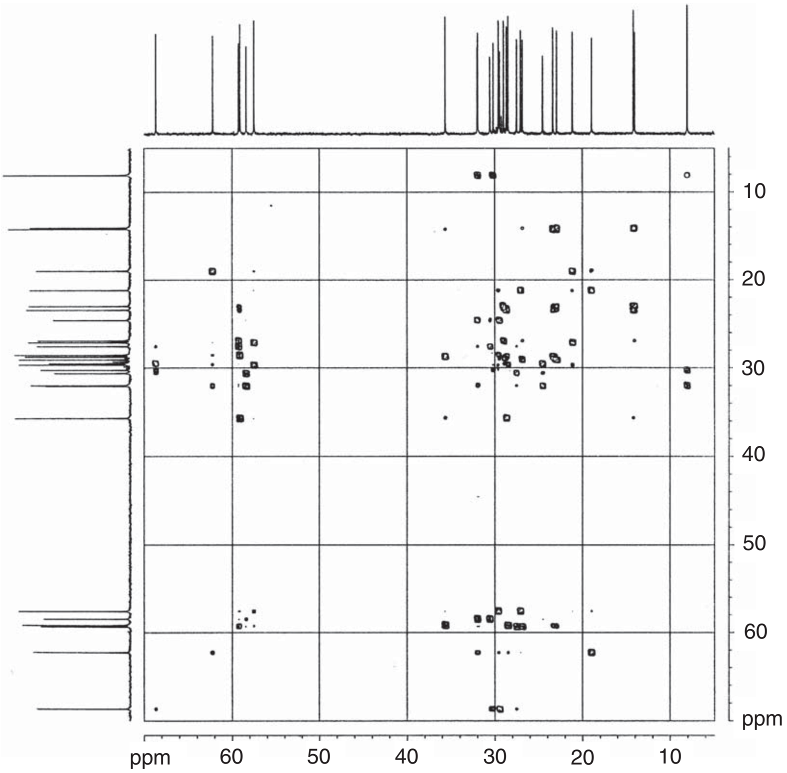


Figure 9 COSY-like (^{13}C,^{13}C)-2D-INADEQUATE of a 220 mg sample of a 1:1 mixture of myrmicarin 237A (13) and 237B (14), acquired over 54 h.[50]

assembly follows an irregular scheme, such as complex glycosides or lipidated natural product derivatives. For the latter groups of compounds, some partial degradation and/or derivatization may still be required in order to enable NMR spectroscopic analysis.

9.06.3 Complex Mixtures

NMR spectroscopy evolved primarily as a tool for the characterization of pure compounds or simple well-defined mixtures (see Section 9.06.1.2), whereas strategies for the NMR spectroscopic identification of compounds from complex mixtures have been developed only recently. As several examples have shown, using NMR spectroscopy for the analysis of complex mixtures can open up new perspectives and may enable new lines of inquiry in both natural products chemistry and metabolomics. Before discussing some of these examples in greater detail, it is useful to consider what developments spawned the recent surge in applications of NMR spectroscopy to mixtures and what prevented earlier uses of NMR spectroscopy for this purpose.

Following the initial observation in the early 1950s that the resonance frequency of a nucleus is influenced by its chemical environment,[51] and that the fine structure of a resonance could be influenced by other nuclei through intervening chemical bonds,[51,52] chemists quickly seized upon the enormous potential of NMR spectroscopy for structure determination. As a result, NMR spectroscopy became one of the most important spectroscopic tools of organic chemists. Combining NMR spectroscopic with mass spectrometric analyses proved particularly useful, with MS providing information about molecular weight and atomic composition, and NMR spectroscopy contributing information about chemical environment and, importantly, connectivity and spatial configuration. Until about 1980, NMR spectroscopic structure elucidation was largely based on one-dimensional spectra, providing information about chemical shift (suggesting a specific chemical environment), relative signal intensity (indicating the number of a specific type of nuclei in the molecule), and multiplicity (suggesting connectivity between individual groups of nuclei in the molecule).[34,53] Extracting signal intensity and multiplicity information from one-dimensional NMR spectra depended crucially on sample purity, because signals of impurities could easily skew signal intensities or obstruct important splitting patterns. Furthermore, based on one-dimensional spectra it is often impossible to determine whether two signals represent nuclei that are part of the same molecule or whether they represent two (or more) separate structures. As a result, NMR spectroscopy was deemed largely unsuitable for the analysis of complex mixtures such as crude natural products extracts, and NMR spectroscopic analysis was usually initiated only after pure or almost pure samples of the compound(s) of interest had been obtained, generally as the endproduct of extensive chromatographic fractionation.

The eventual realization that NMR spectroscopy can nonetheless be applied most advantageously to the characterization of mixtures then depended on at least two separate developments. First, the advent of 2D-(and subsequently, multidimensional) NMR spectroscopy enabled much better access to connectivity information than could be obtained from the analysis of multiplets in one-dimensional spectra. 2D spectra such as COSY, HSQC, or HMBC yield correlations that correspond to connectivity through one or more chemical bonds.[34,54] Importantly, dispersion of signals along a second (or third) chemical shift dimension almost always removes any ambiguity resulting from overlap of signals in one dimension, and therefore enables recognition and identification of partial structures that may belong to several different compounds.[55]

However, even the advent of multidimensional spectroscopy alone did not yet suffice to make NMR spectroscopic analysis of complex mixtures broadly applicable. In the early days of 2D-spectroscopy, the capabilities of spectrometers and processing hardware limited resolution and dynamic range of the spectra severely. For example, processing of a very low-resolution COSY spectrum on a Bruker AC250P (250 MHz proton) spectrometer console in 1990 could take as much as 30 min. Moreover, early 2D spectra, especially the most useful inversely detected HSQC and HMBC, were prone to artifacts, making it nearly impossible to unambiguously discern signals representing minor components.

The advent of improved data acquisition systems and greatly increased computing power fundamentally changed the scope of multidimensional NMR spectroscopy, and today very low-artifact COSY, TOCSY, or HSQC spectra can be obtained whose resolution approximates that of one-dimensional spectra.[35,36] Along with increases in sensitivity and resolution derived from higher magnetic field strength and improved probe design (see Section 9.06.4), these developments have set the stage for a broad exploration of the utility of NMR spectroscopy for characterizing complex mixtures. Sections 9.06.3.1 and 9.06.3.2 describe recent examples for using 2D NMR spectroscopy for the characterization of new natural products from complex biological extracts, whereas Section 9.06.3.3 describes a method for computational deconvolution of 2D-spectra of complex mixtures into one-dimensional subspectra that represent partial structures of individual components. Computational approaches have also been applied to ensembles of one-dimensional spectra of complex small molecule mixtures for the purpose of biomarker identification. Corresponding applications in metabolomics are discussed in Section 9.06.3.4.

9.06.3.1 NMR Spectroscopic Analysis of Complex Natural Products Mixtures

The idea to use 2D-NMR spectroscopy for a systematic characterization of crude or unfractionated natural products mixtures was first conceived in connection with research on the chemical ecology of arthropods.[20,21,25,26,46,56–58] During studies of the chemical composition of various arthropod secretions, several cases were encountered for which conventional analytical methodology based on fractionation of the secretions aiming at the isolation of individual components failed to identify the biologically active principles. It was concluded that the chromatography-based fractionation of these secretions had resulted in destruction or loss of the active components. As a consequence, the use of NMR spectroscopy for the characterization of native, entirely unfractionated materials was considered. As one of the first examples, the unfractionated defensive secretion of a Ladybird beetle pupa, *Epilachna borealis*, was subjected to 2D NMR spectroscopic analysis including dqfCOSY, NOESY, HSQC, and HMBC spectra, for which acquisition parameters were somewhat modified in order to obtain higher-resolution spectra.[58] This approach quickly resulted in the identification of a previously overlooked group of compounds that made up more than 50% of the secretion, a new family of macrocyclic lactone alkaloids, the polyazamacrolides, such as **15 and 16** (**Figure 10**).

These insect secretions presented a perfect starting point for exploring the utility of direct NMR spectroscopic analysis of crude mixtures, because the secretions consisted of mixtures of only one to three structurally

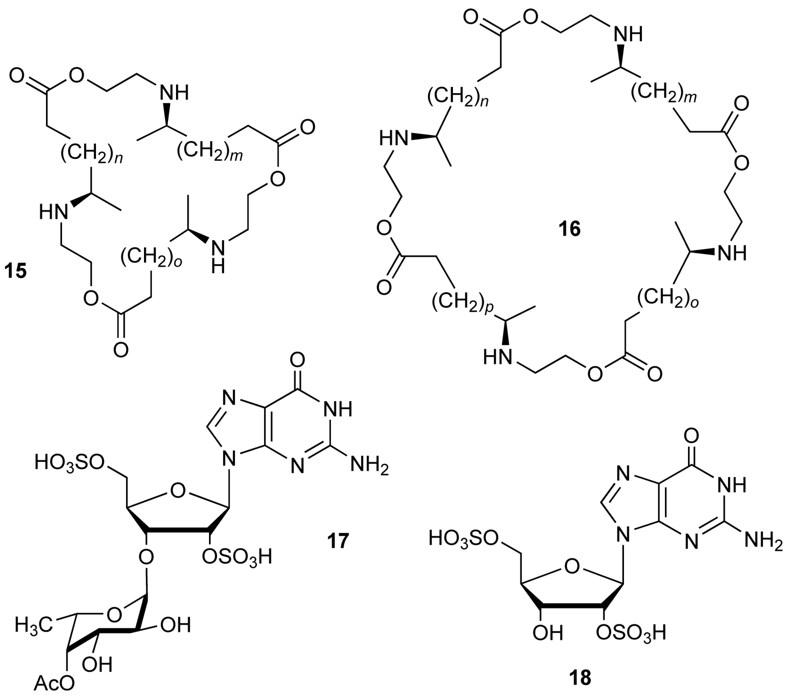


Figure 10 Polyazamacrolides from pupae of the ladybird beetle *Epilachna borealis*[58] and sulfated nucleosides (**17**) and (**18**) identified from spider venom.[20,21,59]

distinct groups of small molecules. However, as recent analyses of crude spider venom have shown that even much more complicated mixtures of small molecules are amenable to NMR spectroscopic analysis.[20,21] NMR spectroscopic studies of crude spider venom were motivated by the earlier identification of a bis-sulfated nucleoside, HF-7 (**17**), a selective and potent kainate receptor antagonist, from the venom of the grass spider, *Hololena curta*.[59] The discovery of this entirely unexpected natural product suggested that spider venoms might harbor interesting new classes of neurotoxins. Moreover, it seemed unlikely that HF-7 is the only spider venom component of its kind. The question remained why sulfated nucleosides had previously escaped detection, even though spider venoms had been subject to intensive chemical scrutiny, which had led to identification of hundreds of proteins, peptides, acylated polyamines, and various small molecule neurotransmitters. Given this very high degree of complexity, it is not surprising that most previous studies of spider venom chemistry applied some form of chromatographic fractionation as a first step. Because sulfated nucleosides are somewhat susceptible to hydrolysis, it was suspected that sulfated nucleosides may have been overlooked in some earlier analyses as a result of decomposition during chromatographic fractionation. Building on experience gathered from characterization of the polyazamacrolides, it was thus attempted to characterize entire, unfractionated spider venom samples using 2D NMR spectra, including dqfCOSY, HMQC, and HMBC. This approach led to the identification of sulfated nucleosides such as **18** as important components in the venoms of several spider species, including previously well-studied species such as the hobo spider, *Tegenaria agrestis*, and the brown recluse spider, *Loxosceles recluse*.[20,21] Effectively, these 2D NMR spectroscopic analyses provided a largely undistorted and impartial view of spider venom composition, without any skewing of the results stemming from chromatographic separation.

It is important to note that such 'direct' NMR spectroscopic analyses of complex natural product mixtures may not always permit assigning complete structures. In many cases, a full or near-complete characterization will only be possible for a few major components, whereas more or less extensive partial structures will be obtained for minor components. However, any partial structures elucidated will provide important information that may be (1) used to search natural product databases for similar compounds, (2) combined with results from GC–MS or LC–MS analysis to develop better hypotheses about their structures, (3) used to develop a fractionation scheme tailored to the isolation of specific compounds of interest, and (4) used to design syntheses for the proposed structures.

Direct NMR spectroscopic analyses are particularly well-suited to examine natural product extracts for the presence of novel or unanticipated compounds. However, 2D spectra of natural product mixtures are often extremely complex, which limits the feasibility of using 2D NMR as a first-line tool for the characterization of large numbers of natural product extracts. In some cases, this concern can be addressed by focusing only on specific features in the spectra, for example, groups of crosspeaks that correlate with a certain biological activity or genotype. A method that facilitates recognition of 2D NMR signals relevant within a specific biological context, DANS ('differential analysis through 2D NMR spectroscopy'), is discussed in Section 9.06.3.2. Even if detailed spectral interpretation is not pursued, 2D spectra obtained for a crude natural products sample can be useful as a largely unbiased record of its original composition against which the results from any subsequent fractionation can be compared. Such comparisons can aid in recognizing artifacts or detecting loss of some components of the original mixture.

9.06.3.1.1 Comparing NMR spectroscopic characterization of proteins and small molecule mixtures

In some regard, NMR spectra obtained from complex mixtures of small molecules resemble those of biological macromolecules such as large peptides, proteins, or oligonucleotides.[55] NMR spectra of both biological macromolecules and small molecule mixtures are similar in that they feature a very large number of overlapping signals, which through the use of a variety of two- or three-dimensional experiments can be assigned to individual substructures. In case of macromolecules such as proteins, these substructures may represent individual amino acids residues, whereas in the case of crude natural products extracts these substructures constitute fragments of the various secondary metabolites it contains. However, there are significant differences in the strategies for NMR spectroscopic analysis of crude natural products mixtures and biological macromolecules. Analysis of sets of NMR spectra from proteins or nucleic acids is primarily based on *template recognition*, and thus NMR spectroscopic analysis of biological macromolecules usually consists of sets of 2D and

3D experiments addressing specific structural features of these templates. For example, NMR spectroscopic analysis of proteins is based on a series of specialized NMR-pulse sequences designed to identify amino acid residues, the sequence of amino acids, and spatial proximity within the chain(s). Many macromolecular NMR pulse sequences are highly specific, such as the HNCO experiment to detect the repeating covalent structure of peptide bonds in proteins.[60] While analyzing crude natural products extracts, NMR spectroscopic experiments cannot be tailored in this way, because these crude mixtures usually contain a very large variety of components featuring highly diverse structures. Furthermore, the structural properties of these compounds will vary considerably between extracts, in a largely unpredictable manner. Therefore, NMR-based analysis of crude natural products extracts has to rely on experiments that focus on the most basic common features of organic molecules as frameworks of carbon and hydrogen. These experiments are primarily versions and combinations of (^{1}H,^{1}H)-COSY/TOCSY, (^{1}H,^{1}H)-NOESY/ROESY, and (^{1}H,^{13}C)-HSQC/HMQC/HMBC.[36]

When dealing with spectra of complex mixtures, signal overlap, especially in the proton dimensions, becomes a serious problem, which in some cases may necessitate some form of pre fractionation. Using high-field spectrometers can help increase spectral dispersion and thus reduce overlap, as the relative size of crosspeaks decreases approximately as $(1/F)^n$, with F being the field strength of the magnet and n the dimensionality of the experiment (neglecting line shape effects). Interference by overlap can be alleviated further by taking advantage of the much longer relaxation times of small molecules compared to those of macromolecules. The slower relaxation of small molecules allows for longer acquisition times especially for directly and indirectly detected proton magnetization, which results in correspondingly higher resolution of the spectra and permits detection of smaller scalar couplings. For example, the initial NMR spectroscopic characterization of crude natural products extracts in the studies discussed here was largely based on very high-resolution dqfCOSY spectra.[21,23,61]

9.06.3.1.2 NMR spectroscopy versus mass-spectrometry-based approaches for characterizing crude mixtures

Traditionally, efforts to characterize unfractionated small-molecule mixtures have relied primarily on combinations of MS with HPLC or gas chromatography. As MS is extremely sensitive and typical GC–MS and LC–MS analyses are fast, can be automated easily, and thus can accommodate large numbers of samples, these techniques would seem extremely well suited for the purpose of characterizing libraries of unfractionated natural product extracts,[62] and in fact, various LC–MS-based approaches are being pursued to characterize fungal or bacterial metabolomes.[33] However, there are important drawbacks to the exclusively LC–MS-based approaches. One major disadvantage of using MS as the primary analytical tool is that most mass-spectrometric techniques are strongly biased toward the detection of a few specific compound classes.[63] For example, positive ion electrospray ionization MS is by orders of magnitude more sensitive for basic amines, amino acids, or peptides than for nonbasic polyketides or terpenoids. Alternative ionization techniques such as APCI, MALDI, and so on are biased in different ways and to varying degrees, and no single spectrometric approach is sufficient to provide an unbiased snapshot of a small molecule mixture of unknown composition. Regardless of the ionization technique chosen, the structural information available from mass spectrometric analyses is often insufficient for detailed structural assignments.[64] Although a few compound classes, notably peptides, can be characterized extremely well by MS[65,66] in most classes of small molecule metabolites, MS can provide only limited structural information beyond a tentative molecular formula. In the context of analyzing natural product extracts of diverse origins, this lack of structural information is particularly problematic, as decisions over further fractionation of an extract depend entirely on assumptions as to whether a specific extract is likely to contain new, interesting chemotypes or not. Therefore, the detailed structural information available through 2D NMR spectroscopy of small molecule mixtures can represent an invaluable addition to mass spectroscopic results. Of course, any NMR-based characterization of natural product mixtures normally will have to be complemented by HPLC–MS or GC–MS analyses.

A significant disadvantage of NMR-based approaches for the characterization of natural product mixtures is represented by the much lower sensitivity and dynamic range of NMR spectra compared to MS. Furthermore, the often high complexity of 2D NMR spectra obtained for mixtures can make their interpretation challenging.

The latter disadvantage could be overcome through the use of computational analysis, or through approaches based on graphical comparison of sets of 2D NMR spectra (DANS), as described in the following section.

9.06.3.2 Differential Analysis through 2D NMR Spectroscopy

One of the big remaining challenges in natural products chemistry is to develop better methods for connecting newly identified small molecule structures with their biological functions, including knowledge of the mechanisms regulating their biosynthesis and of their molecular targets. The traditional armamentarium of natural products chemistry appears ill-suited for this purpose, given the complexity of most organism's metabolomes and the scope of assigning functions to hundreds, if not thousands, of individual components, many of which represent previously undescribed chemical structures.[33] Efforts aimed at determining the structure of biologically relevant small molecules have traditionally relied on bioassay-guided fractionation, usually based on highly time-consuming multistep chromatographic fractionation schemes that require extensive biological assays at every stage in the process. As a result, approaches based on bioassay-guided fractionation often take years to tease out and identify the biologically active component(s), and the need for fractionation poses great difficulty in cases of synergism, that is, cases where more than one compound is required to elicit the monitored activity.[22,23] Importantly, for chemically unstable compounds chromatographic fractionation may be unsuitable all together.

Several recent studies have shown that 2D NMR analyses of natural product extracts can be highly effective for associating small molecules with specific biological properties, most significantly phenotype and genotype of the producing organism(s). These studies are based on differential analyses of 2D NMR spectra (DANS), a method for graphic comparison 2D NMR spectra representing different biological states, for example different phenotypes or genotypes.

9.06.3.2.1 DANS for screening of a fungal extract library

DANS was first used for the detection of differential expression of natural products in a small library of fungal extracts.[67] This library was derived from a *Tolypocladium cylindrosporum* strain that was cultured under a variety of 'stress' conditions, aiming to elicit the production of secondary metabolites from otherwise inactive biosynthetic pathways. The resulting unfractionated metabolite extracts were used to acquire dqfCOSY spectra with very high resolution in both dimensions. dqfCOSY was chosen for these studies because dqfCOSY crosspeaks feature highly regular fine structures and are thus particularly information rich. In addition, dqfCOSY spectra offer fairly good dynamic range, which often permits detailed characterization of spin systems representing even very minor components, as had been demonstrated with the examples described in Section 9.06.3.1. For differential analysis, dqfCOSY spectra corresponding to different extracts were superimposed onto each other, using a specific algorithm that suppressed signals common to all extracts, but highlighted signals unique to individual spectra. The algorithm chosen for this overlay allowed suppression of signals even in cases where compounds occurred at significantly different concentrations in different fungal extracts. As a result, only signals representing compounds whose expression was very strongly dependent on the culturing conditions were highlighted in the overlay. The DANS algorithm can be fine-tuned to reveal less severe differences as well, though it is not suitable for accurate quantitative measurements (**Figure 11**).

Application of DANS enabled fast screening of the *Tolypocladium* extract library for proton spin systems representing chemotypes that are produced only under specific conditions, and led to the identification of two new terpenoid indole alkaloids that are expressed under certain nutrient-deficient conditions, but do not get produced using standard culturing protocols. The structures of the two new indole alkaloids, TC-705A (**19**) and TC-705B (**20**), were proposed on the basis of NMR spectra obtained for the unfractionated extracts and subsequently confirmed through additional spectroscopic analyses of isolated samples.[67] In addition to TC-705A and TC-705B, differential expression of several known compounds was observed. These known compounds were identified based on comparison of NMR spectroscopic data obtained from DANS with literature data, in conjunction with results from additional mass spectrometric analyses.

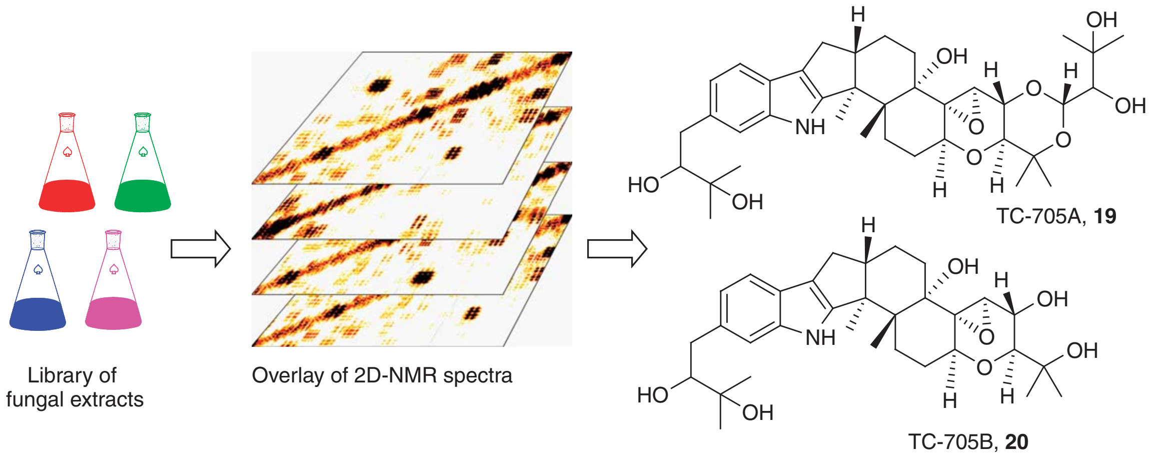


Figure 11 Identification of new fungal natural products through DANS (schematic).[67]

9.06.3.2.2 DANS-based identification of bacillaene

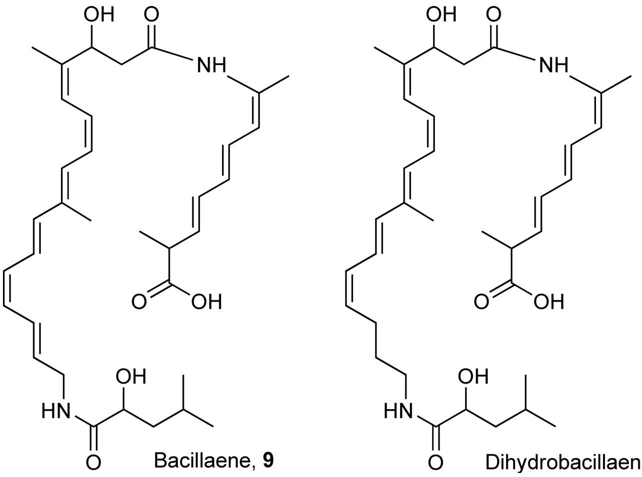
In a second example, DANS was used to determine the structure of the elusive product of the polyketide gene cluster pksX in *B. subtilis*.[24] The ~80 kb pksX gene cluster encodes an unusual hybrid polyketide/nonribosomal peptide synthase that had been linked to the production of the uncharacterized antibiotic bacillaene. Multiple copies of this synthase – each similar in size to the ribosome – assemble into a single organelle-like complex with a mass of tens to hundreds of megadaltons. The resource requirements of the assembled megacomplex suggest that bacillaene serves important biological functions. However, the unconventional domain organization of the PksX synthase and the presence of multiple enzymes that act in *trans* rather than in the standard assembly-line mode that is characteristic of polyketide and nonribosomal peptide biosynthesis precluded bioinformatic prediction of bacillaene's structure. Furthermore, isolation of bacillaene using traditional activity-based fractionation could not be accomplished due to the molecule's chemical instability.

Therefore, identification of bacillaene based on NMR spectra of largely unfractionated bacterial extracts was pursued. DANS-based comparison of a bacillaene-producing *B. subtilis* strain and a corresponding knockout strain clearly identified distinct proton spin systems present in the bacillaene producer but absent in the knockout. Acquisition of additional (^{1}H,^{13}C)-HMQC, (^{1}H,^{13}C)-, and (^{1}H,^{15}N)-HMBC, and ROESY spectra for the bacillaene-producing strain subsequently permitted full identification of the two main products of PksX, bacillaene (**9**) and dihydrobacillaene (**21**), along with several double-bond stereoisomers. The biosynthesis of bacillaene by the PksX synthase was subsequently investigated by Moldenhauer *et al.*[68] Small molecules like bacillaene, which link genotype (the *pksX* gene cluster) with phenotype (antibiotic and likely other activities), are central to chemical biology, and as this example demonstrates, comparative NMR-based approaches such as DANS should be generally useful for their characterization (**Figure 12**).

9.06.3.2.3 Identification of signaling molecules in Caenorhabditis elegans *through DANS*

The utility of DANS for the identification of signaling molecules in eukaryotes was recently demonstrated with the identification of a mating pheromone in the nematode *C. elegans*.[23] *C. elegans* is an important model organism for biomedical research, and a systematic characterization of structures and functions of small molecules in *C. elegans* will be critical for advancing our understanding of many biological processes.[69]

Earlier work had shown that three glycosides of the dideoxysugar ascarylose are part of a male-attracting pheromone that is produced by *C. elegans* hermaphrodites.[22,43,70] These compounds, the ascarosides ascr#2, ascr#3, and ascr#4, showed strong synergism as mating signals: mixtures of ascarosides were potently active at concentrations at which individual components effected no response.[22] Although biologically fascinating, the ascarosides' synergistic properties resulted in tremendous logistical challenges for their identification through activity-guided fractionation, as this required to combinatorially recombine chromatographic fractions in order to assess activity. Despite these efforts, biological testing of mixtures of ascr#2, ascr#3, and ascr#4 at



Bacillaene, **9**

Dihydrobacillaene, **21**

Figure 12 Compounds identified through DANS *Bacillus subtilis*.

physiological concentrations did not fully reproduce activity of the original pheromone extracts, and it seemed likely that important components of the mating pheromones remained to be identified (**Figure 13**).

For the purpose of identifying missing components of the mating pheromone, the *C. elegans* mutant strain *daf-22* offered a unique opportunity. *daf-22*-derived metabolite extracts had been shown to have little dauer-inducing activity and are not significantly active in the male attraction assay. Therefore, a careful comparison of the *daf-22* metabolome with that of wild-type worms should reveal the missing *daf-22*-dependent pheromone components among compounds present in wild-type worms but absent in *daf-22*. As in the examples described in the preceding sections, this comparison was accomplished through DANS based on largely unfractionated metabolite extracts that represent highly complex mixtures of many hundred metabolites.[23] For differential analysis of the dqfCOSY spectra, the *daf-22*-derived spectrum was superimposed onto the wild-type spectrum, again using an algorithm that suppressed signals present in both mutant and wild-type spectra. As a result, only signals present in the wild-type spectrum but entirely absent from the *daf-22* spectrum remained unaltered in the overlay (**Figure 14**).

DANS-based comparisons of *daf-22* and wild-type metabolite extracts revealed several partial structures representing compounds produced only by wild type but not *daf-22* worms, including several previously unknown compounds. These compounds represented far <0.1% of the entire metabolite mixture and therefore further characterization through HSQC or HMBC was not possible based on spectra of the unfractionated metabolite extracts. However, the differentially produced compounds were easily identified after partial

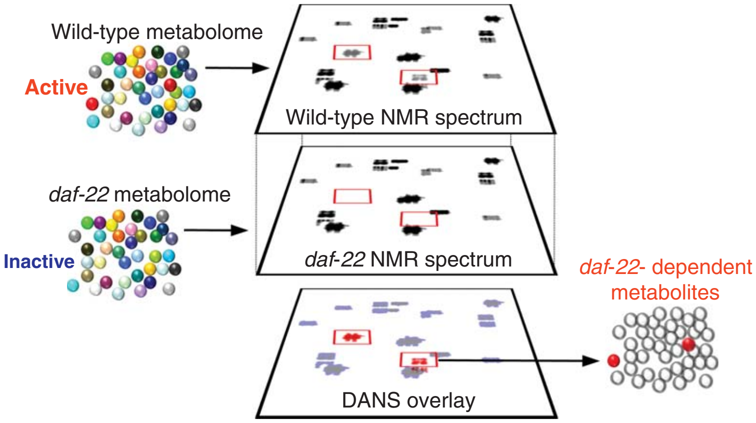


Figure 13 DANS-based comparison of *Caenorhabditis elegans* wild-type and *daf-22* mutant metabolomes.[23]

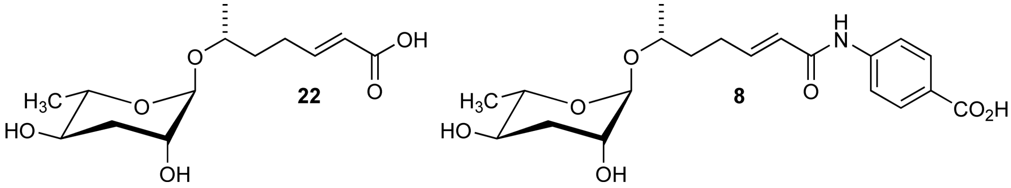


Figure 14 Components of the *Caenorhabditis elegans* mating signal identified through DANS.[23]

chromatographic purification, using additional 2D NMR spectroscopy and MS, as the ascarosides ascr#7 (**22**) and ascr#8 (**8**). In total, the DANS-based comparison of *C. elegans* wild-type and *daf-22* metabolite extracts led to the identification of four novel ascarosides, three of which were shown to function as mating pheromones or regulators of developmental timing. Ultimately, these investigations allowed to fully reconstitute the male-attracting activity of wild-type pheromone extract to that of a *daf-22* mutant.[23]

One significant problem for any comparison of metabolite mixtures is that metabolism is strongly dependent on environmental conditions, and even small changes in temperature, nutrient conditions, or other factors can induce significant changes in relative concentrations of compounds. To minimize the impact of such variations, the algorithm used for DANS in this study was chosen in such a way that it would highlight only cases where a compound is completely absent (given the detection limit of the NMR spectroscopic equipment) from the *daf-22* spectra. Increase of NMR-spectroscopic sensitivity, or consideration of metabolites whose biosynthesis is less strongly *daf-22*-dependent, could reveal additional compounds relevant for phenotypic differences between wild-type and *daf-22* worms.

This study showed that comparative NMR spectroscopic methods such as DANS can be used to dissect changes in small molecule production in response to genetic manipulation, and that this approach could complement or replace activity-guided fractionation for identifying biologically relevant small molecules. The primary benefit of DANS lies in the ability to quickly obtain structural information for metabolites that may represent good candidates for further evaluation in a specific biological context.

9.06.3.3 Complex Mixture Analysis by NMR

Covariance NMR data processing developed by the Bruschweiler laboratory leads to high-resolution symmetric 2D datasets, even with relatively low-resolution acquisition in the indirect dimension.[71,72] Bruschweiler's group has developed an efficient approach called COLMAR to identify individual components in complex biological covariance NMR spectra.[73] COLMAR is freely available through a Web Portal developed and maintained by the Bruschweiler laboratory (http://spin.magnet.fsu.edu/). The input dataset for COLMAR is a covariance processed 2D NMR spectrum. Originally, COLMAR was developed for homonuclear TOCSY spectra, but the Bruschweiler laboratory is adding other options for the analysis of 2D ^{13}C-HSQC–TOCSY datasets. The heart of COLMAR is an algorithm called DemixC, which deconvolutes covariance TOCSY spectra and extracts 1D spectral traces that represent individual spin systems with minimal likelihood of overlap and thus, individual compounds.[13,74] Although they are a probabilistic measure of nonoverlapping spin systems, the 1D traces from DemixC look like 1D NMR spectra and can be analyzed similar to 1D NMR spectra of pure compounds. The final component of COLMAR is an efficient database matching algorithm called COLMAR Query.[75,76] Chemical shifts from the DemixC traces are screened against the BMRB or other metabolomics spectral database.[77] The output of COLMAR Query is a ranked list of the highest scoring compounds with the best matches to known compounds in the database. COLMAR represents an efficient way to semiautomatically identify known compounds from a complex mixture, because it only requires a single 2D TOCSY spectrum as input.

As mixture analysis by NMR is developed, it is increasingly critical to improve NMR small molecule databases. There are currently three main publically accessible small-molecule databases available with NMR data, the Biological Magnetic Resonance Data Bank (BMRB: http://www.bmrb.wisc.edu/),[78,79] the Madison Metabolomics Consortium Database (MMCD: http://mmcd.nmrfam.wisc.edu/),[80] and the Human Metabolome Database (HMDB: http://www.hmdb.ca/).[81] These databases each support searching experimental NMR databases for matches to experimental spectra. The MMCD and HMDB both extensively link to other databases, and MMCD has chemical shift prediction protocols that can aid identification. Importantly, they contain experimental NMR spectra that can be downloaded and compared with experimental mixtures. New tools such as MetaboMiner[82] are being developed to analyze experimental NMR data of unknown mixtures with library spectra from all of the databases. The BMRB database has extensive datasets with raw time-domain data that can be freely downloaded and analyzed. The MMCD directly utilizes the BMRB experimental data, and efforts are being made to put the experimental data from the HMDB into the BMRB database. The BMRB accepts referenced and assigned NMR data from users, so the database is steadily growing. NMR databases are less developed than their GC–MS counterparts, and there are several technical issues related to referencing, solution conditions, scalar couplings, and specific types of NMR experiments and detected nuclei that make NMR more complex than mass-spectrometry databases. As NMR small molecule databases develop, complex mixture analysis by NMR will become more and more important and routine.

9.06.3.4 Metabolomics/Metabonomics

Another very powerful approach to complex mixture analysis by NMR has been developed for biomarker discovery. The Nicholson group has developed computational tools and approaches to identify small-molecule metabolites that change in response to some perturbation such as the use of a drug or from disease.[83] Statistical correlation spectroscopy (STOCSY) is a powerful approach that utilizes the natural variation found in all biological samples to find biomarkers.[27] STOCSY is based on a very simple concept: standard 1D NMR spectra are recorded on a large number of samples, and the NMR signals in these spectra are then statistically correlated by comparing their amplitudes between samples. By statistically correlating the amplitudes of chemical shifts from individual spectra, resonances that are from the same compound or biosynthetic pathway can be identified. Furthermore, the Nicholson group has developed approaches that can be used to discriminate between correlations from the same compound versus correlations from the same metabolic pathway.[84]

Much of the development and most applications of STOCSY concerned very complex mixtures such as human urine or blood plasma. For example, STOCSY can be used to compare groups of control versus diseased or drug-treated individuals in order to discover compounds that are unique biomarkers of the condition of interest.[85–87] A relatively simple extension of STOCSY, called statistical heterospectroscopy (SHY), allows for correlating different types of datasets, such as NMR and MS or even microarray data collected on the same samples.[28] SHY could aid in structure identification from complex mixtures by correlating NMR and mass-spectrometry data to assign molecular weights to compounds with known chemical shifts. One great advantage of the STOCSY approach to complex mixture analysis is that key resonances can be efficiently identified as potential biomarkers. In other words, it can help to find the proverbial needle in a haystack from a large mixture of compounds.

9.06.4 Methods to Improve Sensitivity

Much of the technical development of NMR over the past half century has focused on improving sensitivity. The fundamental problem is the low starting Boltzmann polarization that arises from the low energies of nuclear spin transitions. Several methods have been developed to improve the sensitivity or S/N in NMR. One major approach is through pulse sequence development to optimize the efficiency and information content of NMR spectra through manipulating the spin physics; some of the more important experiments for small molecules were described above.

NMR frequencies are directly proportional to the magnetic field by the basic equation, $\omega_0 = -\gamma B_0$, which relates the frequency (ω_0) to the applied field (B_0) by the gyromagnetic ratio (γ). This simple equation drives the development and purchase of larger and larger magnets, because the S/N goes up as the resonance frequency goes up. The exact increase in S/N depends on many factors, especially differential rates of relaxation at different field strengths, but it is commonly accepted that the S/N increases approximately as ${B_0}^{1.5}$–${B_0}^{1.75}$.[88] Unfortunately, the price of big magnets also increases significantly as the field strength increases.

For example, a 950 MHz (22.3 T) superconducting system costs around \$8 million whereas a 500 MHz (11.7 T) is closer to \$500 000. The biggest magnets also require considerable physical infrastructure and space, making the highest field systems difficult for most users to acquire, maintain, and operate. In the future, NMR facilities might become more like X-ray synchrotron facilities with very large magnets at a few major sites that can provide remote access to users. Some major magnet facilities, such as the National High Magnetic Field Laboratory, also have resistive or hybrid (resistive plus superconducting) magnets that currently can reach field strengths up to 45 T (for hybrid resistive and superconducting magnets) and require large power supplies and other infrastructure. While these low homogeneity magnets are not yet suitable for routine NMR, there is a possibility that methods will be developed to better utilize these for high-resolution studies.[89]

A far more practical solution to improve S/N for most natural products chemists is with the NMR probe. Through the radio frequency (RF) coil, the NMR probe is the interface between the sample and the spectrometer, and it is used to both excite nuclear spins and detect the electrical signals generated by precessing spins. For a fraction of the cost required to purchase a magnet, a fairly routine 500 or 600 MHz system can provide outstanding S/N for small molecules with the right choice of probe. The basic requirements for a probe are that they have an electrical conductor oriented to deliver a magnetic field B_1 that is perpendicular to the static field B_0, and there are several ways to do this. Standard commercial probes that are sold with virtually every NMR system have coils that are made from copper wire and wound in a geometry to deliver a horizontal B_1 magnetic field while accommodating a 5 mm vertically loaded NMR tube. This system has been used successfully for many years, because 5 mm tubes allow for approximately 600 μl of liquid for analysis and provide good S/N for samples with concentrations of about 1 mmol l^{-1} on modern instruments using a standard probe. For a molecule with a molecular weight of 500 Da, an investigator would need about 300 μg to get good results with a standard 5 mm NMR probe using a 600 MHz spectrometer.

For challenging studies, it is helpful to consider two different types of sample limitations, mass and solubility limited. Natural product studies are often mass limited because of challenges associated with the collection and isolation of samples. In contrast, they are often not solubility limited, because a wide range of organic solvents can be employed. In contrast, studies with proteins or other biological macromolecules often are solubility limited, but relatively large quantities of samples can often be produced. The worst scenario for NMR is when a sample is both mass and solubility limited. Although there is some overlap, the type of sample will often dictate the choice in NMR probe technology that can best solve the problem.

Many natural product samples are severely mass limited, and it is difficult or impossible to isolate enough material to achieve the necessary concentration in a 5-mm tube. Several methods can be worked out to improve the situation with mass limited samples. Perhaps the simplest is to use 5-mm NMR tubes with susceptibility matched plugs that reduce the need for excess sample outside of the active volume of the probe. Samples need to be long enough to extend beyond the coil to avoid edge effects that severely degrade the homogeneity of the field. Susceptibility plugs allow most of the sample to be positioned in the center of the probe, but they require careful loading and positioning to avoid air bubbles that will degrade line shapes. This will bring the sample requirements of a 500 Da compound to about 150 μg in a standard probe at 600 MHz. Although susceptibility matched NMR tubes are useful in optimizing the use of the available sample, they can be difficult to shim and, as a result, the line shapes are sometimes compromised, which can lower overall S/N, especially in HMBC spectra. Of all 2D spectra routinely used for small molecule structure elucidation, HMBC spectra generally have the lowest S/N, and, unfortunately, signal strength in HMBC spectra is also strongly dependent on ^{1}H line shapes.

9.06.4.1 Specialized NMR Probes

For studies at common magnetic field strengths such as 11.7 T (500 MHz) or 14.1 T (600 MHz), there are three main ways to improve S/N beyond a standard 5-mm probe. The simplest is to make the coils smaller, as the mass sensitivity of an NMR measurement increases roughly in inverse proportion to the diameter of the coil. A second very popular approach is to cool the entire coil and preamplifier in order to reduce the noise, thus increasing the S/N. A third approach utilizes material that conducts electricity more efficiently than copper wire.

9.06.4.2 Signal-to-Noise Issues

S/N values are routinely used in NMR, especially when shopping for a new spectrometer or probe. One would think that this ratio of two numbers would be an unambiguous and objective way to compare systems, but unfortunately, it is not so straightforward. First, major NMR vendors use different algorithms to estimate noise, and several additional definitions of noise are used in the literature. Second, the thickness of the walls of NMR tubes can influence S/N measurements, especially as the tube diameter decreases. Not all probe and spectrometer manufacturer's use the same standards. It is most common among conventional top-loading tube systems to use 0.1% ethylbenzene in $CDCl_3$, but solenoidal flow systems typically report S/N values using 10 mmol l^{-1} sucrose in D_2O. Finally, when working with very small volumes, solvent volatility can play a role in manufacturing consistent sealed standards. For example, when evaluating the performance of a 1-mm probe, we found differences as large as 10% between two factory-sealed samples of 0.1% ethylbenzene in $CDCl_3$. In short, S/N is a useful guide but needs to be interpreted with great care, especially when informing decisions on major purchases.

9.06.4.3 Small Coils

The sensitivity of an NMR coil is defined as the B_1 field per unit current, and this is inversely proportional to the diameter of the coil,[88] so smaller coils have greater mass sensitivity. Any type of coil can be made smaller, and standard saddle coil or Helmholtz designs that accommodate vertically loaded tubes are commercially available as small as 1 mm in diameter. However, depending on details of the coil geometry, solenoid coils can provide between 2 and 3 times greater sensitivity than a standard saddle coil.[88] Solenoid coils pose two challenges, both of which have been nicely addressed. First, the horizontal orientation of the solenoid in the main B_0 field causes severe distortions to the field. Uncorrected, this leads to poor NMR line shape and big losses in sensitivity. Andrew Webb, Jonathan Sweedler, and colleagues solved this problem by surrounding the coil in a fluid that has the same magnetic susceptibility as copper wire.[90] Using this approach, high-quality NMR spectra can be obtained from extremely small volumes of sample by using small coil diameters.[90–92] The second problem involves sample handling; the horizontal orientation of the coil makes standard NMR tubes and sample loading impossible. In principle, samples can be placed in sealed capillary tubes and inserted into the coils, but this is cumbersome and requires that the probe be removed from the magnet for each sample change. Moreover, sealing the tubes without introducing air bubbles is difficult. A much more practical solution is to connect tubing to flow the samples into and out of the coil. This loading scheme can be as simple as a syringe attached to tubing or as complex as the output of a chromatographic separation.

Integrated systems that utilize 1-mm solenoidal microcoil probes and various sample-loading methods are available commercially from Protasis. The utility of commercially available solenoidal microprobes for the analysis of mass-limited natural products has recently been reviewed.[10] Examples for natural products applications include the identification of 13 new steroids from only 50 specimens of the firefly *Lucidota atra* (e.g., **4** in **Figure 2**).[4] These analyses were carried out on only partially purified samples, each containing 20–100 μg of up to three steroids. In direct comparison to using a 5-mm inverse-detection room temperature probe and susceptibility plugs (Shigemi tubes), the use of a solenoidal microprobe provided an up to threefold gain in S/N while maintaining very high spectral quality.

These small-volume systems are good choices for either high-throughput semiautomated analysis or in environments with multiple users, because the probes can be easily switched with other standard probes.[10] S/N comparisons between conventional tube systems and flow solenoids are especially problematic. However, based on comparisons between a 1-mm cryogenic HTS probe[93] and values in the literature from a 1-mm room temperature solenoid,[94] about 30 μg of a 500 Da sample would give comparable results to a 1 mmol l^{-1} sample in a 5-mm warm copper probe.

9.06.4.4 Cooling the Electronics

A second approach to increasing S/N through NMR probe design involves reducing the coil and receiver noise. Significant advances have been made during the past decade in cryogenically cooling the coils, electrical circuits, and preamplifiers in order to reduce the thermal noise associated with the measurement.[95] Major commercial NMR vendors offer cryogenically cooled probes, and these are very effective, even with standard copper wire and coil geometries that allow top-loading samples. All of these probes thermally isolate the sample from the coils, which are cooled to about 20 K. Sample temperatures can be regulated in modest ranges around room temperature, so biological samples are easily analyzed. Although the increase in S/N is dependent upon the dielectric properties of the sample, an increase of about 4× is not unusual. Most cryogenic probes accommodate 5-mm tubes, and with sample tubes that are large, the conductivity of the solvent can have a significant influence on S/N with these probes. For organic solvents and low-salt aqueous buffers, cryogenic probes deliver the best results. However, even moderate salt concentrations can seriously degrade their performance. The salt dependence worsens with increasing field strengths and with larger diameter samples. A 500 Da sample in an organic solvent would require roughly 75 μg for good NMR spectra in a 5-mm cryoprobe using a standard 600 MHz spectrometer. One should note that the required sample concentration would only be about 250 μmol l^{-1}, whereas in order to get similar performance with a room-temperature probe one would have to use a much higher concentration of about 1 mmol l^{-1}. Large-volume cryogenic probes are excellent choices for samples with limited solubility, explaining their widespread use in biomolecular NMR.

The primary disadvantages of cryogenic probes are that they are very expensive, require more physical infrastructure like chilled water lines than a conventional probe, and are difficult to install and remove. Most facilities with cryogenic probes keep them in dedicated instruments and only remove them for maintenance or repair. This works well for groups with similar samples and needs. Cryogenic probes typically have fixed frequencies and cannot replace the flexibility of broadband probes for unusual nuclei.

9.06.4.5 High-Temperature Superconducting Coils

Copper-based material is the most common conductor for NMR probes. However, there are other choices, which can provide better sensitivity through improved current carrying capacity. High-temperature superconducting material, specifically YBCO (yttrium barium copper oxide), has been used since the early 1990s in NMR coils. The first HTS probe was designed and built at Conductus (Sunnyvale, CA) in the 1990s.[96] HTS coils are constructed by depositing YBCO onto planar surfaces and inductively coupling them to a copper RF circuit.[97] These coils have a much higher quality factor (*Q*) than cooled copper coils and as a result have been shown to provide significantly higher S/N in NMR than achieved by cold copper coils. The drawbacks and challenges to HTS coils include poor filling factors due to the flat coil geometry, difficulty and cost in construction, and difficulty in tuning flat wafers to multiple frequencies required for biological NMR. However, the benefits of HTS probe technology are significant: for the same temperature and coil diameter, planar HTS coils can increase the S/N by up to a factor of 2 over copper wire.

Combining HTS materials with cryogenic cooling and small coil size can result in very sensitive NMR probes. The National High Magnetic Field Laboratory, Bruker Biospin, and University of Florida recently collaborated to design and build a 1-mm cryogenic probe with HTS coils.[93] This probe uses top-loading glass tubes with sample volumes between 5 and 10 μl, depending on the wall thickness. This probe has an S/N value of close to 300 for 0.1% ethylbenzene, which is about 20× more mass sensitive than a commercial 5-mm warm copper probe (S/N ~ 1000 for 0.1% ethylbenzene). Thus, a 500 Da sample would require about 10 μg for good results. However, the concentration would increase to about 2 mmol l^{-1}, so the smallest probes are not appropriate for concentration-limited samples. This 1-mm HTS probe has been used in several natural products studies, including the analysis of insect defensive secretions with a single or very few insects,[12,13,98–100] identification of a component of the *C. elegans* mating pheromone,[22] identification of glycosylated pheromones suspensoside A (**23**) and suspensoside B from male Caribbean fruit flies,[101] and several marine natural product identifications (**Figure 15**).[7,102,103]

A 1.7 mm cryogenic probe is now available commercially, and this appears to provide excellent results. Dalisay *et al.*[5,6] reported the identification of several new natural products of very low abundance from marine sponges of the genus *Phorbas*, including the tetrachloro polyketide muironolide A (**24**) and another polyketide, hemi-phorboxazol A (**2**). These structures were determined based on samples of only 90 μg of muironolide and 16.5 μg of hemiphorboxazol A.

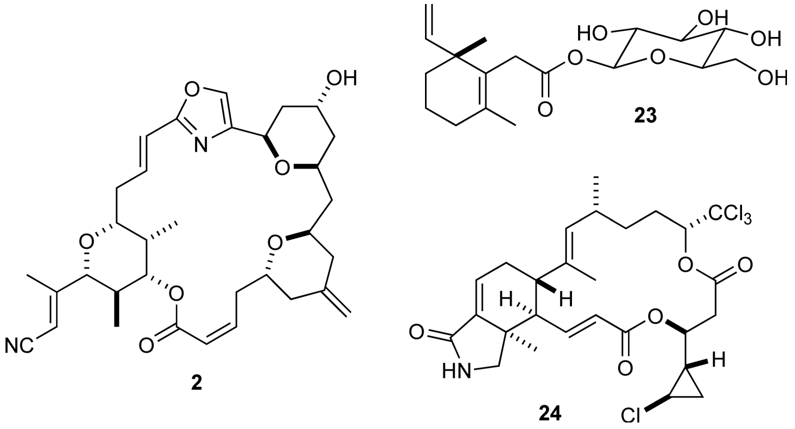


Figure 15 Examples for natural products identified using small-volume cryogenic probes.

The limitations of small-volume cryogenic probes are similar to standard cryogenic probes. If the coils are made from HTS material, there are additional challenges related to the fact that the coils need to be on planar surfaces that are not fully optimized to the geometry of cylindrical sample tubes and the glass vacuum tubes needed to isolate the cryogenic coil temperatures from the sample at room temperature.

9.06.4.6 Probe Summary

There are many choices of probes, and the best one depends on the amount of sample, the solubility of the sample, the number of different types of users of the system, and the budget. The most flexible probes in terms of accommodating a wide range of samples and users are standard 5-mm room temperature probes. Smaller diameter probes have higher mass sensitivity but because their sample volumes are dramatically smaller, they are not optimal for samples with limited solubility. The highest sensitivity probes are small, cryogenically cooled, and utilize high-conducting HTS materials, but these are expensive and suffer from general limitations of cumbersome cryogenic systems and small volumes. A good bet for general biomolecular NMR and some natural product work is a 5-mm cryogenically cooled probe; however, all cryogenic probes are expensive and more difficult to change with other probes such as broadband for nonproton-detected studies.

9.06.4.7 Dynamic Nuclear Polarization

Dynamic nuclear polarization (DNP) is a rapidly developing technique that achieves significantly higher S/N than conventional NMR spectroscopy by transferring the very large polarization of electrons to nuclei at low temperatures. Much of the development in DNP focused on solid state samples and frozen liquid samples,[104] and techniques to directly polarize solutions in high magnetic fields are now being developed.[105] Under the right conditions, DNP enhancements can reach several orders of magnitude,[104,106] but there are some limitations of DNP that need to be overcome before it could be widely applied to natural product studies. First, DNP only lasts as long as the T_1 of the polarized nucleus, so most applications are short 1D ^{13}C experiments, although rapid acquisition 2D methods have been demonstrated on hyperpolarized samples.[107] Second, nuclei with short T_1 times are difficult or impossible to detect, and therefore identification of unknown compounds could be challenging without additional analyses using conventional NMR spectroscopy. Third, radicals need to be added to samples to provide a source of electrons, and investigators might be unwilling to add these to very precious natural product samples that took considerable time to isolate and purify. Finally, most solution DNP studies employ samples that were polarized in the frozen state and then thawed,[106] and most studies have used aqueous solutions. Modified protocols and polarizing agents would need to be developed for natural products analysis in organic solvents. Despite these current limitations, the future of DNP for major enhancements of NMR S/N is very promising, and natural products chemists should follow the developments of this field.

9.06.5 Outlook

It is an exciting time for natural products chemistry. Analytical tools are now available that significantly reduce the amount of sample needed for structure determination. Studies that required heroic efforts and years to isolate enough material for NMR spectroscopic analysis a few decades ago can now be done with two, three, or more orders of magnitude less material today. This not only makes natural products research much more efficient, more importantly, it opens up possibilities for entirely new lines of scientific inquiry, involving individual variation, population chemical biology, and much more extensive examination of the influence of genetic or environmental factors in natural product expression levels.

Because of the advances in analytical technology, natural products chemists now have tremendous opportunities to take full advantage of 'omics'-type approaches. Genomics and proteomics technologies and databases make it much more feasible to both study and manipulate the biosynthesis of important natural products. This will allow a more complete understanding of basic biological processes but also will enable more efficient drug development that uses natural products as a starting point.

Computational power and databases constantly improve and will allow the design of more and more comprehensive approach to biological problems. Metabolomics has recently emerged as a key component of 'systems biology', and with new analytical, computational, and information technology, metabolomics is evolving into a central hub that connects many of the other 'omics' to better understand biology, and consequently, human health. As analytical and computational tools improve, the distinction between natural products chemistry and metabolomics is becoming less and less clear. Natural products represent both downstream products and upstream regulators of metabolic pathways, and as we learn more about these interactions, we learn more about function and possible applications of natural products. One of the current challenges in metabolomics studies is 'biomarker identification', a new term for natural products chemistry. As traditional natural products studies become more integrated with the 'omics' and as metabolomics becomes more focused on identifying key metabolites, the two fields will become less and less distinct.

Abbreviations

2D-NMR	two-dimensional NMR
COLMAR	complex liquid mixture analysis by NMR
COSY	correlation spectroscopy
DANS	differential analysis of 2D NMR spectra
DECODES	diffusion-encoded spectroscopy
DNP	dynamic nuclear polarization
DOSY	diffusion-ordered spectroscopy
dqfCOSY	double-quantum-filtered correlation spectroscopy
E.COSY	exclusive correlation spectroscopy
gCOSY	gradient correlation spectroscopy
HMBC	heteronuclear multiple-bond correlation
HMQC	heteronuclear multiple-quantum correlation
HOHAHA	homonuclear Hartmann–Hahn correlation
HSQC	heteronuclear single-quantum correlation
INADEQUATE	incredible natural abundance double quantum transfer experiment
MS	mass spectrometry
NMR	nuclear magnetic resonance
NOESY	nuclear Overhauser effect spectroscopy
ROESY	rotating-frame Overhauser effect spectroscopy
SHY	statistical heterospectroscopy
STOCSY	statistical correlation spectroscopy
TOCSY	total correlation spectroscopy

References

1. P. F. Wiley; K. Gerzon; E. H. Flynn; M. V. Sigal; O. Weaver; U. C. Quarck; R. R. Chauvette; R. Monahan, *J. Am. Chem. Soc.* **1957**, *79*, 6062–6070.
2. P. F. Wiley; M. V. Sigal; O. Weaver; R. Monahan; K. Gerzon, *J. Am. Chem. Soc.* **1957**, *79*, 6070–6074.
3. P. F. Wiley; R. Gale; C. W. Pettinga; K. Gerzon, *J. Am. Chem. Soc.* **1957**, *79*, 6074–6077.
4. M. Gronquist; J. Meinwald; T. Eisner; F. C. Schroeder, *J. Am. Chem. Soc.* **2005**, *127*, 10810–10811.
5. D. S. Dalisay; T. F. Molinski, *Org. Lett.* **2009**, *11*, 1967–1970.
6. D. S. Dalisay; B. I. Morinaka; C. K. Skepper; T. F. Molinski, *J. Am. Chem. Soc.* **2009**, *131*, 7552–7553.
7. D. S. Dalisay; E. W. Rogers; A. S. Edison; T. F. Molinski, *J. Nat. Prod.* **2009**, *72*, 732–738.
8. T. F. Molinski, *Curr. Opin. Drug Discovery Dev.* **2009**, *12*, 197–206.
9. J. Meinwald; D. F. Wiemer; T. Eisner, *J. Am. Chem. Soc.* **1979**, *101*, 3055–3060.
10. F. C. Schroeder; M. Gronquist, *Angew. Chem. Int. Ed. Engl.* **2006**, *45*, 7122–7131.
11. M. Yoshida; M. Murata; K. Inaba; M. Morisawa, *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 14831–14836.
12. A. T. Dossey; S. S. Walse; J. R. Rocca; A. S. Edison, *ACS Chem. Biol.* **2006**, *1*, 511–514.
13. F. Zhang; A. T. Dossey; C. Zachariah; A. S. Edison; R. Bruschweiler, *Anal. Chem.* **2007**, *79*, 7748–7752.
14. H. Barjat; G. A. Morris; S. Smart; A. G. Swanson; S. C. R. Williams, *J. Magn. Reson. Ser. B* **1995**, *108*, 170–172.
15. K. Bleicher; M. F. Lin; M. J. Shapiro; J. R. Wareing, *J. Org. Chem.* **1998**, *63*, 8486–8490.
16. M. F. Lin; M. J. Shapiro, *J. Org. Chem.* **1996**, *61*, 7617–7619.
17. M. F. Lin; M. J. Shapiro; J. R. Wareing, *J. Org. Chem.* **1997**, *62*, 8930–8931.
18. M. F. Lin; M. J. Shapiro; J. R. Wareing, *J. Am. Chem. Soc.* **1997**, *119*, 5249–5250.
19. M. G. Lin; M. J. Shapiro, *Anal. Chem.* **1997**, *69*, 4731–4733.
20. A. E. Taggi; J. Meinwald; F. C. Schroeder, *J. Am. Chem. Soc.* **2004**, *126*, 10364–10369.
21. F. C. Schroeder; A. E. Taggi; M. Gronquist; R. U. Malik; J. B. Grant; T. Eisner; J. Meinwald, *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 14283–14287.
22. J. Srinivasan; F. Kaplan; R. Ajredini; C. Zachariah; H. T. Alborn; P. E. Teal; R. U. Malik; A. S. Edison; P. W. Sternberg; F. C. Schroeder, *Nature* **2008**, *454*, 1115–1118.
23. C. Pungaliya; J. Srinivasan; B. W. Fox; R. U. Malik; A. H. Ludewig; P. W. Sternberg; F. C. Schroeder, *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 7708–7713.
24. R. A. Butcher; F. C. Schroeder; M. A. Fischbach; P. D. Straight; R. Kolter; C. T. Walsh; J. Clardy, *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 1506–1509.
25. F. Schroder; V. Sinnwell; H. Baumann; M. Kaib, *Chem. Commun.* **1996**, 2139–2140.
26. F. Schroder; V. Sinnwell; H. Baumann; M. Kaib; W. Francke, *Angew. Chem. Int. Ed. Engl.* **1997**, *36*, 77–80.
27. E. Holmes; O. Cloarec; J. K. Nicholson, *J. Proteome Res.* **2006**, *5*, 1313–1320.
28. D. J. Crockford; E. Holmes; J. C. Lindon; R. S. Plumb; S. Zirah; S. J. Bruce; P. Rainville; C. L. Stumpf; J. K. Nicholson, *Anal. Chem.* **2006**, *78*, 363–371.
29. D. B. Kell; M. Brown; H. M. Davey; W. B. Dunn; I. Spasic; S. G. Oliver, *Nat. Rev. Microbiol.* **2005**, *3*, 557–565.
30. J. K. C. Nicholson; L. John; C. John; E. Holmes, *Nat. Rev. Drug Discov.* **2002**, *1*, 153–161.
31. B. J. Blaise, *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 19808.
32. Y. Wang; J.r. Utzinger; J. Saric; J. V. Li; J. Burckhardt; S. Dirnhofer; J. K. Nicholson; B. H. Singer; R. Brun; E. Holmes, *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 6127–6132.
33. S. Rochfort, *J. Nat. Prod.* **2005**, *68*, 1813–1820.
34. A. E. Derome, *Nat. Prod. Rep.* **1989**, *6*, 111–141.
35. W. F. Reynolds; R. G. Enriquez, *J. Nat. Prod.* **2002**, *65*, 221–244.
36. E. E. Kwan; S. G. Huang, *Eur. J. Org. Chem.* **2008**, 2671–2688.
37. T. D. W. Claridge, *High-Resolution NMR Techniques in Organic Chemistry*, 1st ed.; Pergamon: Amsterdam: New York, 1999.
38. S. L. Robinette; F. Zhang; L. Bruschweiler-Li; R. Bruschweiler, *Anal. Chem.* **2008**, *80*, 3606–3611.
39. F. Zhang; R. Bruschweiler, *ChemPhysChem* **2004**, *5*, 794–796.
40. S. P. Rucker; A. J. Shaka, *Mol. Phys.* **1989**, *68*, 509–517.
41. A. Bax; D. G. Davis, *J. Magn. Reson.* **1985**, *65*, 355–360.
42. A. Bax, *Methods Enzymol.* **1989**, *176*, 151–168.
43. R. A. Butcher; M. Fujita; F. C. Schroeder; J. Clardy, *Nat. Chem. Biol.* **2007**, *3*, 420–422.
44. C. Griesinger; O. W. Sorensen; R. R. Ernst, *J. Magn. Reson.* **1987**, *75*, 474–492.
45. B. Vogeli; L. Yao; A. Bax, *J. Biomol. NMR* **2008**, *41*, 17–28.
46. F. C. Schroeder; T. Tolasch, *Tetrahedron* **1998**, *54*, 12243–12248.
47. T. D. Claridge; I. Perez-Victoria, *Org. Biomol. Chem.* **2003**, *1*, 3632–3634.
48. D. Neuhaus; M. P. Williamson, *The Nuclear Overhauser Effect in Structural and Conformational Analysis*, 2nd ed.; Wiley-VCH: New York, 2000.
49. A. Bax; R. Freeman; T. A. Frenkiel, *J. Am. Chem. Soc.* **1981**, *103*, 2102–2104.
50. W. Francke; F. Schroeder; F. Walter; V. Sinnwell; H. Baumann; M. Kaib, *Liebigs Ann.* **1995**, *0*, 965–977.
51. J. T. Arnold; S. S. Dharmatti; M. E. Packard, *J. Chem. Phys.* **1951**, *19*, 507.
52. M. E. Packard; J. T. Arnold, *Phys. Rev.* **1951**, *83*, 210–211.
53. R. R. Ernst, *Biosci. Rep.* **1992**, *12*, 143–187.
54. N. Bross-Walch; T. Kuhn; D. Moskau; O. Zerbe, *Chem. Biodivers.* **2005**, *2*, 147–177.
55. P. L. Rinaldi, *Analyst* **2004**, *129*, 687–699.
56. F. Schroeder; S. Franke; W. Francke; H. Baumann; M. Kaib; J. M. Pasteels; D. Daloze, *Tetrahedron* **1996**, *52*, 13539–13546.
57. F. C. Schroeder; S. R. Smedley; L. K. Gibbons; J. J. Farmer; A. B. Attygalle; T. Eisner; J. Meinwald, *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 13387–13391.

58. F. C. Schroder; J. J. Farmer; A. B. Attygalle; S. R. Smedley; T. Eisner; J. Meinwald, *Science* **1998**, *281*, 428–431.
59. J. McCormick; Y. Li; K. McCormick; H. I. Duynstee; A. K. van Engen; G. A. van der Marel; B. Ganem; J. H. van Boom; J. Meinwald, *J. Am. Chem. Soc.* **1999**, *121*, 5661–5665.
60. S. Grzesiek; H. Dobeli; R. Gentz; G. Garotta; A. M. Labhardt; A. Bax, *Biochemistry* **1992**, *31*, 8180–8190.
61. F. C. Schroeder; D. M. Gibson; A. C. Churchill; P. Sojikul; E. J. Wursthorn; S. B. Krasnoff; J. Clardy, *Angew. Chem. Int. Ed. Engl.* **2007**, *46*, 901–904.
62. F. E. Koehn, *Prog. Drug Res.* **2008**, *65*, *175,* 177–210.
63. K. Biemann, *J. Am. Soc. Mass Spectrom.* **2002**, *13*, 1254–1272.
64. L. Smith; J. Novak; J. Rocca; S. McClung; J. D. Hillman; A. S. Edison, *Eur. J. Biochem.* **2000**, *267*, 6810–6816.
65. V. H. Wysocki; K. A. Resing; Q. Zhang; G. Cheng, *Methods* **2005**, *35*, 211–222.
66. K. Adermann; H. John; L. Standker; W. G. Forssmann, *Curr. Opin. Biotechnol.* **2004**, *15*, 599–606.
67. F. C. Schroeder; D. M. Gibson; A. C. L. Churchill; P. Sojikul; E. J. Wursthorn; S. B. Krasnoff; J. Clardy, *Angew. Chem. Int. Ed. Engl.* **2007**, *46*, 901–904.
68. J. Moldenhauer; X. H. Chen; R. Borriss; J. Piel, *Angew. Chem. Int. Ed. Engl.* **2007**, *46*, 8195–8197.
69. F. C. Schroeder, *ACS Chem. Biol.* **2006**, *1*, 198–200.
70. P. Y. Jeong; M. Jung; Y. H. Yim; H. Kim; M. Park; E. Hong; W. Lee; Y. H. Kim; K. Kim; Y. K. Paik, *Nature* **2005**, *433*, 541–545.
71. R. Bruschweiler; F. Zhang, *J. Chem. Phys.* **2004**, *120*, 5253–5260.
72. R. Bruschweiler, *J. Chem. Phys.* **2004**, *121*, 409–414.
73. F. Zhang; L. Bruschweiler-Li; S. L. Robinette; R. Bruschweiler, *Anal. Chem.* **2008**, *80*, 7549–7553.
74. F. L. Zhang; R. Bruschweiler, *ChemPhysChem* **2004**, *5*, 794–796.
75. S. L. Robinette; F. Zhang; L. Brüschweiler-Li; R. Brüschweiler, *Anal. Chem.* **2008**, *80*, 3606–3611.
76. D. A. Snyder; F. Zhang; S. L. Robinette; L. Brüschweiler-Li; R. Brüschweiler, *J. Chem. Phys.* **2008**, *128*, 052313.
77. B. R. Seavey; E. A. Farr; W. M. Westler; J. L. Markley, *J. Biomol. NMR* **1991**, *1*, 217–236.
78. E. L. Ulrich; H. Akutsu; J. F. Doreleijers; Y. Harano; Y. E. Ioannidis; J. Lin; M. Livny; S. Mading; D. Maziuk; Z. Miller; E. Nakatani; C. F. Schulte; D. E. Tolmie; R. Kent Wenger; H. Yao; J. L. Markley, *Nucleic Acids Res.* **2008**, *36*, D402–D408.
79. J. L. Markley; M. E. Anderson; Q. Cui; H. R. Eghbalnia; I. A. Lewis; A. D. Hegeman; J. Li; C. F. Schulte; M. R. Sussman; W. M. Westler; E. L. Ulrich; Z. Zolnai, *Pac. Symp. Biocomput.* **2007**, *12*, 157–168.
80. Q. L. Cui; A. Ian; A. D. Hegeman; M. E. Anderson; J. Li; C. F. Schulte; W. M. Westler; H. R. Eghbalnia; M. R. Sussman; J. L. Markley, *Nat. Biotechnol.* **2008**, *26*, 162–164.
81. D. S. Wishart; D. Tzur; C. Knox; R. Eisner; A. Guo; N. Young; D. Cheng; K. Jewell; D. Arndt; S. Sawhney; C. Fung; L. Nikolai; M. Lewis; M. Coutouly; I. Forsythe; P. Tang; S. Shrivastava; K. Jeroncic; P. Stothard; G. Amegbey; D. Block; D. Hau; J. Wagner; J. Miniaci; M. Clements; M. Gebremedhin; N. Guo; Y. Zhang; G. E. Duggan; G. D. Macinnis; A. M. Weljie; R. Dowlatabadi; F. Bamforth; D. Clive; R. Greiner; L. Li; T. Marrie; B. D. Sykes; H. J. Vogel; L. Querengesser, *Nucleic Acids Symp. Ser.* **2007**, *35*, D521–D526.
82. J. Munger; B. D. Bennett; A. Parikh; X.-J. Feng; J. McArdle; H. A. Rabitz; T. Shenk; J. D. Rabinowitz, *Nat. Biotechnol.* **2008**, *26*, 1179–1186.
83. J. K. Nicholson; J. C. Lindon, *Nature* **2008**, *455*, 1054–1056.
84. A. Couto Alves; M. Rantalainen; E. Holmes; J. K. Nicholson; T. M. Ebbels, *Anal. Chem.* **2009**, *81*, 2075–2084.
85. A. D. Maher; O. Cloarec; P. Patki; M. Craggs; E. Holmes; J. C. Lindon; J. K. Nicholson, *Anal. Chem.* **2009**, *81*, 288–295.
86. A. D. Maher; D. Crockford; H. Toft; D. Malmodin; J. H. Faber; M. I. McCarthy; A. Barrett; M. Allen; M. Walker; E. Holmes; J. C. Lindon; J. K. Nicholson, *Anal. Chem.* **2008**, *80*, 7354–7362.
87. E. Holmes; R. L. Loo; O. Cloarec; M. Coen; H. R. Tang; E. Maibaum; S. Bruce; Q. Chan; P. Elliott; J. Stamler; I. D. Wilson; J. C. Lindon; J. K. Nicholson, *Anal. Chem.* **2007**, *79*, 2629–2640.
88. D. I. Hoult; R. E. Richards, *J. Magn. Reson.* **1976**, *24*, 71–85.
89. B. Shapira; K. Shetty; W. W. Brey; Z. H. Gan; L. Frydman, *Chem. Phys. Lett.* **2007**, *442*, 478–482.
90. D. L. Olson; T. L. Peck; A. G. Webb; R. L. Magin; J. V. Sweedler, *Science* **1995**, *270*, 1967–1970.
91. D. Raftery, *Anal. Bioanal. Chem.* **2004**, *378*, 1403–1404.
92. A. G. Webb, *J. Pharm. Biomed. Anal.* **2005**, *38*, 892–903.
93. W. W. Brey; A. S. Edison; R. E. Nast; J. R. Rocca; S. Saha; R. S. Withers, *J. Magn. Reson.* **2006**, *179*, 290–293.
94. D. L. Olson; J. A. Norcross; M. O'Neil-Johnson; P. F. Molitor; D. J. Detlefsen; A. G. Wilson; T. L. Peck, *Anal. Chem.* **2004**, *76*, 2966–2974.
95. H. Kovacs; D. Moskau; M. Spraul, *Prog. Nuclear Magn. Reson. Spectrosc.* **2005**, *46*, 131–155.
96. W. W. Brey; W. Anderson; W. H. Wong; L. F. Fuks; V. Y. Kotsubo; R. S. Withers, Nuclear Magnetic Resonance Probe Coil. U.S. Patent 5,565,778, 1996.
97. W. A. Anderson; W. W. Brey; A. L. Brooke; B. Cole; K. A. Delin; L. F. Fuks; H. D. W. Hill; M. E. Johanson; V. Kotsubo; R. Nast; R. S. Withers; W. H. Wong, *Bull. Magn. Reson.* **1995**, *17*, 98–102.
98. A. T. Dossey; S. S. Walse; A. S. Edison, *J. Chem. Ecol.* **2008**, *34*, 584–590.
99. A. T. Dossey; S. S. Walse; O. V. Conle; A. S. Edison, *J. Nat. Prod.* **2007**, *70*, 1335–1338.
100. B. Wang; A. T. Dossey; S. S. Walse; A. S. Edison; K. M. Merz, Jr., *J. Nat. Prod.* **2009**, *72*, 709–713.
101. S. S. Walse; F. Lu; P. E. Teal, *J. Nat. Prod.* **2008**, *71*, 1726–1731.
102. S. Matthew; P. J. Schupp; H. Luesch, *J. Nat. Prod.* **2008**, *71*, 1113–1116.
103. J. C. Kwan; J. R. Rocca; K. A. Abboud; V. J. Paul; H. Luesch, *Org. Lett.* **2008**, *10*, 789–792.
104. A. B. Barnes; G. D. Paepe; P. C. van der Wel; K. N. Hu; C. G. Joo; V. S. Bajaj; M. L. Mak-Jurkauskas; J. R. Sirigiri; J. Herzfeld; R. J. Temkin; R. G. Griffin, *Appl. Magn. Reson.* **2008**, *34*, 237–263.
105. M. J. Prandolini; V. P. Denysenkov; M. Gafurov; B. Endeward; T. F. Prisner, *J. Am. Chem. Soc.* **2009**, *131*, 6090–6092.
106. J. H. Ardenkjaer-Larsen; B. Fridlund; A. Gram; G. Hansson; L. Hansson; M. H. Lerche; R. Servin; M. Thaning; K. Golman, *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 10158–10163.
107. L. Frydman; D. Blazina, *Nat. Phys.* **2007**, *3*, 415–419.

Biographical Sketches

Arthur S. Edison obtained a B.S. in chemistry from the University of Utah, where he studied monoterpenes isolated from southern Utah sagebrush by NMR. He completed his Ph.D. in biophysics from the University of Wisconsin, Madison, where he developed and applied NMR methods for peptide and protein structural studies under the supervision of John Markley and Frank Weinhold. In 1993, Dr. Edison joined the laboratory of Anthony O. W. Stretton at the University of Wisconsin as a Jane Coffin Childs postdoctoral fellow where he investigated the role of neuropeptides in the nervous system of the parasitic nematode *Ascaris suum.* He joined the faculty at the University of Florida and the National High Magnetic Field Laboratory in 1996 and is currently the Director of Chemistry & Biology at the NHMFL. Dr. Edison's current research is in technology development for high-sensitivity NMR and natural product discovery in nematodes and other invertebrates. Dr. Edison is the recipient of the 1997 American Heart Association Robert J. Boucek Award, a CAREER Award from the National Science Foundation in 1999, and, with his postdoctoral scientist Aaron Dossey, the Beal award for the best publication of the year in the *Journal of Natural Products* in 2007.

Frank C. Schroeder studied chemistry and physics at the University of Hamburg, where he worked under the guidance of Wittko Francke. He received his doctorate in 1998 for studies on structures and functions of insect-derived natural products, which included the serendipitous discovery of a group of structurally complex ant alkaloids, the myrmicarins. During his graduate studies, he developed a deep appreciation for NMR spectroscopy as a tool in natural products chemistry and metabolomics. He continued to develop new analytical methodology for characterizing structures and functions of small molecule metabolites as a postdoc and later research associate with Jerrold Meinwald at Cornell University and Jon Clardy at Harvard Medical School. In August 2007, he joined the faculty of Cornell

University's Boyce Thompson Institute and the Cornell Department of Chemistry and Chemical Biology.

Dr. Schroeder's research aims to develop NMR spectroscopy-based approaches that complement or enhance traditional methodology by enabling detailed characterization of small molecule metabolites in complex biological samples, with regard to both chemical structure and biological function. His current work focuses on a comprehensive structural and functional annotation of the metabolome of the model organism *Caenorhabditis elegans*.

9.07 Biomolecular Recognition by Oligosaccharides and Glycopeptides: The NMR Point of View

Katalin E. Kövér, László Szilágyi, and Gyula Batta, University of Debrecen, Debrecen, Hungary

Dušan Uhrín, University of Edinburgh, Edinburgh, UK

Jesús Jiménez-Barbero, Centro de Investigaciones Biológicas, Madrid, Spain

9.07.1 Scalar and Residual Dipolar Coupling Constants in the Structure Determination of Carbohydrates by NMR

9.07.1.1 Introduction

Interactions of the oligosaccharide moieties of glycolipids and glycoproteins with various receptors are fundamental processes in many biological events, particularly those related to immunology. For such interactions to be efficient, it is vital that the three-dimensional (3D) structure (conformation) of the carbohydrate ligand matches the spatial requirements of the receptor's binding site. Hence, knowledge of the conformational behavior is indispensable for understanding these processes on a molecular level. NMR is the main source of structural information for biomolecules in solution. However, a serious impediment with conformational analysis based on NMR observables (such as chemical shifts, coupling constants, NOEs, and relaxation times) is that the measured values represent averages in the case of flexible molecules showing rapidly interconverting conformations.

Recent developments in NMR spectroscopy, along with advances in computational techniques, have produced new approaches to the interpretation of spin–spin coupling constants extracted from biomolecules. Quantum chemical studies of useful accuracy are now becoming more routine, and are increasingly being used in conjunction with experimental data to map out the expected structural patterns for oligosaccharides, as well as other biomolecules.

An understanding of how structure influences coupling constants is still of paramount importance in interpreting the NMR data. During the past 10 years, a variety of systematic attempts have been made to develop databases of coupling constants, allowing the development of empirical rules for their interpretation in terms of molecular structure. At the same time, progress in the field of quantum chemistry has allowed accurate and reliable calculations of these coupling constants using biomolecular fragments containing dozens of atoms. In more propitious instances, a combination of empirical and theoretical studies can now provide valuable information. A recent example includes interpretation of spin–spin couplings involving hydroxyl and hydroxylmethyl groups.

9.07.1.2 NMR Methodology

Carbohydrate NMR spectroscopy has developed rapidly during the last few years and has been reviewed several times.[1–9] In this review, predominantly the latest techniques and applications will be reported. The focus is on the measurement and application of coupling constants in the conformational analysis of flexible regions of carbohydrates, such as glycosidic linkages, exocyclic hydroxymethyl, and hydroxyl groups.

9.07.1.2.1 Assignment of resonances

Evidently, the first step in any NMR study of carbohydrates involves assigning proton and carbon resonances. This can be done on the basis of scalar and through space connectivities, using 2D/3D homo- and heteronuclear correlation experiments.[1,4–6,10]

Traditionally, homonuclear 2D double quantum filtered correlation spectroscopy (DQF-COSY) and total correlated spectroscopy (TOCSY) spectra are valuable in the identification of resonances of individual monosaccharide units. In the presence of small couplings, through space connectivities detected by NOESY/ROESY (nuclear Overhauser effect spectroscopy/ rotational nuclear Overhauser effect spectroscopy) experiments are also useful in completing the resonance assignment. When the ^{1}H NMR spectra of complex oligosaccharides are too crowded to fully elucidate the structure by homonuclear correlation methods, it is efficient to use 2D heteronuclear correlation methods, such as heteronuclear single quantum correlation

(HSQC) or heteronuclear multiple quantum correlation (HMQC). The combined experiments such as 2D HSQC(HMQC)-TOCSY experiments are powerful tools for the assignment of the ^{13}C and ^{1}H resonances belonging to the same sugar residue providing enhanced dispersion of TOCSY correlations in the carbon dimension. More recently, different carbon multiplicity editing methods, for example, DEPT (distortionless enhanced polarization transfer)-HMQC and E-HSQC, have been developed to reduce the complexity of proton–carbon correlation spectra and to enhance the resolution by narrowing the applied spectral window.[11]

To obtain information about the glycosidic linkage, ^{1}H–^{1}H NOESY/ROESY and/or long-range ^{1}H–^{13}C correlated spectra, heteronuclear multiple bond correlation (HMBC) or CT (constant time)-HMBC,[12] are recorded. The combined 2D HSQC(HMQC)-NOESY(ROESY) experiments could also be helpful, but have limited applications due to their low sensitivity in samples with natural abundance of ^{13}C.

Several homonuclear 3D NMR experiments, such as TOCSY–NOESY, ROESY–COSY, or TOCSY–COSY, have been employed to reduce spectral overlap, however, with limited success due to the generally poor proton chemical shift dispersion of complex oligosaccharides.[3,5,10,13,14]

The increased chemical shift dispersion of ^{13}C resonances makes the 3D heteronuclear methods more attractive; however, such experiments on natural abundance samples require very long measurement times. Despite this, several heteronuclear 3D spectra such as HSQC-TOCSY, HMQC-NOESY, and HSQC-HMBC have been reported on natural abundance samples.[15–17] With the availability of uniformly ^{13}C-enriched carbohydrates, ^{13}C-edited homonuclear ^{1}H correlation experiments, such as 3D HSQC(HMQC)-COSY(TOCSY, NOESY), are the techniques of choice that can overcome spectral overlap. Another strategy for carbohydrate NMR assignment of ^{13}C-labeled samples relies on $^{1}\mathcal{J}_{CC}$ ($\sim$40 Hz)-mediated magnetization transfer. It has been used in the form of 3D HCCH-COSY (3D experiment correlating H_a, C_a, and H_b in an ($-H_aC_a\ldots C_bH_b-$) segment), HCCH-TOCSY (3D experiment correlating H_a, C_a, and H_b in an ($-H_aC_a\ldots C_bH_b-$) segment), and their constant time variants in the study of high-molecular-weight glycoconjugates.[18] Many of these 3D homo- and heteronuclear experiments and their applications for structural characterization of carbohydrates have been recently reviewed, and therefore will not be described here in further detail.[5]

Selective and/or double-selective analogues of 2D and 3D homo- and heteronuclear experiments are particularly valuable for carbohydrates. Nonoverlapping resonances of anomeric protons (or anomeric carbons) offer a convenient starting point and are ideal for selective excitation with soft-shaped pulses. In addition, recently developed gradient-enhanced chemical-shift-selective filters enable selective excitation of overlapping resonances, which nevertheless differ in their chemical shift by a few hertz.[19] This widens the applicability of 1D NMR techniques to the studies of carbohydrates. The resulting reduced dimensionality 1D or 2D spectra show good signal-to-noise and high digital resolution, facilitating the extraction of important information. An arsenal of these selective/double-selective 1D and 2D experiments has been implemented using pulsed field gradients, and their usefulness in carbohydrate structure elucidation has been demonstrated.[19–25]

9.07.1.2.2 Measurement of proton–proton coupling constants

The simple 1D proton spectrum of small- to medium-sized carbohydrates is often quite suitable for extracting the proton–proton coupling constants. In some circumstances, however, spectrum simulations become necessary to derive the accurate values due to second-order effects.[26] Signal overlap in the spectra of more complex carbohydrate molecules may hamper the analysis of 1D spectra and therefore also the measurement of coupling constants. Acquisition of pure phase proton multiplets through the gradient-enhanced 1D TOCSY spectra via selective excitation of well-resolved (e.g., anomeric proton) resonances may allow the extraction of coupling constants.[27–29] However, in the case of severe overlap of resonances, 2D proton–proton correlation spectra, such as DQF-COSY, TOCSY and exclusive correlation spectroscopy (E.COSY), and/or 2D $\mathcal{J}$-resolved spectra,[30,31] recorded with sufficient data points to define the peaks properly, can give the values of such coupling constants. It is important to emphasize that in analyzing the peak separation of the resulting antiphase (in DQF-COSY) or in-phase (TOCSY) multiplets caution should be taken as the apparent splitting may not always correspond to the true value of the coupling constant. As a general rule, the gradient versions of NMR experiments provide enhanced sensitivity and fewer spectral artifacts, require shorter measurement time, and allow for better solvent suppression.

More recently, an intensity-based, quantitative J method has been reported, which enables the determination of all endocyclic ^{1}H homonuclear couplings in natural abundance carbohydrates of any molecular size.[32] The proposed 2D ^{13}C-COSMO-HSQC (cosine modulated heteronuclear single quantum correlation) relies on cosine modulation of ^{1}H magnetization with respect to all active homonuclear couplings as a preparation of HSQC. As a result, scalar couplings smaller than the natural line widths can be determined even in the presence of strong signal overlap.

9.07.1.2.3 Measurement of proton–carbon coupling constants

Several NMR pulse sequences have been reported for the accurate measurement of $^{1,n}J_{\mathrm{C,H}}$ values of carbohydrates, including frequency/line-separation-based techniques or quantitative J spectroscopy[33] using either carbon or proton detection. These techniques have developed rapidly over the years with the introduction of gradient spectroscopy and with the advent of improved selective excitation schemes. Such methodological advancements have been comprehensively reviewed recently.[34] In this section we will concentrate mainly on those methods that are typically applied to carbohydrates.

Heteronuclear coupling constants ($^{1,n}J_{\mathrm{C,H}}$) are most commonly measured from heteronuclear 2D experiments. The $^1J_{\mathrm{C,H}}$ couplings can be easily extracted from J-resolved spectra as well as from F_1 or F_2 proton coupled HSQC spectra. The undesired evolution of $^nJ_{\mathrm{C,H}}$ during t_1 can be eliminated with use of an appropriate bilinear rotation decoupling (BIRD) pulse, such as BIRD$^{\mathrm{d,X}}$ in J-resolved spectroscopy[35] and BIRD^r in F_1-coupled HSQC.[36] Spin-state selective excitation techniques, S^{3}E and S^{3}CT[37,38] (spin-state-selective coherence transfer), can also be used for the measurement of $^1J_{\mathrm{C,H}}$.

Long-range heteronuclear coupling constants within the same monosaccharide residue can be conveniently extracted from the E.COSY multiplets of the hetero (ω_1) half-filtered TOCSY (HETLOC (pulse sequence for determination of heteronuclear long-range couplings))[39–42] and also from the proton–carbon correlation-based HECADE[43] spectra. The sensitivity-enhanced gradient versions of these experiments allow for an additional scaling in the F_1 dimension to avoid accidental overlap of the E.COSY multiplets. The signal displacement of cross-peak doublets (resolved by $^1J_{\mathrm{CH}}$ in F_1) measured along the F_2 dimension provides the corresponding $^1J_{\mathrm{C,H}}$ or $^nJ_{\mathrm{C,H}}$ heteronuclear coupling constants. In addition to the magnitude of J, the relative sign of long-range versus one-bond couplings can also be determined from the tilt of the E.COSY pattern. Alternatively, comparison of the widths of cross-peak multiplets in the sensitivity- and gradient-enhanced coupled/decoupled HSQC-TOCSY spectra yields the same coupling information, but requires two experiments to be acquired.[44] More recently, spin-state-edited variants of HSQC-TOCSY experiment have been proposed to separate the components of E.COSY multiplets into two subspectra, allowing an easy and straightforward measurement of couplings even in the case of crowded spectra of complex oligosaccharides.[45] Typically, the J values are measured from direct multiplet analysis, or in the case of complex multiplets, the use of a fitting procedure becomes necessary. In all of the above experiments, the presence of a small proton–proton coupling weakens the signal intensity radically due to inefficiency of the TOCSY transfer. Another inherent limitation of these TOCSY-based experiments is that neither long-range coupling constants between protons and carbons separated by a quaternary carbon or heteroatom, nor those of quaternary carbons can be obtained. Phase-sensitive HMBC[46,47] as well as J-resolved HMBC and their variants[34,48] can be used as complementary experiments, particularly suited for the measurement of conformationally dependent *trans*-glycosidic $^3J_{\mathrm{CH}}$ coupling constants. Phase-sensitive HMBC experiment can be designed in such a way that the coupling constants are evaluated from the intensity of cross-peaks, while in the J-resolved HMBC spectra the cross-peaks are split by $^nJ_{\mathrm{CH}}$ multiplied with a suitable scaling factor in the F_1 dimension.

A very nice collection of spin-state-edited heteronuclear cross-polarization (HCP) experiments has been reported for the measurement of heteronuclear coupling constants of both protonated and nonprotonated carbons.[49,50]

Alternatively, 1D analogues of 2D HSQC, HSQC-TOCSY, and HETLOC experiments and a selective version of J-resolved spectroscopy using selective excitation and/or chemical shift filtering of proton or carbon resonances may be used for the measurement of coupling constants.[49,51–59] Band-selective decoupling of some of the protons during acquisition leads to reduced multiplicity, and so facilitates the multiplet analysis.[56]

More recently, a gradient-enhanced heteronuclear single quantum multiple bond correlation (HSQMBC) experiment and its variants[34,47,60–62] have been proposed for the measurement of heteronuclear coupling constants from the antiphase splitting of cross-peak multiplets. Unfortunately, proton couplings that generate additional in-phase splittings, may cause some complications during the analysis. A proper peak fitting is required for the accurate measurement of these heteronuclear coupling constants.

The availability of ^{13}C-labeled carbohydrates has allowed the use of 3D NMR experiments for the measurement of both proton–proton and proton–carbon coupling constants.[10,63,64]

9.07.1.2.4 Measurement of carbon–carbon coupling constants

^{13}C-labeled carbohydrates have also been used to measure the carbon–carbon coupling constants in carbohydrates. If carbohydrates are labeled at only a few selected positions, 1D INADEQUATE (incredible natural abundance double quantum transfer experiment)[65,66] or simple proton-decoupled 1D ^{13}C spectra can provide an array of one-bond and long-range carbon–carbon coupling constants.[67–70] Uniform ^{13}C enrichment necessitates the use of more sophisticated NMR techniques, as illustrated by the ^{1}H-detected long-range ^{13}C–^{13}C correlation[71,72] or CT ^{13}C–^{13}C COSY experiments[73] using carbohydrates. In addition the signs of long-range ^{13}C–^{13}C coupling constants can be determined by a ^{13}C–^{13}C COSY-45 experiment.[74]

Increased sensitivity of cryoprobes will likely to lead to a more frequent measurement of carbon–carbon coupling constants of carbohydrates using samples with the natural abundance of ^{13}C. The use of pulsed field gradients or sophisticated phase-cycling in combination with double-quantum filtration provide excellent suppression of the main ^{13}C signal in either ^{1}H- or ^{13}C-detected experiments. Measurement of $^{1}\mathcal{J}_{\mathrm{CC}}$ coupling constants is straightforward; problems arise when the resolution of long-range ^{13}C–^{13}C doublets is compromised by the small sizes of coupling constants or fast spin–spin relaxation. Several approaches have recently been proposed and applied to carbohydrates with the aim of overcoming these limitations.[75–78] These include the acquisition of in-phase, rather than antiphase doublets,[75] acquisition of doubly $\mathcal{J}$-modulated spectra,[76] or the combined use of in-phase and antiphase doublets[77,78] as a means of increasing the accuracy of the measured coupling constants.

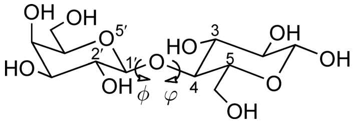
9.07.1.3 Conformational Analysis of Carbohydrates in Solution

9.07.1.3.1 Conformational analysis of glycosidic linkages in carbohydrates

Until recently, conformational analysis of the exocyclic glycosidic linkages has been based largely on the observation of interresidue ^{1}H–^{1}H nuclear Overhauser effects (NOEs). The information obtained about linkage conformation using NOE data alone may not be sufficient due to the considerable mobility of most oligosaccharides. The difficulties arise from the limited number of available NOEs and their nontrivial interpretation due to the r^{-6} dependence on ^{1}H–^{1}H internuclear distances.[3,79] A means of increasing the number of conformational constraints is to use NOEs involving hydroxyl protons. These interresidue NOEs are particularly useful in conformational analysis since they are very sensitive to conformational changes of glycosidic linkage. However, due to rapid exchange with the protons of water the use of hydroxyl protons for structural studies in aqueous solution is still an experimental challenge (see Section 9.07.1.3.3).

To overcome this problem, the measurement of different types of experimental parameters that are sensitive to linkage conformation and are more easily interpreted in flexible systems is required. The use of scalar spin–spin coupling constants for both configurational and conformational analysis of rigid and flexible molecules is well established. The reader is referred to some of the most recent reviews on this topic.[80–86]

In oligosaccharides the two *trans*-glycosidic $^{3}\mathcal{J}_{\mathrm{HCOC}}$ spin–spin coupling constants provide a direct measure of the torsion angles phi (ϕ) and psi (φ), respectively. The ϕ torsion angle is defined as O5′–C1′–O*n*–C*n*, where *n* is the linkage carbon number, while the φ angle is related to C1′–O*n*–C*n*–C(*n*–1) (**Scheme 1**).



Couplings sensitive to ϕ
$^{3}J_{\mathrm{C2',C4}}$, $^{3}J_{\mathrm{H1',C4}}$, $^{2}J_{\mathrm{C1',C4}}$
Couplings sensitive to φ
$^{3}J_{\mathrm{C1',C3}}$, $^{3}J_{\mathrm{C1',C5}}$, $^{3}J_{\mathrm{H4,C1'}}$

Scheme 1 Schematic diagram indicating the glycosidic torsion angles in carbohydrates.

Karplus-type correlation curves that relate these vicinal couplings with the glycosidic dihedral angles are available (Equation (1)).[87–91]

$$^{3}\mathcal{J}_{\mathrm{COCH}} = 5.7\cos^{2}(\Theta) - 0.6\cos(\Theta) + 0.5 \tag{1}$$

The angles determined by this approach are not unique due to the ambiguous nature of the Karplus equation;[92] as a consequence, a single $^{3}\mathcal{J}$ value may correspond to up to four different torsion angles. Even in rigid structure, this ambiguity can only be resolved with the use of molecular modeling, where a single pair of $\phi\varphi$ angles can be identified from 16 possible combinations. In flexible systems, however, the measurement of additional interglycosidic coupling constants, which are sensitive to linkage conformation, is required in order to calculate the population of individual low-energy conformers. Serianni and coworkers demonstrated that the interglycosidic C–C scalar coupling constants ($^{2}\mathcal{J}_{\mathrm{COC}}$ and $^{3}\mathcal{J}_{\mathrm{COCC}}$) provide particularly valuable information on linkage conformation.[68,90,93] The recent advances of heteronuclear NMR spectroscopy and a wider availability of ^{13}C-labeled saccharides have drawn increasing attention to these experimental parameters. The interglycosidic vicinal $^{3}\mathcal{J}_{\mathrm{COCC}}$ coupling constants show the expected Karplus dependence on the $\phi\varphi$ dihedral angles. A $^{3}\mathcal{J}_{\mathrm{COCC}}$ Karplus-type equation (Equation (2)) has been deduced based on experimental and computational studies and parameterized using a wide range of conformationally restricted carbohydrates with particular ^{13}C labels.[68,70, 89–91]

$$^{3}\mathcal{J}_{\mathrm{COCC}} = 3.49\cos^{2}(\Theta) + 0.16 \tag{2}$$

Nevertheless, it has become recently apparent[94] that the two types of *trans*-glycosidic coupling pathways ($\mathrm{C_aOCC}$ related to φ and $\mathrm{CC_aOC}$ related to ϕ where $\mathrm{C_a}$ is the anomeric carbon) are not equivalent. Therefore they cannot be treated using a single, generalized Karplus equation. The effect of an internal electronegative substituent on the $^{3}\mathcal{J}_{\mathrm{CC}}$ value in the $\mathrm{CC_aOC}$ pathway should be properly handled and taken into account in the quantitative analysis of linkage conformations. In addition, it is important to consider the effect of terminal electronegative substituents when they lie in the coupling plane and have positive contribution to the observed coupling constant. Another study has shown that the magnitude and sign of the geminal $^{2}\mathcal{J}_{\mathrm{COC}}$ coupling can also be correlated to the glycosidic linkage conformation.[91,95,96] An approximate correlation, relating $^{2}\mathcal{J}_{\mathrm{COC}}$ with the glycosidic angle ϕ has been derived using the 'projection-resultant' method by Serianni *et al.* It has been shown that $^{2}\mathcal{J}_{\mathrm{COC}}$ depends on the COC glycosidic bond angle as well, with larger angles producing more negative values.[93] Furthermore, $^{1}\mathcal{J}_{\mathrm{CH}}$ coupling constants involving the C–H pairs around the glycosidic linkage have been shown to offer information on the glycosidic dihedral angles.[97,98] Other intraresidue couplings involving anomeric carbon such as $^{1}\mathcal{J}_{\mathrm{C1'C2'}}$ and $^{2}\mathcal{J}_{\mathrm{C1'H2'}}$ are also known to be sensitive to the ϕ angle.[99,100]

The strength of this new strategy for the analysis of linkage conformation in carbohydrates relies on the redundancy of experimental scalar coupling data. The approach includes the combined use of proton–carbon and carbon–carbon coupling constants thus providing sufficient experimental data to deduce the glycosidic linkage conformation even for flexible oligosaccharides in which more than one conformation exists in solution. The calculation of statistical weights of individual conformers, based on all available scalar coupling constants is relatively simple, implying linear averages over an ensemble of conformers. Several recent studies have proved the inherent advantages of this strategy demonstrating that ambiguities of Karplus-type equations can be overcome with the use of multiple coupling data related to the same angle. Although the carbon–carbon coupling constants are mostly measured in ^{13}C-enriched samples, recent progress in biochemical and chemical synthesis of suitably labeled samples encouraged the use of this $\mathcal{J}$ coupling-based approach in conformational studies of flexible oligosaccharides. Moreover, the recent advances in NMR hardware (high field magnets, cryoprobe technology) and methodology make these couplings accessible even in natural abundance samples.

Another recent study has proposed the use of C–H dipolar cross-correlated relaxation rates to resolve the ambiguity of $^{3}\mathcal{J}_{\mathrm{HCOC}}$ coupling data in deducing linkage conformation.[101] The authors developed a new HMBC-type experiment, Γ-HMBC, in which four NMR parameters, including two interglycosidic $^{3}\mathcal{J}_{\mathrm{HCOC}}$ couplings and two $\Gamma_{\mathrm{CiHi,CiHj}}$ cross-correlated relaxation rates are measured and then related to the corresponding torsion angles ϕ and φ. The pulse sequence is shorter and therefore more sensitive than the conventional HMBC experiment, making it feasible to obtain data on natural abundance samples with conventional NMR probes.

9.07.1.3.2 Conformational analysis of hydroxymethyl groups in carbohydrates

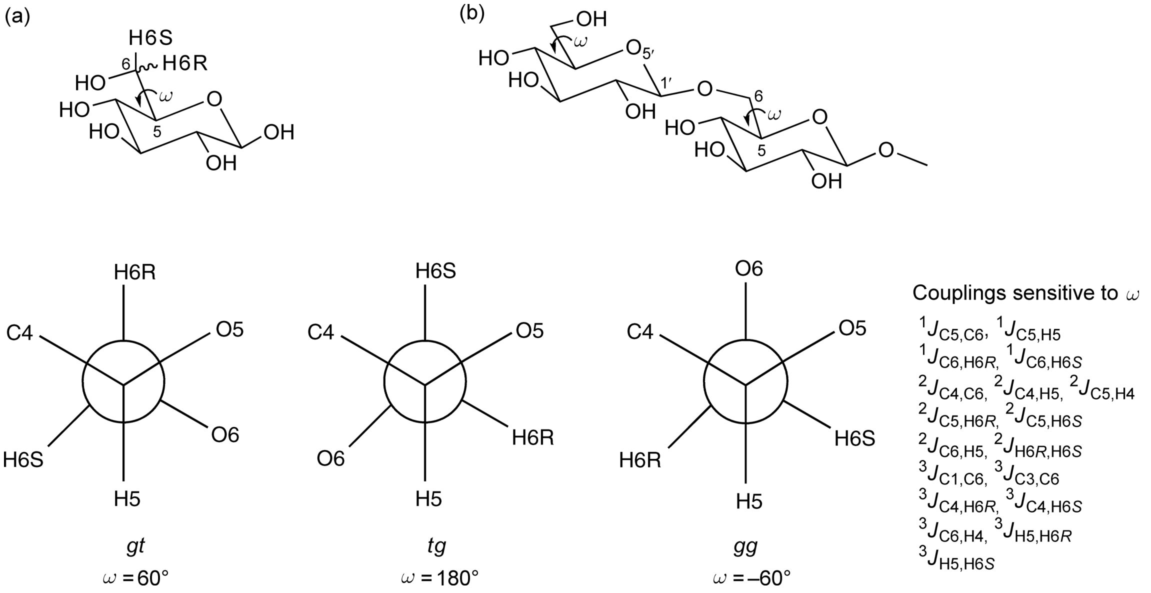
The exocyclic hydroxymethyl (CH_2OH) group in monosaccharides and 1 → 6 glycosidic linkages of oligosaccharides present another important conformational domain of carbohydrates. Rotation about the C5–C6 bond influences both the intra- and intermolecular hydrogen bonding characteristics as well as the dipole moment of the molecules. Therefore, the study of the hydroxymethyl group conformation is essential in the determination of the 3D structures of carbohydrates. Several recent experimental and theoretical studies have focused on the conformational properties of unsubstituted CH_2OH groups in order to identify the factors that influence their rotamer distribution and to derive new Karplus equations for the interpretation of the measured scalar coupling constants.

The conformation of the CH_2OH group about the exocyclic C5–C6 bond can be described by the torsional angle ω (O5–C5–C6–O6), but it is more usual to define it by means of the populations of the three staggered rotamers, gauche–gauche (*gg*), gauche–trans (*gt*), and trans–gauche (*tg*) (see **Scheme 2**). The first letter describes the torsional relationship between O6 and O5, while the second indicates that between O6 and C4.

Most studies use this three-state staggered rotamer model to analyze the coupling constants and thus to deduce the rotamer distribution, but other treatments of J couplings have also been described.[102–104]

Traditional analysis of hydroxymethyl group conformation has relied on vicinal proton–proton scalar couplings, where the rotamer populations are calculated from the measured $J_{\mathrm{H5,H6}R}$ and $J_{\mathrm{H5,H6}S}$ coupling constants.[105,106] The stereochemical assignment of the ^{1}H NMR signals of the prochiral protons at C6, H6*R*, and H6*S*, usually established on the basis of their chemical shifts and vicinal proton–proton coupling constants, is therefore of major importance for this kind of study.[107,108] Although this approach is frequently used, the inappropriate limiting values for the gauche and trans couplings derived from different types of Karplus equations lead to unrealistic negative populations of *tg* rotamers. Serianni and coworkers[109] have recently proposed new limiting values, which provide a more accurate description of the rotamer populations and, in contrast to earlier Karplus equations, generate positive populations for the *tg* rotamer. This approach was applied to several mono- and disaccharides[110–112] and the dependence of the hydroxymethyl rotamer population on nonbonded interactions, stereoelectronic effects, and/or hydrogen bonds were systematically analyzed.

Theoretical and experimental studies have demonstrated that the geminal proton–proton coupling ($^2J_{\mathrm{H6}R,\mathrm{H6}S}$) is influenced by both ω and θ (C5–C6–O6–H6) dihedral angles. The latter dependence is particularly useful in the conformational analysis of the 1 → 6 glycosidic linkage, providing complementary data to other J coupling correlations.[109]



Couplings sensitive to ω

$^1J_{\mathrm{C5,C6}}$, $^1J_{\mathrm{C5,H5}}$
$^1J_{\mathrm{C6,H6}R}$, $^1J_{\mathrm{C6,H6}S}$
$^2J_{\mathrm{C4,C6}}$, $^2J_{\mathrm{C4,H5}}$, $^2J_{\mathrm{C5,H4}}$
$^2J_{\mathrm{C5,H6}R}$, $^2J_{\mathrm{C5,H6}S}$
$^2J_{\mathrm{C6,H5}}$, $^2J_{\mathrm{H6}R,\mathrm{H6}S}$
$^3J_{\mathrm{C1,C6}}$, $^3J_{\mathrm{C3,C6}}$
$^3J_{\mathrm{C4,H6}R}$, $^3J_{\mathrm{C4,H6}S}$
$^3J_{\mathrm{C6,H4}}$, $^3J_{\mathrm{H5,H6}R}$
$^3J_{\mathrm{H5,H6}S}$

Scheme 2 Schematic representation of a hexopyranoses (a) and a 1 → 6 linked disaccharide (b) showing the ω torsion angle. Schematic diagram of the *gt, tg*, and *gg* staggered conformers around the C5–C6 bond.

Vicinal proton–carbon coupling constants ($^3\mathcal{J}_{\mathrm{H6}R,\mathrm{C4}}$ and $^3\mathcal{J}_{\mathrm{H6}S,\mathrm{C4}}$) also provide valuable information about the hydroxymethyl conformation. Theoretical calculations were applied to establish Karplus-type equations that correlate these couplings with ω torsion angle and then these were used to estimate C5–C6 rotamer populations.[88,113,114] The populations obtained from these couplings, however, were significantly different from those deduced from the vicinal proton–proton couplings and from the results of theoretical calculations. This finding may be partly due to the fact that $^3\mathcal{J}_{\mathrm{CH}}$ is less sensitive to conformational changes. A useful application of $^3\mathcal{J}_{\mathrm{CH}}$ is to assist with the stereochemical signal assignment of the diastereotopic C6 methylene protons.[104,113,114]

The introduction of ^{13}C-labeling into saccharides has offered a unique opportunity for the measurement of different proton–carbon and carbon–carbon coupling constants, extending the arsenal of experimental parameters for the conformational analysis of hydroxymethyl fragments. These one-, two-, and three-bond couplings involving C5 and C6 and their attached protons (18 couplings in total) (**Scheme 2**) are known to exhibit dependence on the C5–C6 torsion angle, thus providing complementary conformational constraints. In addition, some of them are also influenced by the C6–O6 angle (θ).[109] Based on theoretical and experimental methods, a set of new Karplus equations have been developed to correlate the magnitudes and signs of these couplings. $\mathcal{J}$ couplings that display dependencies on more than one structural parameter are remarkably useful to probe specific conformational features. For example, two-bond proton–carbon coupling constants ($^2\mathcal{J}_{\mathrm{H6}R,\mathrm{C5}}$ and $^2\mathcal{J}_{\mathrm{H6}S,\mathrm{C5}}$) were shown to be sensitive to both ω and θ. Therefore, together with $^2\mathcal{J}_{\mathrm{H6}R,\mathrm{H6}S}$, correlated conformations about both torsion angles could be studied.[115,116] This approach can be particularly valuable in evaluating linkage conformations of biologically relevant 1 → 6-linked oligosaccharides. One-bond ^{13}C–^{13}C coupling constant $^1\mathcal{J}_{\mathrm{C5,C6}}$ is also influenced by both ω and θ.[115] Likewise, second-order dependence was reported for three-bond ^{13}C–^{13}C coupling constants in aldohexopyranosyl rings.[69,115] These recent studies have unquestionably demonstrated that the collective use of these proton- ($\mathcal{J}_{\mathrm{HH}}$) and carbon-based ($\mathcal{J}_{\mathrm{HC,}}$ $\mathcal{J}_{\mathrm{CC}}$) $\mathcal{J}$ couplings leads to a more detailed understanding of the conformation and mobility of hydroxymethyl group, both free or involved in a glycosidic linkage.

9.07.1.3.3 Conformational analysis of hydroxyl groups in carbohydrates

Intramolecular hydrogen bonds in crystals of carbohydrates are well documented.[117] Hydroxyl groups positioned on the surface of saccharides serve as important recognition sites in saccharide–protein interactions and also mediate the interactions with solvent molecules. More recently, the use of hydroxyl protons in conformation, structure, and interaction studies of carbohydrates in solution by NMR has gained increasing importance. High-resolution NMR offers several different ways to deduce the conformation of hydroxyl groups and to seek evidence for the existence of intra- and/or intermolecular hydrogen bonds.[118,119]

The most important NMR parameters obtained for the hydroxyl protons are chemical shifts (δ), vicinal proton–proton coupling constants ($^3\mathcal{J}_{\mathrm{HC,OH}}$), temperature coefficients ($\Delta\delta/T$), deuterium-induced differential isotope shifts, and exchange rates (k_{ex}).[119–123] These parameters may provide information on hydrogen bond interactions and hydration as well. Moreover NOEs and chemical exchanges involving hydroxyl groups observed by NOESY and ROESY experiments also add to the number of distance restraints used in conformational analysis.

Until a few years ago, the detection of hydroxyl proton resonances was achieved in aprotic solvents, such as dimethyl-sulfoxide (DMSO) or $CDCl_3$, in order to eliminate the problem of fast exchange with the protons of the solvent. However, under these conditions, the influence of the organic solvents on the conformational equilibrium must also be considered. For example, DMSO is known to enhance the ability of hydroxyl groups to participate in intramolecular hydrogen bonding. It has been shown that H-bonds observed in DMSO do not persist significantly in water and are competed out by intermolecular H-bonds involving water molecules. However, strong and persistent H-bonds have been reported to exist for simple and complex oligosaccharides in aqueous solution.[124–126]

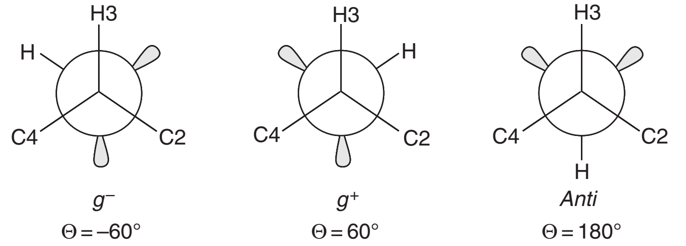
Unfortunately, under normal conditions, in aqueous solution of carbohydrates, the hydroxyl protons are in fast chemical exchange with water, which severely limits the utility of these protons as conformational probes in NMR studies. Recently, several groups have had success in detecting these hydroxyl protons by using binary mixtures of water and different organic solvents (such as water/acetone-d_6, water/methanol-d_4, and water/DMSO-d_6[127,128]). By lowering the temperature, or using dilute aqueous solutions under supercooled conditions[129] the chemical exchange effects are also reduced. However, for all NMR experiments in water careful sample preparation is required, involving the fine adjustment of pH to 6–7, the removal of traces of metal ion impurities to avoid undesired exchange with water protons, and the application of gradient-based pulse schemes for efficient suppression of water signal.

It has been observed that hydrogen bonding causes downfield a change in the chemical shifts of hydroxyl protons and that the magnitude of deshielding is dependent on the strength of such H-bonds.[119,125,130] Since the chemical shifts are influenced by several factors that are difficult to predict, the use of them alone as indicator of H-bonding is not recommended. Solvent accessibility, and thus exchange rates, could be reduced for hydrogen-bonded hydroxyl groups. However, as the exchange rate is very sensitive to pH, temperature, solvent composition, and metal ion impurities as well, this parameter also leaves ambiguity in the probing of the hydrogen bonds. A small temperature coefficient of <5 ppb K^{-1} has been commonly accepted as an indicator of the reduced interaction with the solvent due to participation in intramolecular H-bonds. Further evidence may come from nuclear Overhauser enhancement (NOESY/ROESY) and exchange measurements to determine whether or not a particular H-bond exists in solution. It has been reported that exchange peaks detected between hydroxyl protons can be an indicator of weak, transient hydrogen bonds in aqueous solution.[118,119]

Several studies have demonstrated that more direct and reliable evidence for H-bonds comes from scalar spin–spin couplings of hydroxyl protons. Until recent years, proton–proton vicinal couplings involving hydroxyl protons were mostly measured and used for deducing OH conformation in solution. This is partly due to the sensitivity problem related to the measurement of heteronuclear coupling constants and, additionally, to the lack of appropriate Karplus equation relating the heteronuclear vicinal coupling constant $^3J_{C,OH}$ to the C–O torsion angle. According to the Karplus equation derived for the vicinal proton–proton coupling constant, $^3J_{H,OH}$,[131–135] coupling of the order of approximately 5.5 Hz indicates free rotation of the hydroxyl group around the C–O bond. Deviation from this rotationally averaged value may be a sign of participation in a H-bond. Large $^3J_{H,OH}$ of approximately 8–10 Hz indicates a preferred trans/anti orientation with respect to the ring C–H, while small value of $^3J_{H,OH}$ approximately 2–3 Hz indicates a preferred syn conformation. However, since only one $^3J_{H,OH}$ value is available for each OH group, a more accurate description of OH conformation is not feasible. Calculation of rotamer populations about the C–O bond (**Scheme 3**) requires additional NMR parameters involving heteronuclei.

With the advent of cryoprobe technology and the design of new, gradient- and sensitivity-enhanced pulse sequences, the measurement of heteronuclear coupling constants of hydroxyl protons has become more feasible. As a result, the past several years have witnessed a significant effort to extend the use and interpretation of $^3J_{H,OH}$ and $^3J_{C,OH}$ coupling constants. In particular, their combined use can potentially be a useful approach for the investigation of OH conformation and calculation of the rotamer distribution around the C–O bond. Unfortunately, until recently, only few works have reported the application of $^3J_{C,OH}$ in carbohydrates.[126,136,137] A proper Karplus equation describing the relationship between $^3J_{C,OH}$ and C–O torsion angle was not yet available.

In a recent work,[138] we have attempted to parameterize the Karplus dependence of $^3J_{C,OH}$ couplings taking advantage of the redundant set of J couplings measured for the two anomeric forms of a simple monosaccharide, 4,6-O-benzylidene-1-metoxy-D-glucose (**Figure 1**).



Scheme 3 Definition of hydroxyl rotamers about the C3–O3 bond. The Θ torsion angle is defined by the corresponding H_i–C_i–O_i–H_i angle.

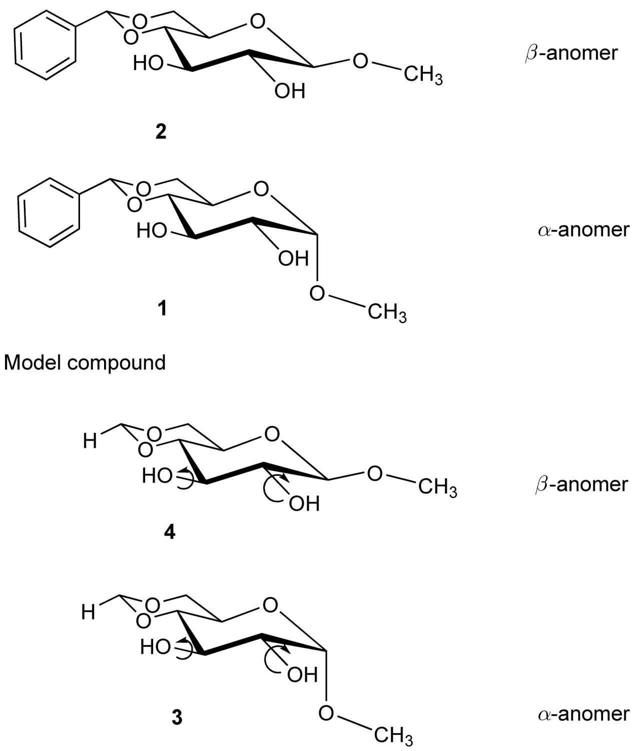


Figure 1 Schematic diagram of the α- and β-anomers of 4,6-*O*-benzylidene-1-metoxy-D-glucose (**1, 2**) and of 4,6-*O*-methylidene-1-metoxy-D-glucose model compounds (**3, 4**).

A complete set of vicinal proton–proton and proton–carbon J couplings involving four OH protons (i.e., three couplings for each OH, altogether 12 pieces of experimental data) were measured for the α- and β-anomers in $CDCl_3$ using natural abundance samples. The data were simultaneously analyzed using a global fit approach to yield the OH rotamer populations and to derive a Karplus equation for the $^3J_{C,OH}$ coupling. In this iterative procedure, the rotamer populations of OH groups (i.e., eight populations in the two molecules) were adjusted together with the three Karplus parameters describing the angular dependence of $^3J_{C,OH}$ couplings (i.e., 11 variables in all) to obtain the best fit between the experimental and calculated coupling constants. The Karplus equation deduced from this fitting procedure is as follows:

$$^3J_{C,OH} = 5.4939\cos^2(\Theta) - 0.5853\cos(\Theta) + 0.1023 \tag{3}$$

It is important to note that to derive a Karplus equation that accounts for the effects of electronegative substituents would require a significantly larger amount of experimental data.

Independently, density functional theory (DFT) calculations have been carried out on both anomers of the model compound 4,6-*O*-methylidene-1-metoxy-D-glucose (**Figure 1**) to investigate the conformational properties of the hydroxyl groups using a standard basis set (B3LYP/6-311++G(d,p)) as well as using functions for accurate description of Fermi contact contribution as implemented in Gaussian 03 rev/D.[139] The calculated populations of the lowest energy conformers and the calculated conformationally averaged coupling constants were in good agreement with the corresponding NMR data. It was found that in all low-energy conformers 'dual-type' hydrogen bonds (**Figure 2**) stabilize the overall structure.

With increasing availability of ^{13}C-enriched saccharides, additional ^{13}C–^1H and ^{13}C–^{13}C coupling constants have provided valuable information to confirm and/or extend structural conclusions based on the traditional proton-based NMR data.

In a recent study, a correlation has been established that relates the values of $^1J_{CH}$ in –CHOH– groups to the strength of H-bonds involving the OH hydrogen.[140] It was also shown that the C–O conformation in solution

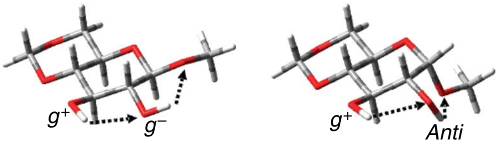


Figure 2 Three-dimensional representation of the ($g-$, $g+$) rotamer of the β- and of the (*anti*, $g+$) rotamer of the α-anomer of 4,6-*O*-methylidene-1-metoxy-D-glucose. Both low-energy conformers contain the 3 → 2 and 2 → 1 type dual hydrogen bond network stabilizing the overall system.

can be evaluated indirectly through complementary coupling constants, such as $^1\mathcal{J}_{CC}$[99] or $^2\mathcal{J}_{C,OH}$. Based on DFT, new Karplus equations have been derived for $^3\mathcal{J}_{H,OH}$ and $^3\mathcal{J}_{C,OH}$ couplings, which also account for the nature and orientation of internal and terminal electronegative substituents.[126] It can be expected that combining $\mathcal{J}$ couplings displaying direct or secondary (indirect) dependence on the C–O torsion angle may provide a more detailed picture of the conformational behavior of hydroxyl groups that have important implications on the chemical and biological reactivity of saccharides.

9.07.1.4 Conformational Analysis of Carbohydrates in Dilute Liquid Crystalline Media

Liquid crystal NMR spectroscopy is a well-established method for obtaining accurate geometries of small and rigid molecules.[141–144] Although this method has been applied during the past three decades to numerous molecules, the route from dipolar couplings to molecular structures is not an easy one. The main complication is that the solutes in liquid crystals normally exhibit complex, second-order spectra. The limitation of this method is that beyond 10 interacting spins spectra acquired in strong liquid crystals usually become too complicated to be analyzed properly.[145] The introduction of dilute liquid crystalline media in the past decade has brought the possibility of imposing very low order on the solute molecules, resulting in significant reduction of dipolar coupling constants, referred to as residual dipolar couplings (RDCs). Preserving the near first-order character of spectra in such media greatly facilitates the extraction of RDCs, which have become a rich source of structural information for large biomolecules.[146] In this chapter we focus on the use of RDCs in the structure elucidation of free carbohydrates. For information on the application of RDCs in the analysis of protein–carbohydrate complexes, see Section 9.07.2.

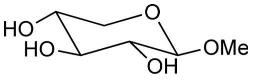
A variety of weak liquid crystalline media have been used to align carbohydrates. Examples include Pf1 phage,[147–150] filamentous bacteriophage fd,[151] $C_{12}E_5$/*n*-hexanol,[77,152–154] C_8E_5/*n*-octanol,[155] cetylpyridinium bromide/*n*-hexanol/NaCl,[73,156–158] DMPC/DHPC (dimyristoyl-phosphatidylcholine/dihexanoyl-phosphatidylcholine) bicelles,[149,159–164] and mineral liquid crystals such as aqueous suspension of V_2O_5 or a lamellar phase composed of covalent rigid planes of $H_3Sb_3P_2O_{14}$ dispersed in water.[165]

Several questions arise when making the transition from the strong to the weak alignment: (1) can RDCs provide accurate structures of small molecules comparable to those obtained in strong liquid crystals? (2) are the existing methods for the measurement of scalar coupling constants suitable for accurate measurement of RDCs? and (3) perhaps most importantly, given the opportunity to study larger systems, can we handle the inevitable flexibility intrinsic to larger molecules and still obtain meaningful information about their conformations?

In this chapter we address the above questions focusing on carbohydrates, noting that parallel efforts are underway in the studies of small, organic molecules.[166–168] We begin by observing that the utilization of RDCs in the analysis of small molecules, including oligosaccharides, is still ‘under construction’ and that there are no definite answers in particular to questions (2) and (3). Because of this, rather than providing a detailed description of one or two approaches, we present here a brief survey of the current state of the field, which will hopefully orient the reader in his/her work with the primary literature.

9.07.1.4.1 Accurate structures of small, rigid molecules from RDCs

A solution structure of a simple methylated monosaccharide, methyl β-D-xylopyranoside (**I**) in a weakly aligned medium, was recently elucidated based on 30 RDCs (15 D_{HH}, 4 $^1D_{CH}$, and 11 $^nD_{CH}$).[153] These were measured using intensity-based methods yielding absolute values of less than 6.5 Hz with an estimated precision of ±0.02 Hz. The structure of I was refined using vibrationally corrected RDCs against a model in which the distances between the directly bonded atoms were fixed at their *ab initio* values, while eight bond angles, eight dihedral angles, and five order parameters were optimized. The refined structure of I is very similar to that obtained by *ab initio* calculations, with 11 bond and dihedral angles differing by 0.8° or less and the remaining five differing by up to 3.3°. Comparison with the neutron diffraction structure showed larger and more numerous differences, which were attributed to crystal packing effects.



I

This study has shown that, providing a large number of RDCs can be obtained with sufficient accuracy, there is no principal reason why small RDCs cannot yield very accurate structures of small, rigid molecules at the level of those obtained in strong liquid crystals. A few aspects of this work deserve a comment here.

1. The largest potential source of error in the accurate determination of RDCs is the higher order effects. These are prevalent in carbohydrates. Fortunately, as the RDCs are determined as the difference between the measured splittings in the aligned and isotropic phases, the errors in the obtained RDCs are reduced substantially. Although affecting the absolute values of splitting in either sample, the strong coupling effects are to some extent reduced in the readout of the RDCs.[151] However, this is only the case when the degree of the higher order is close between the two systems. Hence, it implies that the weakest alignment possible should be used. On the other hand, the induced RDCs should be large enough to allow accurate measurement of as many RDCs as required for a particular study, which usually includes those between more distant nuclei. It is imperative that identical methods are utilized for both isotropic and aligned samples, where the compensation for the higher order effects is most successful.
2. Vibrational corrections have been routinely used to correct RDCs measured in strong liquid[169] while this practice had not been taken up thus far in weakly aligned systems. This work has shown the largest vibrational corrections (~8%) for $^1D_{\mathrm{CH}}$ coupling constants, that is, above the uncertainty of the measurement in this particular instance.
3. The dihedral angles of the neutron diffraction structure differed by less than 5° from those of the refined solution structure of **I**, yet some back-calculated RDCs based on the neutron diffraction structure deviated by up to 0.8 Hz from the experimental values. This illustrates the exquisite sensitivity of RDCs to the molecular geometry and suggests a question: What kind of structures should be used in the interpretation of RDCs? It is the experience of several groups that *ab initio* structures generally yield better agreement between the experimental and back-calculated RDCs than those generated by force fields. It is therefore likely that the use of vibrational corrections when analyzing force field generated structures is superfluous and the main reason why they have been neglected thus far. One should also be aware that there are small but genuine differences between the solid- and liquid-state structures, which can be identified by RDCs.

9.07.1.4.2 Measurement of RDCs

Presented below is a brief outline of the techniques used for the measurement of various types of RDCs in carbohydrates. These methods can be divided into two categories: frequency- and intensity-based. Frequency-based methods determine the splittings from the frequency difference between spectral lines, while the intensity-based methods require a series of spectra to be acquired, each differing in the length of a crucial delay or delays in the pulse sequence. The splitting is then obtained by fitting the intensity of spectral lines to a known function. The latter methods are usually more time consuming, but also work on unresolved multiplets. This, together with higher precision, conditioned by sufficient signal-to-noise ratios, is their main advantage.

The most readily measurable RDCs are those between the directly bonded protons and carbons. However, the internuclear orientations of one-bond CH vectors in monosaccharide rings are degenerate to a large extent (e.g., all axial CH vectors in hexopyranoses point approximately in the same direction). Some methods of interpretation of RDCs require sampling of at least five unique orientations, which must then be provided by additional types of RDCs, such as D_{HH}, $^nD_{\mathrm{CH}}$, and D_{CC}.

Owing to a fourfold degeneracy of the calculations of RDCs measured in the isotropic and aligned samples (RDC $= \pm|\mathcal{J} + D| \pm |\mathcal{J}|$), the signs of RDCs are not always obvious. There are, however, several exceptions. Most notably, the one-bond ^{1}H–^{13}C RDCs – due to the large positive value of corresponding scalar couplings. Based on the same argument, large positive $^3\mathcal{J}_{\mathrm{HH}}$ and large negative $^2\mathcal{J}_{\mathrm{HH}}$ coupling constants also usually allow unambiguous determination of the signs of corresponding RDCs. The E.COSY-based techniques can, albeit mostly for the intraring RDCs, yield the signs of D_{HH} and $^nD_{\mathrm{CH}}$ couplings. In the absence of experimental data, the signs of RDCs can be implied by considering all possibilities (four when $\mathcal{J} \neq 0$ or two when $\mathcal{J} \sim 0$) and using the fit between the experimental and theoretical RDCs as the criterion.[155] As this approach is structure dependent, caution should be exercised and large numbers of couplings need to be used so as not to bias the analysis.

9.07.1.4.2(i) ^{1}H–^{1}H residual dipolar coupling constants *Frequency-based methods.* ^{1}H–^{1}H E.COSY spectra have been used to determine the sizes and, in some cases also the signs, of ^{1}H–^{1}H RDCs. Unfortunately, this does not include the interring dipolar couplings; the corresponding cross-peaks do not show the E.COSY pattern.[170] ^{1}H–^{1}H DQF-COSY in combination with $\mathcal{J}$ doubling was used to extract the RDCs from the spectra of a highly deuterated dodecasaccharide.[157] COSY spectra, analyzed using the ACME procedure,[171] were used to measure RDCs in aligned carbohydrates.[155,161,164] E. COSY style multiplets involving ^{13}C nuclei at natural abundance provide ^{1}H–^{1}H splittings from F$_2$ displacements of the two parts of the multiplet by sampling either ^{1}H[41] or ^{13}C[172] frequencies in F$_1$. The latter approach benefits from larger dispersion of ^{13}C chemical shifts. In addition, both methods provide at the same time the one-bond ^{1}H–^{13}C splittings in F$_1$.[152,156] Alternatively, separation of α/β states into two spectra in S^3-CT-TOCSY[147,173] can be used to reduce the spectral overlap. Geminal ^{1}H–^{1}H couplings of 1 → 6 linkages were measured using COS^3 pulse sequence[148,174] or $\mathcal{J}$-modulated HMQC incorporating a BIRD pulse.[175] Signed COSY experiments[176] with TOCSY or NOESY mixing times were used by Landersjö[164] to determine the signs of some interring RDCs.

Intensity-based methods. ^{1}H–^{1}H RDCs have been obtained by analyzing the intensity ratios of the diagonal and cross-peaks in a series of 2D CT COSY spectra.[177] This method can only be applied to resolved resonances, for example, those of anomeric protons.[149] Similar limitations apply to $\mathcal{J}$-modulated 1D directed COSY,[153,178] which uses selective 180° pulses to produce a series of 1D spectra for each pair of coupled spins. This approach has recently been extended to include additional selection blocks yielding a versatile method for the measurement of coupling constants in compounds with severely overlapping proton resonances such as those found in carbohydrates.[154] The problem of overlapping resonances can also be resolved by involving ^{13}C nuclei, as demonstrated on natural abundance (^{13}C COSMO HSQC)[32] or uniformly ^{13}C isotopically enriched carbohydrates (2D-HSQC-(sel C, sel H)-CT COSY experiment).[73,158]

9.07.1.4.2(ii) One-bond ^{1}H–^{13}C residual dipolar coupling constants *Frequency-based methods.* One-bond ^{1}H–^{13}C splittings can in principle be measured from ^{1}H–^{13}C HSQC-type spectra, where the decoupling is removed either in the F$_2$[150,165,179] or F$_1$[149,156,157,159] domain. Variable[156,157,159] or constant-time sampling[149,158,161] can be employed when sampling in F$_1$.

When strong coupling is present, the line shapes of α and β spin-doublet components are different in F$_2$, which makes the determination of the splitting difficult. To minimize the errors, Almond *et al.*[150] overlaid β resonances in the spectra acquired from the aligned and isotropic samples (which are similar except for line broadening), and then measured the distance between the two α resonances. This distance is equal to $^1D_{\mathrm{CH}}$ but does not require the distance between α and β spin resonances to be measured.

The readout of line frequencies is simpler in F$_1$, but potentially less precise due to limited digital resolution. The digital resolution can be increased if long-range, either proton–carbon or proton–proton, interactions are removed by the action of a BIRD pulse.[36,175,180] If required, overlap reduction can be achieved by separating the α/β states into two spectra as demonstrated in S^3-CT-HSQC[37,147,148] or SPITZE (spin state selective zero overlap)-HSQC.[170]

Intensity-based methods. A 2D $^1\mathcal{J}_{\mathrm{CH}}$-modulated ^{13}C–^{1}H CT-HSQC[181] has been used to measure the proton–carbon RDCs in carbohydrates.[155,164,182] This method samples the CH splitting using carbon magnetization and a narrow range of evolution intervals (25–30 ms) in order to avoid the interference of long-range proton–carbon splittings. Further increase in the precision of the measured splittings was achieved by using longer evolution intervals (~170 ms).[151,153] In this method, the interference of ^{1}H–^{1}H or long-range ^{1}H–^{13}C couplings was removed by the application of BIRD pulses.[180,183] BIRD pulses work best on weakly aligned samples

($^1D_{\mathrm{CH}} < 10\,\mathrm{Hz}$), where the interference of proton–proton RDCs and the variations in the one-bond splitting is minimal. These highly accurate methods were recently used to measure $^1D_{\mathrm{CH}}$ RDCs as a function of the carbohydrate concentration. The observed changes provided evidence for the existence of calcium-mediated interactions between Lewis X-related trisaccharides.[184]

9.07.1.4.2(iii) Long-range ^{1}H–^{13}C residual dipolar coupling constants *Frequency-based methods.* $^2D_{\mathrm{CH}}$ coupling constants have been measured with an E. COSY-type experiment incorporating S^{3}CT pulse element and producing HSQC–like spectra.[149] Homonuclear correlation spectra in the form of ω_1 ^{13}C-filtered DQF-COSY can also provide both one-bond and two-bond proton–carbon splittings.[153] In addition, both methods also yield the signs of $^2D_{\mathrm{CH}}$ coupling constants.

Intensity-based methods. Long-range quantitative *J* spectra have been used to measure $^2D_{\mathrm{CH}}$ coupling constants of sucrose.[161,163] *J*-modulated constant-time HMBC experiments[153] were used to measure the long-range proton–carbon RDCs of methyl β-D-xylopyranoside. The efficiency of the latter experiments, particularly for aligned samples, is improved dramatically when the proton of interest can be selectively inverted allowing refocusing of proton–proton and proton–carbon splittings.[153]

9.07.1.4.2(iv) One-bond and long-range ^{13}C–^{13}C residual dipolar coupling constants *Frequency-based methods.* One-bond ^{13}C–^{13}C RDCs of uniformly ^{13}C isotopically enriched sucrose were measured from 1D spectra simplified through the use of selective ^{13}C decoupling.[161] Increased sensitivity of cryoprobes has allowed the measurement of ^{13}C–^{13}C RDCs at the natural abundance of ^{13}C using tens rather than hundreds of milligrams of compounds. A recent comparison of ^{1}H- and ^{13}C-detection in INADEQUATE experiments[78] showed that ^{13}C detection is a viable alternative, and likely a method of choice for aligned samples.[77] Simultaneous determination of one-bond and long-range ^{13}C–^{13}C RDCs of methyl β-D-xylopyranoside by ^{13}C-detected IPAP (in-phase antiphase)-INADEQUATE illustrated this point.[77]

Intensity-based methods. Intensity-based methods detecting ^{13}C–^{13}C pairs are practical only for ^{13}C isotopically enriched oligosaccharides. ^{13}C–^{13}C CT-COSY was used to measure one-bond ^{13}C–^{13}C RDCs in uniformly ^{13}C isotopically enriched lactose.[73]

9.07.1.4.3 Interpretation of RDCs

Over the past few years, a number of approaches have emerged for the interpretation of RDCs in terms of carbohydrate structure. However, a key feature of all these methods is that they require the alignment tensor to be determined. Order matrix analysis uses experimental RDCs, while some molecular properties, such as molecular shape or mass distribution provide the alignment tensor *a priori* without the need for the experimental RDCs.

Common to all methods, the interpretation of RDCs in terms of structure is down to describing the orientations of internuclear vectors, along which the dipolar couplings are measured. This is usually done in a special Cartesian molecular frame, referred to as the principal order (or alignment) frame (PAF).[185] In this frame of reference the order matrix (or the alignment tensor) is diagonal, characterized by three order parameters, $S_{x'x'}$, $S_{y'y'}$, $S_{z'z'}$ (or axial and rhombic components of the alignment tensor) and three Euler angles, which define the orientation of the PAF in the initial molecular frame.

From this short description it is obvious why RDCs are referred to as a long-range sensor: it is irrelevant how far away the two internuclear vectors are from each other. By establishing their orientation relative to a common molecular axis their mutual orientation can be determined.

In the next section individual approaches for the interpretation of RDCs are briefly outlined.

9.07.1.4.3(i) Order matrix analysis The order matrix describes the residual orientation of the molecule and the strength of the alignment. The subsequent diagonalization of the symmetric order matrix yields the parameters described above: three angles (orientation) and three order parameters (strength). Owing to the fact that an order matrix is traceless ($\Sigma S_{ii} = 0$)[141] a minimum of five, instead of six, unique RDCs are required to calculate it. Singular value decomposition is used to find the best least square solution for the order parameters.[185] The three order parameters can easily be related to two parameters of the alignment tensor taking into account the fact that the order matrix is traceless.

Similar to the orientation of internuclear vectors, the orientation of rigid molecular fragments can be established via the analysis of RDCs. Suppose individual monosaccharide rings are connected by rigid glycosidic linkages and that it was possible to determine the PAF for each ring using five or more unique RDCs. The rings can then be rotated into their principal order frames and translated to make glycosidic linkages. Programs such as REDCAT,[186] have been developed to do precisely this.

Building the oligosaccharide structure in this way is meaningful only if (1) the monosaccharide rings are rigid and (2) share the same order frame. In other words, RDCs interpreted, as outlined above, yield the correct structure of an oligosaccharide only if the whole molecule is rigid.[149] Even then, complications arise from the $\cos^2$ dependency of RDCs; four solutions in all quadrants of the circle are equally valid. Some can be eliminated based on steric clashes; some may be contradicted by additional experimental parameters such as NOEs or coupling constants. Alternatively, the ambiguity can be resolved by measuring RDCs in a complementary aligned medium.[146]

The approximation of rigid oligosaccharides has been used in early interpretation of RDCs in carbohydrates.[179,182] A relaxed grid search has been used to generate structures of human milk oligosaccharides[159] using a consistent valence force field. Individual low-energy structures were then tested against NOEs and RDCs. Even flexible oligosaccharides can be studied in this way by dividing them into rigid sections.[156] In this study, restrained simulated annealing was used to refine the whole molecule, but RDCs were used separately to refine the two flexible parts. Interestingly, small changes of the dihedral angles of the hexapyranose rings were observed accompanied by an improved fit between the experimental and calculated RDCs. Restrained simulated annealing was also used in the study of a trisaccharide from ganglioside Gm_3[182] yielding excellent agreement with the experimental RDCs. The raffinose and sucrose structures have been refined in X-Plore effectively using one alignment tensor for the entire molecule.[170] However, subsequent studies using RDCs found evidence for flexibility in sucrose.[163] It therefore seems possible that the use of restrained MD or simulated annealing can artificially improve the agreement between the experimental and back-calculated RDCs. This fact alone should therefore not be used as evidence for the existence of a single rigid conformer.

In a study of mannose oligosaccharides, a single structure fitted the RDC data; however, a dynamic ensemble of structures was required to predict the experimental relaxation data.[147] The possibility that a single structure could correspond to a virtual conformer should therefore never be ruled out. Its consistency with experimental data sampling motions on different timescales, that is, ^{1}H–^{1}H NOE, $^3\mathcal{J}_{\mathrm{CH}}$ and $^3\mathcal{J}_{\mathrm{CC}}$ scalar couplings, or relaxation data, should always be checked.[147,156]

A promising approach for detecting the signs of conformational averaging in carbohydrates was proposed by Tian *et al.*[149] and illustrated on the analysis of the conformational and motional properties of a trimannoside. This approach is based on the analysis of order matrices of individual monosaccharide ring. A generalized degree of order (GDO) was introduced as $\mathrm{GDO} = \sqrt{(2/3)\sum_{ij} S_{ij}^2}$, where S_{ij} are Saupe order matrix elements. For rigidly connected molecular fragments, GDO values, as well as the individual order parameters, are identical for each monosaccharide ring. In the case of the trimannoside it was found that the GDO for the two rings connected by a $1 \rightarrow 3$ linkage were similar, while the third ring linked via a $1 \rightarrow 6$ linkage had GDO reduced by 40%, indicating flexibility. The advantage of the GDO concept is that it reflects both internal and orientational averaging and it can be expressed in any frame fixed in the molecule.[146] The question that remains to be answered is to what extent can GDO values differ between individual rings for them to be considered rigid. A 1.2-fold difference was deemed sufficient by Tian *et al.*,[149] but was interpreted as a sign of flexibility by Stevensson.[160] Studies of motional averaging of RDCs (e.g., Deschamps[187]) will help to address this problem. The accuracy with which GDO can be determined will also play a role; potential errors can originate in the measurement of RDCs, as well in the monosaccharide structures used to calculate order matrices.

9.07.1.4.3(ii) Determination of the alignment tensor from the molecular shape or mass distribution This methodology does not rely on the calculation of the order matrix from experimental data using a static molecular model. The RDCs are calculated using the alignment tensor derived from a potentially dynamic molecular model and some molecular properties. The agreement between the experimental and calculated RDCs is then used as a criterion for justifying both the model and the approximations used in calculating the alignment tensor. The attraction of this approach is that it does not require five unique RDCs for each monosaccharide ring to be measured. Instead, for rigid structures, a single scaling factor is optimized to obtain a fit between the experimental and theoretical RDCs.

In an early example of the steric approach, the 1 → 6 linkage of the trimannoside, identified as flexible by Tian *et al.*[149] was studied by Almond.[147] To simulate the alignment, it was assumed that the alignment was caused by steric restriction at the phage surface, as described previously in the prediction of the alignment from structure (PALES).[188] The alignment tensor was calculated for every point of the MD trajectory and average RDCs were calculated using the *gg* and the *gt* portions of the simulation. Analysis of a 50 ns trajectory in water and comparison with the experimental data showed that the sampling of major molecular conformers was not correct, likely due to the shortcomings of the force field. A two parameter fit to the experimental RDCs found the best agreement for 55% *gg* and 45% *gt* conformers. This agreed well with the results provided by other experimental techniques[147] and is practically identical to a similar analysis of the same trimannoside performed by Prestegard and Yi.[189] The latter study calculated the alignment based on the rmsd overlay of different conformers of the entire molecule. Another example of the use of steric alignment was provided by the work of Almond *et al.*[147]

The use of an inertia tensor to calculate the alignment of molecules was initially proposed for thermotropic liquid crystals[190] and recently applied to the studies of carbohydrates in weakly ordered media. This approach assumes that the ordering coordinate frame corresponds to the principal axis system of the moment of inertia tensor.[157,191] Assuming the existence of a rigid molecule, it has been used to identify the conformer yielding the best fit between the experimental and theoretical RDCs from a series of Monte Carlo-generated decasaccharide structures.[157] The assumption of a rigid molecule was partly relaxed when a set of structures, generated by molecular dynamics runs, are used.[191] The use of the inertia tensor to calculate the alignment was conveniently formalized by the tracking alignment from the moment of inertia tensor (TRAMITE) program.[162]

As an alternative to the moment of inertia, second moment of atomic distribution, also known as the gyration tensor, was proposed to approximate the molecular alignment.[192] Both tensors share a common PAF, while the latter provides more realistic values of the order parameters, in particular for highly elongated molecular shapes.

The conformation of lactose has been investigated using TRAMITE and PALES.[158] The authors have found that a single syn-Φ, syn-Ψ conformer reproduced the experimental RDCs equally well using either model. The mixing in of the anticonformers for either the Φ or Ψ dihedral angles beyond 3% worsened the fit significantly. This contrasts with the interpretation of NOE data, which required a 10% presence of the anti-Ψ conformer. In this study, only the β-anomer of the reducing glucose was investigated.

An interesting observation was made by Freedberg *et al.*[193] who noticed a significant difference between the RDCs of the β- and α-anomers of glucose in lactose, suggesting that the two molecules have a different 3D structure in aqueous solution. By treating the molecule as a rigid entity and interpreting both the NOE and RDC data, they have concluded that the β-anomer is consistent with the syn conformation also found in the X-ray structure, while the α-anomer is not, suggesting some contribution of anti-Φ conformer.

Although residual dipolar couplings calculated using the radius of gyration tensor, performed for each frame of the MD simulation, were in trend agreement with the experimental data for a pentasaccharide and a hexasaccharide,[150] the correlation was weaker than obtained previously in the study of trimannoside.[148] As the simulations showed the existence of well-defined regions in the Φ/Ψ space, the RDCs were back calculated using one alignment tensor for the entire molecule. A much better agreement with the experimental RDCs was observed from this and the simulated data also agreed with the NOESY data. The authors have therefore concluded that there are inadequacies in prediction of the alignment tensor using the mass distribution methodology. This may be particularly the case for highly anisotropic rod-like carbohydrates with a protruding side sugar.

9.07.1.4.3(iii) Other approaches to the interpretation of RDCs in flexible systems Using order matrix analysis,[185] Freedberg[161] investigated the dynamics of the furanose ring. This involved evaluating the fit of experimentally determined RDCs to 20 possible structures of sucrose's fructofuranosyl ring, which differed only in their pucker phases. Using solely RDCs, the fit indicated that the ring pucker is localized to the NE quadrant of the pseudorotational wheel, most likely within the 20–70° range. Furthermore, the results obtained were in excellent agreement with the data provided by other methods.

The solution structure and dynamics of sucrose were examined using a combination of RDCs and molecular mechanic force fields.[163] It was found that the alignment tensors of the glucose and fructose rings were different, indicating internal dynamics. RDCs were fitted to structural models using order matrix analysis[185] and algorithms available in NMRPipe. Fitting two structures simultaneously using 35 residual dipolar couplings resulted in a substantial improvement compared to using a single rigid structure. This process is dependent on the force field used. As major disagreements between force fields were found, multiple force fields were used to interpret the NMR data.

Assuming rigid monosaccharide rings, the source of flexibility in carbohydrates is the glycosidic linkage characterized by the Φ/Ψ torsion angles. In this approximation the complete information about the conformation of a disaccharide fragment is embodied in the conformational distribution function $P(\Phi,\Psi)$. It has been proposed to construct such a function as a combination of the additive potential AP[194] and maximum entropy ME[195] methods used in the interpretation of RDCs in strong liquid crystals. The authors refer to this method as additive potential maximum entropy (APME) and have shown that it is valid in the low-order limit.[160] This method assumes that each rigid segment (monosaccharide ring) makes its own contribution toward the overall ordering of the molecule and that the conformation-dependent elements of the order matrix are described as the sum of the conformation-independent and conformation-dependent terms. Similar to the singular value decomposition,[185] at least five independent RDCs are required for each monosaccharide ring to characterize the orientational order, while additional interresidue RDCs are required for the construction of the conformational distribution function $P(\Phi,\Psi)$. This function can at the same time incorporate interresidue scalar couplings ($^3\mathcal{J}_{CH}$, $^3\mathcal{J}_{CC}$) and interresidue NOEs.

The APME has thus far been tested on a disaccharide α-L-Rha*p*-(1 → 2)-α-L-Rha*p*-OMe.[155,160] In addition to a global energy minimum at $\Psi = 0°$, the analysis uncovered the existence of a weak local minimum at $\Psi \sim 160°$ corresponding to anti-Ψ conformer. The latter was not found by either the MD or LD molecular simulations, which showed the existence of two minima distributed at around $\pm 40°$ around $\Psi = 0°$ angle. The main APME minimum is a broader one encompassing both MD minima. It is difficult to ascertain whether the differences in the experimental GDO parameters (1.2-fold) for the two rings are due to the transitions between the minima identified by the MD or if these are due to the admixture of the anti-Ψ conformer suggested by the APME analysis.

In order to investigate the motion along the glycosidic linkage of a disaccharide, Yi *et al.*[152] modified a 4-*O*-β-D-galactopyranosyl-α-D-mannopyranoside by attaching *n*-butyl chain to the reducing end of the molecule. This resulted in a significant increase of RDCs compared to the native disaccharide in the $C_{12}E_5$/hexanol/water aligning medium indicating that the modified disaccharide has been immobilized and reoriented through some specific association with the medium. The reduction in GDO by approximately 38% for Gal revealed the existence of a significant internal motion between rigid Gal and Man residues. By transiently anchoring one end of the disaccharide to a aligned bilayer medium using a short alkyl chain, the reference frame was forced to coincide with the frame determined for the reducing end sugar. This eliminated the need for additional data and allowed rigorous interpretation of the differences in sizes of order tensor elements of individual rings.

Two conformers, S1 (17%) and S2 (83%), have been identified by the MD simulation for this disaccharide. To improve the accuracy of the modeling, the MD geometries of S1 and S2 were first submitted to a full geometry optimization using Gaussian 98, which altered the torsion angles slightly compared to the MD and presumably also the ring geometry. Using REDCAT, which can handle multiple state conformational averaging, and the immobilized rigid Man ring as the reference segment, the authors found that the experimental RDCs were satisfied for $15 \pm 10\%$ of S1 and $85 \pm 10\%$ of S2.

9.07.1.4.4 Conclusions

Significant advances have been made in incorporating RDCs into the conformational studies of carbohydrates since the appearance of the first reports at the end of the last decade. It quickly became obvious that the interpretation of RDCs in carbohydrates is not straightforward and model dependent. Inherent flexibility of carbohydrates poses the greatest challenge for the interpretation of RDCs in terms of carbohydrate conformation.

The limited dispersion of internuclear orientations of CH vectors between directly bonded carbon and proton atoms in monosaccharide rings means that additional types of RDCs such as D_{HH}, $^nD_{CH}$, and D_{CC} are required for proper analysis of RDCs. The demand on the number and accuracy of measured RDCs increases for flexible carbohydrates. Attention must be paid to higher order effects caused by the narrow range of ^{1}H chemical shifts of carbohydrates.

The methods for the measurement of RDCs surveyed in this review are those currently used in the field. It therefore is possible that some already published high-resolution NMR methods provide more efficient alternatives. Further development in this area is expected.

RDCs are very sensitive to small changes in molecular geometry. Force field-generated structures therefore introduce certain fuzziness into the interpretation of RDCs. It is expected that the use *ab initio* structures in the interpretation of RDCs will increase. This will generate the need for the incorporation of vibrational corrections for $^1D_{\mathrm{CH}}$ RDCs.

The choice of molecular frames used in the interpretation of RDCs is very important. Currently, order matrix analysis of individual rings or calculations of the alignment tensor from the molecular shape or mass distribution are used. Forcing the reference frame to coincide with the frame determined for a part of the molecule is a promising strategy. More studies are needed in order to evaluate the merits of each individual approach.

When describing conformation of flexible molecules using RDCs, models are needed that take into account both the overall molecular tumbling and internal motions. The simplest approach assumes that the molecule can be described by a single average conformation allowing only for small amplitude motions. This situation can be adequately described by using a single order matrix. However, caution must be exercised, as the existence of a single rigid structure that agrees with the experimental RDCs can hide the presence of conformational averaging, and in fact this structure can turn out to be a virtual.

It is vital that other experimental parameters such as scalar coupling constants, NOEs, relaxation parameters, and potentially other NMR parameters, currently not used in the conformational analysis of carbohydrates, are cross-checked against the RDC-derived structures. Alternatively, these parameters should be incorporated in the process of structure generation as exemplified by the APME approach.

Conformational equilibria can, in principle, be affected by the interaction with the medium. For a pentasaccharide, strong discrepancies between the NOE- and RDC-based structures were attributed to the interactions between the pentasaccharide and the mesogens, shifting the conformational equilibrium.[165] Pure steric alignment is likely to be the safest way of aligning flexible molecules. On the other hand, the electrostatic interactions, or the alignment caused by a transient insertion of a part of the molecule into the oriented phase, should be treated with some caution. It is possible that under these circumstances one could strongly orient a minor member of a preexisting distribution of conformers and the observed RDCs would be heavily weighted by the properties of this conformer.

A combined use of NMR parameters reflecting motions on different timescales will undoubtedly lead to a more accurate description of the conformational space occupied by flexible carbohydrates. The long-range orientation information provided by RDCs will play an increasingly important role.

9.07.2 Conformation of Oligosaccharides in the Free and Bound States

9.07.2.1 Introduction

In recent years, it has been shown that the interactions between carbohydrates and proteins mediate a broad range of biological activities, starting from fertilization and extending to pathological processes such as tumor metastasis.[196] The implications in immunology processes has also been demonstrated.[197] In all these processes, the 3D structures of both molecular entities are of paramount importance.[198,199]

It is obvious that a detailed knowledge of the structure of sugar entities, both free and bound to proteins, is indeed relevant from both basic and applied scientific viewpoints. This information may be extracted by different means, including NMR and different reviews have addressed this topic.[200] X-ray crystallography has also been widely employed for characterizing free and complexed carbohydrate-binding proteins (for instance, Banerji *et al.*[201]). Accordingly, examples of the application of X-ray to the study of these compounds are of prime interest.[202,203] However, carbohydrates are often rather difficult to crystallize, probably because of their inherent flexibility. Furthermore, X-ray basically provides only indirect information on the dynamics of the biomolecules and, moreover, for flexible structures, only one conformation may be analyzed.

This section will focus on recent advances on the application of NMR methods, especially those based on relaxation (mainly NOE) methods to deduce the conformational behavior of saccharides in their free and receptor-bound states. It is not pretended to be exhaustive, and just provides a few key references for each protocol and methodology for further evaluation.

9.07.2.2 The Conformation of Oligosaccharides in Solution

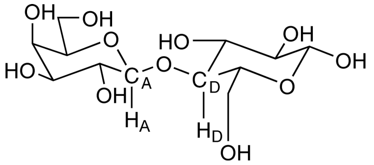
NMR has been widely applied in this field, since it provides both conformational and dynamic information. Because of the particular features of sugars, it is recognized that relaxation NMR parameters[204] should be complemented with computational methods, as molecular mechanics/dynamics calculations,[205] to define the structural and conformational features of the carbohydrate unambiguously. This task is commonly achieved by calculating potential energy surfaces for the glycosidic linkages, using a force field,[206] or *ab initio* methods.[207] It should be kept in mind, when comparing such calculations to experimental data, especially in water solution, that these methods provide just a first estimate of the conformational regions that are energetically accessible, and the possible presence of different conformational families.[208] Protocols based on single-point conformers, restrained[209] or unrestrained molecular dynamics,[210] may also be employed satisfactorily. With molecular mechanics calculations, care needs to be taken when considering the relative energy values provided by the force field. Nevertheless, the calculated geometries are usually very good approximations to those existing in solution[211] and in the solid state.[212]

Obviously, scalar coupling constants gather key conformational information that may be used to access to the distribution of oligosaccharide (or glycomimetics) conformers in solution.[213] Also, residual dipolar couplings can also be used to this end. These methods have been reviewed in Section 9.07.1. A recent example of the combined application of RDC- and NOE-based experiments to deduce the solution conformation of a tetrasaccharide has been reported.[214] It was demonstrated that the inclusion of RDCs permitted to further refine the actual geometries. In a parallel manner, the detailed analysis of the obtained conformer distributions using either NOEs or RDCs for lactose permitted to detect minor, but significant, differences between both methods. The different time-averaging sensitivity of RDCs and NOEs to motional processes may be at the heart of the distinct results.

9.07.2.2.1 The use of NOEs for conformational analysis of oligosaccharide molecules

The relationship between NOEs and proton–proton distances is well established and can be worked out, at least semiquantitatively and also quantitatively, when a full matrix relaxation analysis is considered. The detailed study of the conformation and dynamics of a tetrasaccharide related to the LeX antigen provides a good example of this approach.[215] NOE intensities are sensitive to the respective conformer populations, and that therefore, an indication of the population distribution when these molecules are free in solution and even in the protein-bound state may be obtained by focusing on key interresidual NOEs.[216]

For carbohydrate molecules, the key distances with conformational information are those between proton pairs on either ring.



The distance between hydrogen atoms A and D depends on the torsion angles around the glycosidic linkages Φ (H_A–C_A–O–C_D) and Ψ (H_D–C_D–O–C_A).

Very often, only one or two significant proton–proton distances can be measured.[217] From this viewpoint, when adopting the typical restrained-based approach for generating conformations, as usually adopted for other biomolecules, the possibility of generating virtual conformations[218] is very high and special care should be taken. Restrained simulations include an additional term in the force field to penalize the deviation from the experimentally deduced distance. Depending on the force constant employed to weight the experimental constrain, deviations from ideal geometries could occur, with concomitant distortions of the sugar rings.

In very special cases, for branched sugars, there are a large number of interresidual proton–proton contacts.[219] The careful analysis of these contacts may be used to deduce the existence (or not) of single conformers with molecular motion around well-defined geometries for the different glycosidic torsions of the molecule. Otherwise, when a significant number of restraints do appear, a time-averaged-restrained molecular dynamics protocol may be adopted, using a memory function, in the form of an exponential decay constant.[220]

Simulation lengths of approximately 1 order of magnitude larger than the exponential decay constant should be used to generate reliable estimates of average properties. In this approach, conformational distributions that are able to simultaneously cope with all the deduced proton–proton distances are obtained. NOE-derived distances are included as time-averaged distance constraints and scalar coupling constants (if any) as time-averaged *J* coupling restraints, related to torsion angles (see Section 9.07.1). In principle, $\langle r^{-3}\rangle^{-1/3}$ or $\langle r^{-6}\rangle^{-1/6}$ averages can be used for the distances, while linear averages are used for the coupling constants. The key point is to employ reasonable force constants for both the NOE- and *J*-based terms of the force field to avoid the molecule that can get trapped in high-energy, physically improbable, incorrect minima.[221]

As additional mean to enlarge the number of interresidual NOEs, those involving hydroxyl protons can also be employed (see Section 9.07.1). Technically, the observation of hydroxyl protons is a difficult task, although a variety of methods have been proposed (as reviewed by Siebert *et al.*[222]). Also, regarding the problem of distinguishing pure NOEs from chemical exchange correlations, the best method is to combine NOESY/ROESY data. In ROESY, the interactions due to both processes have a different sign, while in NOESY experiments, at low temperature, both processes give rise to cross-peaks with the same sign as the diagonal peaks. Also, for certain mixing periods, cross-peaks can appear that are mediated by water molecules.

Under all circumstances, and due to the $\langle r^{-6}\rangle$ or $\langle r^{-3}\rangle$ dependence of the NOE, minor populations of conformers can be detected, provided that they show exclusive interresidue proton–proton distances.[223]

In any case, the existence of molecular motion around the glycosidic linkages of oligosaccharides has been firmly established.[224] Even in some special cases, simultaneous negative and positive NOE cross-peaks may be obtained for the same oligosaccharide, thus indicating motion at different timescales in different regions of the molecule. Aminoglycoside antibiotics[225] and a LeX-related saccharide[226] provide two examples of this feature. Since NMR parameters are essentially time averaged, the information deduced from NMR experiments generally corresponds to the time-averaged conformation in solution.[227] Regarding the relaxation timescale, ratios from transverse and longitudinal cross-relaxation rates obtained through off-resonance ROESY experiments[228] and/or through the comparison of data taken from individual NOESY, ROESY, or tilted rotational nuclear Overhauser effect spectroscopy (T-ROESY) experiments[229] may be used to extract local correlation times for different pairs of protons in the oligosaccharide. These ratios are independent of interproton distances and may allow the estimation of specific correlation times. From these correlation times, proton–proton intra- and interresidue distances may be quantitatively extracted. This is probably the best method of choice for quantitative analysis of conformation and molecular motion, while for a semiquantitative analysis, intraresidue signals may also be taken as internal distance references, as a first approximation, to deduce the unknown distances for the target interresidue proton pairs.

^{13}C-NMR relaxation parameters may also be employed to access the rates of overall and internal motions for saccharide molecules.[230] Longitudinal and transversal heteronuclear relaxation times as well as heteronuclear NOEs depend on the molecular motion of the molecule, including overall and internal motions. Thus, careful analysis of these parameters can be employed to demonstrate the presence of conformational heterogeneity and/or dynamics, as well as the restriction to motion and the timescale of the existing fluctuations. Different examples of application of this methodology to trisaccharides,[231,232] tetrasaccharides,[233] pentasaccharides,[234] and polysaccharides[235,236] have been described.

With regard to the monosaccharide moieties, the average shape of the pyranose rings may be deduced from the vicinal proton–proton coupling values, also assisted by NOE values and, if possible, other NMR parameters (i.e., RDC, see Section 9.07.1). A very special case within carbohydrate molecules is the iduronic acid ring within glycosaminoglycans.[237] This ring is usually exchanging between chair and skew boat conformations.[238] [43]NMR spectroscopy is particularly appropriate for analyzing the conformational equilibrium of this flexible ring,[239] as the interconversion between conformers leads to changes in both the dihedral angles between vicinal protons and the intraring H–H distances, which can be monitored by the measurement of scalar spin–spin coupling constants, and NOEs, respectively.[240] Proton–proton scalar coupling constants are very sensitive to changes in the dihedral angles and the empirical relationship between both parameters is well known. In addition, NOESY experiments can clearly reveal the presence of the $^{2}S_{O}$ conformer as only in this conformation, protons at positions 2 and 5 of the ring are close enough as to originate an observable NOE.[241] As in all cases, it should always be remembered that the experimental values correspond to the time-averaged conformation in solution. Obviously, the best approach should combine both *J* and NOE data and the postulated conformation or conformational equilibrium in solution should correlate with all of the NMR data in an unambiguous manner.[242]

The conformation of oligosaccharides under particular experimental conditions has also been elucidated. Thus, several studies have dealt with glycolipids or other amphiphilic sugar moeties embedded into membrane-like environments, such as micelles or lipid aggregates,[243–247] deriving structural, conformational, and dynamic information[248] by using different NMR methods, including relaxation,[249] and/or dipolar coupling measurements (see Section 9.07.1).

The importance of cation binding to modulate the conformational behavior and molecular recognition features of different oligosaccharides, particularly of the aminoglycoside antibiotic family, has also been addressed,[250] together with the concomitant implications for the geometric and thermodynamic features of the interaction processes with nucleic acid and protein receptors.[251]

Also, in the last decade, and following pioneer works in the 1980s,[252] substantial advances on the elucidation of many fine details of the conformational and dynamic features of the conformation of natural[253] and designed nonnatural *O*-glycopeptides[254] have also been elucidated by using NMR and modeling procedures.[255] Conformational differences between peptides glycosylated at either Ser or Thr moieties have been detected.[256] Initial ideas on the principles of mucin architecture due to glycol clustering[257] or to the existence of MUC1 tandem repeats[258] has been shown. Also, within this context, the essential structural motifs for glycoproteins to show antifreeze activity have been deduced.[259] In a parallel manner, the principles underlying the conformational behavior of N-linked glycopeptides in solution has also been studied,[260] emphasizing the observed differences on peptide conformation upon glycosylation,[261] as well as postulating on the molecular basis for the observed glycosylation-induced conformational switchings.[262] Finally, the study of the interactions of some glycopeptides with membranes have shown that glycosyl enkephalin analogues adopt turn conformations in amphipathic media.[263]

9.07.2.3 The Bound State

In some cases, partial[264] or complete information[265] on the 3D structure of the protein–sugar supramolecule has been derived, using modern NMR methods[266] similar to those employed for deriving protein structure in solution, adapted for complexes with carbohydrates. Changes in the dynamic behavior of the protein backbone before and after sugar binding have also been explored using relaxation measurements.[267] The requirements are the standard ones for protein structure determination,[268] with the use of isotope-labeled receptors. Recent examples include, in different occasions, complexes of glycosaminoglycans[269] or sulfated analogues[270] bound to different receptors,[271,272] or neutral sugars to other lectins.[273,274] Regarding protein-bound conformations of carbohydrate molecules, although valuable information may be gained by using X-ray crystallography, transferred (TRNOE) experiments can be used for solution studies, provided that the exchange rate between the bound and the free state is fast.[275] In complexes of large molecules, cross-relaxation rates of the bound compound are opposite in sign to those of the free ligand and produce negative NOEs.[276] Following this methodology, as pioneered by Prestegard[277,278] for studying carbohydrate–protein interactions, many cases have been described.[279] Notably, the conditions required to monitor TR-NOEs appear to be satisfied frequently by sugar receptors.[280] The reason for this favorable situation probably rests in various facts: these interactions are not extremely strong, there is fast exchange between the free and the bound states of the ligand, and the perturbations of the conformational equilibrium of a given oligosaccharide upon binding to a protein are accessible to observation by transferred nuclear Overhauser enhancement (TRNOE).[281] Different mixing times and protein/ligand molar ratios should be systematically used in order to gain quantitative conclusions. Comparison with TR-ROESY[282] and/or QUIET-trNOESY[280] experiments should also be performed to detect spin-diffusion effects.

Depending on the architecture of the protein-binding site, the major conformation of the oligosaccharide existing in solution might be recognized by the receptor, as reported in a variety of cases.[283,284] In other cases, a conformational selection process takes place with exclusive recognition of one of the conformational distribution,[285] which can be drastically different from the major one in water solution.[286] In other cases, the protein can recognize different geometries of the same ligand, provided that no strong ligand–receptor contacts take place to properly define one unique selected conformer.[287]

From the protein perspective, there are also cases in which the protein skeleton is perfectly preorganized to accommodate the carbohydrate ligand, as reported for the hevein family.[288,289] However, cases in which a major reorganization of the protein takes place upon sugar binding have also been reported, as for the CD44 hyaluronan-binding domain[290] or the mannan-specific family 35 carbohydrate-binding module.[291]

The conformation of oligosaccharides bound to other molecular entities, for instance, nucleic acids has also been derived. A paradigmatic case is that of neamine bound to tRNA(Phe), as deduced by trNOE measurements,[292] or the recognition of aminoglycoside analogues by TAR-RNA.[293]

Recently, saturation transfer difference (STD) NMR methods have also been employed to deduce the bound conformation of different ligands and inhibitors into lectins[294,295] and enzymes[296] binding sites. Also, the interaction of different saccharides with large particles, such as viruses[297–299] or living cells has also been monitored by STD methods.[300] Some STD-related NMR-editing methods to discriminate the NMR signals of the ligand and the (isotope-labeled) receptor have also been proposed to efficiently monitor the existence of recognition processes.[301]

When the off-rate of the dissociation process of the ligand from the bound to the free state is slow, trNOE or STD methods are no longer applicable and alternative NMR methods have to be applied. For instance, the bound conformation of a heparin-derived hexasaccharide to fibroblast growth factor 1 has been derived using half-filtered NOESY experiments, by conducting the experiments on a sample containing ^{13}C-labeled protein.[302] In this manner, the protein protons, which are bound to ^{13}C atoms are removed from the NOESY spectrum, which now only contains the saccharide cross-peaks. For different heparin fragments, the trNOE method is still valid to get conformational information in the bound state, using antithrombin III or FGF as receptors.[303,304]

Recently, the possibility of obtaining dihedral angle information from a ligand in the bound state by exchange-transferred cross-correlation spectroscopy has been reported. This method has also been employed in the carbohydrate field with partial success.[305] More examples are still required to further validate this approach to get bound bioactive conformations.

Independently of the use of NOEs, the RDC measured after molecular alignment of the target molecule in a strong magnetic field induced by the use of, for instance, a His-tagged protein with a nickel-chelate-carrying lipid inserted into the lipid bilayer-like[306] or a dilute liquid crystalline medium has been applied to deduce the orientation of oligosaccharides, both free (see Section 9.07.1) in solution, partially modified with ^{13}C-labeled acetyl groups[307] and within protein-binding sites.[308] At this point, it should be mentioned that RDCs are the linear weighted average of those for the free and bound state and, therefore, unless a noticeable percentage of ligand is bound, the changes in RDCs can be difficult to manage to get the bound state conformation. Within this idea[309] recently, the employment of systems containing paramagnetic probes[310] or paramagnetic ion-binding tags, coupled or not[311] to membrane-like environments has also been elegantly employed in a successful manner to deduce binding features of sugar–receptor interactions.[312,313]

9.07.2.4 Conclusions

The conformational study of oligosaccharide molecules in solution is nowadays still a complex problem, but there are a good number of different reported protocols and examples that can be used as models for tackling the problem. The access to high magnetic fields (900 MHz or more in the near future) may allow obtaining better resolved spectra that permit accessing to conformational information in a more straightforward manner.[314] The selection of the method to study the bioactive conformation of the saccharides in the receptor site depends on the kinetics of the dissociation process. Thus, for fast processes, trNOE methods can be successfully employed, while for tight binding complexes more sophisticated approaches are a must to access the fine details of the geometrical features.

9.07.3 Bacterial Cell Wall Peptidoglycans and Fragments: Structural Studies and Functions

9.07.3.1 Introduction

The main task of the immune system of higher organisms is to detect the presence of and combat invading microorganisms. The innate immune system, which is highly conserved, is the only defense in invertebrates and plants. In vertebrates that are endowed with an adaptive mechanism as well the innate system constitutes the first line of immune defense. The innate immune system recognizes pathogens using pattern recognition receptors (PRRs) based on unique molecular signatures, called pathogen-associated molecular patterns (PAMPs), which are absent from the host cells. One of the most important PAMPs is polymeric peptidoglycan (PGN), a supporting constituent present in the cell walls of all bacteria. PGN is, therefore, a main target for the innate immune response. Among the various PRRs, on the surface of (such as CD14[315]) or inside (such as NODs[316]), the immune cells that are able to detect and recognize PGN peptidoglycan recognition proteins (PGRP) constitute an important family. In addition to PRGPs there are other soluble PGN recognition molecules like soluble CD14,[317] C-type lectins (e.g., the mannose-binding lectin[318]), mouse RegIIIγ and human HIP/PAP,[319] and lysozyme, a muramidase, whose bactericidal activity has long been known (for an early history, see Phillips[320]).

PGRPs and some PGN-hydrolyzing enzymes deserve special attention because only their interactions with PGNs have been characterized, to date, at the molecular level. Virtually, all of our present knowledge on PGRPs refer to proteins of insect or mammalian origin. Although their functions and structures are similar, the modes of action are different for these two classes of immune proteins. In insects immune response is usually induced, after the initial binding event, by a complex cascade of intracellular signaling to generate antimicrobial agents, usually small peptides (for an update, see Sahl[321]). Mammalian PGRPs have, in addition, a direct bactericidal activity on interaction with PGN and some of them possess catalytic properties as well.[322–324] The structures of PGNs and the mechanisms of the host's immune response have been described in greater detail in several recent reviews;[322,323,325–327] therefore, only a brief background will be given below.

9.07.3.2 Peptidoglycan Structure

The outer cell wall of Gram-positive bacteria consists of a thick layer of PGN, accounting for approximately 50% of the cell wall mass,[328] intertwined with lipoteichoic acid (LTA), and overlaying the cytoplasmic lipid membrane. The cell wall of Gram-negative bacteria is characterized by a thin PGN layer between two lipid membranes, the outer one harboring the highly immunogenic lipopolysaccharide (LPS) (see **Figure 3**). The main function of PGN is to maintain the integrity of the bacterial cell against external (e.g., antimicrobial agents) or internal (osmosis) challenges. For this reason it is a prime target to antimicrobial chemotherapy. In addition, it participates in 'classical' cell wall functions, such as recognition and regulatory activities, as well.

The backbone of PGN is a linear glycan polymer made up of alternating *N*-acetyl-D-glucosamine (GlcNAc) and *N*-acetylmuramic acid (MurNAc: 3-*O*-(D-2-carboxy)ethyl-GlcNAc) units linked together with β-(1–4)-glycosidic bonds. Short peptides consisting of four or five alternating D- and L-amino acids are attached to MurNAc via the 3-*O*-carboxyethyl group. The glycan backbone is highly conserved in all bacteria (small functional group modifications notwithstanding[329]), the amino acid at position 3 of the peptide chain is, however, variable: being L-lysine in most Gram-positive bacteria and *meso*-diaminopimelic acid (*m*Dap) in most Gram-negative ones (for exception, however, see, e.g., Lee and Hollingsworth[330]). L-lysine is replaced by L-ornithine in *Lactobacillus fermentum.*[331] Cross-links between the pending peptide chains are established either by direct amide bond between *m*Dap in one strand and D-Ala of the other (Gram-negative) or via an oligoglycine interpeptide bridge connecting L-Lys to D-Ala on adjacent strands (Gram-positive) (see **Scheme 4**). Other interpeptide connections may occur, however, depending on the bacterium strain.[332] It is

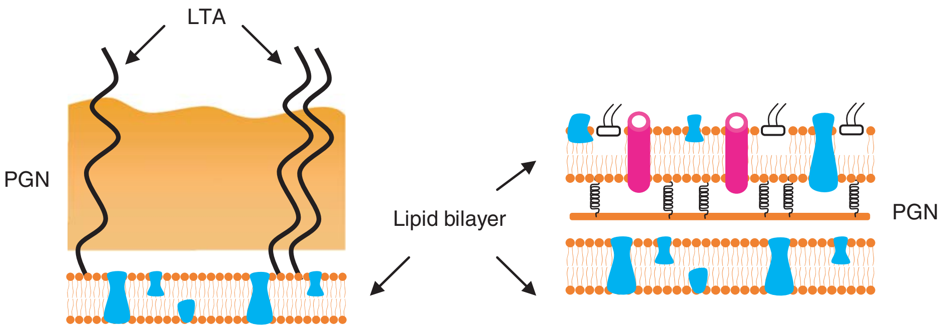
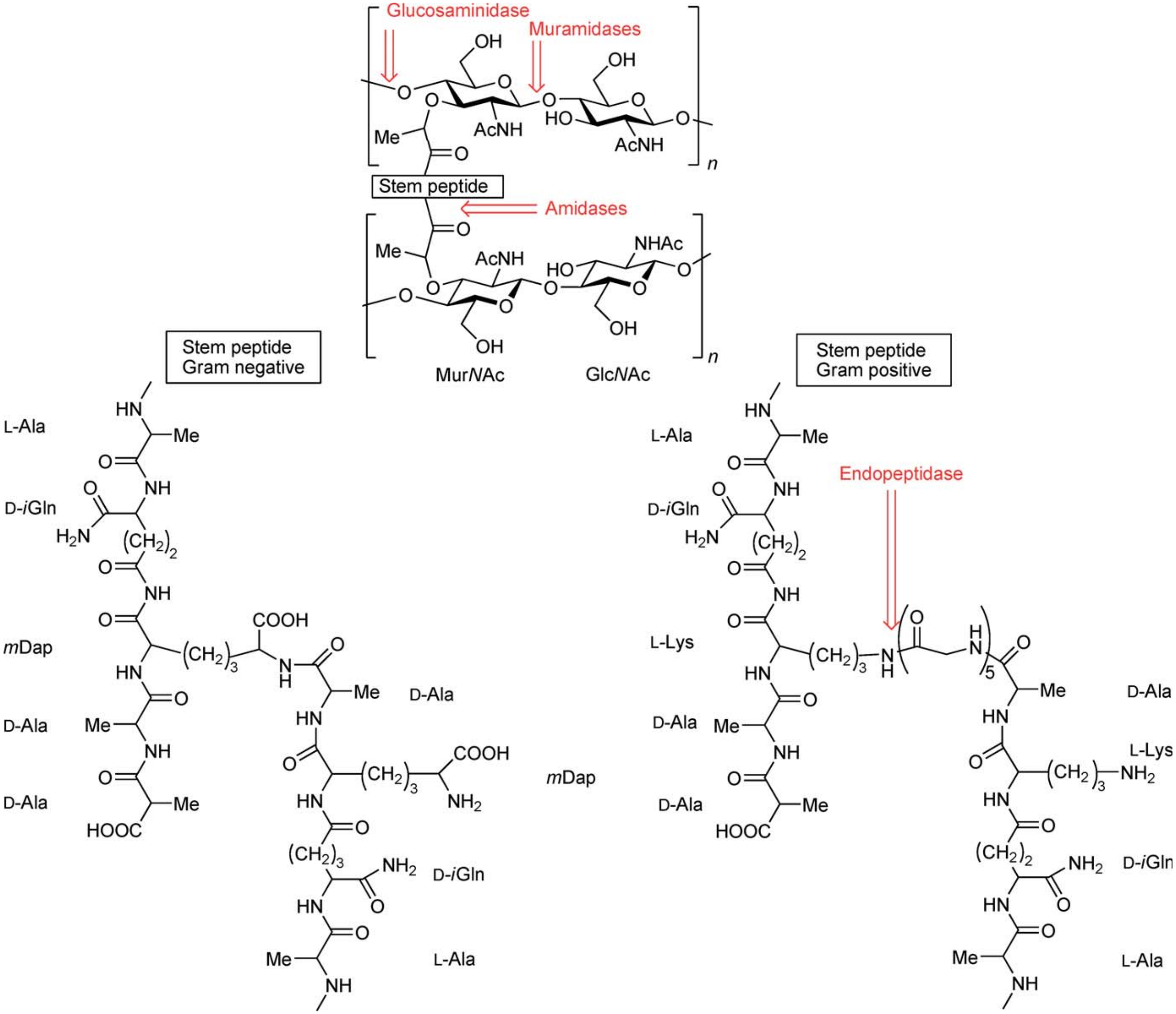


Figure 3 Schematic cross sections of the cell wall of Gram-positive (left) and Gram-negative (right) microorganisms. The light blue and pink features embedded into the lipid bilayers represent integral membrane proteins and lipopolysaccharide (LPS), respectively.



Scheme 4 Chemical structures of the cell wall peptidoglycans and points of attack by different lytic enzymes.

of note that the peptide moieties display several 'unnatural' features such as the nonproteinogenic amino acid *m*Dap or L-Orn, others with D-configurations and, the unusual peptide bond engaging the γ-carboxyl of Gln (therefore it is often labeled as *i*Gln) rather than the α-one as in proteins. Evidently, these nonconventional features are part of the weaponry the microorganisms use to overcome host defense based, for instance, on proteolytic enzymes. Yet, several enzymes are capable to attack and cleave PGN at different sites. Historically, hen egg-white lysozyme, or muramidase, was the first enzyme whose bacteriolytic property was attributed to its hydrolytic activity to cleave the β-glycosidic bond between MurNAc and GlcNAc (for structure and early history, see Phillips[320]). Other enzymes involved in the cleavage of PGN are indicated in **Scheme 4**. For a recent review on PGN hydrolases, see Vollmer.[329] Structural aspects of the interactions of proteins involved in the recognition or cleavage of PGNs will be discussed in Section 9.07.3.5.

The orientation of the PGN chains in the cell wall is the subject of some controversy. In the so-called scaffold arrangement the glycan chains are arranged perpendicular to the cytoplasmic membrane[333] whereas in the horizontal model the glycans and cross-linked peptides are oriented parallel to it.[334] The former would result in a honeycomb-like structure with pores large enough to accommodate proteins that interact with PGN. Recent molecular modeling calculations based on the NMR solution structure of a dimer of the repeating unit of Lys-type PGN (cf. Section 9.07.3.3.2)[335] seem to support the former model. Different models for the architecture of PGN are discussed in a recent review by Vollmer.[336] Solid-state NMR is a powerful emerging technique for structural studies of giant molecules like the PGNs. Very recently, a through-bond ^{13}C-correlation (2D) spectrum of surprisingly good quality could be obtained of a whole, uniformly ^{13}C-labeled, PGN sacculus (estimated molecular weight (MW) 3×10^6 kDa). In addition to chemical shift assignments, various NMR experiments allowed to establish that the glycan strands are more rigid than the peptide stems and to study PGN–protein interactions.[337]

9.07.3.3 Peptidoglycan Fragments

9.07.3.3.1 Syntheses

MurNAc-L-Ala-D-*i*Gln ('muramyl dipeptide': MDP) was found as the minimal structure to possess biological activity.[338–341] Following this discovery hundreds of derivatives have been synthesized and tested *in vitro* and *in vivo* (cf. Section 9.07.3.6). Recent examples of the chemical syntheses of small PGN sequences or analogues include the following. Classical solution phase methods were used to obtain MurNAc-pentapeptide (MPP: MurNAc-L-Ala-D-*i*Gln-L-Lys-D-Ala-D-Ala, in the form of Me- or SPh glycosides).[342] MPP with the anomeric OH phosphorylated was also prepared as an intermediate toward the total synthesis of Lipid I.[343,344] Similarly, synthesis of the complete repeating unit of Lys-type PGN (GMPP: GlcNAc-β(1–4)MurNAc-L-Ala-D-*i*Gln-L-Lys-D-Ala-D-Ala) was reported (with MurNAc OH-1 phosphorylated) in connection with the total synthesis of lipid II, an intermediate of the biosynthesis of PGN.[345,346] GMPP and MurNAc-tetrapeptide were also synthesized by Hesek *et al.*[347] The same group also reported the synthesis of the tetrasaccharide-containing fragment of Lys-type PGN: a dimer of the above unit (GMPP$_2$: GlcNAc-β(1–4)MurNAc-(L-Ala-D-*i*Gln-L-Lys-D-Ala-D-Ala)-GlcNAc-β(1–4)MurNAc-(L-Ala-D-*i*Gln-L-Lys-D-Ala-D-Ala)).[348] Construction of Dap-type muropeptides required procedures for the synthesis of the unusual amino acid *m*Dap.[349] This was accomplished either by a multistep procedure starting from *N*-benzyloxycarbonyl-L-glutamate[350] or by taking advantage of a metathesis cross-coupling reaction.[351] The first syntheses of *m*Dap-containing muramyl- and 1,6-anhydromuramyl peptides (TCT analogues) were then accomplished by Kubasch and Schmidt.[350] MurNAc-tri- to pentapeptides, either with L-Lys or *m*Dap as the third residue in the peptide stem, were also obtained using solid-phase peptide synthesis.[352] For studies of interactions with PGRPs (*vide infra*), a cross-linked Lys-type PGN model was constructed by interconnecting the D-Ala ends of two identical MurNAc tetrapeptides.[353] Di-, tetra-, and octasaccharide fragments containing the full repeating glycan sequence (i.e., GlcNAc-β(1–4)-MurNAc) and truncated peptide chains (i.e., L-Ala-D-*i*Gln) were prepared using a block synthesis approach.[354] The same group recently reported the syntheses of tetra- and octasaccharide fragments of the Lys-type PGN with tri-, tetra-, and pentapeptide chains, respectively, attached to MurNAc.[355,356] The tetrasaccharide tetrapeptides proved to be competitive inhibitors of the melanization cascade, an important immune defense mechanism in arthropods.[357]

9.07.3.3.2 Structural studies

The full repeating unit of a Gram-negative PGN (Dap-GMPP (in earlier literature it was labeled as PGM (for peptidoglycan monomer) but, for consistency, it will be abbreviated as Dap-GMPP henceforth): GlcNAc-β(1–4)-MurNAc-L-Ala-D-*i*Gln-*m*Dap(ωNH$_2$)-D-Ala-D-Ala) was first isolated from the cell wall of *Brevibacterium divaricatum*[358] but, to our knowledge, its chemical synthesis has not, apparently, been reported to date. The conformation of Dap-GMPP was determined in aqueous solution using NMR-restrained molecular modeling.[359] The preferred conformation of the peptide chain is characterized by a well-defined torsion angles around rotatable bonds from C3–O(MurNAc) through N–Cα(Ala^1); the C-terminal part, on the other hand, exhibits more conformational flexibility. The Dap-GMPP molecule was found, furthermore, to exhibit an amphipatic character with well-separated lipophilic side chains of the Lac and Ala residues at one side and a hydrophilic domain constituted by the charged Dap side chain and the disaccharide hydroxylic groups.[359] This property is certainly important in binding to enzymes as indicated by a computational model, based on an X-ray structure, of PGN binding to DD-transpeptidase, an enzyme participating in PGN biosynthesis. The model indicated binding of the hydrophilic glycan strands to hydrophilic grooves on the enzyme surface.[360] Dap-GMPP is a potent, nontoxic, nonpyrogenic immunostimulator (for references, see Halassy *et al.*[361]).

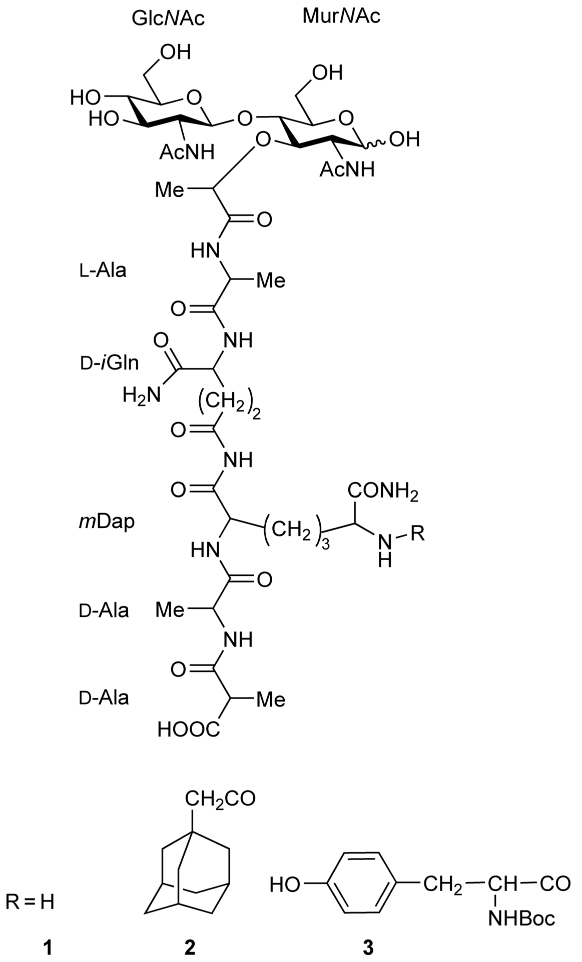
To investigate the influence of lipophilic versus hydrophilic character on biological activity, lipophilic derivatives of Dap-GMPP bearing either (adamant-1-yl)-acetyl- or Boc-Tyr substituents at the ε-amino group of Dap (Ad-Dap-GMPP and BocTyr-Dap-GMPP, respectively, (**Scheme 5**) were synthesized[362] and their solution structures determined by NMR and molecular modeling.[363]

The conformations of these lipophilic derivatives in DMSO exhibited similar characteristics to those found for Dap-GMPP in water,[359] with slightly more disordered C-terminal residues. For Dap-GMPP itself, the C-terminal conformations were different in the two solvent: NOEs measured in DMSO indicated steric proximity between the charged end groups (Ala^5–COO$^-$ and Dap^3 $-$NH$_3$$^+$) because of electrostatic pull. Missing of such ordering in water solution was attributed to the polar water molecules effectively shielding of the explicit charges by hydratation.[363] Recent solid-state NMR results, albeit measured on intact PGN, seem to be in agreement with this solution-state peptide conformation.[364]

The structure of the Lys-type dimer, GMPP$_2$ (*vide supra*) in water solution was recently determined using NMR-restrained molecular modeling computations.[335] The general features of this structure, such as relative rigidity of the glycan backbone, together with the N-terminal peptide part, increased flexibility at the peptide C-terminal, are similar to those of the Dap-GMPP.[359] The glycosidic torsion angles are also very similar, indicating virtually identical glycan backbone conformations in solution for GMPP$_2$ and Dap-GMPP. Surprisingly, however, preferred conformations around the MurNAc-peptide junction (i.e., D-Lac, Ala^1) in GMPP$_2$ were found markedly different from those in Dap-GMPP, both in water solution. The structural significance of this difference will be discussed in Section 9.07.3.5.

9.07.3.4 Peptidoglycan Recognition Proteins

PGRPs are recently discovered[365–367] pattern recognition factors that play important roles in the innate immunity from insects to mammals. The current state of the art regarding PGRPs has been more than adequately covered in recent reviews.[322,323,368] Based on these reviews, only a brief summary will therefore be given here focusing on mammalian PGRPs mainly for reasons mentioned in Section 9.07.3.1. Insect PGRPs are involved in *Toll, Imd*, or *prophenoloxidase* cascade signaling mechanisms to generate immune response. In mammals, four PGRPs have been identified thus far: PGLYRP1 to PGLYRP4 (formerly: PGRP-S, PGRP-L, PGRP-Iα, and PGRP-Iβ, respectively). The main function of mammalian PGRPs seems to be a direct bactericidal action rather than functioning as PRRs, as is the case for their insect counterparts.[326] On the basis of X-ray studies and molecular modeling, it was recently suggested that the bactericidal effect is due to the inhibition of cell wall synthesis by sterically blocking the access of biosynthetic enzymes to the nascent PGN chains and also



Scheme 5 Structures of Dap-GMPP (**1**), the repeating unit of a Gram-negative PGN from the cell wall of *Brevibacterium divaricatum*, Ad-Dap-GMPP (**2**), and BocTyr-Dap-GMPP (**3**). Modified with permission from K. Fehér; P. Pristovšek; L. Szilágyi; D. Ljevaković; J. Tomasić, *Bioorg. Med. Chem.* **2003**, *11* (14), 3133–3140.

by forcing them into a conformation that prevents formation of cross-links between the peptide stems[369] (cf. Section 9.07.3.5). This mechanism is reminiscent of that suggested for glycopeptide antibiotics such as vancomycin, ramoplanin, and actagardine (cf. Section 9.07.4). Nevertheless, several aspects regarding the functions of PGRPs in mammals is still not clear.[322,326,370,371] It was suggested, for instance, that PGLYRP2, a catalytic PGRP with *N*-acetylmuramoyl-L-alanine amidase activity,[372,373] might function to turn off excessive immune response[325] by hydrolizing PGN.

9.07.3.5 Interactions of PGRPs and Other Proteins with PGN Fragments: Structural Studies

Considerable efforts have recently been focused to elucidate the structural basis of PGRPs responsible for PGN binding using X-ray crystallography. At least seven crystal structures for PGRPs of insect or human origin have been reported during the last 5 years.[322,323] All PGRPs contain at least one C-terminal PGN-binding domain, which is approximately 165 amino acids long and shares an approximately 30% sequence homology to bacteriophage T7 lysozyme and other bacterial type 2 amidases (enzymes that hydrolyze the amide bond between the lactyl moiety of MurNAc and L-Ala of PGN).[322,323,368] All amidases, including catalytic PGRPs (such as PGLYRP2, cf. Section 9.07.3.4), have a Zn^{2+} binding site: the metal ion is essential for the catalytic activity. Some insect PGRPs are transmembrane proteins, all mammalian PGRPs are, however, secreted in the form of homo- or heterodimers, which are disulfide linked. Human PGLYRP3 and PGLYRP4 both contain two PGN-binding domains; this might serve to increase the binding affinity to PGN by a multivalent effect. On the other hand, two binding domains may confer multiple binding specificities for these PGRPs to recognize PRRs of different pathogens and, enhancing thereby their antibacterial efficiencies.[322] Insect PGRPs do not form disulfide-linked dimers. All structures exhibit similar overall folds, as determined by X-ray crystallography, comprising a central five-stranded β-sheet and three α-helices, and are cross-linked by 1–3 disulfide bonds. All PGRPs contain, in addition an N-terminal chain of 30–50 residues, which is highly variable with respect to both sequence and conformation; the functional significance of this segment is yet to be elucidated.

To understand the structural basis for the recognition of PGN by PGRPs, complexes of PGRP with PGN would have to be subjected to structure determination methods. Polymeric PGN is, however, too large and noncrystalline and therefore unsuitable for NMR or X-ray studies studies. For this reason complexes of noncatalytic PGRPs with small PGN fragments have been subjected to X-ray crystal determination. The first structure was reported for the C-terminal binding domain of human PGLYRP-3 (PGRP-Iα) in complex with MurNAc-L-Ala-D-*i*Gln-L-Lys (MTP of Gram-positive bacteria), revealing an extensive network of H-bond and van der Waals contacts of MTP with 16 residues of the protein's binding cleft. Most contacts were to the peptide part of MTP and only a few to the MurNAc. No significant conformational change of the protein was detected as a result of the binding.[374] A subsequent structure determination of the same protein complexed, however, with a larger fragment, including the complete peptide stem of PGN, that is, MurNAc-L-Ala-D-*i*Gln-L-Lys-D-Ala-D-Ala (muramyl pentapeptide, MPP), revealed ligand-induced conformational changes in the binding cleft, and this was suggested to occur in many PGRPs.[375] Most of the protein–ligand contacts involve the peptide part of MPP establishing hydrogen bonds between the main chain atoms of the peptide and 20 PGRP residues.

Tracheal cytotoxin (TCT, cf. Section 9.07.3.3.1) induces heterodimerization between PGRP-LCa and PGRP-LCx; this is the first step to trigger immune response in *Drosophila* by activating the *Imd* pathway.[376] PGRP-LCa and PGRP-LCx are both transmembrane proteins with different PGN-binding domains in their extracellular parts. The crystal structure of the ectodomains of PGRP-LCa and PGRP-LCx bridged by TCT shows that TCT binds to LCx in the ternary LCx–TCT–LCa complex through its peptide chain, in a similar way as MPP binds PGLYRP-3, and exposes the disaccharide part for interaction with LCa.[377] This is in line with results obtained by biochemical methods.[332] Both proteins undergo induced-fit conformational changes during this process. Bringing together the PGRP-LCx and PGRP-LCa receptors is necessary to trigger the activation step for the *Imd* pathway.[377] *Drosophila* PGRP-LE, a soluble protein involved in the *Imd* signaling, also binds TCT.[378] A crystallographic study revealed that TCT induces an infinite head-to-tail dimerization in which the disaccharide moiety occupies the dimer interface;[379] this is analogous to the TCT-induced heterodimerization of PGRP-LCx and PGRP-LCa. Dimerization of PGRP-LE increases its binding strength to TCT by approximately 30-fold ($K_D \sim 30\,\text{nmol}\,\text{l}^{-1}$). In both cases the Dap carboxylate group engages in a key electrostatic interaction with the guanidino side chain of an Arg residue of the protein: this provides a basis for the discrimination between Gram-negative and Gram-positive (Dap or Lys, respectively) types of PGN by PGRPs.[377,379] Other studies have, however, suggested Asn-236 and Phe-237 to be the key residues in PGLYRP-3 responsible for the discriminatory recognition of Dap-type or Lys-type PGNs.[352,353] Mutation of just these two residues suffices to change the specificity from one type to the other. This is an indication for the adaptive character of the innate immune system enabling a quick immune response to new microbial challenges.[353]

The crystal and molecular structure of PGLYRP4 C-terminal domain (PGLYRP4-C) in free form and in complex with the complete repeating unit of Gram-positive PGN, GlcNAc–MurNAc-L-Ala-D-*i*Gln-L-Lys-D-Ala-D-Ala (GMPP), was reported recently.[369] The bound conformation of GMPP was compared with two reported solution structures, obtained by NMR-constrained molecular modeling: one for the dimer of GMPP (GMPP$_2$)[335] and another one for a Dap-type PGN monomer (Dap-GMPP, cf. Section 9.07.3.3.2).[359] In the PGLYRP4-C–GMPP complex the glycosidic torsion angles between

GlcNAc and MurNAc change significantly, with respect to the solution conformations of either Dap-GMPP or GMPP$_2$. Particularly, the ψ (C1–O1(GlcNAc)–C4–C3(MurNAc)) value in the crystal structure (161°) indicates a higher energy conformation for the disaccharide part than in solution (118° in the reference Dap-GMPP[359] or 126° in GMPP$_2$[335]). On the other hand, the peptide orientation in the bound state, as indicated by the D-Lac torsions, is very close to that found for Dap-GMPP in solution[359] rather than that for the nonbound GMPP$_2$.[335] The peptide stems of other muropeptides (MPP or TCT) are also locked in this conformation when bound to various PGRPs[374,375,377,379] so these PGN-binding proteins appear to consistently favor one of the two possible conformations for D-Lac. Modeling of the PGLYRP4-C–GMPP$_2$ complex indicated that there would be a steric clash between the protein and the peptide stem of GMPP$_2$ should the latter retain its solution conformation. Based on a computational modeling of the cell wall PGN structure[335] it was suggested that the peptide orientation in this conformation is unfavorable for cross-linking during PGN biosynthesis. On the other hand, distortion of the glycan backbone conformation upon binding to PGRPs is likely to prevent the functions of glycan synthesizing enzymes by steric hindrance. It was hypothesized that the disruptive action of PGRPs on bacterial cell wall formation might be attributed to this double mechanism.[369]

Muramidases or lytic transglycosylases (LTs) are hydrolyzing the MurNAc–GlcNAc glycosidic bond and play a role, just as lysozyme (Section 9.07.3.2), in the turnover and recycling of PGN to facilitate cell growth and division. Their mechanisms of action is, however, different from that of lysozyme. Unlike the latter, and most of other glycoside hydrolyses, which require participation of two catalytic carboxylates at the active site, LTs have a single catalytic residue and may utilize the *N*-acetyl group of MurNAc to provide anchimeric assistance in catalysis similar to the mechanism described for chitinases. As a result, hydrolysis of the interglycosidic bond between MurNAc and GlcNAc by LTs does not produce a reducing MurNAc end with a free glycosidic OH group but rather an 1,6-anhydro bond is formed via nucleophilic attack by the 6-OH group of the same MurNAc residue.[380] Recent crystallographic studies on the structures of protein complexes with PGN or chitin fragments have significantly contributed to our understanding of the reaction mechanisms, PGN dynamics and immunological aspects of these systems.

LTs are generally membrane-anchored proteins, except the 70 kDa Slt70, which is soluble. Its crystal structure in complex with GlcNAc-β(1–4)-MurNAc(1,6-anhydro)-L-Ala-D-*i*Gln-*m*Dap (G(anh)MTP) revealed a shallow groove, adjacent to the PGN-binding site, for the binding of this muropeptide. The structure furthermore confirmed the presence of a specific binding site for the peptide part of G(anh)MTP and it was suggested that Slt70 starts the cleaving reaction at the MurNac end of the PGN chain.[381] A functionally similar enzyme is the 18 kDa lytic transglycosylase from bacteriophage lambda (LaL). Its crystallographic structure, determined in its from complexed with hexa-*N*-acetylchitohexaose, represents the first example of a lysozyme in which all binding subsites are occupied.[382] Slt35 is a fully active, soluble form of the integral membrane transglycosylase MltB. Four sugar-binding sites and two peptide-binding sites were identified in this protein by X-ray crystallography of its complex with the muropeptide GMDP.[383] It is of note in this context that the minimal structure needed to activate the Toll immune pathway in *Drosophila* was found to be a muropeptide dimer, GlcNAc–MurNAc-L-Ala-D-*i*Gln-L-Lys(D-Ala-D-Ala)-(Gly)$_5$-L-Lys[(Gly)$_5$]-D-*i*Gln-L-Ala-MurNAc–GlcNAc, that is, a PGN fragment with four sugar residues.[384] A recent crystal structure of an inactive mutant (D308A) of MltA in complex with chitohexaose has shown that all six sugar residues are bound in the active site of the enzyme and binding induces a large reorientation of two structural domains of the enzyme. Although the natural substrate of MltA is PGN rather than chitin, implications for PGN hydrolysis were drawn from a model of the (GlcNAc–MurNAc)$_3$ complex built on the basis of the MltA(D308A)-chitohexaose crystal structure. Based on this model it was suggested that the cleavage of glycosidic bond is facilitated by a high-energy half-chair conformation of the pyranose ring bound at the active site.[385] This distortion is, interestingly, very similar to that proposed for lysozyme catalysis.[386,387] This protein does not possess, however, unlike other membrane transglycosylases (MLTs), binding sites for the peptide part of PGN.[385] Another enzyme that plays an important role in PGN breakdown in processes like cell wall turnover, cell separation, or sporulation is CwlC amidase from *Bacillus subtilis*, which hydrolyzes the amide bond between the lactyl group of MurNAc and L-Ala. The solution structure of the C-terminal domain of CwlC was determined by 3D and 4D NMR using uniform ^{15}N- and ^{13}C-labeling by taking advantage of $^{h3}\mathcal{J}_{NC'}$ hydrogen bond restraints and $^{1}D_{NH}$ dipolar couplings.[388] The PGN-binding region was explored by following chemical shift changes in the ^{1}H–^{15}N HSQC spectra during titration with non-well-defined soluble PGN digests. Two equivalent, symmetrically located binding sites were proposed and it was suggested that multivalency effects, extensively studied in lectin–carbohydrate interactions, might play a role in the CwlC–PGN interaction as well.[388]

9.07.3.6 Physiological Activities of Muropeptides

It was known long ago that bacterial cells were capable to boost the immune response in certain diseases: 'Freund's adjuvant' to treat pulmonary infection consisted of a suspension in oil of killed mycobacteria.[389] Later it was recognized that the cell wall PGN was responsible for inflammatory syndromes in several diseases like arthritis, meningitis, or septic shock. In a quest to identify the structural basis of immunogenecity, smaller PGN fragments were synthesized and analyzed (for reviews, see Chedid *et al.*,[390] Azuma,[391] and www.curehunter.com/m/keywordSummaryC033575.do). The Dap-type PGN monomer, Dap-GMPP (cf. Section 9.07.3.3.2) isolated from the cell wall of the Gram-negative *Brevibacterium divaricatum* is a nontoxic, nonpyrogenic immunostimulator. Its adjuvant activities on the immune system of mice challenged with ovalbumin (OVA) have been thoroughly investigated. For example, Dap-GMPP enhanced the immunogenicity of peptides of measles virus origin.[361] Lipophilic derivatives of Dap-GMPP bearing either (adamant-1-yl)-acetyl- or Boc-Tyr substituents at the ε-amino group of Dap were shown by NMR and molecular modeling to assume conformations different from that of the parent PGM in solution[363] (cf. Section 9.07.3.3). Their immunostimulating activities were, however, comparable to that of the parent Dap-GMPP.[362]

Tracheal cytotoxin (TCT), a muropeptide derived of Gram-negative PGN, such as the cell wall of *Bordetella pertussis*, elicits immune responses in *Drosophila*.[392] In addition to immunostimulating activity[393] this muropeptide is a very potent somnogenic. This activity may be related to the conformation of the 1,6-anhydro-bridged MurNAc that is very different (${}^{1}C_4$) from that of the nonbridged, monocyclic glucopyranose ring (${}^{4}C_1$). Hundreds of small MW muropeptides, mostly synthetic or synthetically modified PGN fragments, possess multiple biological activities, influencing the immune response from insects to mammals. These aspects have been discussed in detail in a recent review.[394] Growth of the bacteria involves breakdown and resynthesis of the cell wall. Muropeptides released from the cell wall PGN during these processes are mediating a much broader range of interactions, other than immune signaling, between bacteria and other organisms. For instance, PGN fragments, such as TCT, play a role in the pathogenesis of several bacterial infections, such as those caused by the Gram-negative *Helicobacter pylori* or by the Gram-positive *Listeria monocytogenes* and several other bacteria as well (for a review, see Boneca[395]). There are indications that PGN fragments might induce immune responses in plants during host–pathogen interactions and mediate various symbiotic interactions between bacteria themselves and between bacteria and eukaryotes. All these intriguing roles of muropeptides have been summarized in a recent review.[396]

9.07.3.7 Conclusions

Bacterial cell wall peptidoglycans (PGN) and fragments play important roles in the immune response of higher organisms against bacterial infections, and mediate various symbiotic interactions between bacteria themselves and between bacteria and eukaryotes. Recent structural studies of peptidoglycans and smaller fragments thereof have significantly enhanced our understanding of the underlying molecular mechanisms and contributed therefore to develop improved strategies to fight bacterial infections. Experimentally, X-ray diffraction and NMR spectroscopic techniques play major roles in structure elucidation and in studies of the relevant molecular interactions, eminently, those between various recognition proteins and PGN-related carbohydrates. Following characterization of PGN structural features, synthetic approaches to smaller PGN fragments and determination of their 3D structures by NMR and X-ray techniques is discussed in this section. Among proteins interacting with PGNs peptidoglycan recognition proteins (PGRPs) emerged recently as major pattern recognition factors that play important roles in the innate immunity from insects to mammals. Their structures, together with some lytic proteins involved in the cell wall biosynthesis and breakdown, and their complexes with muropeptides were extensively investigated and elucidated by X-ray and NMR methods in the last decade. In view of the intriguing physiological activities of a great number of muropeptides and the new knowledge generated by the structural investigation outlined the prospects for development of efficient therapeutics against the alarmingly increasing bacterial resistance seem promising.

9.07.4 NMR of Glycopeptide (Vancomycin-Type) Antibiotics: Structure and Interaction with Cell Wall Analogue Peptides

9.07.4.1 Introduction

For the pessimist it appears that we may be losers in the fight against the often lethal superbugs. The prevalence of superbugs, such as methicillin-resistant *Staphylococcus aureus* (MRSA) and vancomycin-resistant *Staphylococcus aureus* (VRSA) is increasing rapidly both in hospitals and in the community. Until recently, the glycopeptide vancomycin remained the only antibiotic effective at infections caused by multiresistant Gram-positive bacteria and they have been considered as the last line of defense against MRSA. However, treatments with vancomycin for the past decades has led to the appearance of staphylococcus and enterococcus strains that are resistant to vancomycin and all known glycopeptides. GRE (glycopeptide-resistant enterococci) and GISA (staphylococci with an intermediate resistance to glycopeptides) infections are especially threatening.

This section focuses on the molecular basis of the bacterial cell wall biosynthesis inhibition by glycopeptide antibiotics for two reasons. The first reason is that vancomycin has been in the frontlines of the antibacterial combat for nearly 30 years, which is unusually long before resistance appears. The second reason is that only few biomolecular recognition models could be studied at an atomic resolution to such an extent as glycopeptides and cell wall analogue molecules or stable isotope labeled cell walls. For solution NMR or X-ray studies the cell wall is often mimicked using oligopeptides terminating in D-Ala-D-Ala sequence. The interaction with 'intact' cell wall can be best achieved using solid-state NMR close to *in vivo* conditions and the research on the field has shifted to this direction. However, there are fine details, especially dynamics upon ligand binding, that await for some more experimental and theoretical studies.

9.07.4.2 Basics of Mode of Action

9.07.4.2.1 Structure of the cell wall

Surprisingly, the 3D structure of the cell wall of Gram-positive bacteria has only recently been solved[335] (see also Section 9.07.3). The cell wall's basic structural unit has long been known to be a peptidoglycan, *N*-acetylglucosamine (NAG)–*N*-acetylmuramic (NAM)-pentapeptide. Meroueh *et al.*, who recently synthesized a NAG–NAM(pentapeptide)–NAG–NAM(pentapeptide) dimer, have now succeeded in determining the structure of this 2 kDa peptidoglycan by solution NMR spectroscopy. Their analysis reveals an ordered, right-handed, helical saccharide conformation, with a set of repeated glycosidic torsion angles that lead the authors to suggest an oligomer structure with three NAG–NAM repeats per turn. Computer models of the cell wall structure, assuming such threefold symmetry as well as incomplete cross-linking, show a honeycomb pattern with pores ranging in size from 70 to 120 Å. These pores are large enough to accommodate the cell wall's own catalytic machinery as well as channel proteins and other macromolecules. Furthermore, opposite to beliefs, the orientation of the glycan strand is orthogonal to the membrane and not parallel as had been previously assumed by many, based on the structures of chitin and cellulose, the two other main wall-forming β-1,4-linked glycan biopolymers in nature.

9.07.4.2.2 Main targets of glycopeptide antibiotics

The layer of the bacterial cell wall that ensures strength is this covalently cross-linked peptidoglycan. The larger the fraction of adjacent peptide strands that are connected by action of transpeptidases, the higher the mechanical strength to osmotic lysis. Transglycosylases act on the glycan strands to extend the sugar chains by incorporation of new peptidoglycan units from *N*-acetylglucosamine-1,4-*N*-acetylmuramyl-pentapeptide-pyrophosphoryl-undecaprenol (lipid II). The vancomycin family of glycopeptide antibiotics target the peptidoglycan layer in the cell wall assembly. Briefly, the mode of action of vancomycin antibiotics is through the binding of peptidoglycan cell wall fragments terminating in a D-Ala-D-Ala sequence to the carboxylate anion-binding pocket of the antibiotic as suggested by Nieto and Perkins[397] and Perkins[398] and later confirmed by the Williams group in Cambridge.[399–402] Vancomycin ties up the peptide substrate and thereby prevents it from reacting with either the transpeptidases or the transglycosylases. The net effect is the same: failure to make peptidoglycan cross-links leads to a weaker wall that predisposes the treated bacteria to a killing lysis of the cell wall layer. The concave binding pocket of the vancomycin antibiotic makes five hydrogen bonds to the D-Ala-D-Ala dipeptide terminus of each uncross-linked peptidoglycan pentapeptide side chain, which accounts for the high affinity of the antibiotic for its target, both in partially cross-linked walls and in the lipid II intermediate.

Interestingly, some semisynthetic derivatives of eremomycin, vancomycin, teicoplanin, and some other glycopeptides were found to exhibit activity against viruses,[403] for example, HIV-1, HIV-2, and Malonisarcoma retrovirus.

9.07.4.2.3 *Common structural features of the glycopeptide antibiotics*

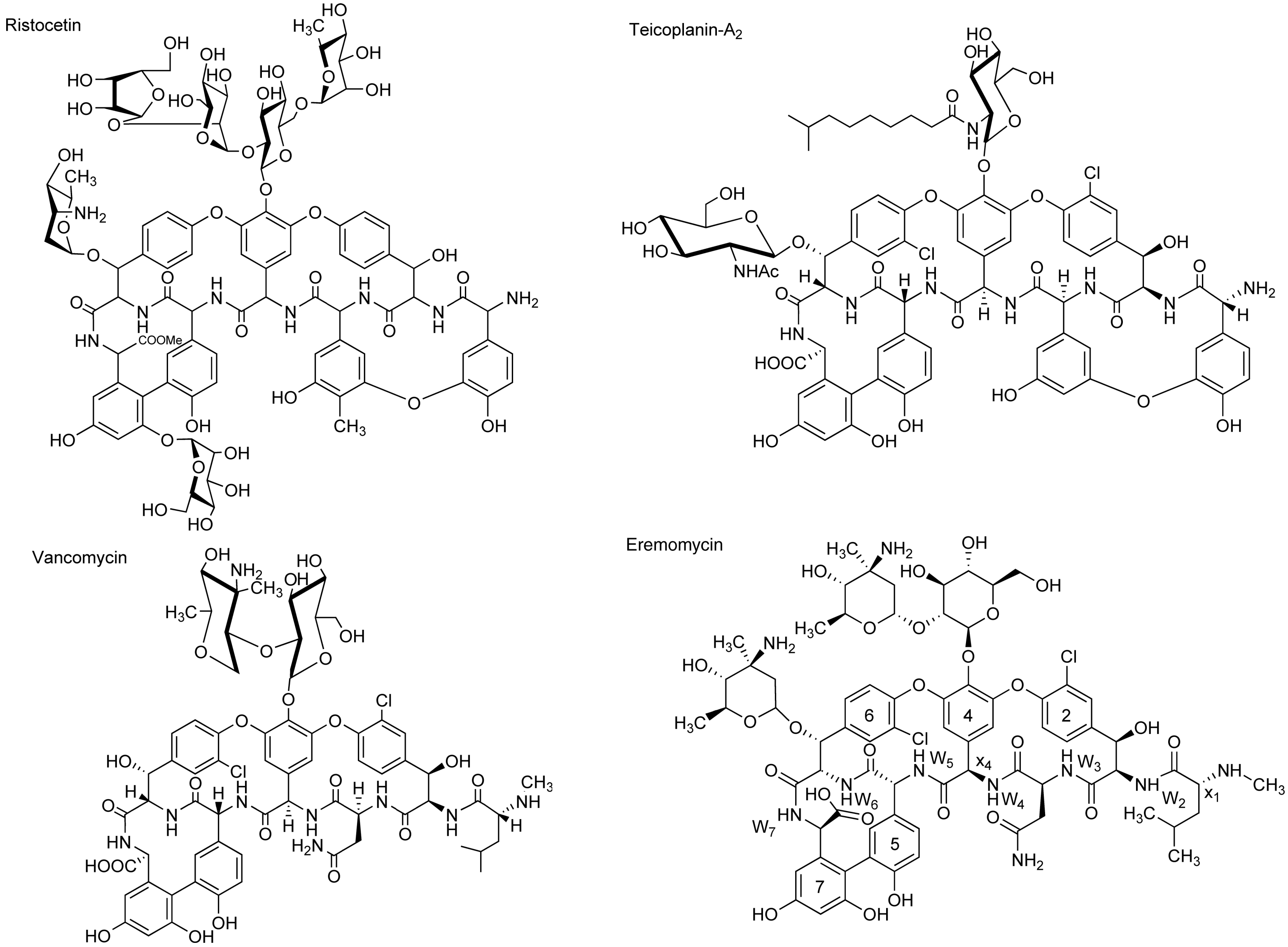
The first known structure within the glycopeptide family of antibiotics was vancomycin. Vancomycin, the clinically most important glycopeptide was discovered in 1956; however, its structure was disclosed only 25 years later, after correcting[404,405] the previous X-ray structure of CDP-1[406] by NMR.[407] Since that time hundreds of related glycopeptides were discovered or prepared as semisynthetic compounds. More than 12 000 papers were published about the vancomycin topic within the last 20 years. Some recent reviews[408–413] give in-depth accounts of the topic. Here we just summarize some basic structural features of the main representatives (vancomycin, eremomycin, teicoplanin, ristocetin-A) in **Scheme 6**.

Hitherto, all known glycopeptides have a heptapeptide aglycon and the residues are numbered starting from the N terminus (1–7). The polypeptide portion of vancomycin consists of an *N*-methyl-leucine at position 1 and substituted phenyl glycines at positions 2, 4, 5, 6, and 7, furthermore an asparagine at position 3. In all tested cases the stereochemical configurations of *R*,*R*,*S*,*R*,*R*,*S*, and *S* were found in vancomycin antibiotics. In many of these glycopeptides, residues 1 and 3 are aliphatic like in vancomycin. In other cases residues 1 and 3 are also aromatic amino acids, for example, in ristocetin and in teicoplanin and not surprisingly, they are substituted phenyl glycines. Aromatic rings 2–4 and 4–6 are cross-linked via etheric oxygens, and if residues 1 and 3 are aromatic, they may be cross-linked similarly. The 5, 7 phenyl rings are directly connected with carbon–carbon bond. The mesh of cross-linked aromatic rings lends an enhanced rigidity to the backbone of the glycopeptides. Moreover, the chlorine-bearing aromatic rings of residues 2 and 6 are prohibited from rotating by other atoms in their cycles (with residue 4). Thus, these rings give rise to stable and distinct rotational isomers, that is, atropisomers.[414,415] Vancomycin is tricyclic with three ring systems, which gives eight possible atropisomers, but the natural product is a single isomer. In vancomycin and other glycopeptides that have one chlorine atom on one or both rings, the chlorine on ring 2 is on the edge facing away from the ligand-binding site, whereas on ring 6 it is on the edge facing toward the ligand-binding site. The peptide bond between residues 5 and 6 exists in the less stable cis configuration; however, the cis arrangement dominates glycopeptide antibiotics.

The role of carbohydrates in glycopeptide antibiotics is somewhat less appreciated: they make the glycopeptides water soluble in spite of the hydrophobic aromatic rings. Glucose units of di- or oligosaccharides are typically bound to the central aromatic ring 4 with β-glycosidic linkage. They are often linked to positively charged amino sugars (e.g., vancosamine, eremosamine = 4-epi-vancosamine) by an $\alpha(1–2)$ glycosidic bond. The 180° conformational flips of these oligosaccharides explain the generation of slowly interconverting asymmetric homodimers[416] in aqueous solution. In the system of eremomycin complexed with an unnatural ligand, the positively charged amino group of the eremosamine monosaccharide at residue-6 forms an 'intradimeric' salt bridge to the bound carboxylate anion of the ligand. Also, carbohydrate–carbohydrate recognition was postulated between the two disaccharides at the two halves of the dimer.

9.07.4.3 NMR Methods for Solution Structure

According to known structures, the molecular weight of the monomeric glycopeptides is in the 1.3–2.1 kDa range. In aqueous solution (*in vitro* conditions) nearly all of the vancomycin antibiotics form stronger or weaker noncovalently bound dimers (except teicoplanin) at concentrations useful for NMR; consequently, the effective molecular weight – especially when they are complexed with ligands – approaches the 4–5 kDa range, which is close to the mass of small proteins. In the beginning, nearly exclusively ^{1}H NMR methods were applied. The early application of negative homonuclear NOEs (nuclear Overhauser effect) in the slow motion regime[399,417] was a landmark to determine intermolecular distances. The NOE experiments were invaluable like the 2D NOESY for protein structures introduced by Ernst[418] and Wüthrich.[419]

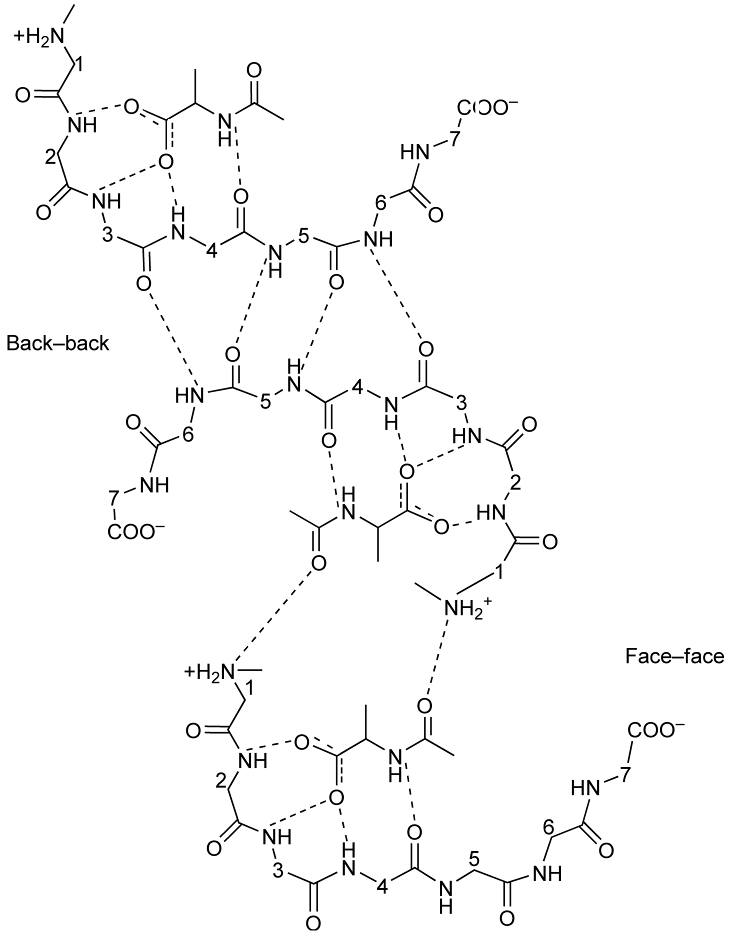


Scheme 6 Structure of some important glycopeptide antibiotics.

NMR studies provided NOE-derived distances between ristocetin and Ac-D-Ala-D-Ala that established the bound position of the dipeptide as lying antiparallel to the antibiotic backbone in the concave exterior of the molecular surface.[400] Since the 1980s were connected with the rapid development of the basic 2D-NMR methods, COSY, TOCSY, NOESY, ROESY, HETCOR, HSQC, and HMBC techniques could be readily implemented to structural works at natural isotope abundance (see also Section 9.07.1). In fact, the general solution NMR methods applicable to oligopeptides and peptide–peptide interactions are useful to glycopeptides. A few textbooks can be cited for more details and practical hints.[420–422] For solution NMR studies isotope labeling was sparse: ^{15}N labeling of eremomycin[423] and ^{13}C incorporation to C1 of Ac-D-Ala-D-Ala[424] since the general sensitivity for assignment purposes is sufficient. Among others, ^{13}C NMR signal assignments of eremomycin,[425] vancomycin,[426] and a few aglycons[427] were published. Application of 3D techniques would be an advantage for congested spectra of dimers in the absence of ligands; however, no applications were reported.

9.07.4.4 Comparison of Crystal and Solution Structures – Dimerization and Ligand Binding

The first two reports of the crystal structure of vancomycin complexed with an acetate anion were independently published[404,405] as milestone contributions. It was concluded that vancomycin dimerized in a 'back-to-back' configuration, with segments of antiparallel structure forming both at the 'back' of each monomer, across the dimer interface between heptapeptide backbones, and at the 'front' of each monomer, between the drug and its ligand. In the crystal,[404] one of the binding pockets is occupied by an acetate ion that mimics the C terminus of the cell wall peptide; the other is closed by the asparagine side chain, which occupies the place of a ligand. The occupied binding pocket exhibits high flexibility but the closed binding pocket is relatively rigid. X-ray crystallography has since verified this back-to-back antiparallel configuration in balhimycin[428,429] crystals, and has also shown that the aromatic rings and the sugar residues in these compounds engage in an extensive array of interactions across the dimer interface. The dimeric structure of ureido-balhimycin was established by NOEs and distance geometry calculations proved the antiparallel orientation of the two monomers.[430] Important reports on vancomycin crystals complexed with NAc-D-Ala suggested a ligand-mediated dimerization mode for vancomycin[431] or with weakly binding ligands[432] were published. Concerning the cooperativity between dimerization and ligand binding in vancomycin, it was suggested[433] that hydrogen bonds across the dimer interface may reduce dynamic fluctuations of the heptapeptide backbone and thereby stabilize hydrogen bonds between the backbone and the ligand. The crystallographic data appear to weigh against this explanation because the shape of the macrocyclic rings in vancomycin and balhimycin is tightly conserved in spite of differing hydrogen bond geometries. Loll first interpreted this as evidence that structural rigidity is an inherent property of these molecules, irrespective of whether their backbones are hydrogen bonded or not. However, higher affinity ligands may induce structural change[432] in the antibiotic if compared to low-affinity ligands. All of these crystal structures show the essential features of unsymmetric homodimers first demonstrated in the NMR study[416] of eremomycin.[425,434] In these structures the H-bond pattern between the two halves of the dimer forms a two-stranded antiparallel β-sheet. The monomers have concave shapes offering the binding site at the exterior of the dimer. NMR structures of aglyco-vancomycin (AGV) were published[435] in D_2O/DMSO (4:1) mixture or neat DMSO.[436] While the ligand binding is demonstrated in the former, there is no dimer formation or ligand binding in the latter. Significant conformational differences exist between the latter NMR structure in neat DMSO and the crystal structure[437]: most notably in the ligand-binding site and in the aromatic rings of residues 2, 4, and 6. The NMR structure shows the peptide backbone bulging outward in the vicinity of residues 2 and 3 to form the so-called β-pleated sheet conformation, which is unfavorable for both ligand binding and dimerization. However, the aglycon crystals were grown from an aqueous solution that supports dimerization. Moreover, the AGV is indeed capable of adopting an active conformation much like to vancomycin, and therefore the inactive conformation is not a result of the removal of the sugar residues. A detailed NMR study[438] with chloroeremomycin (A82846B) and the pentapeptide ligand, Ala-GGlu-Lys-D-Ala-D-Ala, proved that the complex of A82846B and its cell wall pentapeptide form an asymmetrical dimer similar to that seen for eremomycin complexed with the unnatural (pyrrole-2-carboxylate) ligand.[416] Prowse suggested that the carboxylate group may assume more than one orientation in the binding pocket and that the side chain of Asn-3 is an integral part of the hydrogen-bonding network. On the other hand, multiple binding modes were neither found in $^{13}C\{^1H\}$ heteronuclear NOE of vancomycin,[424] nor was a role attributed to Asn-3 eremomycin–ligand complexes.[416]



Scheme 7 The heptapeptide backbones of glycopeptide antibiotics forming back-to-back and face-to-face complexes with Ac-D-Ala.

A new dimeric form of vancomycin has been found[431] in which two monomers are related in a face-to-face configuration, and bound NAc-D-Ala ligands comprise a large portion of the dimer interface. A virtually infinite chain of vancomycin monomers comprise the crystal lattice. These chains are made up of alternating back-to-back and face-to-face contacts between monomers (**Scheme 7**). The biological significance of these new oligomers is not clear yet. Dimerization is believed to promote antimicrobial action because the binding of one monomer to the bacterial cell wall brings a second monomer into proximity with other peptidoglycan ligands, leading to the formation of a 'chelate' with the peptidoglycan. Back-to-back dimerization also increases the affinity between ligands and individual monomers. At least three mechanisms have been suggested to explain this allosteric effect. First, dimerization may induce conformational changes, or suppress thermomolecular deformations, and thereby yield a more favorable site for ligand binding. Second, dimerization may position positively charged groups such that the negatively charged carboxyl terminus of a ligand is attracted into the binding site. Third, antiparallel structure in a dimer may polarize the backbone peptide groups involved in ligand binding and strengthen their hydrogen bonding potential. These mechanisms are mutually compatible, and all may contribute to ligand affinity. In the X-ray study[439] of balhimycin and degluco-balhimycin complexed with di, tri, and pentapeptides, an unexpected variability of the extent of oligomerization an binding modes were observed. Appearance of face-to-face oligomers (tetra, hexa, and octamers), even virtually infinite layers are not due to crystal contacts; they depend on the arrangement of the ligand in the binding pocket and have little impact on the drug backbone conformation.

However, bigger peptide ligands cause enhanced backbone bending of the drug. Face-to-face dimers are formed when the model peptide reaches a critical fraction of the size of the cell wall precursor pentapeptide. The extensive interactions in this interface should enhance the kinetic and thermodynamic stability of the complexes. In the pentapeptide complex, the relative positions of the peptides are close to those required for D-Ala elimination, so this structure may provide a model for the prevention of the enzyme-catalyzed cell wall cross-linking by antibiotic binding. Interesting new results were obtained using more realistic cell-wall mimics. In the absence of drug dimers (vancomycin concentration 0.081 mmol l^{-1}) the binding affinity of the large glycopeptide (NAG–NAM-peptide) dimer was checked[440] to vancomycin by isothermal titration calorimetry (ITC). Two sequential binding events were observed that resulted in a final 2:1 vancomycin/cell wall analogue complex, and the association constant for the second was double of the first. Additional favorable enthalpic increment and a more unfavorable entropic contribution for the second binding step was experimentally observed. Molecular dynamics simulations also displayed reduced motion supporting thereby the ITC results. A challenging review was published on the structural biology aspects of vancomycin antibiotics.[441]

9.07.4.5 Possible Role of Dynamics Upon Ligand Binding

There is a consensus of the main structural motifs of molecular recognition in vancomycin antibiotics complexed with cell wall analogues. However, subtle details, for example, enthalpy–entropy compensation[442–445] and allosterism remained to be further disclosed. Following other studies, recent theoretical molecular dynamics calculations[446,447] suggest that the known cooperativity[433,448–451] between antibiotic dimerization and ligand binding could be explained by the nonadditivity of the entropic costs of dimerization and ligand binding. It was suggested that, in the absence of major conformational changes or other enthalpy-driven processes, enhanced internal molecular dynamics up to the picosecond timescale by themselves may be responsible for the observed cooperativity. NMR order parameters S^2 can be derived from ^{15}N NMR auto- and cross-relaxation, and they sample a similar fraction of internal timescales (picosecond range). For eremomycin, it was proven by ^{15}N relaxation that the two sides of the dimer are dynamically equivalent and the binding pocket is the least protected site for solvent access.[423] Experimental NMR evidences were found on the 180° rotation of the peptide group between residues 2 and 3 in vancomycin.[452] Recent molecular dynamics calculations suggested[453] a breathing mechanism that is able to enhance desolvation of the binding site. According to MD calculations, the barriers to the rotation of two different backbone peptide groups are sufficiently low, and their rotation destabilizes the water captured in the binding pocket. After the water molecule is expelled from the binding cavity – a key first step for molecular recognition – there is a chance for the ligand entry. Sporadic examples on proteins demonstrated increased internal dynamics of the host upon the binding of a small hydrophobic ligand that may outweigh the entropic cost of association.[454] On the other hand, Williams emphasizes the importance of structural tightening upon cooperative ligand binding.[455] According to Williams, the structural model of the ristocetin-A dimer system leads to the conclusion that positively cooperative binding will reduce the dynamic behavior of the receptor system. The importance of structural tightening, as opposed to partially bound states was underlined to explain chemical shift changes upon binding.[456,457] The idea was further extended to support general principles of ligand-induced reduction in motion within receptors and enzymes.[458]

9.07.4.6 Solid-State NMR of Glycopeptide Antibiotics with Bacterial Cell Wall Complexes

Some promising new compounds have been successfully tested against pathologic MRSA strains recently. Among them, the chlorobiphenyl derivatives could be easily transformed to fluoro derivatives. The ^{19}F nucleus is an excellent spy for solid-state NMR, where the detected weak ^{13}C–^{19}F or ^{15}N–^{19}F dipolar interactions can 'see' to long distances, significantly farther than NOEs in solution. A recent review[459] summarizes the potential of contemporary solid-state REDOR[460] (rotational echo double resonance) and TEDOR[461] (transferred echo double resonance) techniques with interesting applications on peptide antibiotics and also with the use of ^{31}P and ^{2}H nuclei. In general, these techniques are capable to measure internuclear distances in between 6 and 20 Å and work on S–I (rare–abundant) spin pairs where the S rare spin is observed. The advantages of REDOR are that it is independent of the chemical shift tensor of the coupled nuclei and does not require the resolution of the S–I dipolar coupling on the S chemical shift scale. On the other hand, in some cases, comparison with unlabeled samples require extra experiments. This problem can be circumvented by the TEDOR[461] method. TEDOR is capable to select the coupled spins from the background of uncoupled nuclei. TEDOR and REDOR can be used in combination, which allows the measurement of the S–X distance in an I–S–X labeled three-spin system. Since the ^{19}F isotope is abundant in some of the new-generation vancomycin antibiotics (e.g., fluorobiphenyl chloroeremomycin, LY329332), the ^{13}C{^{19}F} REDOR can position the single ^{19}F nucleus with respect to the natural abundance ^{13}C (or selectively ^{13}C-labeled Gly and D-Ala and/or [^{15}N] Lys nuclei of cell wall constituents that are resolved on their chemical shift scale.[462] Binding affinity data showed that the enhanced potency of LY329332 and that of the chlorinated analogue (both 1000-fold greater than that of vancomycin against vancomycin-resistant enterococci) are not reflected in an increase in binding affinity for mature peptidoglycan in the tested normal (not resistant) *S. aureus* strain (ATCC 6538P). The binding model assumes that the vancomycin cleft binds to a stem terminating in D-Ala-D-Ala. In the model, the fluorine of the biphenyl moiety is not near the L-Ala of the complexed stem, but rather the L-Ala of a nearest neighbor stem on an adjacent glycan strand. This nearest neighbor stem is situated with a bridge (85% of all stems have bridges), and this arrangement is the source of ^{19}F coupling to the carbonyl carbons of D-Ala and Gly. The complex is presumably stabilized by interactions of the sugars of LY329332 with proximate glycans. The model of mature peptidoglycan-LY329332 complex is putative, since only three distances could be determined with respect to the ^{19}F spy nucleus.

However, a sensible model was built up without the assumption of either dimeric antibiotics or membrane anchoring of the fluorobiphenyl tail. The sample preparation conditions are probably relevant to the picture obtained by solid-state NMR. It must be emphasized that mature, normal (not resistant) *S. aureus* strains were used in this case. The antibiotics concentration in these studies are close to the upper limit in human clinical treatments (100–200 μmol l^{-1}). The lyophilization of cell walls and intact cells comlexed with LY329332, for example, resulted in anhydrous samples where trehalose was applied to mimic water according to the water replacement hypothesis. The results were in contradiction with the generally accepted aqueous solution NMR structures in the sense that neither dimerization nor membrane anchoring was observed in the solid state. Absence of dimers is not too surprising, because reversible dimers are documented exclusively in water solutions (e.g., monomeric vancomycin D-Ala-D-Ala complex was analyzed in DMSO[463]). Furthermore, at 0.1–0.2 mmol l^{-1} antibiotic concentration, the dimer concentration would be low for typical $K_{\text{dim}} = 500$–$1000\ \text{mol}^{-1}\,\text{l}^{-1}$ and one expects 5–15% dimer only, except eremomycin and chloro-eremomycin where the dimer would be over 90% in aqueous solution. These two facts taken together and the use of mature cells may at least in part explain that detection of dimers failed in solid-state NMR studies. Similar investigations[464] determined the effect of vancomycin on cell wall assembly in normal *S. aureus* during active cell division.

It was suggested that at the therapeutic level vancomycin interrupts peptidoglycan synthesis by interference with transglycosylation. In other works[364] of the Scaefer group they studied ^{19}F derivatives of vancomycin, eremomycin, and chloroeremomycin – including modifications at the C terminus – that are active against resistant strains. REDOR technique was used to measure the dipolar couplings between ^{19}F of the drugs and ^{13}C and ^{15}N labels incorporated in peptidoglycan (PG) stems and bridging pentaglycyl segments. They improved their TEDOR/molecular dynamics model of [^{19}F]oritavancin[465] in the following: The pentaglycyl bridge is now believed helical, the pentaglycyl bridging segment is lowered into a protective cleft formed by the 4-eremosamine and the glycopeptide core, the D-*iso*-Gln of the bound stem is moved up toward the 4-eremosamine moiety and, the unbound neighboring stem is moved away from the C-terminus of the glycopeptide core and proximity to the bound stem. These views are supported by some biological evidences showing that sugars on glycopeptides may improve antimicrobial activities without enhancing binding affinities.[466] In the same work using the aid of bioaffinity mass spectrometry, the benefitial impact of drug self-dimerization was questioned. Very recent studies[467] using oritavancin with and without the D-Ala binding pocket suggest that oritavancin has dual mode of action. First, transglycosylation is inhibited via binding to lipid II. Second, correlation of the model structures and antibiotic activity led to the conclusion that the hydrophobic substituent of the drug disaccharide and components of the aglycon structure form a secondary binding site for pentaglycyl segments in *S. aureus*. They proposed that this secondary binding site compensates for the loss of binding affinity to D-Ala-D-Lac stem termini, and thus allows the disaccharide-modified glycopeptides to maintain their activity against VRSA. In similar studies[468] using vancomycin derivatives with damaged D-Ala binding cleft were initiated. REDOR techniques were used to study binding modes of des-*N*-methylleucyl-4-(4-fluorophenyl) benzyl-vancomycin (DFPBV), the fluorinated analogue of des-*N*-methylleucyl-4-(4-chlorophenyl) benzyl-vancomycin (DCPBV), which is active against both vancomycin-susceptible and vancomycin-resistant bacteria. Importantly, lack of hydrophobic side chain in the 4-disaccharide moiety in des-*N*-methylleucyl-vancomycin (DV) causes loss of activity against both types of bacteria. The proposed mode of action of DFPBV is as follows: It binds to the template peptidoglycan as it is positioned to inhibit transpeptidase activity by nonspecific steric interference. This mechanism is independent of the interaction with lipid II and requires no specific binding to enzymes. DFPBV may also bind at nascent peptidoglycan sites even though lattice constraints are only partially formed. In summary, the proposed mode of action of vancomycin and second-generation glycopeptide antibiotics significantly differs from the view suggested until now from interpretation of *in vitro* solution NMR results. Lack of drug dimers and membrane anchoring is surprising, however, the anhydrous environment for solid-state NMR may in part explain these results. The potential of recent solid-state NMR methods can be extended if some more different spy nuclei can be substituted to active drugs, and the biological targets from resistant enterococci and staphylococci will be available.

9.07.4.7 Conclusions

Glycopeptide antibiotics are believed to represent a last line of defense against the often lethal Gram-positive bacterial infections. Nowadays, vancomycin resistance is widespread, and new antibiotics should be developed to cope with multiresistant strains. Although the structures of vancomycin antibiotics and the essence of mode of action has been learned decades ago, fine details of these delicate molecular recognition processes are still not fully understood. Up to date NMR methods (both solution and solid-state), X-ray crystallography, *in silico* calculations, and calorimetry all contribute for better understanding of the mode of antibacterial action. Surprisingly, some recent semisynthetic glycopeptide 'antibiotics' exhibit remarkable antiviral effect extending thereby the scope of glycopeptide research.

Acknowledgments

For Section 9.07.1: K. E. K. thanks the support from the Hungarian Scientific Research Fund (OTKA NK 68578 and T-048713). D. U. acknowledges the support of Wellcome Trust (078780/Z/05/Z). Both authors thank Nicholle G. A. Bell (The University of Edinburgh) for reading the manuscript.

For Section 9.07.3: L. S. wishes to thank financial support from the Hungarian Scientific Research Fund (grants T-048713 and NK-68578).

For Section 9.07.4: G. B. thanks the financial support from the Hungarian Scientific Research Fund (OTKA NK 68578).

Abbreviations

AGV	aglyco-vancomycin
APME	additive potential maximum entropy
BIRD	bilinear rotation decoupling
$C_{12}E_5$	pentaethylene glycol mono-*n*-dodecyl ether
C_8E_5	pentaethylene glycol octyl ether
COSMO-HSQC	cosine modulated heteronuclear single quantum correlation
CT	constant time
DCPBV	des-*N*-methylleucyl-4-(4-chlorophenyl) benzyl-vancomycin
DEPT	distortionless enhanced polarization transfer
DFPBV	des-*N*-methylleucyl-4-(4-fluorophenyl) benzyl-vancomycin
DFT	density functional theory

DHPC	dihexanoyl-phosphatidylcholine
DMPC	dimyristoyl-phosphatidylcholine
DMSO	dimethyl-sulfoxide
DQF-COSY	double quantum filtered correlation spectroscopy
DV	des-*N*-methylleucyl-vancomycin
E.COSY	exclusive correlation spectroscopy
GDO	generalized degree of order
GISA	staphylococci with an intermediate resistance to glycopeptides
GlcNAc	*N*-acetyl-D-glucosamine
GMPP	GlcNAc-β(1-4)MurNAc-pentapeptide
GRE	glycopeptide-resistant enterococci
HCCH-COSY	3D experiment correlating H_a, C_a, and H_b in an ($–H_aC_a...C_bH_b–$) segment
HCCH-TOCSY	3D experiment correlating H_a, C_a, and H_b in an ($–H_aC_a...C_bH_b–$) segment
HCP	heteronuclear cross-polarization
HETLOC	pulse sequence for determination of heteronuclear long-range couplings
HMBC	heteronuclear multiple bond correlation
HMQC	heteronuclear multiple quantum correlation
HSQC	heteronuclear single quantum correlation
HSQMBC	heteronuclear single quantum multiple bond correlation
INADEQUATE	incredible natural abundance double quantum transfer experiment
IPAP	in-phase antiphase
ITC	isothermal titration calorimetry
LPS	lipopolysaccharide
LT	lytic transglycosylase
LTA	lipoteichoic acid
***m*Dap**	*meso*-diaminopimelic acid
MDP	MurNAc-dipeptide
MLT	membrane transglycosylase
MPP	MurNAc-pentapeptide
MRSA	methicillin-resistant *Staphylococcus aureus*
MurNAc	*N*-acetylmuramic acid
NAG	*N*-acetylglucosamine
NAM	*N*-acetylmuramic
NMR	nuclear magnetic resonance
NOD	nucleotide-binding oligomerization domain
NOESY	nuclear Overhauser effect spectroscopy
PAF	principle aligning frame
PALES	prediction of the alignment from structure
PAMP	pathogen-associated molecular pattern
PGM	peptidoglycan monomer
PGN	peptidoglycan
PGRP	peptidoglycan recognition protein
PRR	pattern recognition receptor
RDC	residual dipolar coupling
REDOR	rotational echo double resonance
ROESY	rotational nuclear Overhauser effect spectroscopy
S^3CT	spin-state-selective coherence transfer
SPITZE	spin state selective zero overlap
TCT	tracheal cytotoxin
TEDOR	transferred echo double resonance
TOCSY	total correlated spectroscopy

TRAMITE	tracking alignment from the moment of inertia tensor
TRNOE	transferred nuclear Overhauser enhancement
T-ROESY	tilted rotational nuclear Overhauser effect spectroscopy
VRSA	vancomycin-resistant *Staphylococcus aureus*

References

1. J. Dabrowski, Two-dimensional and Related NMR Methods in Structural Analyses of Oligosaccharides and Polisaccharides. In *Two-Dimensional NMR Spectroscopy. Applications for Chemists and Biochemists*; W. R. Croasmun, R. M. K. Carlson, Eds.; VCH: New York, 1994; pp 741–783.
2. L. E. Lerner, Carbohydrate Structure and Dynamics from NMR Spectroscopy and Its Application to Biomedical Research. In *NMR Spectroscopy and Its Applications to Biomedical Research*; S. K. Sarkar, Ed.; Elsevier Science B.V.: Amsterdam, 1996; pp 313–344.
3. C. A. Bush; M. Martin-Pastor; A. Imbery, *Annu. Rev. Biophys. Biomol. Struct.* **1999**, *28* (1), 269–293.
4. A. S. Serianni, Carbohydrates. In *Bioorganic Chemistry*; S. M. Hecht, Ed.; Oxford University Press: New York, 1999; pp 244–312.
5. J. O. Duus; C. H. Gotfredsen; K. Bock, *Chem. Rev.* **2000**, *100* (12), 4589.
6. W. A. Bubb, *J. Magn. Reson. Part A* **2003**, *19A* (1), 1–19.
7. Y. Kajihara; H. Sato, *Trends Glycosci. Glycotechnol.* **2003**, *15* (84), 197–220.
8. M. Hricovini, *Curr. Med. Chem.* **2004**, *11* (19), 2565–2583.
9. J. L. M. Jansson; A. Maliniak; G. Widmalm, Conformational Dynamics of Oligosaccharides: NMR Techniques and Computer Simulations. In *NMR Spectroscopy and Computer Modeling of Carbohydrates: Recent Advances*; J. F. G. Vliegenthart, R. J. Woods, Eds.; John Wiley & Sons: Chichester, 2006; Vol. 930, pp 20–39.
10. H. van Halbeek, *Curr. Opin. Struct. Biol.* **1994**, *4* (5), 697–709.
11. E. Fukushi, *Biosci. Biotechnol. Biochem.* **2006**, *70* (8), 1803–1812.
12. K. Furihata; H. Seto, *Tetrahedron Lett.* **1998**, *39* (40), 7337–7340.
13. A. Padilla; G. W. Vuister; R. Boelens; G. J. Kleywegt; A. Cave; J. Parello; R. Kaptein, *J. Am. Chem. Soc.* **1990**, *112* (13), 5024–5030.
14. J. N. Breg; R. Boelens; G. W. Vuister; R. Kaptein, *J. Magn. Reson.* **1990**, *87* (3), 646–651.
15. P. Dewaard; R. Boelens; G. W. Vuister; J. F. G. Vliegenthart, *J. Am. Chem. Soc.* **1990**, *112* (8), 3232–3234.
16. G. V. T. Swapna; R. Ramachandran, *J. Magn. Reson.* **1992**, *100* (1), 166–170.
17. D. Uhrin, *J. Magn. Reson.* **2002**, *159* (2), 145–150.
18. L. P. Yu; R. Goldman; P. Sullivan; G. F. Walker; S. W. Fesik, *J. Biomol. NMR* **1993**, *3* (4), 429–441.
19. P. T. Robinson; T. N. Pham; D. Uhrin, *J. Magn. Reson.* **2004**, *170* (1), 97–103.
20. H. Bircher; C. Muller; P. Bigler, *Magn. Reson. Chem.* **1991**, *29* (7), 726–729.
21. L. Poppe; H. Vanhalbeek, *J. Magn. Reson.* **1992**, *96* (1), 185–190.
22. D. Uhrin; J. R. Brisson; D. R. Bundle, *J. Biomol. NMR* **1993**, *3* (3), 367–373.
23. D. Uhrin; J. R. Brisson; G. Kogan; H. J. Jennings, *J. Magn. Reson. B* **1994**, *104* (3), 289–293.
24. M. J. Gradwell; H. Kogelberg; T. A. Frenkiel, *J. Magn. Reson.* **1997**, *124* (1), 267–270.
25. C. Roumestand; C. Delay; J. A. Gavin; D. Canet, *Magn. Reson. Chem.* **1999**, *37* (7), 451–478.
26. R. Laatikainen; M. Niemitz; U. Weber; J. Sundelin; T. Hassinen; J. Vepsalainen, *J. Magn. Reson. Ser. A* **1996**, *120* (1), 1–10.
27. J. J. Titman; J. Keeler, *J. Magn. Reson.* **1990**, *89* (3), 640–646.
28. K. E. Kövér; D. Uhrin; V. J. Hruby, *J. Magn. Reson.* **1998**, *130* (2), 162–168.
29. M. J. Thrippleton; J. Keeler, *Angew. Chem. Int. Ed. Engl.* **2003**, *42* (33), 3938–3941.
30. F. Rastrelli; A. Bagno, *J. Magn. Reson.* **2006**, *182* (1), 29–37.
31. P. Giraudeau; S. Akoka, *J. Magn. Reson.* **2007**, *186* (2), 352–357.
32. C. Zwahlen; S. J. F. Vincent, *J. Am. Chem. Soc.* **2002**, *124* (24), 7235–7239.
33. G. Zhu; A. Bax, *J. Magn. Reson. Ser. A* **1993**, *104* (3), 353–357.
34. B. L. Marquez; W. H. Gerwick; R. T. Williamson, *Magn. Reson. Chem.* **2001**, *39* (9), 499–530.
35. S. Uhrinova; D. Uhrin; T. Liptaj; J. Bella; J. Hirsch, *Magn. Reson. Chem.* **1991**, *29* (9), 912–922.
36. K. Fehér; S. Berger; K. E. Kövér, *J. Magn. Reson.* **2003**, *163* (2), 340–346.
37. M. D. Sorensen; A. Meissner; O. W. Sorensen, *J. Magn. Reson.* **1999**, *137* (1), 237–242.
38. F. Cordier; A. J. Dingley; S. Grzesiek, *J. Biomol. NMR* **1999**, *13* (2), 175–180.
39. M. Kurz; P. Schmieder; H. Kessler, *Angew. Chem. Int. Ed. Engl.* **1991**, *30* (10), 1329–1331.
40. G. Z. Xu; J. S. Evans, *J. Magn. Reson. Ser. A* **1996**, *123* (1), 105–110.
41. D. Uhrin; G. Batta; V. J. Hruby; P. N. Barlow; K. E. Kövér, *J. Magn. Reson.* **1998**, *130* (2), 155–161.
42. G. Z. Xu; B. Zhang; J. S. Evans, *J. Magn. Reson.* **1999**, *138* (1), 127–134.
43. W. Kozminski; D. Nanz, *J. Magn. Reson.* **1997**, *124* (2), 383–392.
44. K. E. Kövér; V. J. Hruby; D. Uhrin, *J. Magn. Reson.* **1997**, *129* (2), 125–129.
45. P. Nolis; T. Parella, *J. Magn. Reson.* **2005**, *176* (1), 15–26.
46. R. A. E. Edden; J. Keeler, *J. Magn. Reson.* **2004**, *166* (1), 53–68.
47. K. Kobzar; B. Luy, *J. Magn. Reson.* **2007**, *186* (1), 131–141.
48. A. Meissner; O. W. Sorensen, *Magn. Reson. Chem.* **2001**, *39* (1), 49–52.
49. T. Parella; J. Belloc; F. Sanchez-Ferrando, *Magn. Reson. Chem.* **2004**, *42* (10), 852–862.
50. P. Nolis; T. Parella, *Curr. Anal. Chem.* **2007**, *3* (1), 47–68.

51. L. Poppe; H. Vanhalbeek, *Magn. Reson. Chem.* **1991**, *29* (8), 848–851.
52. L. Poppe; H. Vanhalbeek, *J. Magn. Reson.* **1991**, *93* (1), 214–217.
53. L. Poppe; S. Q. Sheng; H. Vanhalbeek, *Magn. Reson. Chem.* **1994**, *32* (2), 97–100.
54. D. Uhrin; A. Mele; K. E. Kövér; J. Boyd; R. A. Dwek, *J. Magn. Reson. Ser. A* **1994**, *108* (2), 160–170.
55. K. E. Kövér; D. Jiao; D. Uhrin; P. Forgó; V. J. Hruby, *J. Magn. Reson. Ser. A* **1994**, *106* (1), 119–122.
56. T. Nishida; G. Widmalm; P. Sandor, *Magn. Reson. Chem.* **1995**, *33* (7), 596–599.
57. T. Rundlof; A. Kjellberg; C. Damberg; T. Nishida; G. Widmalm, *Magn. Reson. Chem.* **1998**, *36* (11), 839–847.
58. M. Findeisen; S. Berger, *Magn. Reson. Chem.* **2003**, *41* (6), 431–434.
59. P. Vidal; N. Esturau; T. Parella; J. F. Espinosa, *J. Org. Chem.* **2007**, *72* (9), 3166–3170.
60. R. T. Williamson; B. L. Marquez; W. H. Gerwick; K. E. Kövér, *Magn. Reson. Chem.* **2000**, *38* (4), 265–273.
61. H. Koskela; I. Kilpelainen; S. Heikkinen, *J. Magn. Reson.* **2003**, *164* (2), 228–232.
62. K. E. Kövér; G. Batta; K. Fehér, *J. Magn. Reson.* **2006**, *181* (1), 89–97.
63. R. Gitti; G. X. Long; C. A. Bush, *Biopolymers* **1994**, *34* (10), 1327–1338.
64. Q. W. Xu; S. Mohan; C. A. Bush, *Biopolymers* **1996**, *38* (3), 339–353.
65. J. Wu; A. S. Serianni, *Carbohydr. Res.* **1992**, *226* (2), 209–219.
66. J. M. Duker; A. S. Serianni, *Carbohydr. Res.* **1993**, *249* (2), 281–303.
67. T. E. Walker; R. E. London; T. W. Whaley; R. Barker; N. A. Matwiyoff, *J. Am. Chem. Soc.* **1976**, *98* (19), 5807–5813.
68. B. Bose; S. Zhao; R. Stenutz; F. Cloran; P. B. Bondo; G. Bondo; B. Hertz; I. Carmichael; A. S. Serianni, *J. Am. Chem. Soc.* **1998**, *120* (43), 11158–11173.
69. B. Bose-Basu; T. Klepach; G. Bondo; P. B. Bondo; W. Zhang; I. Carmichael; A. S. Serianni, *J. Org. Chem.* **2007**, *72* (20), 7511–7522.
70. U. Olsson; A. S. Serianni; R. Stenutz, *J. Phys. Chem. B* **2008**, *112* (14), 4447–4453.
71. A. Bax; D. Max; D. Zax, *J. Am. Chem. Soc.* **1992**, *114* (17), 6923–6925.
72. Q. W. Xu; C. A. Bush, *Carbohydr. Res.* **1998**, *306* (3), 335–339.
73. M. Martin-Pastor; A. Canales-Mayordomo; J. Jimenez-Barbero, *J. Biomol. NMR* **2003**, *26* (4), 345–353.
74. S. K. Zhao; G. Bondo; J. Zajicek; A. S. Serianni, *Carbohydr. Res.* **1998**, *309* (2), 145–152.
75. K. E. Kövér; P. Forgó, *J. Magn. Reson.* **2004**, *166* (1), 47–52.
76. T. N. Pham; K. E. Kövér; L. Jin; D. Uhrin, *J. Magn. Reson.* **2005**, *176* (2), 199–206.
77. L. Jin; D. Uhrin, *Magn. Reson. Chem.* **2007**, *45* (8), 628–633.
78. L. Jin; K. E. Kövér; M. R. Lenoir; D. Uhrin, *J. Magn. Reson.* **2008**, *190* (2), 171–182.
79. S. W. Homans, *Prog. Nucl. Magn. Reson. Spectrosc.* **1990**, *22*, 55–81.
80. A. S. Serianni, Nuclear Magnetic Resonance Approaches to Oligosaccharide Structure Elucidation. In *Glycoconjugates*; H. Allen, E. C. Kisalius, Eds.; Marcel Dekker: New York, 1992; pp 71–102.
81. M. Eberstadt; G. Gemmecker; D. F. Mierke; H. Kessler, *Angew. Chem. Int. Ed. Engl.* **1995**, *34* (16), 1671–1695.
82. C. Altona, *Vicinal Coupling Constants and Conformation of Biomolecules*. John Wiley: London, 1996.
83. W. A. Thomas, *Prog. Nucl. Magn. Reson. Spectrosc.* **1997**, *30* (3–4), 183–207.
84. R. H. Contreras; J. E. Peralta; C. G. Giribet; M. C. De Azua; J. C. Facelli, *Annu. Rep. NMR Spectrosc.* **2000**, *41*, 55–184.
85. G. E. Martin, Qualitative and Quantitative Exploitation of Heteronuclear Coupling Constants. In *Annual Reports on NMR Spectroscopy*; G. A. Webb, Ed.; Academic Press: New York, 2002; Vol. 46, pp 37–100.
86. M. Kraszni; Z. Szakacs; B. Noszal, *Anal. Bioanal. Chem.* **2004**, *378* (6), 1449–1463.
87. B. Mulloy; T. A. Frenkiel; D. B. Davies, *Carbohydr. Res.* **1988**, *184*, 39–46.
88. I. Tvaroska; M. Hricovini; E. Petrakova, *Carbohydr. Res.* **1989**, *189*, 359–362.
89. M. J. Milton; R. Harris; M. A. Probert; R. A. Field; S. W. Homans, *Glycobiology* **1998**, *8* (2), 147–153.
90. F. Cloran; I. Carmichael; A. S. Serianni, *J. Am. Chem. Soc.* **1999**, *121* (42), 9843–9851.
91. M. Martin-Pastor; C. A. Bush, *Biochemistry* **1999**, *38* (25), 8045–8055.
92. M. Karplus, *J. Chem. Phys.* **1959**, *30* (1), 11–15.
93. F. Cloran; I. Carmichael; A. S. Serianni, *J. Am. Chem. Soc.* **2000**, *122* (2), 396–397.
94. H. Q. Zhao; I. Carmichael; A. S. Serianni, *J. Org. Chem.* **2008**, *73* (8), 3255–3257.
95. A. S. Serianni; P. B. Bondo; J. Zajicek, *J. Magn. Reson. B.* **1996**, *112* (1), 69–74.
96. T. Church; I. Carmichael; A. S. Serianni, *Carbohydr. Res.* **1996**, *280* (2), 177–186.
97. I. Tvaroska; F. R. Taravel, *J. Biomol. NMR* **1992**, *2* (5), 421–430.
98. I. Tvaroska; F. R. Taravel, *Adv. Carbohydr. Chem. Biochem.* **1995**, *51*, 15–61.
99. I. Carmichael; D. M. Chipman; C. A. Podlasek; A. S. Serianni, *J. Am. Chem. Soc.* **1993**, *115* (23), 10863–10870.
100. T. E. Klepach; I. Carmichael; A. S. Serianni, *J. Am. Chem. Soc.* **2005**, *127* (27), 9781–9793.
101. S. Ilin; C. Bosques; C. Turner; H. Schwalbe, *Angew. Chem. Int. Ed. Engl.* **2003**, *42* (12), 1394–1397.
102. Z. Dzakula; W. M. Westler; A. S. Edison; J. L. Markley, *J. Am. Chem. Soc.* **1992**, *114* (15), 6195–6199.
103. Z. Dzakula; A. S. Edison; W. M. Westler; J. L. Markley, *J. Am. Chem. Soc.* **1992**, *114* (15), 6200–6207.
104. L. Poppe, *J. Am. Chem. Soc.* **1993**, *115* (18), 8421–8426.
105. K. Bock; J. O. Duus, *J. Carbohydr. Chem.* **1994**, *13* (4), 513–543.
106. G. D. Rockwell; T. B. Grindley, *J. Am. Chem. Soc.* **1998**, *120* (42), 10953–10963.
107. Y. Nishida; H. Hori; H. Ohrui; H. Meguro, *J. Carbohydr. Chem.* **1988**, *7* (1), 239–250.
108. H. Hori; Y. Nishida; H. Ohrui; H. Meguro, *J. Carbohydr. Chem.* **1990**, *9* (5), 601–618.
109. R. Stenutz; I. Carmichael; G. Widmalm; A. S. Serianni, *J. Org. Chem.* **2002**, *67* (3), 949–958.
110. A. Roen; J. I. Padron; J. T. Vazquez, *J. Org. Chem.* **2003**, *68* (12), 4615–4630.
111. C. Nobrega; J. T. Vazquez, *Tetrahedron: Asymmetry* **2003**, *14* (18), 2793–2801.
112. C. Mayato; R. Dorta; J. Vázquez, *Tetrahedron: Asymmetry* **2004**, *15* (15), 2385–2397.
113. I. Tvaroska; J. Gajdos, *Carbohydr. Res.* **1995**, *271* (2), 151–162.
114. I. Tvaroska; F. R. Taravel; J. P. Utille; J. P. Carver, *Carbohydr. Res.* **2002**, *337* (4), 353–367.

115. C. Thibaudeau; R. Stenutz; B. Hertz; T. Klepach; S. Zhao; Q. Q. Wu; I. Carmichael; A. S. Serianni, *J. Am. Chem. Soc.* **2004**, *126* (48), 15668–15685.
116. M. Tafazzoli; M. Grhiasi, *Carbohydr. Res.* **2007**, *342* (14), 2086–2096.
117. G. A. S. W. Jeffrey, *Hydrogen Bond in Biological Structures*. Springer-Verlag: Berlin, 1991.
118. H. C. Siebert; M. Frank; C. W. Lieth; J. Jimenez-Barbero; H. J. Gabius, Detection of Hydroxyl Protons. In *NMR Spectroscopy of Glycoconjugates*; J. Jiménez-Barbero, T. Peters, Eds.; Wiley-VCH: Weinheim, 2003.
119. C. Sandstrom; L. Kenne, Hydroxy Protons in Structural Studies of Carbohydrates by NMR Spectroscopy. In *NMR Spectroscopy and Computer Modeling of Carbohydrates: Recent Advances*; J. F. G. Vliegenthart, R. J. Woods, Eds.; John Wiley & Sons: Chichester, 2006; Vol. 930; pp 114.
120. J. Dabrowski; H. Grosskurth; C. Baust; N. E. Nifant'ev, *J. Biomol. NMR* **1998**, *12* (1), 161–172.
121. C. E. Anderson; A. J. Pickrell; S. L. Sperry; T. E. Vasquez; T. G. Custer; M. B. Fierman; D. C. Lazar; Z. W. Brown; W. S. Iskenderian; D. D. Hickstein; D. J. O'Leary, *Heterocycles* **2007**, *72*, 469–495.
122. T. T. Nguyen; T. N. Le; F. Duus; B. K. V. Hansen; P. E. Hansen, *Magn. Reson. Chem.* **2007**, *45* (3), 245–252.
123. P. E. Hansen, *J. Labelled Comp. Radiopharm.* **2007**, *50* (11–12), 967–981.
124. C. Sandstrom; H. Baumann; L. Kenne, *J. Chem. Soc., Perkin Trans. 2* **1998**, (4), 809–815.
125. C. Sandstrom; H. Baumann; L. Kenne, *J. Chem. Soc., Perkin Trans. 2* **1998**, (11), 2385–2393.
126. H. Q. Zhao; Q. F. Pan; W. H. Zhang; I. Carmichael; A. S. Serianni, *J. Org. Chem.* **2007**, *72* (19), 7071–7082.
127. B. Adams; L. Lerner, *J. Am. Chem. Soc.* **1992**, *114* (12), 4827–4829.
128. H. C. Siebert; S. Andre; J. F. G. Vliegenthart; H. J. Gabius; M. J. Minch, *J. Biomol. NMR* **2003**, *25* (3), 197–215.
129. L. Poppe; H. Vanhalbeek, *Nat. Struct. Biol.* **1994**, *1* (4), 215–216.
130. B. Bernet; A. Vasella, *Helv. Chim. Acta* **2000**, *83* (5), 995–1021.
131. R. R. Fraser; M. Kaufman; P. Morand; G. Govil, *Can. J. Chem.* **1969**, *47* (3), 403–409.
132. K. G. R. Pachler, *Tetrahedron* **1971**, *27* (1), 187.
133. C. A. G. Haasnoot; F. Deleeuw; C. Altona, *Tetrahedron* **1980**, *36* (19), 2783–2792.
134. H. Fukui; T. Baba; H. Inomata; K. Miura; H. Matsuda, *Mol. Phys.* **1997**, *92* (1), 161–165.
135. L. Alkorta; J. Elguero, *Theor. Chem. Acc.* **2004**, *111* (1), 31–35.
136. P. Dais; A. S. Perlin, *Can. J. Chem.* **1982**, *60* (13), 1648–1656.
137. G. Batta; K. E. Kövér, *Carbohydr. Res.* **1999**, *320* (3–4), 267–272.
138. K. E. Kövér; A. Lipták; T. Beke; A. Perczel, *J. Comput. Chem.* **2009**, *30* (4), 540–550.
139. M. J. Frisch, *et al., Gaussian03*, rev. D.01.; Gaussian Inc.: Wallingford, CT: 2004.
140. N. C. Maiti; Y. P. Zhu; I. Carmichael; A. S. Serianni; V. E. Anderson, *J. Org. Chem.* **2006**, *71* (7), 2878–2880.
141. A. Saupe, *Angew. Chem. Int. Ed. Engl.* **1968**, *7* (2), 97.
142. J. W. Emsley; J. C. Lindon, *NMR Spectroscopy Using Liquid Crystal Solvents*; Pergamon: Oxford, 1975.
143. R. Y. Dong, *Nuclear Magnetic Resonance of Liquid Crystals*; Springer: New York, 1994.
144. J. W. Emsley, In *Encyclopedia of Nuclear Magnetic Resonance*; D. M. Grant, R. K. Harris, Eds.; Wiley: Chichester, 1996; pp 2788–2799.
145. C. Algieri; F. Castiglione; G. Celebre; G. De Luca; M. Longeri; J. W. Emsley, *Phys. Chem. Chem. Phys.* **2000**, *2* (15), 3405–3413.
146. J. H. Prestegard; H. M. Al-Hashimi; J. R. Tolman, *Q. Rev. Biophys.* **2000**, *33* (4), 371–424.
147. A. Almond; J. Bunkenborg; T. Franch; C. H. Gotfredsen; J. O. Duus, *J. Am. Chem. Soc.* **2001**, *123* (20), 4792–4802.
148. A. Almond; J. O. Duus, *J. Biomol. NMR* **2001**, *20* (4), 351–363.
149. F. Tian; H. M. Al-Hashimi; J. L. Craighead; J. H. Prestegard, *J. Am. Chem. Soc.* **2001**, *123* (3), 485–492.
150. A. Almond; B. O. Petersen; J. O. Duus, *Biochemistry* **2004**, *43* (19), 5853–5863.
151. T. N. Pham; T. Liptaj; K. Bromek; D. Uhrin, *J. Magn. Reson.* **2002**, *157* (2), 200–209.
152. X. B. Yi; A. Venot; J. Glushka; J. H. Prestegard, *J. Am. Chem. Soc.* **2004**, *126* (42), 13636–13638.
153. T. N. Pham; S. L. Hinchley; D. W. H. Rankin; T. Liptaj; D. Uhrin, *J. Am. Chem. Soc.* **2004**, *126* (40), 13100–13110.
154. L. Jin; T. N. Pham; D. Uhrin, *ChemPhysChem* **2007**, *8* (8), 1228–1235.
155. C. Landersjö; J. L. M. Jansson; A. Maliniak; G. Widmalm, *J. Phys. Chem. B* **2005**, *109* (36), 17320–17326.
156. M. Martin-Pastor; C. A. Bush, *J. Biomol. NMR* **2001**, *19* (2), 125–139.
157. K. Lycknert; A. Maliniak; G. Widmalm, *J. Phys. Chem. A* **2001**, *105* (21), 5119–5122.
158. M. Martin-Pastor; A. Canales; F. Corzana; J. L. Asensio; J. Jimenez-Barbero, *J. Am. Chem. Soc.* **2005**, *127* (10), 3589–3595.
159. M. Martin-Pastor; C. A. Bush, *Biochemistry* **2000**, *39* (16), 4674–4683.
160. B. Stevensson; C. Landersjo; G. Widmalm; A. Maliniak, *J. Am. Chem. Soc.* **2002**, *124* (21), 5946–5947.
161. D. I. Freedberg, *J. Am. Chem. Soc.* **2002**, *124* (10), 2358–2362.
162. H. F. Azurmendi; C. A. Bush, *J. Am. Chem. Soc.* **2002**, *124* (11), 2426–2427.
163. R. M. Venable; F. Delaglio; S. E. Norris; D. I. Freedberg, *Carbohydr. Res.* **2005**, *340* (5), 863–874.
164. C. Landersjö; B. Stevensson; R. Eklund; J. Ostervall; P. Soderman; G. Widmalm; A. Maliniak, *J. Biomol. NMR* **2006**, *35* (2), 89–101.
165. P. Berthault; D. Jeannerat; F. Camerel; F. A. Salgado; Y. Boulard; J. C. P. Gabriel; H. Desvaux, *Carbohydr. Res.* **2003**, *338* (17), 1771–1785.
166. R. M. Gschwind, *Angew. Chem. Int. Ed. Engl.* **2005**, *44* (30), 4666–4668.
167. J. L. Yan; E. R. Zartler, *Magn. Reson. Chem.* **2005**, *43* (1), 53–64.
168. C. M. Thiele, *Concepts Magn. Reson. Part A* **2007**, *30A* (2), 65–80.
169. S. Sykora; J. Vogt; H. Bosiger; P. Diehl, *J. Magn. Reson.* **1979**, *36* (1), 53–60.
170. H. Neubauer; J. Meiler; W. Peti; C. Griesinger, *Helv. Chim. Acta* **2001**, *84* (1), 243–258.
171. F. Delaglio; Z. R. Wu; A. Bax, *J. Magn. Reson.* **2001**, *149* (2), 276–281.
172. W. Willker; D. Leibfritz, *J. Magn. Reson.* **1992**, *99* (2), 421–425.
173. M. H. Lerche; A. Meissner; F. M. Poulsen; O. W. Sorensen, *J. Magn. Reson.* **1999**, *140* (1), 259–263.
174. T. S. Untidt; T. Schulte-Herbruggen; O. W. Sorensen; N. C. Nielsen, *J. Phys. Chem. A* **1999**, *103* (45), 8921–8926.
175. K. E. Kövér; K. Fehér, *J. Magn. Reson.* **2004**, *168* (2), 307–313.
176. G. Otting; M. Ruckert; M. H. Levitt; A. Moshref, *J. Biomol. NMR* **2000**, *16* (4), 343–346.

177. F. Tian; P. J. Bolon; J. H. Prestegard, *J. Am. Chem. Soc.* **1999**, *121* (33), 7712–7713.
178. T. N. Pham; T. Liptaj; P. N. Barlow; D. Uhrin, *Magn. Reson. Chem.* **2002**, *40* (11), 729–732.
179. T. Rundlof; C. Landersjo; K. Lycknert; A. Maliniak; G. Widmalm, *Magn. Reson. Chem.* **1998**, *36* (10), 773–776.
180. D. Uhrin; T. Liptaj; K. E. Kövér, *J. Magn. Reson. Ser. A* **1993**, *101* (1), 41–46.
181. N. Tjandra; A. Bax, *J. Magn. Reson.* **1997**, *124* (2), 512–515.
182. G. R. Kiddle; S. W. Homans, *FEBS Lett.* **1998**, *436* (1), 128–130.
183. J. R. Garbow; D. P. Weitekamp; A. Pines, *Chem. Phys. Lett.* **1982**, *93* (5), 504–509.
184. G. Nodet; L. Poggi; D. Abergel; C. Gourmala; D. X. Dong; Y. M. Zhang; J. M. Mallet; G. Bodenhausen, *J. Am. Chem. Soc.* **2007**, *129* (29), 9080–9085.
185. J. A. Losonczi; M. Andrec; M. W. F. Fischer; J. H. Prestegard, *J. Magn. Reson.* **1999**, *138* (2), 334–342.
186. H. Valafar; J. H. Prestegard, *J. Magn. Reson.* **2004**, *167* (2), 228–241.
187. M. Deschamps; I. D. Campbell; J. Boyd, *J. Magn. Reson.* **2005**, *172* (1), 118–132.
188. M. Zweckstetter; A. Bax, *J. Am. Chem. Soc.* **2000**, *122* (15), 3791–3792.
189. J. H. Prestegard; X. B. Yi, Structure and Dynamics of Carbohydrates Using Residual Dipolar Couplings. In *NMR Spectroscopy and Computer Modeling of Carbohydrates: Recent Advances*; J. F. G. Vliegenthart, R. J. Woods, Eds.; John Wiley & Sons: Chichester, 2006; Vol. 930, pp 40–59.
190. E. T. Samulski; R. Y. Dong, *J. Chem. Phys.* **1982**, *77* (10), 5090–5096.
191. C. Landersjö; C. Hoog; A. Maliniak; G. Widmalm, *J. Phys. Chem. B* **2000**, *104* (23), 5618–5624.
192. A. Almond; J. B. Axelsen, *J. Am. Chem. Soc.* **2002**, *124* (34), 9986–9987.
193. D. I. Freedberg; S. O. Ano; S. E. Norris; R. M. Venable, Carbohydrate Structure from NMR Residual Dipolar Couplings: Is There a Correlation between Lactose's Anomeric Configuration and Its Three-Dimensional Structure? In: In *NMR Spectroscopy and Computer Modeling of Carbohydrates: Recent Advances*; J. F. G. Vliegenthart, R. J. Woods, Eds.; John Wiley & Sons: Chichester, 2006; Vol. 930, pp 220–234.
194. J. W. Emsley; G. R. Luckhurst; C. P. Stockley, *Proc. Math. Phys. Eng. Sci.* **1982**, *381* (1780), 117–138.
195. D. Catalano; L. Dibari; C. A. Veracini; G. N. Shilstone; C. Zannoni, *J. Chem. Phys.* **1991**, *94* (5), 3928–3935.
196. H. J. Gabius; H. C. Siebert; S. Andre; J. Jimenez-Barbero; H. Rudiger, *ChemBioChem* **2004**, *5* (6), 740–764.
197. D. B. Moody, *Nature* **2007**, *448* (7149), 36–37.
198. M. A. Johnson; B. M. Pinto, *Carbohydr. Res.* **2004**, *339* (5), 907–928.
199. H. Kogelberg; D. Solis; J. Jimenez-Barbero, *Curr. Opin. Struct. Biol.* **2003**, *13* (5), 646–653.
200. J. Jiménez-Barbero; T. Peters, *NMR Spectroscopy of Glycoconjugates*; Wiley-VCH: Weinheim, 2002.
201. S. Banerji; A. J. Wright; M. Noble; D. J. Mahoney; I. D. Campbell; A. J. Day; D. G. Jackson, *Nat. Struct. Mol. Biol.* **2007**, *14* (3), 234–239.
202. M. A. Walti; P. J. Walser; S. Thore; A. Grunler; M. Bednar; M. Kunzler; M. Aebi, *J. Mol. Biol.* **2008**, *379* (1), 146–159.
203. U. Neu; K. Woellner; G. Gauglitz; T. Stehle, *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105* (13), 5219–5224.
204. M. R. Wormald; A. J. Petrescu; Y. L. Pao; A. Glithero; T. Elliott; R. A. Dwek, *Chem. Rev.* **2002**, *102* (2), 371–386.
205. A. Imberty; S. Perez, *Chem. Rev.* **2000**, *100* (12), 4567–4588.
206. S. Perez; A. Imberty; S. B. Engelsen; J. Gruza; K. Mazeau; J. Jimenez-Barbero; A. Poveda; J. F. Espinosa; B. P. van Eyck; G. Johnson; A. D. French; M. Louise; C. E. Kouwijzer; P. D. J. Grootenuis; A. Bernardi; L. Raimondi; H. Senderowitz; V. Durier; G. Vergoten; K. Rasmussen, *Carbohydr. Res.* **1998**, *314* (3–4), 141–155.
207. O. Coskuner, *J. Chem. Phys.* **2007**, *127* (1), 015101.
208. M. K. Dowd; P. J. Reilly; A. D. French, *Biopolymers* **1994**, *34* (5), 625–638.
209. F. Corzana; I. Cuesta; F. Freire; J. Revuelta; M. Torrado; A. Bastida; J. Jimenez-Barbero; J. L. Asensio, *J. Am. Chem. Soc.* **2007**, *129* (10), 2849–2865.
210. K. N. Kirschner; A. B. Yongye; S. M. Tschampel; J. Gonzalez-Outeirino; C. R. Daniels; B. L. Foley; R. J. Woods, *J. Comput. Chem.* **2008**, *29* (4), 622–655.
211. J. L. Asensio; M. Martin-Pastor; J. Jimenez-Barbero, *Int. J. Biol. Macromol.* **1995**, *17* (3–4), 137–148.
212. A. D. French; A. M. Kelterer; G. P. Johnson; M. K. Dowd; C. J. Cramer, *J. Mol. Graph. Model.* **2000**, *18* (2), 95–107.
213. B. Lopez-Mendez; C. Jia; Y. Zhang; L. H. Zhang; P. Sinay; J. Jimenez-Barbero; M. Sollogoub, *Chem. Asian J.* **2008**, *3* (1), 51–58.
214. A. Silipo; Z. Zhang; F. J. Canada; A. Molinaro; R. J. Linhardt; J. Jimenez-Barbero, *ChemBioChem* **2008**, *9* (2), 240–252.
215. A. Poveda; J. L. Asensio; M. Martin-Pastor; J. Jimenez-Barbero, *Carbohydr. Res.* **1997**, *300* (1), 3–10.
216. J. Jimenez-Barbero; J. L. Asensio; F. J. Canada; A. Poveda, *Curr. Opin. Struct. Biol.* **1999**, *9* (5), 549–555.
217. J. L. Asensio; J. Jimenez-Barbero, *Biopolymers* **1995**, *35* (1), 55–73.
218. D. A. Cumming; J. P. Carver, *Biochemistry* **1987**, *26* (21), 6664–6676.
219. J. L. Asensio; A. Hidalgo; I. Cuesta; C. Gonzalez; J. Canada; C. Vicent; J. L. Chiara; G. Cuevas; J. Jimenez-Barbero, *Chem. Commun. (Camb.)* **2002** (19), 2232–2233.
220. J. L. Asensio; F. J. Canada; X. Cheng; N. Khan; D. R. Mootoo; J. Jimenez-Barbero, *Chemistry* **2000**, *6* (6), 1035–1041.
221. D. A. Cumming; R. N. Shah; J. J. Krepinsky; A. A. Grey; J. P. Carver, *Biochemistry* **1987**, *26* (21), 6655–6663.
222. H.-C. Siebert; M. Frank; C.-W. von der Lieth; J. Jiménez-Barbero; H.-J. Gabius, Detection of Hydroxyl Protons. In *NMR Spectroscopy of Glycoconjugates*; J. Jiménez-Barbero, T. Peters, Eds.; Wiley-VCH: Weinheim, 2002; pp 39–57.
223. J. Dabrowski; T. Kozar; H. Grosskurth; N. E. Nifantev, *J. Am. Chem. Soc.* **1995**, *117* (20), 5534–5539.
224. R. Eklund; K. Lycknert; P. Soderman; G. Widmalm, *J. Phys. Chem. B* **2005**, *109* (42), 19936–19945.
225. J. L. Asensio; A. Hidalgo; I. Cuesta; C. Gonzalez; J. Canada; C. Vicent; J. L. Chiara; G. Cuevas; J. Jimenez-Barbero, *Chemistry* **2002**, *8* (22), 5228–5240.
226. A. Poveda; J. L. Asensio; M. Martin-Pastor; J. Jimenez-Barbero, *Chem. Commun.*421–422.
227. M. Hricovini; R. N. Shah; J. P. Carver, *Biochemistry* **1992**, *31* (41), 10018–10023.
228. A. Poveda; M. Santamaria; M. Bernabe; A. Prieto; M. Bruix; J. Corzo; J. Jimenez-Barbero, *Carbohydr. Res.* **1997**, *304* (3–4), 209–217.
229. M. Mackeen; A. Almond; I. Cumpstey; S. C. Enis; E. Kupce; T. D. Butters; A. J. Fairbanks; R. A. Dwek; M. R. Wormald, *Org. Biomol. Chem.* **2006**, *4* (11), 2241–2246.
230. A. Andersson; A. Ahl; R. Eklund; G. Widmalm; L. Maler, *J. Biomol. NMR* **2005**, *31* (4), 311–320.
231. A. M. Dixon; R. Venable; G. Widmalm; T. E. Bull; R. W. Pastor, *Biopolymers* **2003**, *69* (4), 448–460.

232. C. Hoog; C. Landersjo; G. Widmalm, *Chemistry* **2001**, *7* (14), 3069–3077.
233. A. Poveda; J. L. Asensio; M. Martin-Pastor; J. Jimenez-Barbero, *J. Biomol. NMR* **1997**, *10* (1), 29–43.
234. L. Maler; G. Widmalm; J. Kowalewski, *J. Biomol. NMR* **1996**, *7* (1), 1–7.
235. A. Poveda; M. Martin-Pastor; M. Bernabe; J. A. Leal; J. Jimenez-Barbero, *Glycoconj. J.* **1998**, *15* (3), 309–321.
236. K. Lycknert; G. Widmalm, *Biomacromolecules* **2004**, *5* (3), 1015–1020.
237. B. Mulloy; M. J. Forster, *Glycobiology* **2000**, *10* (11), 1147–1156.
238. J. Angulo; M. Hricovini; M. Gairi; M. Guerrini; J. L. de Paz; R. Ojeda; M. Martin-Lomas; P. M. Nieto, *Glycobiology* **2005**, *15* (10), 1008–1015.
239. D. R. Ferro; A. Provasoli; M. Ragazzi; B. Casu; G. Torri; V. Bossennec; B. Perly; P. Sinay; M. Petitou; J. Choay, *Carbohydr. Res.* **1990**, *195* (2), 157–167.
240. J. Angulo; P. M. Nieto; M. Martin-Lomas, *Chem. Commun.* **2003** (13), 1512–1513.
241. D. Mikhailov; R. J. Linhardt; K. H. Mayo, *Biochem. J.* **1997**, *328* (Pt 1), 51–61.
242. G. Torri; B. Casu; G. Gatti; M. Petitou; J. Choay; J. C. Jacquinet; P. Sinay, *Biochem. Biophys. Res. Commun.* **1985**, *128* (1), 134–140.
243. D. Acquotti; L. Poppe; J. Dabrowski; C. W. Vonderlieth; S. Sonnino; G. Tettamanti, *J. Am. Chem. Soc.* **1990**, *112* (21), 7772–7778.
244. L. Poppe; H. van Halbeek; D. Acquotti; S. Sonnino, *Biophys. J.* **1994**, *66* (5), 1642–1652.
245. B. G. Winsborrow; J. R. Brisson; I. C. Smith; H. C. Jarrell, *Biophys. J.* **1992**, *63* (2), 428–437.
246. K. P. Howard; J. H. Prestegard, *Biophys. J.* **1996**, *71* (5), 2573–2582.
247. J. H. Prestegard; J. Glushka, Residual Dipolar Couplings: Structure and Dynamics of Glycolipids. In *NMR Spectroscopy of Glycoconjugates*; J. Jimenez Barbero, T. Peters, Eds.; Wiley-VCH: Weinheim, 2002; pp 231–245.
248. J. J. Hernandez-Gay; L. Panza; F. Ronchetti; F. J. Canada; F. Compostella; J. Jimenez-Barbero, *Carbohydr. Res.* **2007**, *342* (12–13), 1966–1973.
249. F. Chevalier; J. Lopez-Prados; P. Groves; S. Perez; M. Martin-Lomas; P. M. Nieto, *Glycobiology* **2006**, *16* (10), 969–980.
250. J. Revuelta; T. Vacas; M. Torrado; F. Corzana; C. Gonzalez; J. Jimenez-Barbero; M. Menendez; A. Bastida; J. L. Asensio, *J. Am. Chem. Soc.* **2008**, *130* (15), 5086–5103.
251. F. Freire; I. Cuesta; F. Corzana; J. Revuelta; C. Gonzalez; M. Hricovini; A. Bastida; J. Jimenez-Barbero; J. L. Asensio, *Chem. Commun. (Camb.)*174–176.
252. B. N. Rao; C. A. Bush, *Biopolymers* **1987**, *26* (8), 1227–1244.
253. A. Kuhn; H. Kunz, *Angew. Chem. Int. Ed. Engl.* **2007**, *46* (3), 454–458.
254. F. Corzana; J. H. Busto; S. B. Engelsen; J. Jimenez-Barbero; J. L. Asensio; J. M. Peregrina; A. Avenoza, *Chemistry* **2006**, *12* (30), 7864–7871.
255. F. Corzana; J. H. Busto; G. Jimenez-Oses; M. Garcia de Luis; J. L. Asensio; J. Jimenez-Barbero; J. M. Peregrina; A. Avenoza, *J. Am. Chem. Soc.* **2007**, *129* (30), 9458–9467.
256. F. Corzana; J. H. Busto; G. Jimenez-Oses; J. L. Asensio; J. Jimenez-Barbero; J. M. Peregrina; A. Avenoza, *J. Am. Chem. Soc.* **2006**, *128* (45), 14640–14648.
257. D. M. Coltart; A. K. Royyuru; L. J. Williams; P. W. Glunz; D. Sames; S. D. Kuduk; J. B. Schwarz; X. T. Chen; S. J. Danishefsky; D. H. Live, *J. Am. Chem. Soc.* **2002**, *124* (33), 9833–9844.
258. L. Kinarsky; G. Suryanarayanan; O. Prakash; H. Paulsen; H. Clausen; F. G. Hanisch; M. A. Hollingsworth; S. Sherman, *Glycobiology* **2003**, *13* (12), 929–939.
259. Y. Tachibana; G. L. Fletcher; N. Fujitani; S. Tsuda; K. Monde; S. Nishimura, *Angew. Chem. Int. Ed. Engl.* **2004**, *43* (7), 856–862.
260. S. E. O'Connor; J. Pohlmann; B. Imperiali; I. Saskiawan; K. Yamamoto, *J. Am. Chem. Soc.* **2001**, *123* (25), 6187–6188.
261. C. J. Bosques; S. M. Tschampel; R. J. Woods; B. Imperiali, *J. Am. Chem. Soc.* **2004**, *126* (27), 8421–8425.
262. S. E. O'Connor; B. Imperiali, *Chem. Biol.* **1996**, *3* (10), 803–812.
263. M. M. Palian; V. I. Boguslavsky; D. F. O'Brien; R. Polt, *J. Am. Chem. Soc.* **2003**, *125* (19), 5823–5831.
264. C. Vanhaverbeke; J. P. Simorre; R. Sadir; P. Gans; H. Lortat-Jacob, *Biochem. J.* **2004**, *384* (Pt 1), 93–99.
265. N. Aboitiz; M. Vila-Perello; P. Groves; J. L. Asensio; D. Andreu; F. J. Canada; J. Jimenez-Barbero, *ChemBioChem* **2004**, *5* (9), 1245–1255.
266. D. C. Williams, Jr.; M. Cai; J. Y. Suh; A. Peterkofsky; G. M. Clore, *J. Biol. Chem.* **2005**, *280* (21), 20775–20784.
267. A. Canales-Mayordomo; R. Fayos; J. Angulo; R. Ojeda; M. Martin-Pastor; P. M. Nieto; M. Martin-Lomas; R. Lozano; G. Gimenez-Gallego; J. Jimenez-Barbero, *J. Biomol. NMR* **2006**, *35* (4), 225–239.
268. S. C. Tjong; T. S. Chen; W. N. Huang; W. G. Wu, *Biochemistry* **2007**, *46* (35), 9941–9952.
269. C. D. Blundell; A. Almond; D. J. Mahoney; P. L. DeAngelis; I. D. Campbell; A. J. Day, *J. Biol. Chem.* **2005**, *280* (18), 18189–18201.
270. K. W. Hung; T. K. Kumar; K. M. Kathir; P. Xu; F. Ni; H. H. Ji; M. C. Chen; C. C. Yang; F. P. Lin; I. M. Chiu; C. Yu, *Biochemistry* **2005**, *44* (48), 15787–15798.
271. A. P. Herbert; J. A. Deakin; C. Q. Schmidt; B. S. Blaum; C. Egan; V. P. Ferreira; M. K. Pangburn; M. Lyon; D. Uhrin; P. N. Barlow, *J. Biol. Chem.* **2007**, *282* (26), 18960–18968.
272. A. Canales; R. Lozano; B. Lopez-Mendez; J. Angulo; R. Ojeda; P. M. Nieto; M. Martin-Lomas; G. Gimenez-Gallego; J. Jimenez-Barbero, *FEBS J.* **2006**, *273* (20), 4716–4727.
273. L. M. Koharudin; A. R. Viscomi; J. G. Jee; S. Ottonello; A. M. Gronenborn, *Structure* **2008**, *16* (4), 570–584.
274. C. A. Bewley; S. Kiyonaka; I. Hamachi, *J. Mol. Biol.* **2002**, *322* (4), 881–889.
275. A. A. Bothnerb; R. Gassend, *Ann. N. Y. Acad. Sci.* **1973**, *222* (DEC31), 668–676.
276. J. P. Albrand; B. Birdsall; J. Feeney; G. C. K. Roberts; A. S. V. Burgen, *Int. J. Biol. Macromol.* **1979**, *1* (1), 37–41.
277. V. L. Bevilacqua; D. S. Thomson; J. H. Prestegard, *Biochemistry* **1990**, *29* (23), 5529–5537.
278. V. L. Bevilacqua; Y. Kim; J. H. Prestegard, *Biochemistry* **1992**, *31* (39), 9339–9349.
279. A. Poveda; J. Jimenez-Barbero, *Chem. Soc. Rev.* **1998**, *27* (2), 133–143.
280. J. Jiménez-Barbero; T. Peters, TR-NOE Experiments to Study Carbohydrate-Protein Interactions. In *NMR Spectroscopy of Glycoconjugates*; J. Jiménez-Barbero, T. Peters, Eds.; Wiley-VCH: Weinheim, 2002; pp 289–309.
281. J. Angulo; C. Rademacher; T. Biet; A. J. Benie; A. Blume; H. Peters; M. Palcic; F. Parra; T. Peters, *Meth. Enzymol.* **2006**, *416*, 12–30.

282. J. L. Asensio; F. J. Canada; J. Jimenez-Barbero, *Eur. J. Biochem.* **1995**, *233* (2), 618–630.
283. A. Poveda; J. L. Asensio; J. F. Espinosa; M. Martin-Pastor; J. Canada; J. Jimenez-Barbero, *J. Mol. Graph. Model.* **1997**, *15* (1), 9–17, 53.
284. A. Germer; C. Mugge; M. G. Peter; A. Rottmann; E. Kleinpeter, *Chemistry* **2003**, *9* (9), 1964–1973.
285. H. C. Siebert; S. Andre; S. Y. Lu; M. Frank; H. Kaltner; J. A. van Kuik; E. Y. Korchagina; N. Bovin; E. Tajkhorshid; R. Kaptein; J. F. Vliegenthart; C. W. von der Lieth; J. Jimenez-Barbero; J. Kopitz; H. J. Gabius, *Biochemistry* **2003**, *42* (50), 14762–14773.
286. T. Weimar; B. O. Petersen; B. Svensson; B. M. Pinto, *Carbohydr. Res.* **2000**, *326* (1), 50–55.
287. K. Lycknert; M. Edblad; A. Imberty; G. Widmalm, *Biochemistry* **2004**, *43* (30), 9647–9654.
288. M. I. Chavez; C. Andreu; P. Vidal; N. Aboitiz; F. Freire; P. Groves; J. L. Asensio; G. Asensio; M. Muraki; F. J. Canada; J. Jimenez-Barbero, *Chemistry* **2005**, *11* (23), 7060–7074.
289. J. Jimenez-Barbero; Javier F. Canada; J. L. Asensio; N. Aboitiz; P. Vidal; A. Canales; P. Groves; H. J. Gabius; H. C. Siebert, *Adv. Carbohydr. Chem. Biochem.* **2006**, *60*, 303–354.
290. M. Takeda; S. Ogino; R. Umemoto; M. Sakakura; M. Kajiwara; K. N. Sugahara; H. Hayasaka; M. Miyasaka; H. Terasawa; I. Shimada, *J. Biol. Chem.* **2006**, *281* (52), 40089–40095.
291. R. B. Tunnicliffe; D. N. Bolam; G. Pell; H. J. Gilbert; M. P. Williamson, *J. Mol. Biol.* **2005**, *347* (2), 287–296.
292. R. Szilaghi; S. Shahzad-ul-Hussan; T. Weimar, *ChemBioChem* **2005**, *6* (7), 1270–1276.
293. D. Raghunathan; V. M. Sanchez-Pedregal; J. Junker; C. Schwiegk; M. Kalesse; A. Kirschning; T. Carlomagno, *Nucleic Acids Res.* **2006**, *34* (12), 3599–3608.
294. A. Bhunia; V. Jayalakshmi; A. J. Benie; O. Schuster; S. Kelm; N. R. Krishna; T. Peters, *Carbohydr. Res.* **2004**, *339* (2), 259–267.
295. C. Sandstrom; O. Berteau; E. Gemma; S. Oscarson; L. Kenne; A. M. Gronenborn, *Biochemistry* **2004**, *43* (44), 13926–13931.
296. J. Angulo; B. Langpap; A. Blume; T. Biet; B. Meyer; N. R. Krishna; H. Peters; M. M. Palcic; T. Peters, *J. Am. Chem. Soc.* **2006**, *128* (41), 13529–13538.
297. A. J. Benie; R. Moser; E. Bauml; D. Blaas; T. Peters, *J. Am. Chem. Soc.* **2003**, *125* (1), 14–15.
298. C. Rademacher; N. R. Krishna; M. Palcic; F. Parra; T. Peters, *J. Am. Chem. Soc.* **2008**, *130* (11), 3669–3675.
299. T. Haselhorst; H. Blanchard; M. Frank; M. J. Kraschnefski; M. J. Kiefel; A. J. Szyczew; J. C. Dyason; F. Fleming; G. Holloway; B. S. Coulson; M. von Itzstein, *Glycobiology* **2007**, *17* (1), 68–81.
300. S. Mari; D. Serrano-Gomez; F. J. Canada; A. L. Corbi; J. Jimenez-Barbero, *Angew. Chem. Int. Ed. Engl.* **2004**, *44* (2), 296–298.
301. K. E. Kövér; P. Groves; J. Jimenez-Barbero; G. Batta, *J. Am. Chem. Soc.* **2007**, *129* (37), 11579–11582.
302. A. Canales; J. Angulo; R. Ojeda; M. Bruix; R. Fayos; R. Lozano; G. Gimenez-Gallego; M. Martin-Lomas; P. M. Nieto; J. Jimenez-Barbero, *J. Am. Chem. Soc.* **2005**, *127* (16), 5778–5779.
303. M. Guerrini; M. Hricovini; G. Torri, *Curr. Pharm. Des.* **2007**, *13* (20), 2045–2056.
304. M. Guerrini; S. Guglieri; D. Beccati; G. Torri; C. Viskov; P. Mourier, *Biochem. J.* **2006**, *399* (2), 191–198.
305. S. Ravindranathan; J. M. Mallet; P. Sinay; G. Bodenhausen, *J. Magn. Reson.* **2003**, *163* (2), 199–207.
306. R. D. Seidel, 3rd; T. Zhuang; J. H. Prestegard, *J. Am. Chem. Soc.* **2007**, *129* (15), 4834–4839.
307. F. Yu; J. J. Wolff; I. J. Amster; J. H. Prestegard, *J. Am. Chem. Soc.* **2007**, *129* (43), 13288–13297.
308. H. Shimizu; A. Donohue-Rolfe; S. W. Homans, *J. Am. Chem. Soc.* **1999**, *121* (24), 5815–5816.
309. N. U. Jain; S. Noble; J. H. Prestegard, *J. Mol. Biol.* **2003**, *328* (2), 451–462.
310. C. Tang; C. D. Schwieters; G. M. Clore, *Nature* **2007**, *449* (7165), 1078–1082.
311. S. Liu; A. Venot; L. Meng; F. Tian; K. W. Moremen; G. J. Boons; J. H. Prestegard, *Chem. Biol.* **2007**, *14* (4), 409–418.
312. T. Zhuang; H. Leffler; J. H. Prestegard, *Protein Sci.* **2006**, *15* (7), 1780–1790.
313. T. Zhuang; H. S. Lee; B. Imperiali; J. H. Prestegard, *Protein Sci.* **2008**, *17* (7), 1220–1231.
314. C. D. Blundell; M. A. Reed; M. Overduin; A. Almond, *Carbohydr. Res.* **2006**, *341* (12), 1985–1991.
315. R. Dziarski; R. I. Tapping; P. S. Tobias, *J. Biol. Chem.* **1998**, *273* (15), 8680–8690.
316. L. Franchi; C. McDonald; T. D. Kanneganti; A. Amer; G. Nunez, *J. Immunol.* **2006**, *177* (6), 3507–3513.
317. R. Dziarski; S. Viriyakosol; T. N. Kirkland; D. Gupta, *Infect. Immun.* **2000**, *68* (9), 5254–5260.
318. J. Nadesalingam; A. W. Dodds; K. B. M. Reid; N. Palaniyar, *J. Immunol.* **2005**, *175* (3), 1785–1794.
319. H. L. Cash; C. V. Whitham; C. L. Behrendt; L. V. Hooper, *Science* **2006**, *313* (5790), 1126–1130.
320. D. C. Phillips, *Proc. Nat. Acad. Sci. U.S.A.* **1967**, *57*, 484.
321. H. G. Sahl, *Chem. Biol.* **2006**, *13*, 1015–1016.
322. R. J. Guan; R. A. Mariuzza, *Trends Microbiol.* **2007**, *15* (3), 127.
323. J. Royet; R. Dziarski, *Nat. Rev. Microbiol.* **2007**, *5* (4), 264–277.
324. X. F. Lu; M. H. Wang; J. Qi; H. T. Wang; X. N. Li; D. Gupta; R. Dziarski, *J. Biol. Chem.* **2006**, *281* (9), 5895–5907.
325. H. Steiner, *Immunol. Rev.* **2004**, *198* (1), 83–96.
326. R. Dziarski; D. Gupta, *J. Endotoxin Res.* **2005**, *11* (5), 304–310.
327. B. Fournier; D. J. Philpott, *Clin. Microbiol. Rev.* **2005**, *18* (3), 521.
328. R. Dziarski, *Cell. Mol. Life Sci.* **2003**, *60* (9), 1793–1804.
329. W. Vollmer, *FEMS Microbiol. Rev.* **2008**, *32*, 287–306.
330. J. Lee; R. I. Hollingsworth, *Carbohydr. Res.* **1997**, *303* (1), 103–112.
331. E. Simelyte; M. Rimpilainen; L. Lehtonen; X. Zhang; P. Toivanen, *Infect. Immun.* **2000**, *68* (6), 3535–3540.
332. P. Mellroth; J. Karlsson; J. Hakansson; N. Schultz; W. E. Goldman; H. Steiner, *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102* (18), 6455–6460.
333. B. A. Dmitriev; O. Holst; E. T. Rietschel; S. Ehlers, *J. Bacteriol.* **2004**, *186* (21), 7141–7148.
334. W. Vollmer; J. V. Holtje, *J. Bacteriol.* **2004**, *186* (18), 5978–5987.
335. S. O. Meroueh; K. Z. Bencze; D. Hesek; M. Lee; J. F. Fisher; T. L. Stemmler; S. Mobashery, *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103* (12), 4404–4409.
336. W. Vollmer; B. Joris; P. Charlier; S. Foster, *FEMS Microbiol. Rev.* **2008**, *32* (2), 259–286.
337. T. Kern; S. Hediger; P. Muller; C. Giustini; B. Joris; C. Bougault; W. Vollmer; J. P. Simorre, *J. Am. Chem. Soc.* **2008**, *130* (17), 5618.

338. F. Ellouz; A. Adam; R. Ciobaru; E. Lederer, *Biochem. Biophys. Res. Commun.* **1974**, *59*, 1317–1325.
339. S. Kotani; Y. Watanabe; F. Kinoshita; T. Shimono; I. Morisaki; T. Shiba; S. Kusumoto; Y. Tarumi; K. Ikenaka, *Biken J.* **1975**, *18* (2), 105–111.
340. I. Azuma; K. Sugimura; T. Taniyama; M. Yamawaki; Y. Yamamura; S. Kusumoto; S. Okada; T. Shiba, *Infect. Immun.* **1976**, *14* (1), 18–27.
341. L. Chedid; F. Audibert; P. Lefrancier; J. Choay; E. Lederer, *Proc. Natl. Acad. Sci. U.S.A.* **1976**, *73* (7), 2472–2475.
342. D. J. Silva; C. L. Bowe; A. A. Branstrom; E. R. Baizman; R. J. Sofia, *Bioorg. Med. Chem. Lett.* **2000**, *10* (24), 2811–2813.
343. S. Ha; E. Chang; M. C. Lo; H. Men; P. Park; M. Ge; S. Walker, *J. Am. Chem. Soc.* **1999**, *121* (37), 8415–8426.
344. M. S. VanNieuwenhze; S. C. Mauldin; M. Zia-Ebrahimi; J. A. Aikins; L. C. Blaszczak, *J. Am. Chem. Soc.* **2001**, *123* (29), 6983–6988.
345. B. Schwartz; J. A. Markwalder; Y. Wang, *J. Am. Chem. Soc.* **2001**, *123* (47), 11638–11643.
346. M. S. VanNieuwenhze; S. C. Mauldin; M. Zia-Ebrahimi; B. E. Winger; W. J. Hornback; S. L. Saha; J. A. Aikins; L. C. Blaszczak, *J. Am. Chem. Soc.* **2002**, *124* (14), 3656–3660.
347. D. Hesek; M. J. Lee; K. I. Morio; S. Mobashery, *J. Org. Chem.* **2004**, *69* (6), 2137–2146.
348. D. Hesek; M. Suvorov; K. Morio; M. Lee; S. Brown; S. B. Vakulenko; S. Mobashery, *J. Org. Chem.* **2004**, *69* (3), 778–784.
349. R. J. Cox; A. Sutherland; J. C. Vederas, *Bioorg. Med. Chem.* **2000**, *8* (5), 843–871.
350. N. Kubasch; R. R. Schmidt, *Eur. J. Org. Chem*. **2002**, (16), 2710–2726.
351. A. R. Chowdhury; G. J. Boons, *Tetrahedron Lett.* **2005**, *46* (10), 1675–1678.
352. S. Kumar; A. Roychowdhury; B. Ember; Q. Wang; R. J. Guan; R. A. Mariuzza; G. J. Boons, *J. Biol. Chem.* **2005**, *280* (44), 37005–37012.
353. C. P. Swaminathan; P. H. Brown; A. Roychowdhury; Q. Wang; R. J. Guan; N. Silverman; W. E. Goldman; G. J. Boons; R. A. Mariuzza, *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103* (3), 684–689.
354. S. Inamura; K. Fukase; S. Kusumoto, *Tetrahedron Lett.* **2001**, *42* (43), 7613–7616.
355. S. Inamura; Y. Fujimoto; A. Kawasaki; Z. Shiokawa; E. Woelk; H. Heine; B. Lindner; N. Inohara; S. Kusumoto; K. Fukase, *Org. Biomol. Chem.* **2006**, *4* (2), 232–242.
356. Y. Fujimoto; S. Inamura; A. Kawasaki; Z. Shiokawa; A. Shimoyama; T. Hashimoto; S. Kusumoto; K. Fukase, *J. Endotoxin Res.* **2007**, *13* (3), 189–196.
357. J. W. Park; B. R. Je; S. Piao; S. Inamura; Y. Fujimoto; K. Fukase; S. Kusumoto; K. Soderhall; N. C. Ha; B. L. Lee, *J. Biol. Chem.* **2006**, *281* (12), 7747–7755.
358. D. Keglević; B. Ladešić; J. Tomašić; Z. Valinger; R. Naumski, *Biochim. Biophys. Acta* **1979**, *585* (2), 273–281.
359. H. Matter; L. Szilágyi; P. Forgó; Z. Marinić; B. Klaić, *J. Am. Chem. Soc.* **1997**, *119* (9), 2212–2223.
360. W. Lee; M. A. McDonough; L. P. Kotra; Z. H. Li; N. R. Silvaggi; Y. Takeda; J. A. Kelly; S. Mobashery, *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98* (4), 1427–1431.
361. B. Halassy; S. Mateljak; F. B. Bouche; M. M. Putz; C. P. Muller; R. Frkanec; L. Habjanec; J. Tomašić, *Vaccine* **2006**, *24* (2), 185–194.
362. D. Ljevaković; J. Tomašić; V. Šporec; B. H. Špoljar; I. Hanzl-Dujmović, *Bioorg. Med. Chem.* **2000**, *8* (10), 2441–2449.
363. K. Fehér; P. Pristovšek; L. Szilágyi; D. Ljevaković; J. Tomašić, *Bioorg. Med. Chem.* **2003**, *11* (14), 3133–3140.
364. S. J. Kim; L. Cegelski; M. Preobrazhenskaya; J. Schaefer, *Biochemistry* **2006**, *45* (16), 5235–5250.
365. H. Yoshida; K. Kinoshita; M. Ashida, *J. Biol. Chem.* **1996**, *271* (23), 13854–13860.
366. D. W. Kang; G. Liu; A. Lundstrom; E. Gelius; H. Steiner, *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95* (17), 10078–10082.
367. C. Liu; Z. J. Xu; D. Gupta; R. Dziarski, *J. Biol. Chem.* **2001**, *276* (37), 34686–34694.
368. R. Dziarski; D. Gupta, *Genome Biol.* **2006**, *7* (8), 232.
369. S. Cho; Q. Wang; C. P. Swaminathan; D. Hesek; M. Lee; G. J. Boons; S. Mobashery; R. A. Mariuzza, *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104* (21), 8761–8766.
370. S. E. Girardin; D. J. Philpott, *Immunity* **2006**, *24* (4), 363–366.
371. A. E. Myhre; A. O. Aasen; C. Thiemermann; J. E. Wang, *Shock* **2006**, *25* (3), 227–235.
372. Z. M. Wang; X. N. Li; R. R. Cocklin; M. H. Wang; M. Wang; K. Fukase; S. Inamura; S. Kusumoto; D. Gupta; R. Dziarski, *J. Biol. Chem.* **2003**, *278* (49), 49044–49052.
373. P. Mellroth; H. Steiner, *Biochem. Biophys. Res. Commun.* **2006**, *350* (4), 994–999.
374. R. J. Guan; A. Roychowdhury; B. Ember; S. Kumar; G. J. Boons; R. A. Mariuzza, *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101* (49), 17168–17173.
375. R. J. Guan; P. H. Brown; C. P. Swaminathan; A. Roychowdhury; G. J. Boons; R. A. Mariuzza, *Protein Sci.* **2006**, *15* (5), 1199–1206.
376. C. R. Stenbak; J. H. Ryu; F. Leulier; S. Pili-Floury; C. Parquet; M. Herve; C. Chaput; I. G. Boneca; W. J. Lee; B. Lemaitre; D. Mengin-Lecreulx, *J. Immunol.* **2004**, *173* (12), 7339–7348.
377. C. I. Chang; Y. Chelliah; D. Borek; D. Mengin-Lecreulx; J. Deisenhofer, *Science* **2006**, *311* (5768), 1761–1764.
378. T. Kaneko; T. Yano; K. Aggarwal; J. H. Lim; K. Ueda; Y. Oshima; C. Peach; D. Erturk-Hasdemir; W. E. Goldman; B. H. Oh; S. Kurata; N. Silverman, *Nat. Immunol.* **2006**, *7* (7), 715–723.
379. J. H. Lim; M. S. Kim; H. E. Kim; T. Yano; Y. Oshima; K. Aggarwal; W. E. Goldman; N. Silverman; S. Kurata; B. H. Oh, *J. Biol. Chem.* **2006**, *281* (12), 8286–8295.
380. A. Taylor; B. C. Das; J. Vanheijenoort, *Eur. J. Biochem.* **1975**, *53* (1), 47–54.
381. E. J. van Asselt; A. Thunnissen; B. W. Dijkstra, *J. Mol. Biol.* **1999**, *291* (4), 877–898.
382. A. K. W. Leung; H. S. Duewel; J. F. Honek; A. M. Berghuis, *Biochemistry* **2001**, *40* (19), 5665–5673.
383. E. J. van Asselt; K. H. Kalk; B. W. Dijkstra, *Biochemistry* **2000**, *39* (8), 1924–1934.
384. S. R. Filipe; A. Tomasz; P. Ligoxygakis, *EMBO Rep.* **2005**, *6* (4), 327–333.
385. K. E. van Straaten; T. R. M. Barends; B. W. Dijkstra; A. Thunnissen, *J. Biol. Chem.* **2007**, *282* (29), 21197–21205.
386. N. C. J. Strynadka; M. N. G. James, *J. Mol. Biol.* **1991**, *220* (2), 401–424.
387. N. C. J. Strynadka; M. N. G. James, *FASEB J.* **1992**, *6* (1), A11.
388. M. Mishima; T. Shida; K. Yabuki; K. Kato; J. Sekiguchi; C. Kojima, *Biochemistry* **2005**, *44* (30), 10153–10163.

389. J. Freund, *Adv. Tuberc. Res.* **1956**, *1*, 130–148.
390. L. Chedid; F. Audibert; M. Jolivet, *Dev. Biol. Stand.* **1986**, *63*, 133–140.
391. I. Azuma, *Int. J. Immunopharmacol.* **1992**, *14* (3), 487–496.
392. T. Kaneko; W. E. Goldman; P. Mellroth; H. Steiner; K. Fukase; S. Kusumoto; W. Harley; A. Fox; D. Golenbock; N. Silverman, *Immunity* **2004**, *20* (5), 637–649.
393. W. H. A. Dokter; A. J. Dijkstra; S. B. Koopmans; B. K. Stulp; W. Keck; M. R. Halie; E. Vellenga, *J. Biol. Chem.* **1994**, *269* (6), 4201–4206.
394. L. Szilágyi; P. Pristovšek, *Mini Rev. Med. Chem.* **2007**, *7* (8), 861–870.
395. I. G. Boneca, *Curr. Opin. Microbiol.* **2005**, *8* (1), 46–53.
396. K. A. Cloud-Hansen; S. B. Peterson; E. V. Stabb; W. E. Goldman; M. J. McFall-Ngai; J. Handelsman, *Nat. Rev. Microbiol.* **2006**, *4* (9), 710–716.
397. M. Nieto; H. R. Perkins, *Biochem. J.* **1971**, *123*, 773.
398. H. R. Perkins, *Biochem. J.* **1969**, *111*, 195–205.
399. D. H. Williams, *Acc. Chem. Res.* **1984**, *17* (10), 364–369.
400. J. R. Kalman; D. H. Williams, *J. Am. Chem. Soc.* **1980**, *102* (3), 906–912.
401. D. H. Williams; M. P. Williamson; D. W. Butcher; S. J. Hammond, *J. Am. Chem. Soc.* **1983**, *105* (5), 1332–1339.
402. J. C. J. Barna; D. H. Williams; M. P. Williamson, *J. Chem. Soc. Chem. Commun*. **1985**, 254–256.
403. J. Balzarini; C. Pannecouque; E. De Clereq; A. Y. Pavlov; S. S. Printsevskaya; O. V. Miroshnikova; M. I. Reznikova; M. N. Preobrazhenskaya, *J. Med. Chem.* **2003**, *46* (13), 2755–2764.
404. M. Schafer; T. R. Schneider; G. M. Sheldrick, *Structure* **1996**, *4* (12), 1509–1515.
405. P. J. Loll; A. E. Bevivino; B. D. Korty; P. H. Axelsen, *J. Am. Chem. Soc.* **1997**, *119* (7), 1516–1522.
406. G. M. Sheldrick; P. G. Jones; O. Kennard; D. H. Williams; G. A. Smith, *Nature* **1978**, *271* (5642), 223–225.
407. M. P. Williamson; D. H. Williams, *J. Am. Chem. Soc.* **1981**, *103* (22), 6580–6585.
408. K. C. Nicolaou; C. N. C. Boddy; S. Brase; N. Winssinger, *Angew. Chem. Int. Ed. Engl.* **1999**, *38* (15), 2097–2152.
409. D. L. Boger, *Med. Res. Rev.* **2001**, *21* (5), 356–381.
410. D. Kahne; C. Leimkuhler; L. Wei; C. Walsh, *Chem. Rev.* **2005**, *105* (2), 425–448.
411. M. N. Preobrazhenskaya; E. N. Olsufyeva, *Expert Opin. Ther. Pat.* **2004**, *14* (2), 141–173.
412. E. N. Olsuf'eva; M. N. Preobrazhenskaya, *Russ. J. Bioorganic Chem.* **2006**, *32* (4), 303–322.
413. D. H. Williams; B. Bardsley, *Angew. Chem. Int. Ed. Engl.* **1999**, *38* (9), 1173–1193.
414. D. L. Boger; S. Miyazaki; S. H. Kim; J. H. Wu; S. L. Castle; O. Loiseleur; Q. Jin, *J. Am. Chem. Soc.* **1999**, *121* (43), 10004–10011.
415. D. L. Boger; J. H. Weng; S. Miyazaki; J. J. McAtee; S. L. Castle; S. H. Kim; Y. Mori; O. Rogel; H. Strittmatter; Q. Jin, *J. Am. Chem. Soc.* **2000**, *122* (41), 10047–10055.
416. P. Groves; M. S. Searle; J. P. Mackay; D. H. Williams, *Structure* **1994**, *2* (8), 747–754.
417. J. R. Kalman; D. H. Williams, *J. Am. Chem. Soc.* **1980**, *102* (3), 897–905.
418. A. Kumar; R. R. Ernst; K. Wuthrich, *Biochem. Biophys. Res. Commun.* **1980**, *95* (1), 1–6.
419. K. Wüthrich, *NMR of Proteins and Nucleic Acids*; Wiley: New York, 1986.
420. J. Keeler, *Understanding NMR Spectroscopy*; John Wiley & Sons: 2005.
421. J. Cavanagh; W. J. Fairbrother; I. A. G. Palmer; N. J. Skelton; M. J. Rance, *Protein NMR Spectroscopy: Principles and Practice*; Acad Press/Elsevier: Burlington, MA, 2006.
422. N. E. Jacobsen, *NMR Spectroscopy Explained: Simplified Theory, Applications and Examples for Organic Chemistry and Structural Biology*; Wiley-Interscience: Hoboken, NJ, 2007.
423. G. Batta; F. Sztaricskai; M. O. Makarova; E. G. Gladkikh; V. V. Pogozheva; T. F. Berdnikova, *Chem. Commun.* **2001** (5), 501–502.
424. G. Batta; K. E. Kövér; Z. Székely; F. Sztaricskai, *J. Am. Chem. Soc.* **1992**, *114* (7), 2757–2758.
425. G. Batta; F. Sztaricskai; K. E. Kövér; C. Rudel; T. F. Berdnikova, *J. Antibiot.* **1991**, *44* (11), 1208–1221.
426. C. M. Pearce; D. H. Williams, *J. Chem. Soc. Perkin Trans.* **1995**, *2* (1), **1995** (1),153–157.
427. F. Sztaricskai; G. Batta; P. Herczegh; A. Balázs; J. Jekő; E. Roth; P. T. Szabó; S. Kardos; F. Rozgonyi; Z. Boda, *J. Antibiot.* **2006**, *59* (9), 564–582.
428. G. M. Sheldrick; E. Paulus; L. Vertesy; F. Hahn, *Acta Crystallogr. B* **1995**, *51*, 89–98.
429. M. Schafer; G. M. Sheldrick; T. R. Schneider; L. Vertesy, *Acta Crystallogr. D Biol. Crystallogr.* **1998**, *54*, 175–183.
430. M. Eberstadt; W. Guba; H. Kessler; H. Kogler; D. F. Mierke, *Biopolymers* **1995**, *36* (4), 429–437.
431. P. J. Loll; R. Miller; C. M. Weeks; P. H. Axelsen, *Chem. Biol.* **1998**, *5* (5), 293–298.
432. P. J. Loll; J. Kaplan; B. S. Selinsky; P. H. Axelsen, *J. Med. Chem.* **1999**, *42* (22), 4714–4719.
433. J. P. Mackay; U. Gerhard; D. A. Beauregard; M. S. Westwell; M. S. Searle; D. H. Williams, *J. Am. Chem. Soc.* **1994**, *116* (11), 4581–4590.
434. G. F. Gause; M. G. Brazhnikova; N. N. Lomakina; T. F. Berdnikova; G. B. Fedorova; N. L. Tokareva; V. N. Borisova; G. Y. Batta, *J. Antibiot.* **1989**, *42* (12), 1790–1799.
435. D. Li; U. Sreenivasan; N. Juranic; S. Macura; F. J. Puga, II; P. M. Frohnert; P. H. Axelsen, *J. Mol. Recognit.* **1997**, *10* (2), 73–87.
436. S. G. Grdadolnik; P. Pristovsek; D. F. Mierke, *J. Med. Chem.* **1998**, *41* (12), 2090–2099.
437. J. Kaplan; B. D. Korty; P. H. Axelsen; P. J. Loll, *J. Med. Chem.* **2001**, *44* (11), 1837–1840.
438. W. G. Prowse; A. D. Kline; M. A. Skelton; R. J. Loncharich, *Biochemistry* **1995**, *34* (29), 9632–9644.
439. C. Lehmann; G. Bunkoczi; L. Vertesy; G. M. Sheldrick, *J. Mol. Biol.* **2002**, *318* (3), 723–732.
440. M. Rekharsky; D. Hesek; M. Lee; S. O. Meroueh; Y. Inoue; S. Mobashery, *J. Am. Chem. Soc.* **2006**, *128* (24), 7736–7737.
441. P. J. Loll; P. H. Axelsen, *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 265–289.
442. M. S. Searle; G. J. Sharman; P. Groves; B. Benhamu; D. A. Beauregard; M. S. Westwell; R. J. Dancer; A. J. Maguire; A. C. Try; D. H. Williams, *J. Chem. Soc. Perkin Trans.* **1996**, *1* (23), **1996** (23), 2781–2786.
443. D. McPhail; A. Cooper, *J. Chem. Soc., Faraday Trans.* **1997**, *93* (13), 2283–2289.
444. A. Losi; A. A. Wegener; M. Engelhard; S. E. Braslavsky, *J. Am. Chem. Soc.* **2001**, *123* (8), 1766–1767.
445. D. H. Williams; D. P. O'Brien; B. Bardsley, *J. Am. Chem. Soc.* **2001**, *123* (4), 737–738.
446. S. Jusuf; P. J. Loll; P. H. Axelsen, *J. Am. Chem. Soc.* **2002**, *124* (14), 3490–3491.

447. S. Jusuf; P. J. Loll; P. H. Axelsen, *J. Am. Chem. Soc.* **2003**, *125* (13), 3988–3994.
448. M. S. Searle; P. Groves; D. H. Williams, *Proc. Indian Acad. Sci. Chem. Sci.* **1994**, *106* (5), 937–954.
449. M. F. Cristofaro; D. A. Beauregard; H. S. Yan; N. J. Osborn; D. H. Williams, *J. Antibiot.* **1995**, *48* (8), 805–810.
450. Y. R. Cho; A. J. Maguire; A. C. Try; M. S. Westwell; P. Groves; D. H. Williams, *Chem. Biol.* **1996**, *3* (3), 207–215.
451. B. Bardsley; D. H. Williams, *J. Chem. Soc. Perkin Trans.* **1998**, *2* (9), **1998** (9), 1925–1929.
452. J. P. Waltho; D. H. Williams; D. J. M. Stone; N. J. Skelton, *J. Am. Chem. Soc.* **1988**, *110* (17), 5638–5643.
453. S. Jusuf; P. H. Axelsen, *Biochemistry* **2004**, *43* (49), 15446–15452.
454. L. Zidek; M. V. Novotny; M. J. Stone, *Nat. Struct. Biol.* **1999**, *6* (12), 1118–1121.
455. D. H. Williams; A. J. Maguire; W. Tsuzuki; M. S. Westwell, *Science* **1998**, *280* (5364), 711–714.
456. D. H. Williams; N. L. Davies; J. J. Koivisto, *J. Am. Chem. Soc.* **2004**, *126* (43), 14267–14272.
457. D. H. Williams; N. L. Davies; R. Zerella; B. Bardsley, *J. Am. Chem. Soc.* **2004**, *126* (7), 2042–2049.
458. D. H. Williams; E. Stephens; D. P. O'Brien; M. Zhou, *Angew. Chem. Int. Ed. Engl.* **2004**, *43* (48), 6596–6616.
459. O. Toke; L. Cegelski; J. Schaefer, *Biochi. Biophys. Acta* **2006**, *1758* (9), 1314–1329.
460. T. Gullion; J. Schaefer, *J. Magn. Reson.* **1989**, *81* (1), 196–200.
461. A. W. Hing; S. Vega; J. Schaefer, *J. Magn. Reson.* **1992**, *96* (1), 205–209.
462. S. J. Kim; L. Cegelski; D. R. Studelska; R. D. O'Connor; A. K. Mehta; J. Schaefer, *Biochemistry* **2002**, *41* (22), 6967–6977.
463. H. Molinari; A. Pastore; L. Y. Lian; G. E. Hawkes; K. Sales, *Biochemistry* **1990**, *29* (9), 2271–2277.
464. L. Cegelski; S. J. Kim; A. W. Hing; D. R. Studelska; R. D. O'Connor; A. K. Mehta; J. Schaefer, *Biochemistry* **2002**, *41* (43), 13053–13058.
465. L. Cegelski; D. Steuber; A. K. Mehta; D. W. Kulp; P. H. Axelsen; J. Schaefer, *J. Mol. Biol.* **2006**, *357* (4), 1253–1262.
466. P. J. Vollmerhaus; E. Breukink; A. J. R. Heck, *Chemistry* **2003**, *9* (7), 1556–1565.
467. S. J. Kim; L. Cegelski; D. Stueber; M. Singh; E. Dietrich; K. S. E. Tanaka; T. R. Parr; A. R. Far; J. Schaefer, *J. Mol. Biol.* **2008**, *377* (1), 281–293.
468. S. J. Kim; S. Matsuoka; G. J. Patti; J. Schaefer, *Biochemistry* **2008**, *47* (12), 3822–3831.

Biographical Sketches

Katalin E. Kövér is a distinguished research scientist at the Department of Inorganic and Analytical Chemistry, University of Debrecen. She obtained her M.S. in chemistry in 1979 from the L. Kossuth University, Debrecen, her Univ.D. in chemistry in 1984 from the L. Kossuth University, Debrecen, and her Ph.D. in chemistry in 1988 from the Hungarian Academy of Sciences, Budapest. She was a postdoc fellow (1991–93) in Tucson, Arizona with V. J. Hruby. She was awarded the D.Sc. degree in chemistry in 2002 by the Hungarian Academy of Sciences, Budapest. She has many years of expertise in pulse program development for sensitive and accurate determination of NMR parameters by 1D and 2D methods and in analyzing structural and motional parameters from the measured data for small and large molecules alike. Her current research interests include methodological developments focusing on multidimensional, proton-detected heteronuclear experiments, selective experiments, and gradient-enhanced experiments; NMR structure determination of biologically important oligopeptides/proteins, oligosaccharides, and antibiotics; investigations of receptor–ligand interactions; NMR dynamics study including heteronuclear relaxation, dipole–dipole (DD/DD) and dipole–chemical shift anisotropy (DD/CSA) relaxation interference measurements and their interpretation in terms of the relevant structural and dynamic parameters; application of Transverse relaxation optimized spectroscopy (TROSY) approach; and measurement and application of residual dipolar coupling constants (RDC).

László Szilágyi has been associated all his career with the University of Debrecen (formerly, L. Kossuth University) except for two postdoc periods in Strasbourg (with J.-M. Lehn) and Stanford (with O. Jardetzky). His research interests include application of NMR spectroscopy to the structure elucidation of natural products such as carbohydrates, aminoglycoside and macrolide antibiotics, flavonoids, morphine alkaloids, etc.; conformational studies of (glyco)peptides and proteins by NMR; synthesis of novel carbohydrate scaffolds; and studies of carbohydrate–protein interactions.

Gyula Batta is a distinguished research scientist and professor at the Department of Biochemistry, University of Debrecen. He completed his M.S. in physics in 1976 from the L. Kossuth University, Debrecen and his Ph.D. in chemistry in 1988 from the Hungarian Academy of Sciences (HAS), Budapest. He was a postdoc fellow (1993–94) in Tucson, Arizona with Professor J. Gervay. Under the Go West scholarship, he worked with Professor D. H. Williams in 1994 at Cambridge. He also worked with Professor J. Kowalewski under the STINT scholarship in 1999 at Stockholm. He was awarded the D.Sc. degree in chemistry in 2001 by the HAS, Budapest.

His research interests are in high-resolution NMR, methodological developments, NMR dynamics from relaxation and relaxation interferences, and diffusion and saturation transfer methods; NMR structure determination of calcium binding and antifungal proteins, glycopeptide antibiotics, oligosaccharides, and antibiotics; and molecular recognition by glycopeptide antibiotics and protein–carbohydrate interactions.

Dušan Uhrín received his M.Sc. in chemistry from the Slovak Technical University in Bratislava, Slovakia in 1982. From 1983 he worked as research assistant at the Institute of Chemistry, Slovak Academy of Sciences in Bratislava, where he received his Ph.D. in 1990. In 1991 he was awarded the Soros/FCO Scholarship for East European Scientists and worked under the supervision of Professor Raymond Dwek in the Glycobiology Institute, Oxford University, UK. He was a research associate (1992–95) in the Institute for Biological Sciences, NRC, Ottawa, Canada. He returned to the UK in 1995 as the manager of the Edinburgh Biomolecular NMR unit in the School of Chemistry, University of Edinburgh, where he was appointed to a lectureship in 2000. He is currently head of the NMR facility and a reader at the University of Edinburgh. His research interests are in the development (selective techniques, measurement of scalar and residual dipolar coupling constants) and the application of high-resolution NMR spectroscopy to the structure elucidation of molecules. The systems he studies include small organic molecules, complex carbohydrates, and proteins and their complexes with carbohydrates.

Jesús Jiménez-Barbero was born in Madrid in 1960. He studied chemistry at the University of Madrid (UAM) and received his B.Sc. degree in 1982. After serving in the military service, he started his Ph.D. work with Manuel Martín-Lomas and Manuel Bernabé at the Institute of Organic Chemistry of the Higher Research Council of Spain (CSIC), Madrid, working on the synthesis and conformational analysis of sugar derivatives. In the last year of his Ph.D. work, he worked with NMR, especially with those methods applied to the measurements of long-range carbon–proton coupling constants. He received his Ph.D. degree in 1987, after a stay at CERMAV-CNRS, Grenoble, working with Serge Perez on the application of molecular mechanics calculations to polysaccharide molecules. In 1988, after a short period at the University of Zürich, he moved as a postdoctoral fellow to the National Institute for Medical

Research at Mill Hill, UK to work with Jim Feeney and Berry Birdsall on NMR of proteins, in particular, with dihydrofolate reductase. After returning to Madrid he got a tenure scientist position, although he was allowed to move to Carnegie Mellon University, Pittsburg, USA to work with Aksel Bothner-By on NMR methodology and then with Miguel Llinás on protein NMR (1990–92). After returning to Madrid again, he worked on molecular recognition, especially on protein–carbohydrate interactions, with particular emphasis on the application of NMR methods, but also using a variety of other techniques, from organic synthesis to modeling protocols and other biophysical techniques. In 1996, he was promoted to senior research scientist of CSIC at the Institute of Organic Chemistry and in 2002 to research professor of CSIC. Soon after that, he moved to the Centre for Biological Research (CIB-CSIC), Madrid, where he is working at the Protein Science Department.

He has coauthored more than 280 publications in international journals, has delivered more than 120 lectures at symposia and institutions, and despite not working at the university, has tutored 14 Ph.D. students. He was awarded the Janssen-Cilag Prize in Organic Chemistry of the Royal Society of Chemistry of Spain (RSEQ) in 2003 and is serving as the Secretary General of this Institution (RSEQ) since 2004. He is a member of the editorial boards of *Chemistry – A European Journal* (from 2001 to date), *Organic & Biomolecular Chemistry* (from 2007 to date), *Glycoconjugate Journal* (from 2008 to date), *Carbohydrate Research* (from 2001 to date), *Journal of Carbohydrate Chemistry* (from 2002 to date), and *European Journal of Organic Chemistry* (starting in 2009).

9.08 Determination of Three-Dimensional Structures of Nucleic Acids by NMR

Nikolai B. Ulyanov and Thomas L. James, University of California, San Francisco, San Francisco, CA, USA

9.08.1 Introduction

The majority of structures of biological macromolecules are being solved in solid state using X-ray crystallography. Nevertheless, nuclear magnetic resonance (NMR) has been established as a routine alternative method for high-resolution structure determination. Until May 2008, among almost 4000 nucleic acid structures (3888) deposited in the Protein Data Bank (PDB),[1] a formidable 24% have been determined by solution NMR methods (compared to 13% for proteins). Both methods have pros and cons, a detailed comparison of which is beyond the scope of this chapter but has been discussed elsewhere.[2–4] A choice of the method is dictated by many factors, not the least of which is the specific expertise in a particular research group, but also a success or failure in growing crystals suitable for X-ray diffraction versus the availability of significant time and resources required for structure determination by NMR. A major advantage of solid-state structures determined by crystallography is a defined spatial resolution, which can be used to assess the overall quality of structure determination. Solution structures determined by NMR do not have an intrinsic spatial resolution; their quality can be instead assessed by a number of approaches, some of which are discussed below. On the other hand, solution conditions often approximate the physiological state of a functional biomolecule in a better manner, while the solid state structures are sometimes distorted by crystal packing forces–a problem that can be especially severe for nucleic acids. In this chapter, we outline typical approaches to solve a DNA or RNA structure in solution, including sample preparation, resonance assignments, extracting structural information, and refinement.

9.08.2 Sample Preparation

To determine a three-dimensional (3D) structure in solution, one typically needs to prepare several samples of a nucleic acid in milligram quantities. A successful structure determination often depends on a careful design of strategies for labeling of the molecule with isotopes ^{13}C, ^{15}N, and sometimes ^{2}H. Isotopic labeling is necessary to overcome a severe overlap of ^{1}H resonances in larger molecules and facilitate the assignments of NMR resonances and extracting structural information by acquiring heteronuclear multidimensional NMR spectra.[5–8] Deoxyoligoribonucleotides (DNA oligonucleotides) or oligoribonucleotides (RNA oligonucleotides) can be synthesized either chemically or enzymatically, both unlabeled and isotopically labeled.[7,9,10]

9.08.2.1 Chemical Synthesis of Oligonucleotides

The established and widely used method for oligonucleotide synthesis is based on phosphoramidite chemistry.[11] Phosphoramidites are nucleotides (nts) with protection groups attached to each reactive group, amines, hydroxyls, and phosphates. In the case of RNA, the additional 2′-hydroxyl group is also protected. Starting with a 3′-terminal nucleoside attached to an insoluble polymeric support, phosphoramidite monomers are sequentially added to the growing oligonucleotide chain in the 3′→5′ direction. The 5′-terminal protection group is specifically removed before the addition of each new monomer. In the end of the synthesis, the oligonucleotide is cleaved from the support and all protection groups are removed. The oligonucleotide is then purified either with polyacrylamide gel electrophoresis (PAGE), high-performance liquid chromatography (HPLC), or both. Using the solid support allows automation of each step, adding the reagents and washing out the reactants, which is controlled by the computer in modern DNA/RNA synthesizers. All reagents for the automated synthesis of unlabeled oligonucleotides are readily available. Furthermore, many companies offer a reasonably priced custom synthesis of unlabeled DNA oligonucleotides of up to 50–60 nts long, while chemical RNA synthesis, although also feasible, is almost an order of magnitude more expensive. The amount of a DNA oligonucleotide synthesized at the 1-μmol scale should be sufficient to prepare an NMR sample with the concentration of 0.5 m mol l^{-1} in a volume of 0.25 ml.

Isotopically labeled oligonucleotides can be prepared in a similar way using automated synthesizers with labeled phosphoramidites. This approach has serious advantages for NMR applications over the enzymatic synthesis (see below), because it allows incorporation of labeled residues in specific positions in DNA or RNA sequence. However, not all labeled phosphoramidites are commercially available, and even the ones that are available are quite expensive. Labeled phosphoramidites can be prepared chemically from labeled nucleosides, which in turn can be either harvested from bacteria grown on labeled media or by chemical or enzymatic coupling of ribose and nucleobases (see a review by Kojima *et al.*[10] and references therein). One additional advantage of this approach is the possibility of introducing isotopes in specific positions instead of uniform labeling of a nucleoside, which can simplify NMR spectra of such samples significantly. Practically, any position on the ribose moiety can be individually labeled with ^{13}C,[10,12] and also certain positions on nucleobases can be labeled with ^{13}C or ^{15}N.[13–15]

9.08.2.2 Enzymatic Synthesis of RNA

DNA-dependent RNA polymerases from bacteriophages T3, T7, or SP6[16–18] are a family of homologous relatively small (~100 kDa) single-subunit RNA polymerases that do not require additional protein factors for any stages of transcription, that is, initiation, elongation, or termination. These polymerases are easy to overexpress in *Escherichia coli*; they are very active, terminate less frequently (compared to the *E. coli* RNA polymerase) and initiate the transcription very stringently from their own promoters (reviewed in Tabor[19]). Such properties make these polymerases a very convenient tool for *in vitro* RNA synthesis using a variety of experimental strategies.[20] RNA polymerase from bacteriophage T7 (T7 RNAP) is perhaps the one used most commonly. Although T7 RNAP is available commercially, it is more cost-effective to purify it in-house for the synthesis of large quantities of RNA; a number of protocols for the purification of T7 RNAP overexpressed in *E. coli* have been published.[9,21–23] Nucleoside triphosphates (NTPs) required for the *in vitro* transcription are available commercially, both unlabeled and labeled, including uniformly ^{15}N-labeled, doubly ^{13}C/^{15}N-labeled, and ^{2}H-labeled. It is also possible to produce labeled NTPs in-house from ribosomal RNA extracted from bacterial cells grown in appropriately labeled media.[24,25] The procedure involves hydrolyzing RNA down to nucleotide monophosphates (NMPs) and then phosphorylating them to NTPs. A number of strategies have been worked out by using variously labeled media for growing *E. coli* cells that produce isotopic labels incorporated into specific positions in nucleosides. For example, using ^{13}C-formate and ^{12}C-glucose as carbon sources, produces ^{13}C isotopes incorporated specifically into the C8 positions of purines with more than 85% efficiency, see a review by Latham *et al.*[8] and references therein.

Normally, bacteriophage RNAP initiates and terminates RNA transcription at specific sequences (with certain efficiency).[26] The most straightforward experimental setup for the *in vitro* preparative RNA production, however, is a so-called run-off transcription, when the RNA synthesis starts at a specific T7 promoter and ends when RNAP falls off from the physical end of the DNA template.[20,27] This setup allows for a convenient preparation of DNA promoter and template sequences, which can be chemically synthesized. Furthermore, it has been found that the DNA template does not have to be fully base paired; most or even the entire coding strand can remain single-stranded.[27] The minimum base-paired region spanning positions from −15 to −3 (where +1 denotes the start of transcription residue) is still sufficient for the fully efficient transcription with T7 RNAP.[27] This is consistent with the notion of forming the transcription bubble as a part of the process of promoter recognition: in the crystal structure of T7 RNAP with the open promoter, the template and nontemplate strands are unwound downstream starting with position –4; reviewed in Cheetham and Steitz.[28] This property allows preparing a single universal DNA top strand for the transcription of all RNA sequences; only the bottom coding DNA strand needs to be redesigned each time (**Figure 1**).

Natural T7 RNAP promoter sequences are strongly conserved from position −17 to position +6.[29] Nevertheless, RNA is also transcribed from promoters with altered positions +1 to +6, though with a possibly decreased yield. As a rule, the most efficient yield is achieved for RNA sequences starting with GG or GA.[27] To improve the yield, the transcription reaction conditions need to be optimized for each RNA sequence, which includes varying $MgCl_2$ concentration and relative amounts of T7 RNAP, template, and NTPs.[9] In addition to the correct RNA fragment and several shorter abortive products produced by the *in vitro* transcription, T7 RNAP also often incorporates a nontemplate nucleotide at the 3′-end of the main transcript,[27] creating a so-called '$n+1$ product'. The desired RNA product then needs to be separated from the incorrect-size transcripts and also from the DNA template, unused NTPs, and RNAP. The RNAP is sometimes removed from the reaction mixture by phenol–chloroform extraction; then the RNA is purified most often by denaturing PAGE at a single-nucleotide resolution, but also sometimes using HPLC with anion-exchange[30] or gel-filtration[31] columns. These methods can be advantageous because PAGE purification is the most time-consuming step of RNA preparation, but also because acrylamide oligomers often contaminate the final RNA sample and complicate the NMR spectrum.[31] Also, using gel-filtration chromatography does not require denaturing of RNA, which may be critical in some cases. However, column purification does not separate the n from $n+1$ transcripts, which necessitates different approaches to avoid 3′-end heterogeneity. The amount of acrylamide oligomers after the elution of RNA from the polyacrylamide gel can be significantly reduced by repetitive (5–6 times) ethanol precipitation after the elution of RNA from the gel (Z. Du, personal communication).

The relative amounts of unwanted products strongly depend on RNA sequence; sometimes the yield of the $n+1$ product can be greater than that of the main transcript. The mechanism of the non-nucleotide addition by bacteriophage RNAP is poorly understood; it is likely that the stability of the hybrid between the template DNA and the nascent RNA is implicated. Sometimes, the redesign of the sequence at the 3′-end of RNA can significantly reduce the amount of the $n+1$ product. A very useful strategy for reducing the nontemplate nucleotide addition is modification of deoxyribonucleotides near the 5′-end of the template DNA. Kao *et al.*[32] found that introducing 2′-*O*-methyl groups in the ribose in the penultimate or the last two positions of the template dramatically reduces the amount of the $n+1$ transcripts. An alternative approach to produce

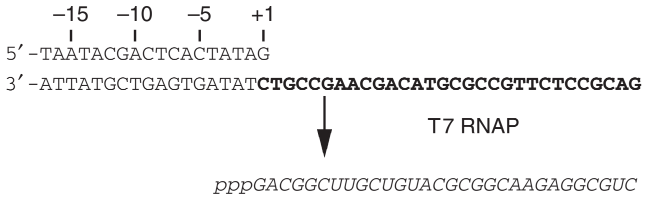

–15 –10 –5 +1
5′-TAATACGACTCACTATAG
3′-ATTATGCTGAGTGATATCTGCCGAACGACATGCGCCGTTCTCCGCAG
T7 RNAP
pppGACGGCUUGCUGUACGCGGCAAGAGGCGUC

Figure 1 Design of the promoter and template DNA sequences for the *in vitro* transcription with T7 RNAP. The double-stranded portion of the promoter is numbered relative to the transcription start site (+1). The coding portion of DNA is shown in bold; the RNA product is shown in italics.

homogenous termini in RNA is to chemically synthesize short chimeric oligonucleotides consisting of DNA residues and residues with 2′-*O*-methyl ribose modifications complementary to the RNA region just upstream of the desired 3′-terminus. When such chimeric oligonucleotide is hybridized to RNA, it directs a site-specific cleavage of RNA by RNase H, producing a precise 3′-terminus.[33,34]

An alternate method to avoid the 3′-end heterogeneity is by using a hammerhead ribozyme (HHR) cleaving RNA at a specific site.[35] This method is gaining popularity especially for larger RNA, see, for example, Kim *et al*.,[36] Tzakos *et al*.,[37] and Easton and Lukavsky.[38] HHR folds in a three-way junction structure.[39] Its catalytic center has conserved unpaired residues at the junctions, but there is very little sequence requirement for base pairs in the three stems. The autocatalytic cleavage depends on the presence of Mg^{2+} and occurs downstream of the nucleotide denoted H in **Figure 2(a)**. This could be any residue except G, although the highest cleavage rates are observed in ribozymes with sequences GUG, GUA, AUA, and AUC just upstream of the cleavage site.[40] The 5′-fragment produced as a result of the cleavage has a 2′–3′ cyclic phosphate group at its 3′-terminus, and the 3′-fragment has a hydroxyl group at its 5′-terminus.[39,41] The RNA substrate can be either added in *trans* or designed attached to the ribozyme. To produce a homogeneous 3′-end after the cleavage, the HHR can be designed in *cis* at the 3′-end of the transcript[35] (**Figure 2(b)**).

RNA with practically any length and sequence can be prepared via *in vitro* transcription with T7 RNAP.[20] However for larger RNA, using chemically synthesized DNA templates becomes less practical, because of the exponential decrease in the yield of the template with the oligonucleotide length. For templates greater than 50–100 nt, an alternative method is to use a fully double-stranded DNA template designed within a linearized high-copy DNA plasmid, example, pUC18.[31] Another potential complication for preparation of large RNA is the denaturation that RNA undergoes during the PAGE purification. RNA requires to be refolded into its native conformation after such purification, which sometimes may be problematic, see, for example, Uhlenbeck.[42] To alleviate this potential problem, several nondenaturing methods of RNA purification have been proposed, including using gel-filtration columns[31] and various affinity tag purification strategies.[43–45] As an example of the latter, one affinity tag purification strategy included the Ffh M-domain protein from the signal recognition particle (SRP) of *Thermotoga maritima* coupled to an Affigel-10 matrix.[44] The designed RNA included at its 3′-terminus a duplicated *T. maritima* SRP RNA, which forms a high-affinity complex with the M-domain protein. The SRP RNA was separated from the RNA of interest (at the 5′-terminus of the transcript) by the C75U mutant hepatitis delta virus ribozyme that is activated by imidazole. For the purification, the transcription reaction mixture was loaded onto the M-domain protein affinity column, washed, and then the RNA of interest was released by adding an imidazole-containing buffer. In another variant of affinity tag purification,[45] the specific RNA–protein interaction was achieved by using a coat protein of bacteriophage MS2 that binds with high affinity to a short RNA hairpin.[46] The MS2 coat protein was fused with a histidine-tagged maltose binding protein, so that a traditional Ni^{2+}-affinity column could be used for the immobilization of the RNA transcript. The cleavage and release of the RNA of interest was achieved by using another ribozyme that is activated by a small molecule, glucosamine-6-phosphate.[47]

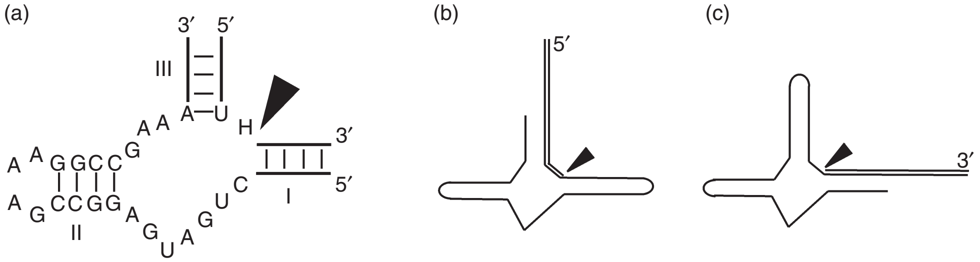


Figure 2 Hammerhead ribozyme. The filled triangle denotes the cleavage site. (a) An example of sequence in the catalytic center of the HHR. Stems I, II, and III are numbered; thin lines show Watson–Crick base pairs. 'H' stands for any nucleotide except G. (b) Design of the HHR at the 3′-end of the RNA transcript. After the self-cleavage, the RNA of interest (shown in a double line) has a homogenous 3′-terminus with the 2′–3′ cyclic phosphate. (c) HHR at the 5′-end of the RNA transcript producing a homogenous 5′-terminus with the terminal OH group in the RNA of interest.

Finally, certain abundant RNA molecules, such as tRNA, can be overexpressed and purified in quantities sufficient for NMR samples directly from *E. coli* cells; the RNA can be prepared unlabeled or isotopically labeled when grown on appropriate media.[8,48]

9.08.2.3 Enzymatic Synthesis of DNA

While enzymatic synthesis of RNA is cost effective compared to chemical synthesis even for unlabeled molecules, enzymatic synthesis of DNA is more expensive because of the higher costs of deoxyribonucleotide triphosphates (dNTPs), and it is used almost exclusively to produce isotopically labeled DNA samples. Several methods have been proposed and used to enzymatically synthesize labeled DNA for NMR studies. These methods can be divided into two groups, *in vitro* primer extension methods, and growing bacterial cells with a plasmid containing the DNA fragment of interest on an isotopically labeled minimal media. The most common setup for the primer extension reaction makes use of the Klenow fragment.[49–53] The Klenow fragment is a fragment of the *E. coli* DNA polymerase I devoid of the 5′–3′ exonuclease activity but retaining the 5′–3′ polymerase and 3′–5′ exonuclease activities.[54] The polymerization reaction requires a single-stranded DNA template and either a DNA or RNA primer; the two can be combined in a single chemically synthesized hairpin construct (**Figure 3**). In contrast to the RNA transcription, the DNA product remains covalently attached to the primer. If the 3′-terminal residue of the primer is ribonucleotide, then the DNA product can be easily cleaved off from the primer by incubation at alkaline conditions. Zimmer and Crothers[49] have found that the DNA yield is higher when using a mutant Klenow fragment that is additionally devoid of the 3′–5′ exonuclease activity,[55] however, this enzyme can produce longer DNA products beyond the template for certain sequences. To remove the nontemplate residues, the wild-type Klenow fragment with the intact 3′–5′ exonuclease activity can be added before the alkaline cleavage of the DNA product from the primer.[56] Other DNA polymerases can be used instead of the Klenow fragment with a similar setup, such as Taq DNA polymerase[56] and murine mammary leukemia virus reverse transcriptase.[57] The primer extension methods produce single-stranded DNA product; to prepare a double-stranded DNA, each strand needs to be synthesized separately. The amount of the DNA product is limited by the amount of the template and primer introduced into the reaction, so they also need to be prepared in milligram quantities.

In a different setup, a fragment of double-stranded DNA is amplified by DNA polymerase in a polymerase chain reaction (PCR) during repeated thermal cycling.[58–61] The DNA sequence must be designed with flanking sites for a restriction enzyme. The procedure starts with preparing a chemically synthesized double-stranded DNA of interest directly repeated two times. Most often, this tandem repeat is used both as a template and self-primers in PCR, which leads both to the amplification of the quantity of DNA and amplification of the number of repeats in a process called endonuclease-sensitive repeat amplification (ESRA).[58] Bidirectional primers can also be used.[59] The PCR is run in two steps, with the concentration of dNTPs increased for the second step. Also, adding single-repeat DNA as additional primers for the second step can increase the final DNA yield by twofold.[60] In the end, the multi-repeat product is cut with the restriction enzyme to produce single-repeat double-stranded DNA (possibly with overhangs, depending on the restriction enzyme used).

The above methods require one of the DNA polymerase enzymes and labeled dNTPs; both are available commercially or can be prepared in-house. In any case, the costs of making labeled DNA sample are higher

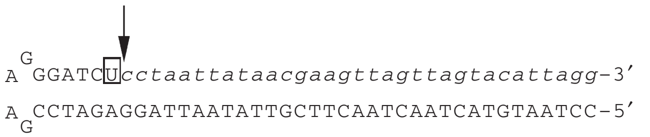

G
A GGATCUcctaattataacgaagttagttagtacattagg-3′
A CCTAGAGGATTAATATTGCTTCAATCAATCATGTAATCC-5′
G

Figure 3 Design of the DNA template and primer for the *in vitro* primer extension synthesis of DNA with the Klenow fragment. The product DNA (shown in lower case italics) is covalently attached to the 3′-terminus of the primer. For convenience, the template and primer are combined in a single monomolecular hairpin construct in this example. The primer ends with a single RNA residue (boxed); the site of the alkaline cleavage is shown with an arrow. If the RNA residue is placed several positions upstream in the primer, the portion of the primer downstream of that ribonucleotide will remain attached to the product DNA after the alkaline cleavage, and will therefore remain unlabeled.

than RNA, because of the higher costs of labeled dNTPs compared to NTPs. For example, Feigon and coworkers reported that growing *Methylophilus methylotropus* bacteria on labeled media at optimized conditions yielded 1.5 g NMPs and 0.5 g dNMPs from 31 l of culture.[56]

Finally, double-stranded DNA can be directly amplified by cloning it in a high-copy number plasmid and growing *E. coli* cells in a medium with ^{13}C-labeled glucose and $^{15}NH_4Cl$ as the only source of carbon and nitrogen, respectively.[58,62] Prior to growing bacteria on the labeled media, the repeat number of the DNA is amplified using the ESRA procedure (see above), and stable clones with multiple repeats are selected. It has been noted that stable cloning of sequences with multiple repeats may not be easily achieved.[58] This method does not require preparation or purchasing labeled dNTPs.

9.08.2.4 Segmental Isotopic Labeling

The main purpose of isotopic labeling is overcoming severe resonance overlap in larger molecules by conducting multidimensional heteronuclear experiments.[63] However, for highly repetitive sequences and with increase in molecular weight, the overlap in resonances catches up with these methods and again becomes severe. Preparing multiple samples where different parts of the molecule are labeled one at a time can significantly simplify NMR spectra, because it allows acquiring NMR signals only from the labeled portions. Alternatively, parts of the molecule could be deuterated to make them ‘invisible’ to NMR. Chemical synthesis of oligonucleotides (see above) is the most flexible approach in this respect, because it allows incorporation of labeled residues in arbitrary positions,[64,65] although, for large oligonucleotides this method can be prohibitively expensive. With the enzymatic synthesis, a straightforward approach is a type-specific isotopic labeling, for example, preparing nucleic acid molecules with all G’s labeled but the rest of the residues unlabeled, or with only A’s and U’s labeled.[66–68] However, this approach only partially solves the problem, because residues of the same type tend to have overlapping resonances.

Segmental labeling of enzymatically synthesized RNA molecules involves ligation of two (or potentially more) fragments, one isotopically labeled, and another unlabeled. Two strategies of RNA ligation have been reported for preparing milligram quantities of the product. In one, T4 DNA ligase was used to ligate two RNA fragments annealed to a continuous complementary DNA.[34] Creating oligomeric RNA molecules, circularization or joining RNA molecules in incorrect orientation is prevented in this approach, because precise base pairing to the cDNA at the junction is critical for the RNA–RNA ligation with T4 DNA ligase.

In another approach, T4 RNA ligase is used to join together two RNA oligonucleotides.[36] T4 RNA ligase catalyzes formation of the 3′,5′ phosphodiester linkage between one RNA fragment with a monophosphate at the 5′-terminus and another fragment with a hydroxyl group at the 3′-terminus.[69] Preparing the correct termini on RNA fragments is possible when using the HHR.[36] To prepare the 5′-fragment, the HHR is placed at the 3′-end of the transcript (**Figure 2**(**c**)). After the cleavage, this fragment has a hydroxyl group at the 5′-terminus, which is not a substrate for the T4 RNA ligase. The 3′-terminus of this fragment, produced by T7 RNAP, also has a hydroxyl group, which is a valid substrate for T4 RNA ligase. To prepare the 3′-fragment, the HHR is placed in the 5′-end of the transcript (**Figure 2**(**b**)). After the cleavage, this fragment has a cyclic 2′–3′ phosphate at its 3′-terminus, which is not a substrate for T4 RNA ligase. The 5′-terminus of this fragment normally has a triphosphate, produced by T7 RNAP, however, by priming the transcription reaction with GMP, the 5′-terminus is replaced with a monophosphate[70] and becomes a valid substrate for T7 RNA ligase.

An alternative method to prepare correct termini for the ligation with T7 RNA ligase is to dephosphorylate both ends of the 5′-fragment with *E. coli* alkaline phosphatase and phosphorylate both ends of the 3′-fragment with T4 polynucleotide kinase.[71,72]

For the segmental isotopic labeling of DNA, a variant of the primer extension method with the Klenow fragment (see above) has been used.[51,52] For example, to label only the 3′-part of the DNA molecule, the ribonucleotide within the chemically synthesized primer is placed not at the 3′-terminus but several residues upstream, such that after the alkaline cleavage, part of the primer (unlabeled) is included in the DNA product (see **Figure 3** legend). To label the 5′-part of DNA, the DNA is produced in two steps. At first, the labeled part is synthesized as usual with the primer extension method, cleaved off, and purified from the primer. Then, it is annealed to another, longer DNA template and again is extended with the Klenow fragment using unlabeled

NTPs. Because the product and the template have exactly the same length after the second step, the template can be biotinylated in the case if it needs to be separated from the product.[51]

9.08.3 Resonance Assignments

Assigning resonances of nuclei to specific frequencies is a critical step in structure determination. Obviously, errors in resonance assignments lead to errors in the resulting structures, which can be sometimes severe, such as incorrectly folded structures, and sometimes subtle. Unfortunately, there are presently no robust tools for finding possible errors in assignments. Because of that, structure determination must be an iterative process: after the structures are calculated, it is necessary to not only calculate average figures of merit (see below), but also to examine individual violations of structural restraints as some of them may be due to mis-assignments. Even though shorter nucleic acids can be solved by purely homonuclear NMR methods, isotopic labeling, and heteronuclear NMR methods allow for more reliable and more complete assignments, allow measurements of a greater number of and qualitatively different kinds of experimental structural restraints, which in its turn improves the accuracy of the solution structure. Assignment strategies for nucleic acids have been discussed in great detail in numerous reviews,[7,8,63,73–75] so we will cover them briefly with some emphasis on lesser discussed topics.

9.08.3.1 Spin System Assignments

The assignments process involves identification of spin systems within each residue and sequential assignments. A number of experiments are available for identifying the spin systems. Proton pairs directly connected via scalar through-bond $\mathcal{J}$-coupling interactions can be revealed in homonuclear two-dimensional (2D) COSY spectra;[76] multi-step $\mathcal{J}$-coupling interactions can be detected in 2D TOCSY experiments.[77–79] In contrast to antiphase multiplet structure of COSY peaks, TOCSY peaks have a simple in-phase structure, and therefore they have a better signal-to-noise ratio. While many intra-sugar proton correlations can be potentially detected in 2D COSY and TOCSY spectra, the resonance dispersion is particularly favorable in two spectral regions. One correlates base H6 and H5 protons for cytosines in DNA and for cytosines and uracils in RNA. Both H6 and H5 resonances have good dispersion, so this spectral region is also often used to assess the general quality of a sample, identify possible impurities, conformational species, and so on. Another spectral region correlates anomeric H1′ protons with the rest of the sugar protons. This is especially useful for DNA, where most sugars have predominantly C2′-*endo* puckers with both H1′–H2′ and H1′–H2″ $\mathcal{J}$-couplings in the 5–10 Hz range.[80,81] In contrast, sugars in helical regions of RNA have C3′-*endo* puckers with $\mathcal{J}_{\mathrm{H1'\text{-}H2'}}$ below 3 Hz; therefore, these peaks are only observable for flexible residues and nonstandard conformation (**Figure 4**(**a**)). A complete set of such correlations (**Figure 4**(**b**)) can be observed in ^{13}C-labeled RNA molecules by taking advantage of relatively large C–H and C–C $\mathcal{J}$-couplings. Experiments HCCH-COSY, HCCH-COSY-RELAY, and HCCH-TOCSY with a single, double, and multi-step COSY-type transfer of magnetization, respectively, between neighboring ^{13}C nuclei can be run either in a 2D version showing only correlations between the protons, or in a 3D version with ^{13}C selection.[84–87]

The HCCH-TOCSY experiment can also be used to correlate H8 and H2 protons in adenines taking advantage of relatively small (8–10 Hz) two-bond carbon–carbon couplings,[88–91] and its variant, HCCCH-TOCSY, can be used to correlate aromatic H6 protons with methyl groups in thymines in DNA.[92] The correlations between H6 and methyl groups in thymines are also usually observed in homonuclear 2D TOCSY spectra (**Figure 5(a)**), even though this four-bond $\mathcal{J}$-coupling is very small, approximately 1 Hz.[92] It is likely that these cross-peaks are observed via residual rotating frame cross-relaxation (ROESY) rather than through-bond $\mathcal{J}$-coupling interactions, despite the fact that the ROESY effect is minimized in modern ‘clean’ TOCSY pulse sequences.[79] Indeed, cross-peaks in 2D TOCSY can also be observed between H1′ and H8 in residues with the *syn* conformation around the glycosidic bond (**Figure 5(b)**) and even occasionally for sequential H2′(i)–H6/H8($i+1$) correlations in RNA (**Figure 5(c)**), where the corresponding interproton distance is very short (see below). The signal-to-noise ratio for such peaks is very low; therefore, it is not recommended to rely solely on such data during the assignments, but rather use them in combination with other assignment methods.

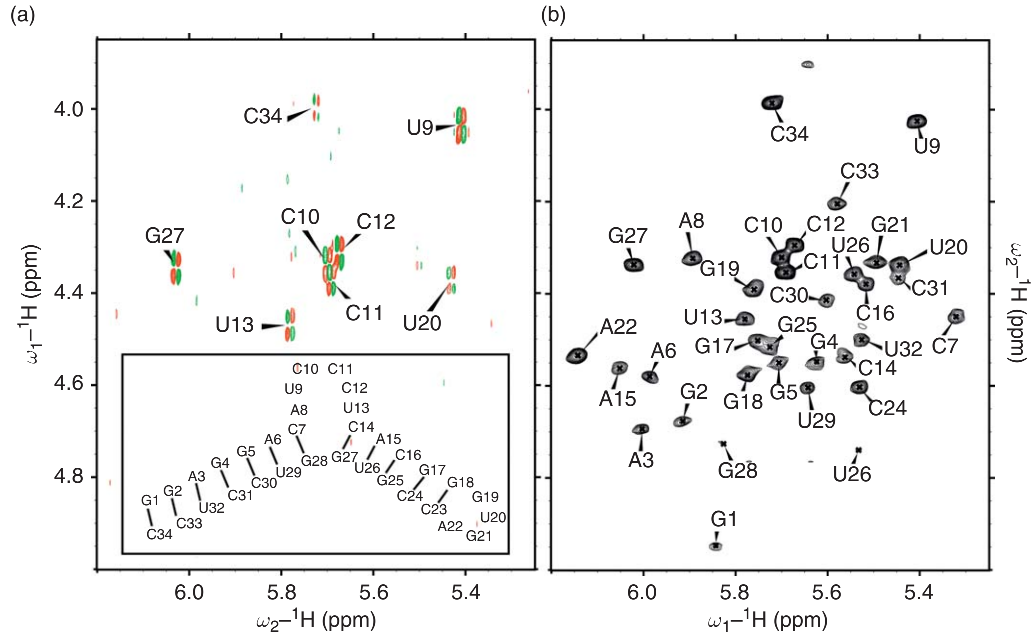


Figure 4 2D double quantum-filtered homonuclear COSY (a) and 2D version of HCCH-COSY (b) spectra of a 34-nt RNA from the stem-loop IV domain of the *Enterovirus* internal ribosome entry site;[68] the nucleotide sequence is shown in the inset. Positive components of the multiplet peaks are shown in red and negative components are shown in green. COSY H2′–H1′ cross-peaks, labeled in (a), are only observed for highly flexible residues associated with a 6-nt internal loop, for the 3′-terminal residue G34 and for U20 from the tetraloop GUGA. In contrast, all residues show H1′–H2′ cross-peaks in the HCCH-COSY spectrum. The spectrum was acquired with the spectral width of 1800 Hz in both dimensions; the symmetric region of the spectrum additionally contains aliased H6–H5 cross-peaks (not shown). All spectra shown in this chapter were acquired on a 600 MHz Varian Inova spectrometer, processed with the NMRPipe/NMRDraw[82] and annotated with the Sparky program.[83]

A series of through-bond experiments has been developed to correlate exchangeable protons with aromatic base protons: imino H1 proton with H8 proton in guanines, imino H3 proton with H5 and H6 in uracils, amino protons with H6 in cytosines and amino protons with H2 and H8 in adenines in uniformly ^{13}C,^{15}N-labeled RNA;[90,94–99] see also discussion of these experiments in Furtig *et al.*[7] Finally, aromatic protons H6 and H8 can be correlated with anomeric H1′ protons within the same residue (**Figure 6**) by establishing H6/H8–C6/C8–N1/N9 and H1′–C1′–N1/N9 connectivities via triple-resonance HCN experiments.[100,102–106] These experiments can be run either in 2D or 3D versions; in many cases, the N1/N9 resonance dispersion is sufficient to establish unambiguous H6/H8–H1′ correlations by acquiring the ^{1}H,^{15}N plane in the 2D version. The aromatic-to-anomeric proton correlations are very important for establishing sequential assignments using NOESY spectra (see below).

9.08.3.2 Sequential Assignments

Rigorous sequential assignments, that is, correlating nuclei in neighboring residues, require transferring magnetization along the backbone, including the phosphorus nucleus. Various variants of triple-resonance HCP experiments have been developed for this purpose.[107–111] The connections between sequential residues are established in these experiments by correlating the C4′ and H4′ nuclei with phosphorus in the same (n) and the 3′-neighboring ($n+1$) residues. Observing these correlations may be sometimes problematic, especially for helical regions, because of a limited resonance dispersion for C4′ and H4′. However, for residues in 'unusual' conformations, such as in internal loops, the dispersion of these nuclei is markedly better. In addition, a related HCP-CCH-TOCSY experiment extends the magnetization transfer to the C1′ and H1′ nuclei, which are much better resolved.[112] Besides, sequential connectivities can be established by correlating phosphorus with protons in unlabeled nucleic acids, using such experiments as

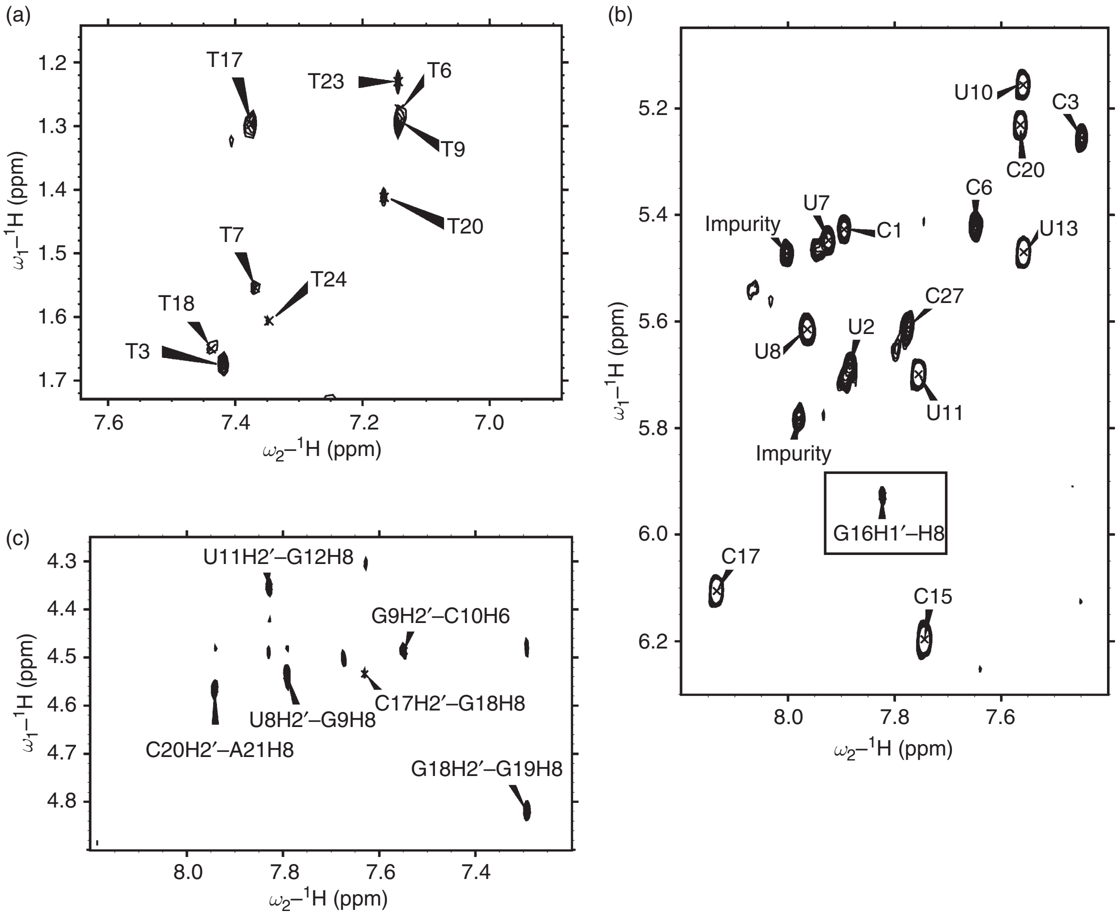


Figure 5 Cross-peaks in homonuclear 2D TOCSY spectra arising due to ROESY effects. 'Clean' TOCSY spectra were acquired with the MLEV-17 spin-lock sequence. (a) Base proton H6-to-methyl correlations in a 27-nt AT-rich DNA stem-loop structure;[93] the spectrum was recorded with the 50-ms mixing sequence. (b) and (c) TOCSY spectra acquired for a 31-nt stem-loop RNA (unpublished data). (b) H5–H6 cross-peaks in pyrimidines and a H1′–H8 cross-peak (boxed) in the syn guanine from the tetraloop UACG; the spectrum was recorded with the 30-ms mixing sequence. (c) Sequential H2′–H6/H8 cross-peaks; the spectrum was recorded with the 90-ms mixing sequence.

HETCOR[113] or hetero-TOCSY;[114,115] see also a review by Pardi[87] and references therein. Unfortunately, these methods are not routinely used for structure determination of nucleic acids, except for relatively short oligonucleotides, because of the relatively small chemical shift dispersion of ^{31}P and its fast relaxation via the chemical shift anisotropy mechanism.

The main method for establishing sequential assignments still remains the one based on the through-space dipolar interactions between protons within a short distance of each other. Such interactions give rise to nuclear Overhauser effects (NOE), which can be recorded in 2D or 3D NOESY experiments.[116–118] This method deals with nonexchangeable protons; therefore, the spectra are recorded with the sample in D_2O. If possible, the sample needs to be lyophilized and dissolved in high-grade D_2O to avoid the necessity of any water suppression, which can lead to disappearance of cross-peaks for protons resonating close to the water proton resonance frequency (H3′ in DNA, H2′ and H3′ in RNA).

Strictly speaking, the NOE-based method of sequential assignments is not rigorous, because it requires some assumptions about the structure of DNA or RNA. In the worst-case scenario, incorrect structural assumptions may lead to seemingly self-consistent, but still erroneous assignments. Fortunately, structured nucleic acids have the majority of residues in right-handed helical conformations, where the presence or absence of certain NOE cross-peaks does not depend on the details of the structure. NOE-based sequential

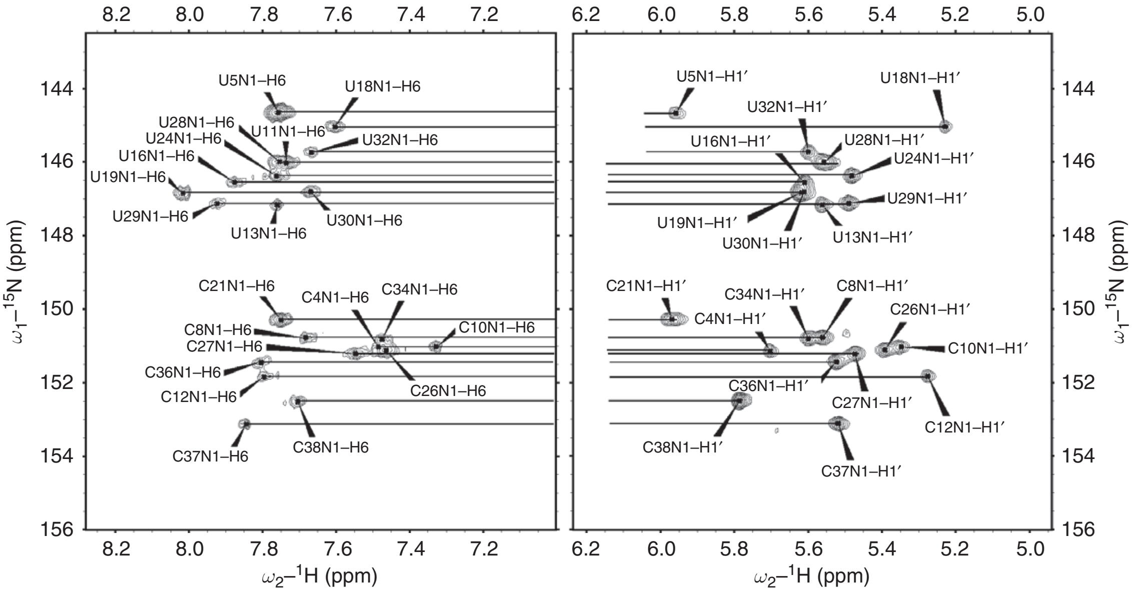


Figure 6 Correlations between H6 and H1′ protons established via common N1 nitrogens in a 38-nt RNA construct from the consensus stem D of the cloverleaf domain of 5′-untranslated region of enteroviruses; the nucleotide sequence is shown in the inset of **Figure 11**. Two 2D (^{1}H,^{15}N) versions of multiple-quantum HCN experiments[100] were acquired with optimization for the H6/H8–C6/C8–N1/N9 transfer (left panel) and for the H1′–C1′–N1/N9 transfer (right panel). Correlations for pyrimidine residues are shown; purine N9 nuclei resonate downfield between 168 and 172 ppm (not shown). Reproduced with permission from Z. Du; J. Yu; N. B. Ulyanov; R. Andino; T. L. James, *Biochemistry* **2004**, *43*, 11959–11972, Copyright (2004) American Chemical Society.

assignments for helical regions are based on the fact that many sugar protons are within the NOE distance from H6 or H8 aromatic protons from the same residue and from the downstream residue (**Table 1**), but not from the upstream residue. Anomeric H1′ protons are most useful for this purpose, because of better resonance dispersion for these protons; this spectral region is sometimes called a fingerprint region for nucleic acids. **Figure 7** shows an outline of a fragment of the assignment 'walk' (H1′A4, H8A4)–(H1′A4, H8G5)–(H1′G5, H8G5), and so forth. Establishing this walk is greatly facilitated by the HCN correlations (see above), which help distinguish intra- and inter-residue H1′–H6/H8 NOE cross-peaks. These connectivities can be interrupted between neighboring nts lacking stacking interactions or with nonstandard stacking interactions, such as in internal and apical loops. Therefore, it is useful to have several starting points for the assignment walk. In addition to 5′- and 3′-ends of the oligonucleotide, starting points for assignments can be found based on specific features of the nucleotide sequence that give rise to specific patterns in NOESY spectra. For example, all instances of two neighboring pyrimidines can be located with the help of H5 protons. Indeed, in addition to the cross-peaks H1′(n)–H6(n), H1′(n)–H6($n+1$), and H1′($n+1$)–H6($n+1$), common to all residues, the consecutive pyrimidines exhibit two strong cross-peaks H5(n)–H6(n) and H5($n+1$)–H6($n+1$) and a medium cross-peak H5($n+1$)–H6(n) in the same spectral region; two medium-to-strong cross-peaks H5(n)–H5($n+1$) and H1′(n)–H5($n+1$) are observed in the region of the anomeric diagonal, and a weak-to-medium cross-peak H6(n)–H6($n+1$) is observed in the region of the aromatic diagonal. In all cases, the diagonal regions have plenty of NOE cross-peaks that are useful not only for assignments but also for extraction of structural information (see below). It is important to remember that modern spectrometers are more sensitive, and larger molecules tumble in solution more slowly, which leads to more effective spin diffusion. Because of that, NOE cross-peaks can be sometimes observed at an interproton distance well above 5 Å. For example, at higher mixing times, sequential H1′–H1′ cross-peaks and even cross-strand H1′–H1′ cross-peaks can be observed in the region

Table 1 Typical interproton distances (Å) in helical regions

First proton	*Second proton*	*DNA*	*RNA*
H1′ (*n*)	H6/H8 (*n*)	3.6–3.9	3.5–3.9
H1′ (*n*)	H6/H8 (*n* + 1)	3.2–4.4	4.4–4.9
H2′ (*n*)	H6/H8 (*n*)	2.0–2.9	3.7–4.1
H2′ (*n*)	H6/H8 (*n* + 1)	2.4–3.7	2.0–2.2
H2″ (*n*)	H6/H8 (*n*)	3.5–4.2	n/a
H2″ (*n*)	H6/H8 (*n* + 1)	2.1–2.6	n/a
H3′ (*n*)	H6/H8 (*n*)	3.6–4.5	2.7–3.2
H3′ (*n*)	H6/H8 (*n* + 1)	4.0–5.0	3.1–3.6
H4′ (*n*)	H6/H8 (*n*)	4.5–5.0	4.0–4.6
H5′ (*n*)	H6/H8 (*n*)	4.0–6.5	3.6–4.1
H5″ (*n*)	H6/H8 (*n*)	4.5–5.4	4.0–4.4
H1′ (*n*)	H5 (*n*)	5.3–5.4	5.3–5.4
H1′ (*n*)	H5 (*n* + 1)	4.9–5.2	5.3–5.7
H2′ (*n*)	H5 (*n*)	4.3–4.7	5.2–5.5
H2′ (*n*)	H5 (*n* + 1)	2.3–3.0	3.6–3.8
H2″ (*n*)	H5 (*n* + 1)	3.7–4.1	n/a
H3′ (*n*)	H5 (*n*)	(6.0–6.6)[a]	4.6–5.1
H3′ (*n*)	H5 (*n* + 1)	4.0–4.5	3.5–4.1
H1′ (*n*)	H2′ (*n* – 1)	(5.3–7.0)[a]	4.0–4.5
H1′ (*n*)	H4′ (*n*)	2.9–3.4	3.2–3.6
H1′ (*n*)	H4′ (*n* + 1)	4.4–5.4	(5.6–6.2)[a]
H1′ (*n*)	H5′ (*n*)	4.5–4.8	4.6–4.8
H1′ (*n*)	H5′ (*n* + 1)	2.8–4.0	4.7–4.9
H1′ (*n*)	H5″ (*n*)	4.9–5.1	5.1–5.3
H1′ (*n*)	H5″ (*n* + 1)	4.6–5.5	(6.3–6.6)[a]
H6/H8 (*n*)	H5 (*n* + 1)	3.3–3.9	3.8–4.3
H2 A (*n*)	H1′ (*n*)	4.3–4.5	4.5–4.7
H2 A (*n*)	H2′ (*n*)	(6.5–6.8)[a]	4.7–5.1
H2 A (*n*)	H1′ (*n* + 1)	3.5–4.8	2.8–3.7
H2 A (*n*)	H6/H8 (*n* + 1)	(5.4–6.0)[a]	4.7–5.3
H2 A	H1′[b]	4.6–5.2	5.2–5.4
H2 A	H1′[c]	3.8–5.2	3.7–4.3
H2 A (*n*)	H2 A (*n* + 1)	3.5–3.7	4.2–4.7
H6/H8 (*n*)	H6/H8 (*n* + 1)	4.5–5.2	4.6–5.4
H5 (*n*)	H5 (*n* + 1)	3.8–4.3	3.8–4.0
H1′ (*n*)	H1′ (*n* + 1)	4.5–5.4	(5.3–5.8)[d]
H1′	H1′[e]	(6.0–7.7)[a]	(5.6–6.7)[d]

[a] These NOE cross-peaks are typically not observed.
[b] H1′ from the residue base-paired to the adenine.
[c] H1′ downstream from the residue base-paired to the adenine.
[d] These NOE cross-peaks can be observed at larger mixing times.
[e] H1′ two residues downstream in the opposite strand.
n/a, not applicable.

of the anomeric diagonal (see Figure 8 and Table 1). A higher mixing time is recommended for this observation not only to make the cross-peaks stronger (Figure 9), but also to decrease the intensity of the diagonal peaks via spin diffusion, which otherwise could mask weak cross-peaks.

Similar connectivities can also be established for the (H2′, H6/H8) and (H3′, H6/H8) cross-peaks, and, in the case of DNA, also for cross-peaks entailing H2″ and H6/H8, although the chemical shift dispersion is less favorable for H2′ and H3′ protons in RNA. Nevertheless, these spectral regions can also be very useful when used in combination with the walk in the anomeric-to-aromatic region. Figure 10 shows examples of 1D slices of 2D NOESY spectra through frequencies of aromatic protons for DNA and RNA, and Table 1 lists most common cross-peaks expected for residues in helical regions.

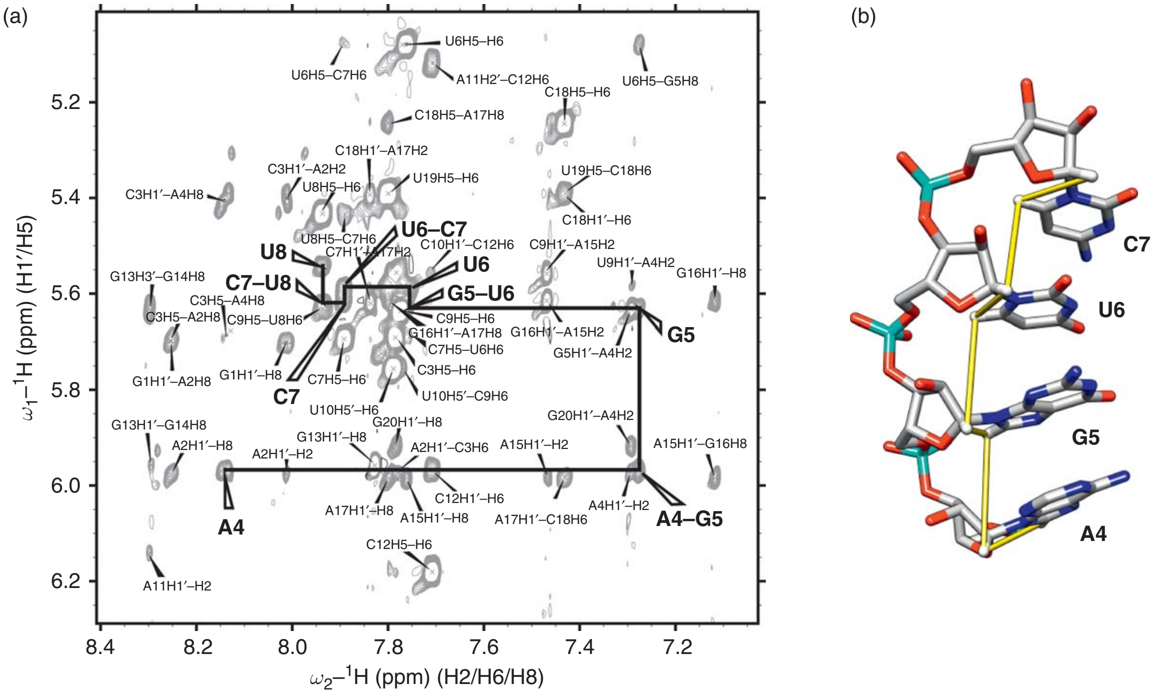


Figure 7 Fingerprint region of a 2D NOESY spectrum (a) of a 20-nt RNA hairpin from U4 snRNA acquired with a mixing time of 400 ms.[119] The H1′–H6/H8 'walk' is shown for the A4–G5–U6–C7 helical segment; the NMR structure of this segment is shown in (b). Yellow lines connect H1′ protons with aromatic H6 or H8 protons.

9.08.4 Extracting Structural Information

9.08.4.1 Detection of Hydrogen Bonds

Establishing base pairing patterns in nucleic acids gives perhaps the most important structural information – the one about the secondary structure of the molecule. Although base pairing can be predicted reliably in simple cases of duplex DNA or RNA, it is less obvious in an arbitrary case, because nucleobases can pair in a great variety of geometries (see, e.g., Leontis and Westhof[121]) or even form triples or quadruples. The simplest way to detect Watson–Crick AT, AU, GC, and wobble GU pairs is via a 2D NOESY spectrum recorded in a 90/10% H_2O/D_2O solvent.[122] To slow down the rapid exchange of imino and amino protons with solvent, this spectrum is often acquired at a slightly lower temperature (5–15 °C) and in slightly acidic buffer (~pH 6). However, one needs to be careful when refining structures based on data acquired at different temperatures, because change in temperature usually leads to some changes in chemical shifts and sometimes even to structural alterations (see, e.g., Lefevre *et al.*[123] and Ulyanov *et al.*[124]). The AU and AT Watson–Crick base pairs are easily distinguished[122] by a strong NOE cross-peak between the hydrogen-bonded imino proton of U or T and the H2 proton of A. GC pairs are distinguished by two strong peaks between the imino proton of G and resolved amino protons of C; the two amino protons of C are also correlated with each other by a strong NOE cross-peak.

It has been found that *J*-scalar couplings can be observed for nuclei connected by hydrogen bonds;[125] the nature of such couplings is similar to the *J*-couplings of nuclei connected by covalent bonds, that is, via interaction of nuclear spins with electron spins.[126] A series of heteronuclear experiments have been developed in the past decade for detecting hydrogen bonds between nucleobases by observing scalar coupling across the hydrogen bonds (see reviews of Furtig *et al.*,[7] Latham *et al.*,[8] Grzesiek *et al.*,[127]Cornish *et al.*,[128] and Dingley *et al.*,[129] and references therein). Hydrogen bonds NH–N involving imino protons can be detected in the HNN-COSY experiment (**Figure 11**) due to the *J*-coupling between the two ^{15}N nuclei, $^{2h}J_{NN}$ (symbol 'h' in the superscript denotes that out of the two bonds separating the ^{15}N nuclei one is actually a hydrogen bond). The

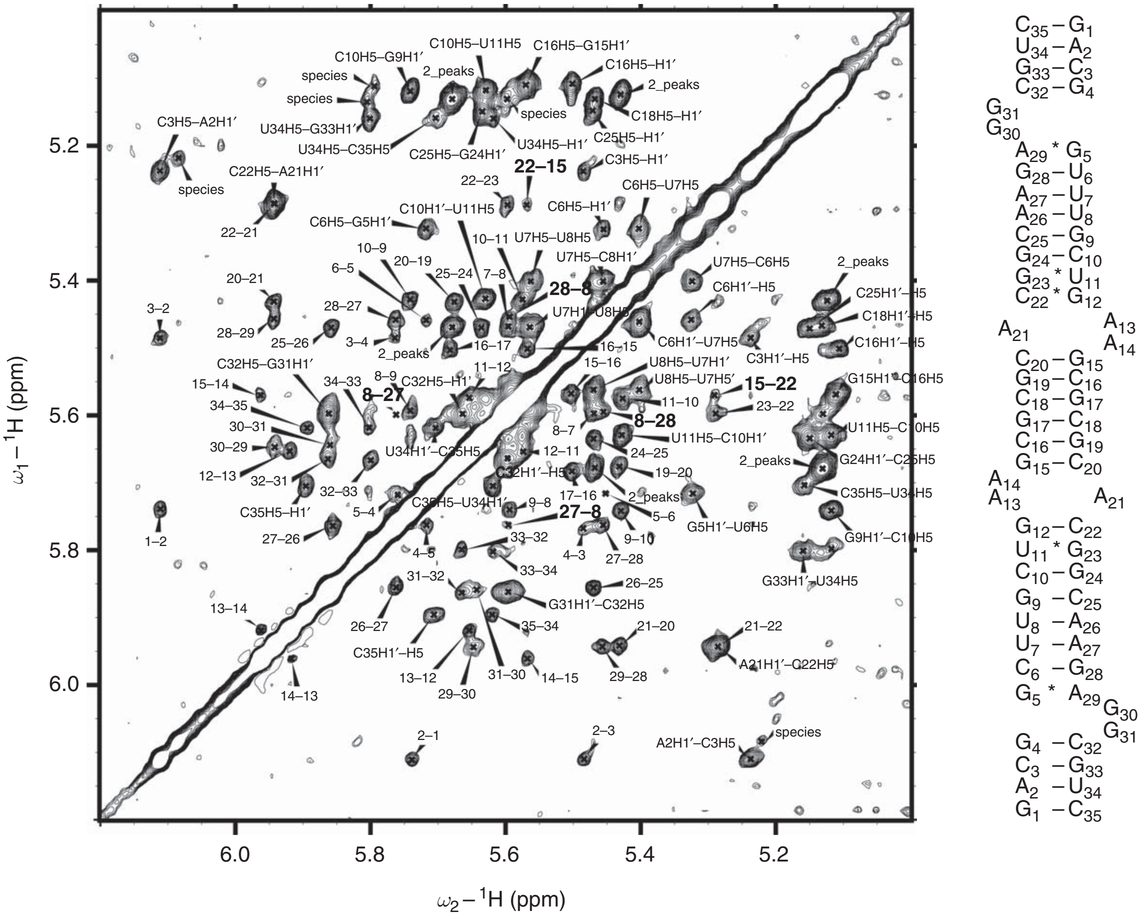


Figure 8 Region of the anomeric diagonal of the 300-ms 2D NOESY spectrum of the 35-nt extended dimer stem-loop SL1 RNA from HIV-1. Peaks labeled only with numbers denote residue numbers for H1′–H1′ cross-peaks; cross-strand cross-peaks are labeled in italics. Reproduced with permission from N. B. Ulyanov; A. Mujeeb; Z. Du; M. Tonelli; T. G. Parslow; T. L. James, *J. Biol. Chem.* **2006**, *281*, 16168–16177. Copyright © 2006 American Society for Biochemistry and Molecular Biology.

$^{2h}\mathcal{J}_{NN}$ was found to be in the range of 5–10 Hz for the Watson–Crick and Hoogsteen hydrogen bonds.[130–132] In hydrogen bonds involving amino groups NH_2–N, the ^{15}N frequencies on donor and acceptor groups are separated by approximately 150 ppm. A pseudo-heteronuclear variant of the HNN-COSY experiment with selective ^{15}N pulses has been developed for the detection of such hydrogen bonds, which are present, for example, in sheared A–A and other purine–purine base pairs.[133,134] The N–H–O═C hydrogen bonds can be detected with the selective long-range H(N)CO experiment.[134]

A H(CN)N(H) pulse sequence has been developed and used to detect the NH_2–N7 hydrogen bonds in base tetrads via the H8–(N2,N6) correlations despite the fact that the amino protons were not observed due to conformational exchange broadening.[135] The NH–N hydrogen bonds can also be detected in the absence of an observable imino proton, by correlating the imino ^{15}N nucleus with the nonexchangeable proton on the paired residue, adenine H2 for Watson–Crick interactions, or purine H8 for Hoogsteen interactions.[136] This experiment, the quantitative $^{2}\mathcal{J}_{HN}$ HNN-COSY, can be conducted even in a D_2O solvent, because the magnetization originates on nonexchangeable H2 or H8 protons and it is also detected on nonexchangeable H2 or h8 protons.

A hydrogen bond involving the ribose 2′-hydroxyl group, OH–N, have been detected for the stable tetraloop in the $^{1}H,^{15}N$ CPMG HSQC experiment,[137] even though hydroxyl protons are typically not observed due to the rapid exchange with water. Finally, intra- and inter-molecular hydrogen bonds in symmetric dimers

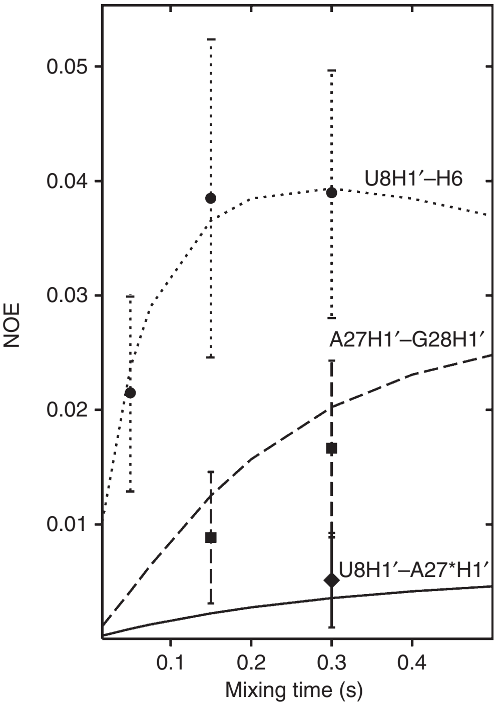


Figure 9 NOE intensities were simulated for the extended dimer stem-loop SL1 RNA (PDB 2GM0, first structure in the ensemble[120]) via CORMA using an effective correlation time of 31 ns for a series of mixing times (unpublished data). The curves (solid, dashed, and dotted) show calculated NOE intensities, and the symbols (diamond, squares, and circles) show normalized experimental NOE intensities for the cross-peaks U8H1′–A27*H1′, A27H1′–G28H1′, and U8H1′–H6, respectively (the asterisk denotes that the residue is from the symmetric strand). The vertical bars show estimated experimental errors in the intensities. The corresponding three distances in the structure used for the simulations are 6.74, 5.45, and 3.59 Å, and the lower and upper distance bounds calculated with RANDMARDI from the experimental data are 4.4–7.0, 4.0–7.7, and 3.1–4.5 Å, respectively. The experimental cross-peaks U8H1′–A27*H1′ and A27H1′–G28H1′ can be seen in **Figure 8** at a mixing time of 300 ms. Note that despite the relatively large size of the dimer (22.6 kDa), the intensity of weak cross-peaks can still benefit from further increase in the mixing time, while the medium-strength cross-peak, U8H1′–H6, starts decaying after 300 ms.

can be discriminated by comparing intensities of the $^{2h}\mathcal{J}_{\mathrm{NN}}$ HNN-COSY cross-peaks for fully ^{15}N-labeled samples and for the 1-to-1 mixtures of labeled and unlabeled samples;[138] this approach is similar to the asymmetric isotope labeling in combination with NOE measurements.[139–143]

Once the base pairing pattern is established, the structural restraints are generated for each hydrogen bond, usually as hydrogen bond length, and sometimes as hydrogen bond angle as well. The length and angle parameters for each type of hydrogen bond are usually derived from crystal structures of nucleic acids.

9.08.4.2 Nuclear Overhauser Effects and Interproton Distances

Interproton distances derived from the NOE data are very important types of structural information, in fact, generally, the most important basis for rigorous structure determination. It has become possible to solve high-resolution structures of proteins and short DNA duplexes (see, e.g., Wüthrich[73] andKaptein *et al.*,[144] and references therein) only after the introduction of an experimental technique, homonuclear 2D NOE spectroscopy, or NOESY, allowing measurements of many NOE cross-peaks.[116,145] The off-diagonal cross-peaks in the NOESY spectrum arise due to the exchange of magnetization between the nuclei during the mixing period of the experiment via dipole–dipole cross-relaxation. In short, NOE cross-peaks are observed only for relatively short interproton distances, the NOE intensity builds up with increasing mixing time, and the build-up is more efficient for larger molecules. A mathematical framework for calculations of interproton NOEs has been

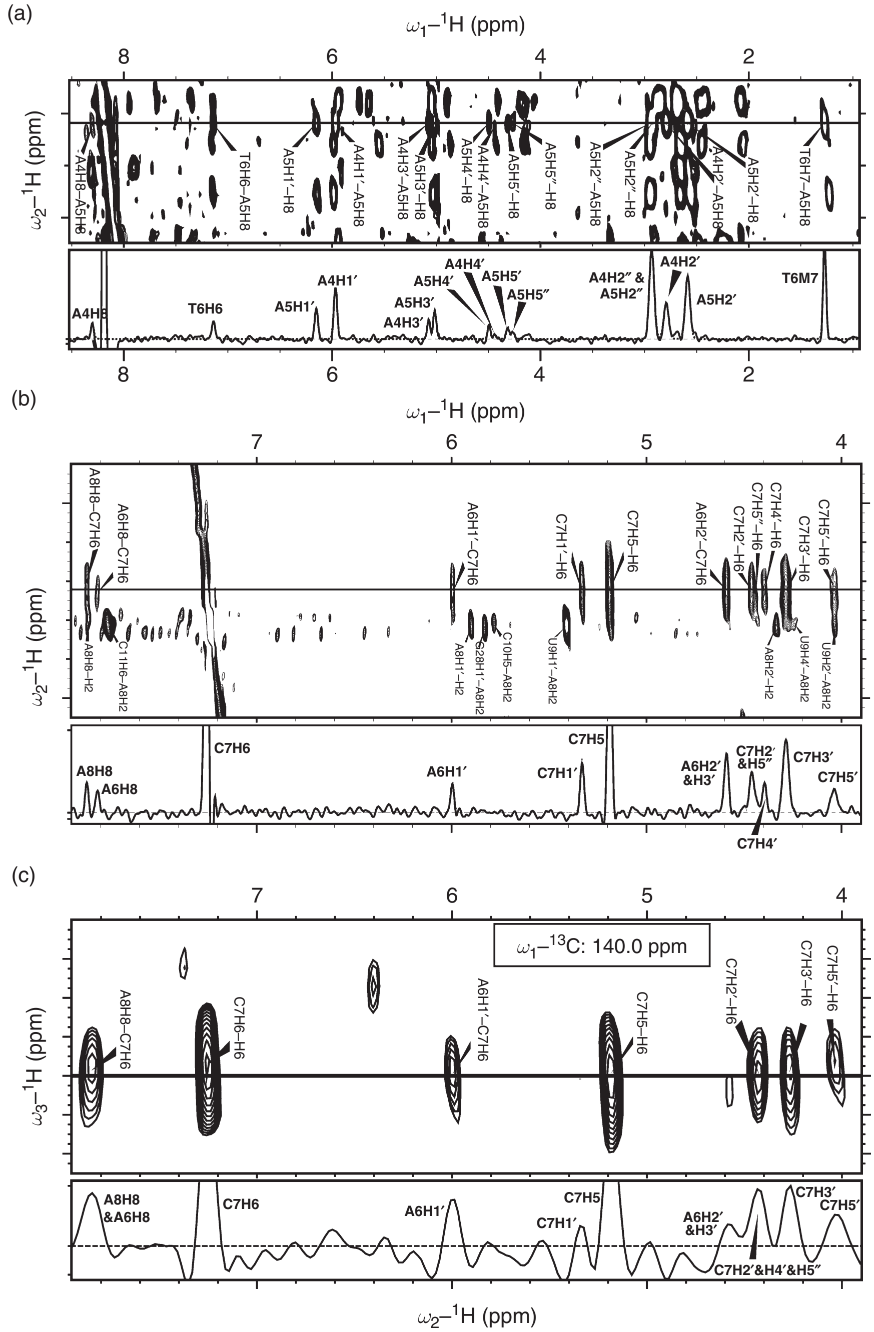


Figure 10 Portions of NOESY spectra and 1D slices through the frequencies of aromatic protons. (a) A 150-ms 2D NOESY spectrum of a 27-nt DNA stem-loop;[93] a slice through the frequency of A5H8 is shown. (b) A 200-ms 2D NOESY spectrum of a 34-nt RNA stem-loop;[68] a slice through the frequency of C7H6 is shown. Assignments of H5′ and H5″ protons are tentative. Note that some of the cross-peaks partially overlap with cross-peaks in another slice through the frequency of A8H2. (c) A 150-ms 3D ^{13}C-edited NOESY-HMQC spectrum of the same molecule shown in (b). A slice of the proton and carbon frequencies of H6 and C6 in residue C7 are shown. Note a significantly lower digital resolution in the indirect ω_2 dimension in this spectrum compared to the indirect ω_1 dimension in the 2D NOESY spectrum shown in (b).

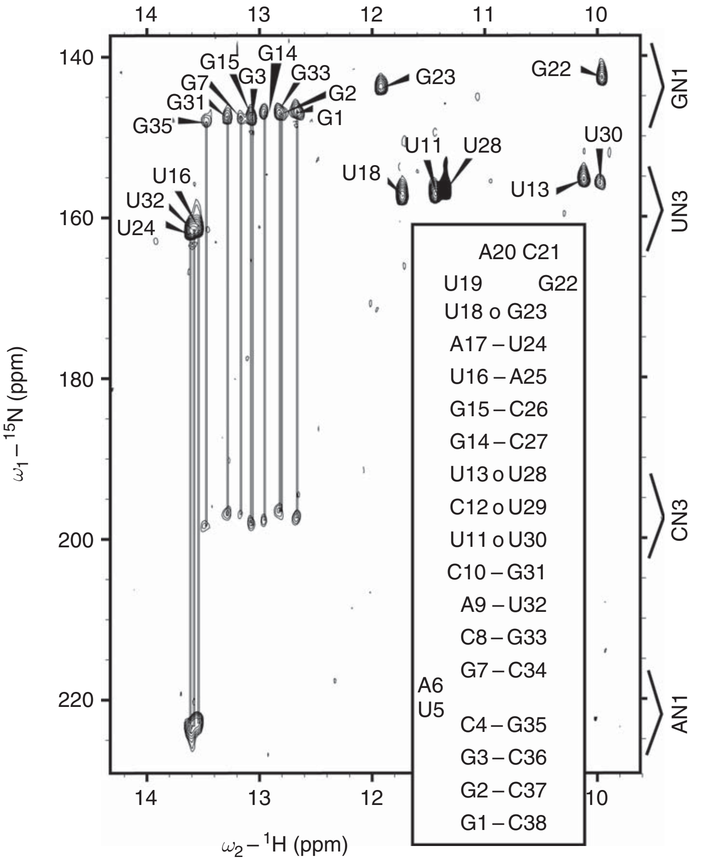


Figure 11 2D HNN-COSY spectrum of the 38-nt RNA construct from the consensus stem D of the cloverleaf domain of enteroviruses; sequence is shown in the inset. Labeled peaks in the upper part of the spectrum arise from the one-bond correlations of the NH imino groups. Cross-peaks in the lower part of the spectrum (not labeled) arise due to the scalar coupling (${}^2J_{\mathrm{NN}}$ of 5–7 Hz) between ^{15}N nuclei across the NH–N hydrogen bonds in Watson–Crick AU and GC pairs. They have the opposite phase compared with the diagonal NH peaks. Reproduced with permission from Z. Du; J. Yu; N. B. Ulyanov; R. Andino; T. L. James, *Biochemistry* **2004**, *43*, 11959–11972, with permission from the American Chemical Society. Copyright (2004) American Chemical Society.

established subsequently.[146–148] A matrix of NOE intensities **A** is related to a matrix of dipolar relaxation rates **R** by an exponential matrix expression:

$$\mathbf{A}(\tau_{\mathrm{m}}) = \exp(-\mathbf{R}\tau_{\mathrm{m}}) \times \mathbf{A}(0) \tag{1}$$

where τ_{m} is the experimental mixing time, the length of the mixing period in the three-pulse 2D NOE experiment. The off-diagonal terms in matrix **R**, the dipolar cross-relaxation rates R_{ij} between protons i and j, are inversely proportional to the sixth power of the interproton distances. The proportionality coefficients depend on the motional characteristics of the molecule; they increase with the rotational correlation time τ_{c}, that is, with the size of the molecule. A complete set of expressions for the matrix in Equation (1) for an isotropically tumbling rigid molecule is given in Keepers and James,[148] and the rate expressions for spins in rapidly rotating methyl groups are given in Liu *et al.*[149] CORMA[148] is a computer program used in our lab to evaluate Equation (1). Although the cross-relaxation *rates* depend only on the corresponding interproton distances, the resulting NOE *intensities* depend on the full relaxation network because of the matrix nature of the exponential equation (Equation (1)). This gives rise to the so-called spin diffusion, or indirect magnetization transfer, an effect when the observed NOE intensity is affected by the surrounding protons. Similarly to the direct magnetization transfer, the spin diffusion is also more effective at higher mixing times τ_m and higher correlation times τ_c, that is, for larger molecules.

Using Equation (1), it is possible to calculate theoretical NOE intensities for various molecular models and compare them with observed NOE data, see, for example, Keepers and James,[148] Massefski and Bolton,[150] and Suzuki *et al.*[151] To refine the molecular structure using NOE data, conceptually the most straightforward approach is to incorporate the NOE calculations directly into the refinement program.[152–160] However, with available computers at the time when these techniques were introduced, almost two decades ago, this was a computationally challenging task. An approach used instead in many labs was to estimate interproton distances from the NOE data at first, and then use distance restraints to refine the solution structure. This is still the most frequently used approach, although one might anticipate revisiting methods of direct refinement against NOE because of the dramatically increased power of modern computers.

The interproton distances can be estimated approximately, using a so-called isolated spin-pair approximation (ISPA), by ignoring the full relaxation network. This can be done by either using the initial slopes of the NOE build-up curves[161–163] (i.e., NOE measured at a series of mixing times), or qualitatively categorizing NOE intensities into weak, medium, and strong groups.[2,4] Both work well for the determination of solution structures of globular proteins, because even approximate distances extracted from the long-range NOEs (i.e., coming from non-neighboring residues) efficiently help define the protein fold during the refinement. On the other hand, such approximate distances have a diminished utility for the determination of extended structures with relatively few long-range NOEs, such as DNA or RNA duplexes. To a large extent, it was the interest in the sequence-dependent conformation of DNA in solution (see, e.g., Schmitz and James[164] and Ulyanov and James[165]) that motivated the development of full relaxation matrix methods for calculation of interproton distances from NOE data.

In the 'modified ISPA' approach,[166,167] the distances are calculated from the NOE intensities using special calibration curves, which take into account that short distances are typically overestimated and long distances underestimated in the classical ISPA due to the spin diffusion, but ignores individual differences between specific pairs of protons. The curves are calibrated based on NOE data for interproton distances with fixed values. Model calculations for a short DNA duplex showed that this approach produces good results,[166] however, it is expected that the errors should increase for larger molecules with more prominent spin diffusion.

Nevertheless, to calculate the interproton distances rigorously from the 2D NOESY cross-peaks, one needs to invert Equation (1), which is mathematically possible only when the *complete* matrix **A** of NOE intensities is available. In a typical NOESY experiment, however, many cross-peaks remain not quantified due to spectral overlap and incomplete resonance assignments; in addition, many cross-peaks are not detected because they are below the noise level. To solve this problem approximately, various iterative algorithms have been developed, including MARDIGRAS,[168,169] IRMA,[170] and MORASS.[171] For example, in MARDIGRAS, an algorithm developed in our lab, the iterations start with substituting all missing NOE intensities with intensities calculated from an arbitrary molecular model. Then, the relaxation rates are calculated from the hybrid NOE matrix using the inverted Equation (1), and the rates corresponding to the experimentally observed cross-peaks are substituted with the ideal rates calculated from the model. The hybrid rate matrix is then used to calculate the next approximation of the NOE matrix, and the process is iterated until convergence, after which the interproton distances corresponding to the observed NOE's are calculated from the relaxation rates.

The method was found to be relatively insensitive to the model structure.[169,172–175] Still, there is some residual dependence of calculated interproton distances on the model structure;[176] such dependence is expected to grow for larger molecules with higher correlation time τ_c. This dependence can be minimized by calculating the distances and refining the molecule in two or more iterations: using a starting model to calculate first set of distances and refining a preliminary structure, and then using this preliminary structure to calculate the final set of distances.[120,143] To obtain meaningful and accurate distances, it is important to accurately integrate the intensities of NOE cross-peaks, which could be achieved, for example, with the line-fitting procedures incorporated in SPARKY software.[83] Only nonoverlapped or successfully deconvoluted peaks should be used for distance calculation. Equally important is to account for random errors in experimental NOE intensities. For this purpose we developed a procedure, RANDMARDI, which calculates lower and upper distance bounds based on repeated MARDIGRAS calculations for randomly perturbed experimental intensities.[177] The extent of random perturbations must correspond to realistic estimates of experimental errors.

If the experimental errors are underestimated, it can lead to tight but inaccurate distance bounds. Conversely, the overestimated errors can lead to unnecessarily imprecise distance bounds. RANDMARDI takes into account two types of experimental errors: relative integration errors and absolute errors due to spectral noise. The first kind can be estimated, for example, by comparing intensities of symmetric peaks below and above the diagonal, and the second type can be estimated as 50–200% of the lowest quantifiable peak, depending on the spectrum quality.

In addition to the arbitrary model, distance calculations with MARDIGRAS require isotropic rotational correlation time τ_c as input parameter. Effective rotational correlation time can be estimated by a number of experimental approaches.[176] An approach that usually produces self-consistent results is to estimate τ_c based on the same NOESY data that are used for distance calculations. MARDIGRAS can be run at a series of correlation times, and a τ_c range can be selected that reproduces best fixed interproton distances and distances with limited variation, see, for example, Ulyanov *et al.*[120] For that purpose, the experimental NOE intensities (which are integrated in arbitrary units) must be normalized based on the total sum of all observed intensities; if possible, intensities of diagonal peaks must also be integrated and included to make the dependence of calculated distances on τ_c more apparent, see a discussion in Tonelli.[176] Fixed interproton distances and distances with limited variation in nucleic acids are listed in **Table 2**.

The full relaxation matrix approach for distance determination from 2D NOESY data has been used in many groups to refine solution structures of DNA, including mismatched duplexes, variously chemically modified duplexes, complexes with small molecules, and so on.[93,174,176,178–195] However, there are fewer RNA structures

Table 2 Fixed distances and distances with limited variation between nonexchangeable protons in nucleic acids[a]

Protons	*Lower bound*	*Upper bound*
H1′–H2′	2.7	3.0
H1′–H2″[b]	2.2	2.4
H1′–H3′	3.8	4.0
H1′–H4′	2.9	4.0
H2′–H2″[b]	1.8	1.8
H2′–H3′	2.3	2.5
H2′–H4′	3.8	3.9
H3′–H4′	2.6	3.1
H2″–H3′[b]	2.7	3.1
H2″–H4′[b]	2.8	4.2
H1′–H5′/H5″[c]	4.0	5.3
H2′–H5′/H5″[c]	2.3	5.4
H3′–H5′/H5″[c]	2.2	3.9
H4′–H5′/H5″[c]	2.1	3.0
H2′–H5′/H5″[b,c]	3.8	5.4
H5–H6[d]	2.4	2.5
H6–M7[e]	2.9	2.9
H2–H8[f]	6.4	6.4

[a] Interproton distances (Å) are calculated assuming the aliphatic and aromatic C–H distances of 1.09 and 1.08 Å, respectively. The distance variations correspond to the sugar conformation variation covering the range of pseudo-rotation phase angle between −30° and 210° with amplitude of pseudo-rotation between 36° and 41°, that is, excluding only the most unfavorable sugar puckers. The calculations have been carried out with the miniCarlo program.

[b] DNA only.

[c] To determine the ranges of distances involving H5′/H5″ protons, the backbone torsion angle gamma was additionally varied from 0° to 360°.

[d] Cytosine or uracil.

[e] Third-root averaged distance between H6 and methyl protons in thymine.

[f] Adenine.

determined using this method,[66–68,119,120,196–201] where approximate methods of distance determination are used more often. There may be several reasons for that. One reason is that for many RNAs, the main scientific interest is in its global fold and not in relatively subtle structural features, such as sequence-dependent bending of DNA duplexes, thus justifying the elimination of a time-consuming process of accurate NOE integration and distance calculation. The other reason is that because of the relative ease of synthesizing labeled RNA molecules, accurate interproton distances can be substituted with other types of structural information derived from heteronuclear NMR, such as residual dipolar couplings (RDC) (see below), while most of the DNA structures have been determined using homonuclear data. Finally, for larger RNA molecules, the resonance overlap in homonuclear 2D NOESY spectra can preclude accurate integration of NOE intensities. Nevertheless, for moderate size RNA, extracting accurate distances from NOE data acquired at a series of mixing times can be beneficial for better defining the structure; MARDIGRAS combined with the random error analysis allows accurate if not very precise estimates of bounds for distances well above 6 Å. For example, intra-residue adenine H2–H8 cross-peaks have been observed at 300 ms for the extended dimer of SL1 RNA from HIV-1 for adenines A13, A14, A21, and A26;[120] these cross-peaks correspond to a fixed distance of 6.4 Å (**Table 2**). The RANDMARDI procedure produced distance bounds of 5.3–7.8, 5.4–7.4, 5.7–7.9, and 4.8–7.4 Å for these four cross-peaks, respectively, thus justifying the use of distance restraints for such long interproton distances. While the intra-residue H2–H8 distances are of no use for structure determination, there are several inter-residue H2–H8 and H2–H6 cross-peaks observed at higher mixing times in the region of the aromatic diagonal of the 2D NOESY spectra (not shown) and H1′–H1′ cross-peaks in the region of the anomeric diagonal (**Figure 8**). For example, the distance restraints of 4.0–7.7 Å for the sequential distance A27–H1′–G28H1′ and especially 4.4–7.0 Å for the cross-strand distance U8H1′–A27*H1′ (**Figure 9**) are very helpful for better defining the RNA conformation, even despite the relatively large error bars. Indeed, even though the sequential H1′–H1′ distance is typically within the range of 5.3–5.8 Å for helical regions of RNA (**Table 1**), it can be beyond 10 Å for certain RNA conformations, and there is no theoretical upper limit for cross-strand distances.

9.08.4.3 Scalar Coupling Data

The magnitude of scalar (*J*) couplings between nuclei separated by rotatable bonds depends on the value of the dihedral (torsion) angle,[202,203] which can be used in refinements as structural information. In addition to classical *J*-correlated spectroscopy (COSY)[76] a number of experimental techniques have been developed for observing and measuring homonuclear and heteronuclear *J*-coupling constants, including E.COSY[204] and quantitative *J* correlation,[205] reviewed in Bax *et al.*[206] Specific applications of these techniques to nucleic acids are discussed in a very comprehensive review by Wijmenga and van Buuren;[167] see also applications of the constant time HSQC technique for measuring $^3J_{\mathrm{CP}}$ and constant time COSY for $^3J_{\mathrm{HP}}$ couplings[207,208] and a discussion of line-fitting of homonuclear COSY peaks using the ACME program for measurements of small proton–proton couplings.[209] A discussion of an older approach for measurement of $^3J_{\mathrm{HH}}$ couplings using the SPHINX and LINSHA programs[210] can be found.[164]

To briefly summarize these approaches, the backbone beta torsion angle (O5′–C5′) can be estimated from the $^3J_{\mathrm{H5'/5''P5}}$ or $^3J_{\mathrm{C4'P5}}$ couplings, and the epsilon torsion (C3′–O3′) can be estimated from $^3J_{\mathrm{H3'P3}}$, $^3J_{\mathrm{C4'P3}}$, or $^3J_{\mathrm{C2'P3}}$. The parameterizations for the generalized Karplus equations for these couplings are given in Mooren *et al.*[211] Also, the conformations for beta and epsilon torsions can be established qualitatively based on intensities of cross-peaks in the triple-resonance 3D HCP experiment.[75,108] The gamma torsion angle (C4′–C5′) can be estimated using a properly parameterized Karplus equation for the $^3J_{\mathrm{H4'H5'}}$ and $^3J_{\mathrm{H4'H5''}}$ couplings, or qualitatively based on $^3J_{\mathrm{H5'C3'}}$, $^3J_{\mathrm{H5''C3'}}$, $^2J_{\mathrm{H5'C4'}}$, $^2J_{\mathrm{H5''C4'}}$; the same couplings can help establish stereospecific assignments for H5′ and H5″ protons.[167,212] The glycosidic torsion angle can be estimated from the $^3J_{\mathrm{H1'C4/2}}$ and $^3J_{\mathrm{H1'C8/6}}$ couplings; the corresponding Karplus equation parameters have been derived.[213]

There are many homonuclear and heteronuclear couplings that are sensitive to the conformation, that is, pucker, of the five-membered sugar ring. The parameters for generalized Karplus equations for $^3J_{\mathrm{HH}}$ couplings have been given,[214,215] and parameters for $^3J_{\mathrm{H3'C1'}}$, $^3J_{\mathrm{H2'C4'}}$, $^2J_{\mathrm{H2'C1'}}$, $^2J_{\mathrm{H3'C2'}}$, $^2J_{\mathrm{H2'C3'}}$, and $^2J_{\mathrm{H3'C4'}}$ have been reported.[213]

It is possible to use scalar coupling data directly during refinement of NMR structures using appropriately parameterized Karplus equations

$$\mathcal{J}(\varphi) = A\cos^2\varphi + B\cos\varphi + C \tag{2}$$

where φ is the torsion angle, and A, B, C are parameters specific for the $\mathcal{J}$-coupling (φ). However, the functional form of generalized Karplus equations for $\mathcal{J}$-couplings for the sugar ring is different from Equation (2),[214,215] which is typically not implemented in refinement programs. More commonly, torsion angles are first estimated from the experimental $\mathcal{J}$-couplings[167] and then used as restraints during refinement. Sugar ring conformational parameters, pseudo-rotation phase angle and pseudo-rotation amplitude, can be estimated from the experimental $\mathcal{J}$-couplings using the PSEUROT program,[216] and then used directly as restraints or further converted into exocyclic torsion angles, depending on the refinement program.

In RNA, most residues have C3′-*endo* sugar puckers (N-conformations) with small $^3\mathcal{J}_{\mathrm{H1'H2'}}$ coupling (<2–3 Hz) and with the corresponding H1′–H2′ cross-peaks not observed in homonuclear 2D COSY or TOCSY spectra. These cross-peaks are observed only for flexible residues and residues locked in the S-conformations; the presence or absence of TOCSY H1′–H2′ cross-peaks is often used for qualitative estimation of sugar puckers.

9.08.4.4 Residual Dipolar Couplings

Direct through-space interactions between magnetic dipoles depend on the dipole orientations in such a way that the interactions average to zero for molecules tumbling freely in isotropic solutions, giving rise to sharp NMR signals. However, if a molecule is oriented relative to the magnetic field, such interactions no longer average to zero, and the RDC can be observed.[217–220] The orientation can be achieved by magnetic field after dissolving the molecule in dilute liquid crystalline media, such as phospholipid bicelles,[221,222] filamentous phage Pf1,[223,224] *n*-alcyl-PEG mixture with *n*-hexanol,[225] or even by magnetic field alone.[226,227] The RDC values depend on the nature of the nuclei, the distance between the nuclei, the average orientation of the internuclear vector relative to the magnetic field, and the degree of orientation. To maintain the NMR signals sharp, the degree of orientation must be very small, such that the dipolar interactions are on the order of 0.1% of their full values.[228] Mathematically, the RDC value D can be calculated for the internuclear vector in the matrix form, using a symmetric alignment tensor with zero trace (with five independent matrix elements), see, for example, Tsui *et al.*[229] Alternatively (see, e.g., Clore *et al.*[230] and Bax *et al.*[231]), it can be written as

$$D(\theta,\varphi) = D_{\mathrm{a}}\left(3\cos^2\theta - 1 + \frac{3}{2}R\sin^2\theta\cos 2\varphi\right) \tag{3}$$

where angles θ and φ describe the orientation of the internuclear vector in the frame of the diagonalized alignment tensor, and D_{a} and R are the alignment tensor characteristics called the magnitude of the residual dipolar coupling tensor and the rhombicity. Three more parameters not included explicitly in Equation (3) are the three Euler angles defining the alignment tensor orientation. The magnitude D_{a} depends not only on the alignment tensor, but also on the nature of the two nuclei; in some refinement programs this parameter is always scaled for the N–H dipolar interactions, so the experimental RDC values must be scaled accordingly.[232] Although it is possible to extract the orientations of chemical bonds from the RDC data,[233] usually the RDC data are used directly in refinements, either via the explicit matrix expression or via Equation (3).

The alignment tensor fitting experimental RDC values can be calculated for a given molecular structure using the singular-value decomposition method.[234–236] For an unknown structure, parameters D_{a} and R can be estimated based on analysis of the distribution of experimental RDC values, see, for example, Bax *et al.*[231] However, this method works best only for proteins, where the distribution of internuclear vectors with measured RDCs is relatively uniform. For nucleic acids, D_{a} and R can be estimated using grid search and preliminarily refined structures.[221,237] Alternatively, the alignment tensor can be kept unconstrained and optimized together with molecular conformation during structure refinement.[120,229] The latter approach has an advantage of not artificially restricting the conformational variability in the ensemble of refined structures by fixing the alignment tensor; also see Tjandra *et al.*[221] for a discussion of this problem.

Practically, RDCs are measured using the same methods as for *J*-couplings, specifically as a difference in couplings observed in the aligned media and couplings observed under isotropic conditions. It is easier to measure RDC for C–H and N–H vectors with large one-bond couplings. However, RDCs can be measured for a variety of one- and two-bond C–C, N–C, N–H, C–H and also for P–H and homonuclear H–H interactions that may not even be coupled under isotropic conditions (reviewed in Latham *et al.*[8]).

Owing to the orientation dependence, the RDC restraints define the global orientation of internuclear vectors, making them fundamentally different from the distance and torsion angle restraints, which define relative positions of nuclei. This property makes RDC restraints especially valuable for elongated nucleic acid molecules, where the experimental errors in relative restraints tend to propagate along the polynucleotide chains. Incorporation of RDC data in refinement of nucleic acid structures substantially improves the accuracy and precision of global conformations.[238–240] During the past decade, using RDC data became routine in structure determination of nucleic acids, both for DNA[52,53,191,221,241–247] and RNA.[68,101,120,248–278]

9.08.4.5 Other Structural Restraints

In this section we briefly mention the rarely used or newly emerging types of structural information that can aid in refinements of solution structures. Only applications to nucleic acids will be listed, even though many of these methods have been previously applied to protein structure determination.

When a paramagnetic molecule is present in solution, the magnetic dipoles of its unpaired electrons cause strong perturbations of chemical shifts of surrounding nuclei, called pseudo-contact shifts. A structure of a DNA duplex in complex with chromomycin A3 and a divalent metal was solved based on pseudo-contact shifts measured as difference in chemical shifts of Co^{2+} and Zn^{2+} complexes.[279,280] The structure was refined together with the magnetic susceptibility tensor, the knowledge of which is necessary to calculate the pseudo-contact shifts. Because of the long range nature of pseudo-contact shifts, the structure was defined to a much higher degree than is typical for NOE-based refinements.

Chemical shift anisotropic (CSA) tensor of each nucleus is reduced to its isotropic average value for molecules in isotropic solutions. However, if a molecule is partially aligned in a liquid crystalline solution (see above), the incomplete CSA tensor averaging leads to a difference in the chemical shifts observed under isotropic and aligned conditions ($\Delta\delta$) on the order of a few to tens of parts per billion for the degrees of alignment typically used for RDC measurements. These changes in chemical shifts can be calculated for a given structure based on the molecular alignment tensor (see above), provided that the principal components of the CSA tensor and their orientation relative to the molecular frame are known.[281] The $\Delta\delta$ data can provide efficient orientation restraints. ^{31}P $\Delta\delta$ data were used in refinement of a DNA duplex[242] using the ^{31}P CSA tensor measured by single-crystal NMR for the phosphodiester diethyl phosphate.[282] The only caveat for this approach is an assumption that the CSA tensor does not depend on molecular conformation. Nevertheless, magnitudes and orientations of CSA tensors for ^{31}P and sugar carbons have fairly uniform values for the helical regions. This has been demonstrated by fitting observed $\Delta\delta$ data to the helical residues of a stem-and-loop RNA structure previously refined using RDC data.[283,284] Still there may exist a conformation-dependent variability of CSA tensors, even though relatively small; there remains a concern that fixing the CSA tensors to particular values may artificially narrow the conformational envelope of refined structures. The ^{13}C CSA tensor magnitudes are more conformation-dependent for some base carbons, for example, for pyrimidine C6 carbons in B-form DNA vs. A-form RNA.[285] Such a dependence is likely even greater for nonhelical residues.

When the positions of the downfield components of ^{1}H–^{13}C TROSY HSCQ cross-peaks[286,287] are compared under the isotropic and aligned conditions, both RDC and CSA effects contribute to the observed difference: $\Delta\delta' = \Delta\delta + \mathrm{RDC}/2$. These values, referred to as pseudo-CSA, can be used directly in molecular refinements.[288] The reason for using combined $\Delta\delta'$ values, rather than $\Delta\delta$ and RDC separately, is that it is easier to measure accurately the positions of downfield TROSY components for larger molecules, because of the optimized line width of these components.

In addition to using *J*-scalar coupling data (see above), several other methods have been proposed to estimating sugar conformations in RNA: based on the ^{13}C–^{1}H dipole–dipole cross-correlated relaxation,[289] based on the cross-correlated relaxation rates involving ^{13}C CSA and ^{13}C–^{1}H dipolar interactions,[290] and based on the ^{13}C chemical shifts of sugar carbons.[291]

Isotropic chemical shifts of protons are very sensitive to the environment and as a result are very conformation-dependent. It is possible to calculate ^{1}H chemical shifts from the nucleic acid structure.[292–295] The observed chemical shift values are usually represented as a sum of 'random coil', or reference values, and a conformation-dependent part: $\delta_{\text{obs}} = \delta_{\text{ref}} + \delta_{\text{conf}}$. The conformation-dependent δ_{conf} is calculated from a structure as a sum of two components, the ring current effect and the contribution of the electric field created by partial atomic charges of the molecule. The electric field contribution was shown to be minor for nucleic acids,[293,294] but this conclusion may be structure-dependent to some extent. The reference values δ_{ref} are calibrated based on a set of reference structures. The proton chemical shifts could be used as restraints during structural refinements, but more typically they are back-calculated from the refined structures for validation purposes, see, for example, Flodell *et al.*[266]

Finally, NMR data can be used in combination with data of other methods to determine solution structures of nucleic acids. For example, the homology model of *E. coli* tRNA$^{\text{Val}}$ based on the X-ray structure of yeast tRNA$^{\text{Phe}}$ was refined based on experimental RDC data and small angle X-ray scattering (SAXS) data.[296]

9.08.5 Three-Dimensional Structure Refinement

The details of computational approaches for determination of NMR structures have been extensively reviewed,[164,237,297–302] so only some general considerations will be discussed here. Refinement of nucleic acid structures based on experimental NMR data can be carried out with any molecular simulation or molecular modeling software that have options for calculating NMR parameters for simulated structures. Some examples of such general-purpose programs are AMBER,[303] GROMOS,[304] XPLOR,[305] CNS,[306] NIH version of XPLOR,[232,307] and DYANA.[308] Some programs specialized for nucleic acids modeling, such as miniCarlo[309] or JUMNA[310] are also capable of refining structures against NMR data.[300,311] Most of these programs are not especially user-friendly, so the selection of software is often dictated by expertise present in a particular lab. However, this choice also depends on the availability of options for calculating specific NMR parameters acquired in the experiment. From this perspective, the NIH version of XPLOR is arguably one of the most advanced for NMR refinement of structure. In addition to the traditional XPLOR interface, it also has a Python wrapper, which allows great flexibility in designing refinement protocols and developing custom potentials. The downside is that it takes a Python programmer to fully utilize this program. Nevertheless, many sample Python scripts are distributed together with the software, which help a novice learn this program.

The purpose of the refinement is to find a stereochemically sound structure or a set of structures that satisfy all experimental restraints. This is achieved by optimizing the total energy of the system defined as a sum of conformational energy of the molecule and the pseudo-energy of restraints: E_{conf}. The conformational energy E_{conf} is calculated according to a general-purpose force field or one specialized for nucleic acids (for a review, see Orozco *et al.*[312]). The role of the restraint energy E_{restr} is to enforce the experimental restraints; it can be either in the form of a simple harmonic potential or a flat-well potential[313,314]

$$E_{\text{restr}} = \begin{cases} k(x - x_{\text{lower}})^2, & \text{for } x < x_{\text{lower}} \\ 0, & \text{for } x_{\text{lower}} \le x \le x_{\text{upper}} \\ k(x - x_{\text{upper}})^2, & \text{for } x > x_{\text{upper}} \end{cases} \tag{4}$$

where x is an observable NMR parameter calculated for the structure, and x_{lower} and x_{upper} are the experimentally determined bounds for this parameter; k is a user-defined force constant. The form of potential described by Equation (4) does not penalize the molecule when the NMR parameter remains within the experimental uncertainty, but quickly builds up when it deviates from the observed values. Sometimes this form is further modified by making the potential linear when the calculated parameters deviate too far either from x_{lower} or x_{upper}[178] to avoid an overly strong build-up of energy, which could interfere with some refinement engines, such as molecular dynamics (MD). When all observed parameters are self-consistent, the flat-well potential has a single (within experimental uncertainty) region of global minimum; a simple harmonic form of the penalty function is intended to simplify the otherwise rugged potential surface of the molecule.

By far the most popular method of optimizing the total energy E is via simulated annealing (SA) protocols within restrained MD.[315] In MD simulations, the Newtonian equation of motion is solved for a molecule coupled to a thermal bath.[316] During SA, the molecule is simulated at first at high temperature, and then the temperature is slowly reduced; this procedure helps avoid entrapment in local energy minima. The Metropolis Monte Carlo method[317] can be used to generate the Boltzmann distribution of molecular conformation at a given temperature; this method is also applicable for setting up the SA procedure.[300] The molecule is often refined in the space of atomic Cartesian coordinates. However, it is also possible to refine NMR structures in the space of internal coordinates, either torsion angles,[301,308] or helicoidal parameters.[300] The helicoidal internal coordinates are only applicable for nucleic acids, while torsion angles can be used for refinement of any molecule. The internal variables module (IVM) in XPLOR allows a flexible setup of simulations using an arbitrary mixture of rigid-body, torsion angles, and Cartesian coordinates.[307,318] Using internal coordinates has an advantage of significantly reducing the degrees of freedom in the system, allowing for a more efficient search for the global minimum energy. Even more important, it effectively prevents unintentional distortion of nucleic acid geometry, such as bond lengths and angles and planarity of aromatic bases. When using Cartesian coordinates-based methods, a sufficient number of improper torsion angle restraints must be used to enforce base planarity and the SHAKE algorithm to constrain N–H and C–H bond lengths.[319] Still, the presence of NOE and especially RDC restraints may create strong forces distorting the bond angles involving N–H and C–H bonds; special care must be taken to prevent such distortions.[320]

Another computational method for the structure refinement is restrained energy minimization; this method is used less frequently and mostly in combination with internal coordinates, because it is less efficient in overcoming energy barriers between local minima, especially in the Cartesian coordinate space. The restrained energy minimization is often used, however, at the end of a SA protocol. All three methods, restrained MD, Metropolis Monte Carlo, and minimization require an initial structure. Such a structure can be either modeled, generated in an extended or random conformation, or calculated using experimental distance restraints utilizing the Distance Geometry algorithm.[321]

The conformational energy E_{conf} is defined by a chemical force field; its role in the refinement of NMR structures is to make sure that the resulting structures are physically reasonable, that is, they do not contain inter-atomic clashes or unfavorable electrostatic interactions. It is necessary to use this term during the energy optimization, because experimental structural restraints alone are never sufficient to uniquely determine the solution conformation, even with the most complete NMR data. Force fields are often used even during X-ray refinements of high-resolution crystal structures.[315] Because of this, there is always a possibility that the resulting structures are somewhat biased toward the force field used. In particular, it is still a matter of substantial controversy if the electrostatic component of E_{conf} should be used during the NMR refinements of nucleic acids, see, for example, Zhou *et al.* and Brünger *et al.*[237,315] With few exceptions (see, e.g., Aramini *et al.*,[191,323] and Schmitz *et al.*[322]), the simulations of nucleic acids during the refinements are carried out *in vacuo*, without explicit water molecules, with the effect of solvent modeled using an effective dielectric constant that scales down the electrostatic interactions. The electrostatic interactions are sometimes scaled down even further during the high-temperature stage of the SA procedure to lower energy barriers between local minima. Sometimes the electrostatic interactions are omitted entirely, and the van der Waals potential is replaced with a simplified repulsion term in an attempt to make the resulting structures less biased toward a particular force field choice. An alternative approach is to choose the force field as realistic as possible in an attempt to compensate for the always insufficient number of experimental restraints. An extreme of this approach is to supplement the chemical force field with a mean force potential describing relative positions of bases derived from a database of crystal structures.[324] The question of the force field influence can be addressed directly by comparing structures refined using different force fields;[174,237,240,300] this question is interconnected with the number and type of experimental structural restraints used in the refinement. It has been acknowledged that the degree of definition of nucleic acid structures can be rather low when only the NOE-based restraints are used[325] and especially when important cross-strand restraints are missing.[158] However, the degree of definition can improve when more NOE-based distance restraints are available.[180,326] Also it improves dramatically when long-range RDC restraints are used.[238–240] In particular, it has been shown that the force field dependence decreases for structures refined with RDC restraints.[240]

The question of force field dependence is a part of two more general issues, precision and accuracy of structure determination by NMR, that is, the degree of definition of the structure and how far the determined structure is from the 'true' solution structure. An accepted method to determine the precision of structure determination is to repeat the computations with different initial conformations, typically randomly generated. Usually, the resulting conformations are ranked according to either total or restraint energy; the 10–20 best structures are selected to represent the final 'NMR ensemble'. The differences between the conformations are usually assessed using atomic root-mean-square deviation (RMSD) after the structures are superimposed onto each other, although other measures have also been proposed that are independent of molecular size.[327] The precision is expressed as an average RMSD, either an average pair-wise RMSD or an average RMSD between calculated structures and a structure with averaged coordinates.

Assessing the accuracy of structure determination is a much more difficult problem, because the 'true' structure is not known. Still, certain things can be done to assess the quality of the refined structures. First, the quality of the conformations must be examined, for example, by comparing the conformational energy E_{conf} of refined structure with the energy of the structure minimized in the absence of any restraints, or at least by verifying the absence of van der Waals clashes. Unfortunately, because of the higher intrinsic complexity, there is no convenient equivalent of a protein Ramachandran map for nucleic acids. Nevertheless, some validation tools are available with structure deposition in the PDB.[1] Next, the degree to which all experimental restraints are satisfied by the refined structures must be examined. If possible, not only derived structural restraints (NOE-derived distances, torsion angles), but the raw NMR data as well (NOE intensities, J-couplings) must be analyzed. Several figures of merits, R- and Q-factors have been proposed for this purpose (reviewed in James[328]). Most often, average deviations are calculated for distance restraints, sixth-root weighted R-factor for NOE intensities, and average deviations or RMSD for RDC and J-couplings. Also, individual large deviations need to be examined separately, as they may potentially indicate problems with experimental data, such as integration errors or even mis-assignments. A good indicator of accuracy is the free R-factor,[329] which is calculated by repeating the refinement with excluding 10% randomly chosen experimental data, and then calculating the R-factor only for these 10% of data.

The accuracy of a procedure for structure determination can also be assessed using NMR data simulated for model structures.[238–240,326] Using this approach, a possible bias in refined structures due to a particular choice of computational procedure, force field, number and type of experimental restraints, and so on can be investigated. However, the source of the bias may also be due to experimental data, such as mis-assignments, incorrectly estimated experimental errors, or conformational averaging.

Since nucleic acids are flexible in solution at room temperature, structural restraints derived from NMR data are averaged over the measurement time and over the ensemble of accessible conformations, sometimes with complicated averaging rules (e.g., for NOE-derived distances). Therefore, structures determined by one of the methods outlined above represent average structures, or more exactly, model structures satisfying average experimental restraints. The conformational variations in the 'NMR ensembles' must be regarded as reflecting a degree of indetermination of such an average structure by available experimental data, and not the true variability of solution conformations, although sometimes there may be some correlation between the two. When all solution conformations belong to the same energy minimum, the average structure will be close to this minimum and have low energy. However, when distinct conformers contribute to the observed NMR signal, the NMR-derived structural restraints may have intrinsic contradictions, and the resulting refined 'average' structures may have a relatively high energy. The best documented example of such a situation is sugar repuckering in DNA. Sugar rings in solution DNA exist in two rapidly interconverting conformations; the major conformer is S, and the minor conformer is N; the minor conformation is often more pronounced for pyrimidine residues, see, for example, Schmitz and James,[164] Rinkel *et al.*,[330] Celda *et al.*,[331] and Ulyanov *et al.*[332] Many observable NMR parameters, including interproton distances and J-couplings, depend on the exact sugar conformation; therefore, values for the corresponding experimental restraints are averaged taking into account populations of these two conformers. This leads to conformational averaging artifacts, whenever one attempts to satisfy such restraints during refinement, thus explaining why high-resolution NMR structures of DNA duplexes tend to have sugars with lower pseudo-rotation phase angle than crystal structures (see, e.g., Ulyanov and James[165]).

Several approaches have been developed that allow the time- and ensemble-averaged nature of NMR restraints and determining individual solution conformers. They include MD with time-averaging of restraints (MDtar),[333] calculating the populations of individual solution conformers either with quadratic programming algorithm (PDQPRO)[332] or genetic algorithm (FINGAR),[334] and various variants of multiple-copy refinement.[335–338] Applications to nucleic acids have been mostly limited to either MDtar or MDtar/PDQPRO combination.[322,323,339–341] The main impediment for successful application of these methods has been a paucity of experimental restraints, because a significantly greater number of restraints are required to define several solution conformers than a single average conformation. However, with many new types of structural restraints introduced recently, we can expect a renewed interest in application of such methods, see, for example, Schwieters and Clore.[342]

In conclusion, NMR has become a routine and reliable technique for the determination of average solution structures of moderately sized nucleic acids of up to about 30–40 nt; the challenges for structure determination increase more than linearly as molecular size increases. It is expected that development of experimental and computational techniques will lead to rapid progress in two directions: a better understanding of flexibility and dynamics of nucleic acids in solution,[342,343] and increasing the size limit of nucleic acids amenable to structure determination.[254,259,261,265,275,296]

Abbreviations

1D	one-dimensional
2D	two-dimensional
3D	three-dimensional
COSY	classical *J*-correlated spectroscopy
CSA	chemical shift anisotropy
DNA	deoxyribonucleic acid
dNTP	deoxynucleoside triphosphate
ESRA	endonuclease-sensitive repeat amplification
HHR	hammerhead ribozyme
HPLC	high-performance liquid chromatography
ISPA	isolated spin-pair approximation
IVM	internal variables module
MD	molecular dynamics
MDtar	MD with time-averaging of restraints
NMP	nucleoside monophosphate
NMR	nuclear magnetic resonance
NOE	nuclear Overhauser effect
NTP	nucleoside triphosphate
PAGE	polyacrylamide gel electrophoresis
PCR	polymerase chain reaction
PDB	Protein Data Bank
RDC	residual dipolar coupling
RMSD	root-mean square deviation
ROESY	residual rotating frame cross-relaxation
PEG	polyethylene glycol
RNA	ribonucleic acid
RNAP	RNA polymerase
SA	simulated annealing
SAXS	small-angle X-ray scattering
SRP	signal recognition particle
tRNA	transfer RNA

Nomenclature

Hz	hertz
kDa	kiloDalton
ms	millisecond
ns	nanosecond
nt	nucleotide
ppm	parts per million

References

1. H. M. Berman; J. Westbrook; Z. Feng; G. Gilliland; T. N. Bhat; H. Weissig; I. N. Shindyalov; P. E. Bourne, *Nucleic Acids Res.* **2000**, *28*, 235–242.
2. G. Wagner; S. G. Hyberts; T. F. Havel, *Annu. Rev. Biophys. Biomol. Struct.* **1992**, *21*, 167–198.
3. M. Billeter, *Q. Rev. Biophys.* **1992**, *25*, 325–377.
4. K. Wüthrich, *Acta Crystallogr. D Biol. Crystallogr.* **1995**, *51*, 249–270.
5. E. P. Nikonowicz; A. Pardi, *J. Mol. Biol.* **1993**, *232*, 1141–1156.
6. G. Varani; F. Aboulela; F. H. T. Allain, *Prog. Nucl. Magn. Reson. Spectrosc.* **1996**, *29*, 51–127.
7. B. Furtig; C. Richter; J. Wohnert; H. Schwalbe, *Chembiochem.* **2003**, *4*, 936–962.
8. M. P. Latham; D. J. Brown; S. A. McCallum; A. Pardi, *Chembiochem.* **2005**, *6*, 1492–1505.
9. J. F. Milligan; O. C. Uhlenbeck, *Methods Enzymol.* **1989**, *180*, 51–62.
10. C. Kojima; A. Ono; M. Kainosho, *Methods Enzymol.* **2001**, *338*, 261–283.
11. M. H. Caruthers; A. D. Barone; S. L. Beaucage; D. R. Dodds; E. F. Fisher; L. J. McBride; M. Matteucci; Z. Stabinsky; J. Y. Tang, *Methods Enzymol.* **1987**, *154*, 287–313.
12. P. Wenter; L. Reymond; S. D. Auweter; F. H. Allain; S. Pitsch, *Nucleic Acids Res.* **2006**, *34*, e79.
13. A. Ono; S. Tate; Y. Ishido; M. Kainosho, *J. Biomol. NMR* **1994**, *4*, 581–586.
14. J. Santalucia; L. X. Shen; Z. P. Cai; H. Lewis; I. Tinoco, Jr., *Nucleic Acids Res.* **1995**, *23*, 4913–4921.
15. A. J. Shallop; B. L. Gaffney; R. A. Jones, *J. Org. Chem.* **2003**, *68*, 8657–8661.
16. G. A. Kassavetis; E. T. Butler; D. Roulland; M. J. Chamberlin, *J. Biol. Chem.* **1982**, *257*, 5779–5788.
17. J. J. Dunn; F. W. Studier, *J. Mol. Biol.* **1983**, *166*, 477–535.
18. C. E. Morris; J. F. Klement; W. T. McAllister, *Gene* **1986**, *41*, 193–200.
19. S. Tabor, *Curr. Protoc. Mol. Biol.* **2001**, *16*, 16.2.1–16.2.11.
20. G. Krupp, *Gene* **1988**, *72*, 75–89.
21. P. Davanloo; A. H. Rosenberg; J. J. Dunn; F. W. Studier, *Proc. Natl. Acad. Sci. U.S.A.* **1984**, *81*, 2035–2039.
22. J. R. Wyatt; M. Chastain; J. D. Puglisi, *Biotechniques* **1991**, *11*, 764–769.
23. B. He; M. Rong; D. Lyakhov; H. Gartenstein; G. Diaz; R. Castagna; W. T. McAllister; R. K. Durbin, *Protein Expr. Purif.* **1997**, *9*, 142–151.
24. R. T. Batey; J. L. Battiste; J. R. Williamson, *Methods Enzymol.* **1995**, *261*, 300–322.
25. E. Nikonowicz, *Methods Enzymol.* **2001**, *338*, 320–341.
26. S. T. Jeng; J. F. Gardner; R. I. Gumport, *J. Biol. Chem.* **1992**, *267*, 19306–19312.
27. J. F. Milligan; D. R. Groebe; G. W. Witherell; O. C. Uhlenbeck, *Nucleic Acids Res.* **1987**, *15*, 8783–8798.
28. G. M. Cheetham; T. A. Steitz, *Curr. Opin. Struct. Biol.* **2000**, *10*, 117–123.
29. D. Imburgio; M. Rong; K. Ma; W. T. McAllister, *Biochemistry* **2000**, *39*, 10419–10430.
30. T. P. Shields; E. Mollova; L. Ste Marie; M. R. Hansen; A. Pardi, *RNA* **1999**, *5*, 1259–1267.
31. P. J. Lukavsky; J. D. Puglisi, *RNA* **2004**, *10*, 889–893.
32. C. Kao; M. Zheng; S. Rudisser, *RNA* **1999**, *5*, 1268–1272.
33. Y. Hayase; H. Inoue; E. Ohtsuka, *Biochemistry* **1990**, *29*, 8793–8797.
34. J. Xu; J. Lapham; D. M. Crothers, *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 44–48.
35. C. A. Grosshans; T. R. Cech, *Nucleic Acids Res.* **1991**, *19*, 3875–3880.
36. I. Kim; P. J. Lukavsky; J. D. Puglisi, *J. Am. Chem. Soc.* **2002**, *124*, 9338–9339.
37. A. G. Tzakos; L. E. Easton; P. J. Lukavsky, *J. Am. Chem. Soc.* **2006**, *128*, 13344–13345.
38. J. M. Carothers; J. H. Davis; J. J. Chou; J. W. Szostak, *RNA* **2006**, *12*, 567–579.
39. C. J. Hutchins; P. D. Rathjen; A. C. Forster; R. H. Symons, *Nucleic Acids Res.* **1986**, *14*, 3627–3640.
40. G. Ferbeyre; V. Bourdeau; M. Pageau; P. Miramontes; R. Cedergren, *Genome Res.* **2000**, *10*, 1011–1019.
41. G. A. Prody; J. T. Bakos; J. M. Buzayan; I. R. Schneider; G. Bruening, *Science* **1986**, *231*, 1577–1580.
42. O. C. Uhlenbeck, *RNA* **1995**, *1*, 4–6.
43. H. K. Cheong; E. Hwang; C. Lee; B. S. Choi; C. Cheong, *Nucleic Acids Res.* **2004**, *32*, e84.
44. J. S. Kieft; R. T. Batey, *RNA* **2004**, *10*, 988–995.
45. R. T. Batey; J. S. Kieft, *RNA* **2007**, *13*, 1384–1389.
46. K. A. LeCuyer; L. S. Behlen; O. C. Uhlenbeck, *EMBO J.* **1996**, *15*, 6847–6853.
47. W. C. Winkler; A. Nahvi; A. Roth; J. A. Collins; R. R. Breaker, *Nature* **2004**, *428*, 281–286.
48. A. Vermeulen; S. A. McCallum; A. Pardi, *Biochemistry* **2005**, *44*, 6024–6033.
49. D. P. Zimmer; D. M. Crothers, *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 3091–3095.

50. D. E. Smith; J.-Y. Su; F. M. Jucker, *J. Biomol. NMR* **1997**, *10*, 245–253.
51. G. Mer; W. J. Chazin, *J. Am. Chem. Soc.* **1998**, *120*, 607–608.
52. D. MacDonald; K. Herbert; X. Zhang; T. Polgruto; P. Lu, *J. Mol. Biol.* **2001**, *306*, 1081–1098.
53. A. Barbic; D. P. Zimmer; D. M. Crothers, *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 2369–2373.
54. H. Klenow; I. Henningsen, *Proc. Natl. Acad. Sci. U.S.A.* **1970**, *65*, 168–175.
55. V. Derbyshire; N. D. Grindley; C. M. Joyce, *EMBO J.* **1991**, *10*, 17–24.
56. J. E. Masse; P. Bortmann; T. Dieckmann; J. Feigon, *Nucleic Acids Res.* **1998**, *26*, 2618–2624.
57. A. Kettani; S. Bouaziz; E. Skripkin; A. Majumdar; W. Wang; R. A. Jones; D. J. Patel, *Structure* **1999**, *7*, 803–815.
58. J. M. Louis; R. G. Martin; G. M. Clore; A. M. Gronenborn, *J. Biol. Chem.* **1998**, *273*, 2374–2378.
59. X. Chen; S. V. S. Mariappan; J. J. Kelly, III; J. H. Bushweller; E. M. Bradbury; G. Gupta, *FEBS Lett.* **1998**, *436*, 372–376.
60. M. H. Werner; V. Gupta; L. J. Lambert; T. Nagata, *Methods Enzymol.* **2001**, *338*, 283–304.
61. B. Rene; G. Masliah; L. Zargarian; O. Mauffret; S. Fermandjian, *J. Biomol. NMR* **2006**, *36*, 137–146.
62. S. Ramanathan; B. J. Rao; K. V. Chary, *Biochem. Biophys. Res. Commun.* **2002**, *290*, 928–932.
63. J. Cromsigt; B. van Buuren; J. Schleucher; S. Wijmenga, *Methods Enzymol.* **2001**, *338*, 371–399.
64. S.-I. Yamakage; T. V. Maltseva; F. P. Nilson; A. Földesi; J. Chattopadhyaya, *Nucleic Acids Res.* **1993**, *21*, 5005–5011.
65. A. Földesi; S.-I. Yamakage; F. P. R. Nilsson; T. V. Maltseva; J. Chattopadhyaya, *Nucleic Acids Res.* **1996**, *24*, 1187–1194.
66. C. Glemarec; J. Kufel; A. Foldesi; T. Maltseva; A. Sandstrom; L. A. Kirsebom; J. Chattopadhyaya, *Nucleic Acids Res.* **1996**, *24*, 2022–2035.
67. U. Schmitz; S. Behrens; D. M. Freymann; R. J. Keenan; P. Lukavsky; P. Walter; T. L. James, *RNA* **1999**, *5*, 1419–1429.
68. Z. Du; N. B. Ulyanov; J. Yu; R. Andino; T. L. James, *Biochemistry* **2004**, *43*, 5757–5771.
69. P. J. Romaniuk; O. C. Uhlenbeck, *Methods Enzymol.* **1983**, *100*, 52–59.
70. J. R. Sampson; O. C. Uhlenbeck, *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 1033–1037.
71. T. Ohtsuki; G. Kawai; K. Watanabe, *J. Biochem.* **1998**, *124*, 28–34.
72. T. Ohtsuki; G. Kawai; K. Watanabe, *FEBS Lett.* **2002**, *514*, 37–43.
73. K. Wüthrich, *NMR of Proteins and Nucleic Acids*; Wiley: New York, 1986.
74. F. J. M. van de Ven; C. W. Hilbers, *Nucleic Acids Res.* **1988**, *16*, 5713–5726.
75. P. J. Lukavsky, Basic Principles of RNA NMR Spectroscopy. In *Structure and Biophysics – New Technologies for Current Challenges in Biology and Beyond*; J. D. Puglisi, Ed.; Heidelberg: Springer 2007; pp 65–80.
76. W. P. Aue; E. Bartholdi; R. R. Ernst, *J. Chem. Phys.* **1976**, *64*, 2229–2246.
77. L. Braunschweiler; R. R. Ernst, *J. Magn. Reson.* **1983**, *53*, 521–528.
78. A. Bax; S. Subramanian, *J. Magn. Reson.* **1986**, *67*, 565–569.
79. C. Griesinger; G. Otting; K. Wüthrich; R. R. Ernst, *J. Am. Chem. Soc.* **1988**, *110*, 7870–7872.
80. C. A. G. Haasnoot; F. A. A. M. de Leeuw; C. Altona, *Tetrahedron* **1980**, *36*, 2783–2792.
81. L. J. Rinkel; C. Altona, *J. Biomol. Struct. Dyn.* **1987**, *4*, 621–649.
82. F. Delaglio; S. Grzesiek; G. W. Vuister; G. Zhu; J. Pfeifer; A. Bax, *J. Biomol. NMR* **1995**, *6*, 277–293.
83. T. D. Goddard; D. G. Kneller, *SPARKY, Ver. 3.0*; University of California: San Francisco, 1998.
84. L. E. Kay; M. Ikura; A. Bax, *J. Am. Chem. Soc.* **1990**, *112*, 888–889.
85. G. M. Clore; A. Bax; P. C. Driscoll; P. T. Wingfield; A. M. Gronenborn, *Biochemistry* **1990**, *29*, 8172–8184.
86. S. W. Fesik; H. L. Eaton; E. T. Olejniczak; E. R. P. Zuiderweg; L. P. McIntosh; F. W. Dahlquist, *J. Am. Chem. Soc.* **1990**, *112*, 886–888.
87. A. Pardi, *Methods Enzymol.* **1995**, *261*, 350–380.
88. P. Legault; B. T. Farmer; L. Mueller; A. Pardi, *J. Am. Chem. Soc.* **1994**, *116*, 2203–2204.
89. J. P. Marino; J. H. Prestegard; D. M. Crothers, *J. Am. Chem. Soc.* **1994**, *116*, 2205–2206.
90. J.-P. Simorre; G. R. Zimmermann; L. Mueller; A. Pardi, *J. Am. Chem. Soc.* **1996**, *118*, 5316–5317.
91. B. Simon; K. Zanier; M. Sattler, *J. Biomol. NMR* **2001**, *20*, 173–176.
92. V. Sklenàr; J. Masse; J. Feigon, *J. Magn. Reson.* **1999**, *137*, 345–349.
93. N. B. Ulyanov; W. R. Bauer; T. L. James, *J. Biomol. NMR* **2002**, *22*, 265–280.
94. J.-P. Simorre; G. R. Zimmermann; A. Pardi; B. T. Farmer, II; L. Mueller, *J. Biomol. NMR* **1995**, *6*, 427–432.
95. J.-P. Simorre; G. R. Zimmermann; L. Mueller; A. Pardi, *J. Biomol. NMR* **1996**, *7*, 153–156.
96. V. Sklenàr; T. Dieckmann; S. E. Butcher; J. Feigon, *J. Biomol. NMR* **1996**, *7*, 83–87.
97. R. Fiala; F. Jiang; D. J. Patel, *J. Am. Chem. Soc.* **1996**, *118*, 689–690.
98. J. Wohnert; R. Ramachandran; M. Gorlach; L. R. Brown, *J. Magn. Reson.* **1999**, *139*, 430–433.
99. J. Wohnert; M. Gorlach; H. Schwalbe, *J. Biomol. NMR* **2003**, *26*, 79–83.
100. J. P. Marino; J. L. Diener; P. B. Moore; C. Griesinger, *J. Am. Chem. Soc.* **1997**, *119*, 7361–7366.
101. Z. Du; J. Yu; N. B. Ulyanov; R. Andino; T. L. James, *Biochemistry* **2004**, *43*, 11959–11972.
102. V. Sklenàr; R. D. Peterson; M. R. Rejante; J. Feigon, *J. Biomol. NMR* **1993**, *3*, 721–728.
103. B. T. Farmer, II; L. Muller; E. P. Nikononwicz; A. Pardi, *J. Biomol. NMR* **1994**, *4*, 129–134.
104. R. Fiala; J. Czernek; V. Sklenar, *J. Biomol. NMR* **2000**, *16*, 291–302.
105. R. Riek; K. Pervushin; C. Fernandez; M. Kainosho; K. Wüthrich, *J. Am. Chem. Soc.* **2001**, *123*, 658–664.
106. B. Brutscher; J. P. Simorre, *J. Biomol. NMR* **2001**, *21*, 367–372.
107. H. A. Heus; S. S. Wijmenga; F. J. M. van de Ven; C. W. Hilbers, *J. Am. Chem. Soc.* **1994**, *116*, 4983–4984.
108. J. P. Marino; H. Schwalbe; C. Anklin; W. Bermel; D. M. Crothers; C. Griesinger, *J. Am. Chem. Soc.* **1994**, *116*, 6472–6473.
109. G. Varani; F. Aboulela; F. Allain; C. C. Gubser, *J. Biomol. NMR* **1995**, *5*, 315–320.
110. S. Tate; A. Ono; M. Kainosho, *J. Magn. Reson. B* **1995**, *106*, 89–91.
111. R. Ramachandran; C. Sich; M. Grüne; V. Soskie; L. R. Brown, *J. Biomol. NMR* **1996**, *7*, 251–255.
112. J. P. Marino; H. Schwalbe; C. Anklin; W. Bermel; D. M. Crothers; C. Griesinger, *J. Biomol. NMR* **1995**, *5*, 87–92.
113. S. A. Schroeder; J. M. Fu; C. R. Jones; D. G. Gorenstein, *Biochemistry* **1987**, *26*, 3812–3821.
114. G. W. Kellogg; A. A. Szewczak; P. B. Moore, *J. Am. Chem. Soc.* **1992**, *114*, 2727–2728.
115. G. W. Kellogg; B. I. Schweitzer, *J. Biomol. NMR* **1993**, *3*, 577–595.

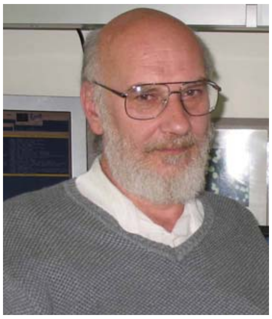
116. J. Jeener; B. H. Meier; P. Bachmann; R. R. Ernst, *J. Chem. Phys.* **1979**, *71*, 4546–4553.
117. L. Mueller; P. Legault; A. Pardi, *J. Am. Chem. Soc.* **1995**, *117*, 11043–11048.
118. C. Zwahlen; P. Legault; S. J. F. Vincent; J. Greenblatt; R. Konrat; L. E. Kay, *J. Am. Chem. Soc.* **1997**, *119*, 6711–6721.
119. L. R. Comolli; N. B. Ulyanov; A. M. Soto; L. A. Marky; T. L. James; W. H. Gmeiner, *Nucleic Acids Res.* **2002**, *30*, 4371–4379.
120. N. B. Ulyanov; A. Mujeeb; Z. Du; M. Tonelli; T. G. Parslow; T. L. James, *J. Biol. Chem.* **2006**, *281*, 16168–16177.
121. N. B. Leontis; E. Westhof, *RNA* **2001**, *7*, 499–512.
122. H. A. Heus; A. Pardi, *J. Am. Chem. Soc.* **1991**, *113*, 4360–4361.
123. J. F. Lefevre; A. N. Lane; O. Jardetzky, *FEBS Lett.* **1985**, *190*, 37–40.
124. N. Ulyanov; M. H. Sarma; V. B. Zhurkin; R. H. Sarma, *Biochemistry* **1993**, *32*, 6875–6883.
125. P. R. Blake; B. Lee; M. F. Summers; M. W. Adams; J. B. Park; Z. H. Zhou; A. Bax, *J. Biomol. NMR* **1992**, *2*, 527–533.
126. A. J. Dingley; F. Cordier; S. Grzesiek, *Concepts Magn. Reson.* **2001**, *13*, 103–127.
127. S. Grzesiek; F. Cordier; A. Dingley, *Methods Enzymol.* **2001**, *338*, 111–133.
128. P. V. Cornish; D. P. Giedroc; M. Hennig, *J. Biomol. NMR* **2006**, *35*, 209–223.
129. A. J. Dingley; L. Nisius; F. Cordier; S. Grzesiek, *Nat. Protoc.* **2008**, *3*, 242–248.
130. A. J. Dingley; S. Grzesiek, *J. Am. Chem. Soc.* **1998**, *120*, 8293–8297.
131. K. Pervushin; A. Ono; C. Fernandez; T. Szyperski; M. Kainosho; K. Wüthrich, *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 14147–14151.
132. A. J. Dingley; J. E. Masse; R. D. Peterson; M. Barfield; J. Feigon; S. Grzesiek, *J. Am. Chem. Soc.* **1999**, *121*, 6019–6027.
133. A. Majumdar; A. Kettani; E. Skripkin, *J. Biomol. NMR* **1999**, *14*, 67–70.
134. A. J. Dingley; J. E. Masse; J. Feigon; S. Grzesiek, *J. Biomol. NMR* **2000**, *16*, 279–289.
135. A. Majumdar; A. Kettani; E. Skripkin; D. J. Patel, *J. Biomol. NMR* **1999**, *15*, 207–211.
136. M. Hennig; J. R. Williamson, *Nucleic Acids Res.* **2000**, *28*, 1585–1593.
137. D. P. Giedroc; P. V. Cornish; M. Hennig, *J. Am. Chem. Soc.* **2003**, *125*, 4676–4677.
138. H. Sotoya; A. Matsugami; T. Ikeda; K. Ouhashi; S. Uesugi; M. Katahira, *Nucleic Acids Res.* **2004**, *32*, 5113–5118.
139. C. H. Arrowsmith; R. Pachter; R. B. Altman; S. B. Iyer; O. Jardetzky, *Biochemistry* **1990**, *29*, 6332–6341.
140. P. J. M. Folkers; R. H. A. Folmer; R. N. H. Konings; C. W. Hilbers, *J. Am. Chem. Soc.* **1993**, *115*, 3798–3799.
141. M. Burgering; R. Boelens; R. Kaptein, *J. Biomol. NMR* **1993**, *3* (6), 709–714.
142. F. Aboul-ela; E. P. Nikonowicz; A. Pardi, *FEBS Lett.* **1994**, *347*, 261–264.
143. N. B. Ulyanov; V. I. Ivanov; E. E. Minyat; E. B. Khomyakov; M. V. Petrova; K. Lesiak; T. L. James, *Biochemistry* **1998**, *37*, 12715–12726.
144. R. Kaptein; R. Boelens; R. M. Scheek; W. F. van Gunsteren, *Biochemistry* **1988**, *27*, 5389–5394.
145. A. Kumar; R. R. Ernst; K. Wüthrich, *Biochem. Biophys. Res. Commun.* **1980**, *95*, 1–6.
146. S. Macura; R. R. Ernst, *Mol. Phys.* **1980**, *41*, 95–117.
147. S. Macura; Y. Huang; D. Suter; R. R. Ernst, *J. Magn. Reson.* **1981**, *43*, 259–281.
148. J. W. Keepers; T. L. James, *J. Magn. Reson.* **1984**, *57*, 404–426.
149. H. Liu; P. D. Thomas; T. L. James, *J. Magn. Reson.* **1992**, *98*, 163–175.
150. W. Massefski; P. H. Bolton, *J. Magn. Reson.* **1985**, *65*, 526–530.
151. E.-I. Suzuki; N. Pattabiraman; G. Zon; T. L. James, *Biochemistry* **1986**, *25*, 6854–6865.
152. G. Gupta; M. H. Sarma; R. H. Sarma, *Biochemistry* **1988**, *27*, 7909–7918.
153. A. N. Lane, *Biochim. Biophys. Acta* **1990**, *1049*, 205–212.
154. J. D. Baleja; J. Moult; B. D. Sykes, *J. Magn. Reson.* **1990**, *87*, 375–384.
155. J. D. Baleja; M. W. Germann; J. H. van de Sande; B. D. Sykes, *J. Mol. Biol.* **1990**, *215*, 411–428.
156. A. M. J. J. Bonvin; R. Boelens; R. Kaptein, *J. Biomol. NMR* **1991**, *1*, 305–309.
157. H. Robinson; A. H.-J. Wang, *Biochemistry* **1992**, *31*, 3524–3533.
158. N. Ulyanov; A. A. Gorin; V. B. Zhurkin; B.-C. Chen; M. H. Sarma; R. H. Sarma, *Biochemistry* **1992**, *31*, 3918–3930.
159. S.-G. Kim; B. R. Reid, *Biochemistry* **1992**, *31*, 12103–12116.
160. M. Foti; S. Marshalko; E. Schurter; S. Kumar; G. P. Beardsley; B. I. Schweitzer, *Biochemistry* **1997**, *36*, 5336–5345.
161. A. Kumar; G. Wagner; R. R. Ernst; K. Wüthrich, *J. Am. Chem. Soc.* **1981**, *103*, 3654–3658.
162. P. Cuniasse; L. C. Sowers; R. Eritja; B. Kaplan; M. F. Goodman; J. A. H. Cognet; M. LeBret; W. Guschlbauer; G. V. Fazakerley, *Nucleic Acids Res.* **1987**, *15*, 8003–8022.
163. B. Reid; K. Banks; P. Flynn; W. Nerdal, *Biochemistry* **1989**, *28*, 10001–10007.
164. U. Schmitz; T. L. James, *Methods Enzymol.* **1995**, *261*, 3–44.
165. N. B. Ulyanov; T. L. James, *Methods Enzymol.* **1995**, *261*, 90–120.
166. F. J. M. van de Ven; M. J. J. Blommers; R. E. Schouten; C. W. Hilbers, *J. Magn. Reson.* **1991**, *94*, 140–151.
167. S. S. Wijmenga; B. N. M. van Buuren, *Prog. Nucl. Magn. Reson. Spectrosc.* **1998**, *32*, 287–387.
168. B. A. Borgias; T. L. James, *Methods Enzymol.* **1989**, *176*, 169–183.
169. B. A. Borgias; T. L. James, *J. Magn. Reson.* **1990**, *87*, 475–487.
170. R. Boelens; T. M. G. Koning; G. A. van der Marel; J. H. van Boom; R. Kaptein, *J. Magn. Reson.* **1989**, *82*, 290–308.
171. C. B. Post; R. P. Meadows; D. G. Gorenstein, *J. Am. Chem. Soc.* **1990**, *112*, 6796–6803.
172. M. Gochin; T. L. James, *Biochemistry* **1990**, *29*, 11172–11180.
173. K. Weisz; R. H. Shafer; W. Egan; T. L. James, *Biochemistry* **1992**, *31*, 7477–7487.
174. U. Schmitz; I. Sethson; W. Egan; T. L. James, *J. Mol. Biol.* **1992**, *227*, 510–531.
175. A. Mujeeb; S. M. Kerwin; W. Egan; G. L. Kenyon; T. L. James, *Biochemistry* **1992**, *31*, 9325–9338.
176. M. Tonelli; E. Ragg; A. M. Bianucci; K. Lesiak; T. L. James, *Biochemistry* **1998**, *37*, 11745–11761.
177. H. Liu; H. P. Spielmann; N. B. Ulyanov; D. E. Wemmer; T. L. James, *J. Biomol. NMR* **1995**, *6*, 390–402.
178. U. Schmitz; D. A. Pearlman; T. L. James, *J. Mol. Biol.* **1991**, *221*, 271–292.
179. A. Mujeeb; S. M. Kerwin; G. L. Kenyon; T. L. James, *Biochemistry* **1993**, *32*, 13419–13431.
180. K. Weisz; R. H. Shafer; W. Egan; T. L. James, *Biochemistry* **1994**, *33*, 354–366.
181. H. P. Spielmann; T. J. Dwyer; J. E. Hearst; D. E. Wemmer, *Biochemistry* **1995**, *34*, 12937–12953.

182. P. V. Sahasrabudhe; R. T. Pon; W. H. Gmeiner, *Biochemistry* **1996**, *35*, 13597–13608.
183. Y. Coppel; N. Berthet; C. Coulombeau; C. Coulombeau; J. Garcia; J. Lhomme, *Biochemistry* **1997**, *36*, 4817–4830.
184. M. Petersen; J. P. Jacobsen, *Bioconjug. Chem.* **1998**, *9*, 331–340.
185. N. B. Ulyanov; V. I. Ivanov; E. E. Minyat; E. B. Khomyakova; M. V. Petrova; K. Lesiak; T. L. James, *Biochemistry* **1998**, *37*, 12715–12726.
186. L. Ayadi; M. Jourdan; C. Coulombeau; J. Garcia; R. Lavery, *J. Biomol. Struct. Dyn.* **1999**, *17*, 245–257.
187. E. V. Bichenkova; D. Marks; M. I. Dobrikov; V. V. Vlassov; G. A. Morris; K. T. Douglas, *J. Biomol. Struct. Dyn.* **1999**, *17*, 193–211.
188. R. J. Isaacs; W. S. Rayens; H. P. Spielmann, *J. Mol. Biol.* **2002**, *319*, 191–207.
189. M. Petersen; K. Bondensgaard; J. Wengel; J. P. Jacobsen, *J. Am. Chem. Soc.* **2002**, *124*, 5974–5982.
190. H. V. Tommerholt; N. K. Christensen; P. Nielsen; J. Wengel; P. C. Stein; J. P. Jacobsen; M. Petersen, *Org. Biomol. Chem.* **2003**, *1*, 1790–1797.
191. J. M. Aramini; S. H. Cleaver; R. T. Pon; R. P. Cunningham; M. W. Germann, *J. Mol. Biol.* **2004**, *338*, 77–91.
192. I. Gomez-Pinto; E. Cubero; S. G. Kalko; V. Monaco; G. van der Marel; J. H. van Boom; M. Orozco; C. Gonzalez, *J. Biol. Chem.* **2004**, *279*, 24552–24560.
193. Q. Zhang; T. J. Dwyer; V. Tsui; D. A. Case; J. Cho; P. B. Dervan; D. E. Wemmer, *J. Am. Chem. Soc.* **2004**, *126*, 7958–7966.
194. H. Baruah; M. W. Wright; U. Bierbach, *Biochemistry* **2005**, *44*, 6059–6070.
195. G. Shanmugam; A. K. Goodenough; I. D. Kozekov; F. P. Guengerich; C. J. Rizzo; M. P. Stone, *Chem. Res. Toxicol.* **2007**, *20*, 1601–1611.
196. D. J. Kerwood; P. N. Borer, *Magn. Reson. Chem.* **1996**, *34*, S136–S146.
197. P. V. Sahasrabudhe; W. H. Gmeiner, *Biochemistry* **1997**, *36*, 5981–5991.
198. A. Mujeeb; T. G. Parslow; A. Zarrinpar; C. Das; T. L. James, *FEBS Lett.* **1999**, *458*, 387–392.
199. U. Schmitz; T. L. James; P. Lukavsky; P. Walter, *Nat. Struct. Biol.* **1999**, *6*, 634–638.
200. D. J. Kerwood; M. J. Cavaluzzi; P. N. Borer, *Biochemistry* **2001**, *40*, 14518–14529.
201. Y. Yuan; D. J. Kerwood; A. C. Paoletti; M. F. Shubsda; P. N. Borer, *Biochemistry* **2003**, *42*, 5259–5269.
202. M. Karplus, *J. Chem. Phys.* **1959**, *30*, 11–31.
203. M. Karplus, *J. Am. Chem. Soc.* **1963**, *85*, 2870–2871.
204. C. Griesinger; O. W. Sorensen; R. R. Ernst, *J. Magn. Reson.* **1987**, *75*, 474–492.
205. G. W. Vuister; A. Bax, *J. Am. Chem. Soc.* **1993**, *115*, 7772–7777.
206. A. Bax; G. W. Vuister; S. Grzesiek; F. Delaglio; A. C. Wang; R. Tschudin; G. Zhu, *Methods Enzymol.* **1994**, *239*, 79–105.
207. C. Sich; O. Ohlenschläger; R. Ramachandran; M. Görlach; L. R. Brown, *Biochemistry* **1997**, *36*, 13989–14002.
208. T. Carlomagno; M. Hennig; J. R. Williamson, *J. Biomol. NMR* **2002**, *22*, 65–81.
209. F. Delaglio; Z. Wu; A. Bax, *J. Magn. Reson.* **2001**, *149*, 276–281.
210. H. Widmer; K. Wüthrich, *J. Magn. Reson.* **1986**, *70*, 270–279.
211. M. M. Mooren; S. S. Wijmenga; G. A. van der Marel; J. H. van Boom; C. W. Hilbers, *Nucleic Acids Res.* **1994**, *22*, 2658–2666.
212. J. P. Marino; H. Schwalbe; S. J. Glaser; C. Griesinger, *J. Am. Chem. Soc.* **1996**, *118*, 4388–4395.
213. J. H. Ippel; S. S. Wijmenga; R. de Jong; H. A. Heus; C. W. Hilbers; E. de Vroom; G. A. van der Marel; J. H. van Boom, *Magn. Reson. Chem.* **1996**, *34*, S156–S176.
214. J. van Wijk; B. D. Huckriede; J. H. Ippel; C. Altona, *Meth. Enzym.* **1992**, *211*, 286–306.
215. C. Altona; R. Francke; R. de Haan; J. H. Ippel; G. H. Daalmans; A. J. H. Westra Hoekzema; J. van Wijk, *Magn. Reson. Chem.* **1994**, *32*, 670–678.
216. Y. T. van den Hoogen; C. M. Hilgersom; D. Brozda; K. Lesiak; P. F. Torrence; C. Altona, *Eur. J. Biochem.* **1989**, *182*, 629–637.
217. J. R. Tolman; J. M. Flanagan; M. A. Kennedy; J. H. Prestegard, *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 9279–9283.
218. N. Tjandra; A. Bax, *Science* **1997**, *278*, 1111–1114.
219. N. Tjandra; J. G. Omichinski; A. M. Gronenborn; G. M. Clore; A. Bax, *Nat. Struct. Biol.* **1997**, *4*, 732–738.
220. R. S. Lipsitz; N. Tjandra, *Annu. Rev. Biophys. Biomol. Struct.* **2004**, *33*, 387–413.
221. N. Tjandra; S. Tate; A. Ono; M. Kainosho; A. Bax, *J. Am. Chem. Soc.* **2000**, *122*, 6190–6200.
222. P. Bayer; L. Varani; G. Varani, *J. Biomol. NMR* **1999**, *14*, 149–155.
223. M. R. Hansen; L. Mueller; A. Pardi, *Nat. Struct. Biol.* **1998**, *5*, 1065–1074.
224. M. R. Hansen; P. Hanson; A. Pardi, *Methods Enzymol.* **2000**, *317*, 220–240.
225. M. Rückert; G. Otting, *J. Am. Chem. Soc.* **2000**, *122*, 7793–7797.
226. R. D. Beger; V. M. Marathias; B. F. Volkman; P. H. Bolton, *J. Magn. Reson.* **1998**, *135*, 256–259.
227. J. Ying; A. Grishaev; M. P. Latham; A. Pardi; A. Bax, *J. Biomol. NMR* **2007**, *39*, 91–96.
228. A. Bax, *Protein Sci.* **2003**, *12*, 1–16.
229. V. Tsui; L. Zhu; T. H. Huang; P. E. Wright; D. A. Case, *J. Biomol. NMR* **2000**, *16*, 9–21.
230. G. M. Clore; A. M. Gronenborn; N. Tjandra, *J. Magn. Reson.* **1998**, *131*, 159–162.
231. A. Bax; G. Kontaxis; N. Tjandra, *Methods Enzymol.* **2001**, *339*, 127–174.
232. C. D. Schwieters; J. J. Kuszewski; N. Tjandra; G. M. Clore, *J. Magn. Reson.* **2003**, *160*, 66–73.
233. W. J. Wedemeyer; C. A. Rohl; H. A. Scheraga, *J. Biomol. NMR* **2002**, *22*, 137–151.
234. J. A. Losonczi; M. Andrec; M. W. Fischer; J. H. Prestegard, *J. Magn. Reson.* **1999**, *138*, 334–342.
235. M. Zweckstetter; A. Bax, *J. Am. Chem. Soc.* **2000**, *122*, 3791–3792.
236. Y. Wei; M. H. Werner, *J. Biomol. NMR* **2006**, *35*, 17–25.
237. H. Zhou; A. Vermeulen; F. M. Jucker; A. Pardi, *Biopolymers* **1999**, *52*, 168–180.
238. A. Vermeulen; H. Zhou; A. Pardi, *J. Am. Chem. Soc.* **2000**, *122*, 9638–9647.
239. O. Mauffret; G. Tevanian; S. Fermandjian, *J. Biomol. NMR* **2002**, *24*, 317–328.
240. K. McAteer; M. A. Kennedy, *J. Biomol. Struct. Dyn.* **2003**, *20*, 487–506.
241. P. Padrta; R. Stefl; L. Kralik; L. Zidek; V. Sklenar, *J. Biomol. NMR* **2002**, *24*, 1–14.
242. Z. Wu; F. Delaglio; N. Tjandra; V. B. Zhurkin; A. Bax, *J. Biomol. NMR* **2003**, *26*, 297–315.
243. K. McAteer; A. Aceves-Gaona; R. Michalczyk; G. W. Buchko; N. G. Isern; L. A. Silks; J. H. Miller; M. A. Kennedy, *Biopolymers* **2004**, *75*, 497–511.

244. R. Stefl; H. Wu; S. Ravindranathan; V. Sklenar; J. Feigon, *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 1177–1182.
245. B. Wu; F. Girard; B. van Buuren; J. Schleucher; M. Tessari; S. Wijmenga, *Nucleic Acids Res.* **2004**, *32*, 3228–3239.
246. J. G. Renisio; S. Cosquer; I. Cherrak; S. El Antri; O. Mauffret; S. Fermandjian, *Nucleic Acids Res.* **2005**, *33*, 1970–1981.
247. F. Alvarez-Salgado; H. Desvaux; Y. Boulard, *Magn. Reson. Chem.* **2006**, *44*, 1081–1089.
248. N. Sibille; A. Pardi; J. P. Simorre; M. Blackledge, *J. Am. Chem. Soc.* **2001**, *123*, 12135–12146.
249. J. J. Warren; P. B. Moore, *J. Biomol. NMR* **2001**, *20*, 311–323.
250. K. Bondensgaard; E. T. Mollova; A. Pardi, *Biochemistry* **2002**, *41*, 11532–11542.
251. T. C. Leeper; M. B. Martin; H. Kim; S. Cox; V. Semenchenko; F. J. Schmidt; S. R. Van Doren, *Nat. Struct. Biol.* **2002**, *9*, 397–403.
252. L. D. Finger; L. Trantirek; C. Johansson; J. Feigon, *Nucleic Acids Res.* **2003**, *31*, 6461–6472.
253. D. C. Lawrence; C. C. Stover; J. Noznitsky; Z. Wu; M. F. Summers, *J. Mol. Biol.* **2003**, *326*, 529–542.
254. P. J. Lukavsky; I. Kim; G. A. Otto; J. D. Puglisi, *Nat. Struct. Biol.* **2003**, *10*, 1033–1038.
255. S. A. McCallum; A. Pardi, *J. Mol. Biol.* **2003**, *326*, 1037–1050.
256. C. A. Theimer; L. D. Finger; L. Trantirek; J. Feigon, *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 449–454.
257. E. O'Neil-Cabello; D. L. Bryce; E. P. Nikonowicz; A. Bax, *J. Am. Chem. Soc.* **2004**, *126*, 66–67.
258. P. Vallurupalli; P. B. Moore, *J. Mol. Biol.* **2003**, *325*, 843–856.
259. V. D'Souza; A. Dey; D. Habib; M. F. Summers, *J. Mol. Biol.* **2004**, *337*, 427–442.
260. D. G. Sashital; G. Cornilescu; C. J. McManus; D. A. Brow; S. E. Butcher, *Nat. Struct. Mol. Biol.* **2004**, *11*, 1237–1242.
261. J. H. Davis; M. Tonelli; L. G. Scott; L. Jaeger; J. R. Williamson; S. E. Butcher, *J. Mol. Biol.* **2005**, *351*, 371–382.
262. T. C. Leeper; G. Varani, *RNA* **2005**, *9*, 394–403.
263. D. W. Staple; S. E. Butcher, *J. Mol. Biol.* **2005**, *349*, 1011–1023.
264. C. A. Theimer; C. A. Blois; J. Feigon, *Mol. Cell* **2005**, *17*, 671–682.
265. Y. Chen; J. Fender; J. D. Legassie; M. B. Jarstfer; T. M. Bryan; G. Varani, *EMBO J.* **2006**, *25*, 3156–3166.
266. S. Flodell; M. Petersen; F. Girard; J. Zdunek; K. Kidd-Ljunggren; J. Schleucher; S. Wijmenga, *Nucleic Acids Res.* **2006**, *34*, 4449–4457.
267. Y. Nomura; M. Kajikawa; S. Baba; S. Nakazato; T. Imai; T. Sakamoto; N. Okada; G. Kawai, *Nucleic Acids Res.* **2006**, *34*, 5184–5193.
268. R. J. Richards; C. A. Theimer; L. D. Finger; J. Feigon, *Nucleic Acids Res.* **2006**, *34*, 816–825.
269. R. J. Richards; H. Wu; L. Trantirek; C. M. O'Connor; K. Collins; J. Feigon, *RNA* **2006**, *12*, 1475–1485.
270. S. J. Headey; H. Huang; J. K. Claridge; G. A. Soares; K. Dutta; M. Schwalbe; D. Yang; S. M. Pascal, *RNA* **2007**, *13*, 351–360.
271. R. J. Marcheschi; D. W. Staple; S. E. Butcher, *J. Mol. Biol.* **2007**, *373*, 652–663.
272. D. G. Sashital; V. Venditti; C. G. Angers; G. Cornilescu; S. E. Butcher, *RNA* **2007**, *13*, 328–338.
273. N. Shankar; T. Xia; S. D. Kennedy; T. R. Krugh; D. H. Mathews; D. H. Turner, *Biochemistry* **2007**, *46*, 12665–12678.
274. C. A. Theimer; B. E. Jady; N. Chim; P. Richard; K. E. Breece; T. Kiss; J. Feigon, *Mol. Cell* **2007**, *27*, 869–881.
275. H. Wu; J. Feigon, *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 6655–6660.
276. J. Zoll; M. Tessari; F. J. Van Kuppeveld; W. J. Melchers; H. A. Heus, *RNA* **2007**, *13*, 781–792.
277. N. J. Reiter; L. J. Maher, III; S. E. Butcher, *Nucleic Acids Res.* **2008**, *36*, 1227–1236.
278. H. Van Melckebeke; M. Devany; C. Di Primo; F. Beaurain; J. J. Toulme; D. L. Bryce; J. Boisbouvier, *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 9210–9215.
279. K. Tu; M. Gochin, *J. Am. Chem. Soc.* **1999**, *121*, 9276–9285.
280. M. Gochin, *Structure* **2000**, *8*, 441–452.
281. Z. Wu; N. Tjandra; A. Bax, *J. Am. Chem. Soc.* **2001**, *123*, 3617–3618.
282. J. Herzfeld; R. G. Griffin; R. A. Haberkorn, *Biochemistry* **1978**, *17*, 2711–2718.
283. E. O'Neil-Cabello; Z. Wu; D. L. Bryce; E. P. Nikonowicz; A. Bax, *J. Biomol. NMR* **2004**, *30*, 61–70.
284. D. L. Bryce; A. Grishaev; A. Bax, *J. Am. Chem. Soc.* **2005**, *127*, 7387–7396.
285. J. Ying; A. Grishaev; D. L. Bryce; A. Bax, *J. Am. Chem. Soc.* **2006**, *128*, 11443–11454.
286. K. Pervushin; R. Riek; G. Wider; K. Wüthrich, *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 12366–12371.
287. B. Brutscher; J. Boisbouvier; A. Pardi; D. Marion; J.-P. Simorre, *J. Am. Chem. Soc.* **1998**, *120*, 11845–11851.
288. A. Grishaev; J. Ying; A. Bax, *J. Am. Chem. Soc.* **2006**, *128*, 10010–10011.
289. C. Richter; C. Griesinger; I. Felli; P. T. Cole; G. Varani; H. Schwalbe, *J. Biomol. NMR* **1999**, *15*, 241–250.
290. J. Boisbouvier; B. Brutscher; A. Pardi; D. Marion; J. P. Simorre, *J. Am. Chem. Soc.* **2000**, *122*, 6779–6780.
291. M. Ebrahimi; P. Rossi; C. Rogers; G. S. Harbison, *J. Magn. Reson.* **2001**, *150*, 1–9.
292. D. A. Case, *J. Biomol. NMR* **1995**, *6*, 341–346.
293. S. S. Wijmenga; M. Kruithof; C. W. Hilbers, *J. Biomol. NMR* **1997**, *10*, 337–350.
294. J. Cromsigt; C. W. Hilbers; S. S. Wijmenga, *J. Biomol. NMR* **2001**, *21*, 11–29.
295. D. S. Wishart; D. A. Case, *Methods Enzymol.* **2001**, *338*, 3–34.
296. A. Grishaev; J. Ying; M. D. Canny; A. Pardi; A. Bax, *J. Biomol. NMR* **2008**, *42*, 99–109.
297. G. M. Clore; A. M. Gronenborn, *Crit. Rev. Biochem. Mol. Biol.* **1989**, *24*, 479–564.
298. A. T. Brünger; M. Karplus, *Acc. Chem. Res.* **1991**, *24*, 54–61.
299. T. L. James; V. J. Basus, *Annu. Rev. Phys. Chem.* **1991**, *42*, 501–542.
300. N. B. Ulyanov; U. Schmitz; T. L. James, *J. Biomol. NMR* **1993**, *3*, 547–568.
301. E. G. Stein; L. M. Rice; A. T. Brünger, *J. Magn. Reson.* **1997**, *124*, 154–164.
302. G. M. Clore; C. D. Schwieters, *Curr. Opin. Struct. Biol.* **2002**, *12*, 146–153.
303. D. A. Case; T. E. Cheatham, III; T. Darden; H. Gohlke; R. Luo; K. M. Merz, Jr.; A. Onufriev; C. Simmerling; B. Wang; R. J. Woods, *J. Comput. Chem.* **2005**, *26*, 1668–1688.
304. J. de Vlieg; R. M. Scheek; W. F. van Gunsteren; R. Kaptein; J. Thomason, *Proteins* **1988**, *3*, 209–218.
305. A. T. Brünger, *XPLOR Manual, Ver. 3.1*; Yale University Press: New Haven, 1993.
306. A. T. Brünger; P. D. Adams; G. M. Clore; W. L. DeLano; P. Gros; R. W. GrosseKunstleve; J. S. Jiang; J. Kuszewski; M. Nilges; N. S. Pannu; R. J. Read; L. M. Rice; T. Simonson; G. L. Warren, *Acta Crystallogr. D Biol. Crystallogr.* **1998**, *54*, 905–921.
307. C. D. Schwieters; J. J. Kuszewski; G. Marius Clore, *Prog. Nucl. Magn. Reson. Spectrosc.* **2006**, *48*, 47–62.

308. P. Güntert; C. Mumenthaler; K. Wüthrich, *J. Mol. Biol.* **1997**, *273*, 283–298.
309. V. B. Zhurkin; N. B. Ulyanov; A. A. Gorin; R. L. Jernigan, *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 7046–7050.
310. R. Lavery; K. Zakrzewska; H. Sklenar, *Comput. Phys. Commun.* **1995**, *91*, 135–158.
311. A. Amir-Aslani; O. Mauffret; F. Sourgen; S. Neplaz; R. G. Maroun; E. Lescot; G. Tevanian; S. Fermandjian, *J. Mol. Biol.* **1996**, *263*, 776–788.
312. M. Orozco; A. Perez; A. Noy; F. J. Luque, *Chem. Soc. Rev.* **2003**, *32*, 350–364.
313. J. D. Baleja; R. T. Pon; B. D. Sykes, *Biochemistry* **1990**, *29*, 4828–4839.
314. D. J. Kerwood; G. Zon; T. L. James, *Eur. J. Biochem.* **1991**, *197*, 583–595.
315. A. T. Brünger; P. D. Adams; L. M. Rice, *Structure* **1997**, *5*, 325–336.
316. H. J. C. Berendsen; J. P. M. Postma; W. F. van Gunsteren; A. Di Nola; J. R. Haak, *J. Chem. Phys.* **1984**, *81*, 3684–3690.
317. N. Metropolis; A. W. Rosenbluth; M. N. Rosenbluth; A. H. Teller; E. Teller, *J. Chem. Phys.* **1953**, *21*, 1087–1092.
318. C. D. Schwieters; G. M. Clore, *J. Magn. Reson.* **2001**, *152*, 288–302.
319. J. P. Ryckaert; G. Cicotti; H. J. C. Berendsen, *J. Comput. Phys.* **1977**, *23*, 327–341.
320. N. B. Ulyanov; Z. Du; T. L. James, Refinement of Nucleic Acid Structures with Residual Dipolar Coupling Restraints in Cartesian Coordinate Space. In *Modern Magnetic Resonance*; G. A. Webb, Ed.; Springer: Netherlands, 2006; pp 665–670.
321. T. F. Havel; K. Wüthrich, *J. Mol. Biol.* **1985**, *182*, 281–294.
322. U. Schmitz; N. B. Ulyanov; A. Kumar; T. L. James, *J. Mol. Biol.* **1993**, *234*, 373–389.
323. J. M. Aramini; A. Mujeeb; N. B. Ulyanov; M. W. Germann, *J. Biomol. NMR* **2000**, *18*, 287–302.
324. J. Kuszewski; C. Schwieters; G. M. Clore, *J. Am. Chem. Soc.* **2001**, *123*, 3903–3918.
325. W. J. Metzler; C. Wang; D. Kitchen; R. M. Levy; A. Pardi, *J. Mol. Biol.* **1990**, *214*, 711–736.
326. F. H. Allain; G. Varani, *J. Mol. Biol.* **1997**, *267*, 338–351.
327. V. N. Maiorov; G. M. Crippen, *Proteins* **1995**, *22*, 273–283.
328. T. L. James, *Methods Enzymol.* **1994**, *239*, 416–439.
329. A. Brünger, *Nature* **1992**, *355*, 472–474.
330. L. J. Rinkel; G. A. van der Marel; J. H. van Boom; C. Altona, *Eur. J. Biochem.* **1987**, *166*, 87–101.
331. B. Celda; H. Widmer; W. Leupin; W. J. Chazin; W. A. Denny; K. Wüthrich, *Biochemistry* **1989**, *28*, 1462–1470.
332. N. B. Ulyanov; U. Schmitz; A. Kumar; T. L. James, *Biophys. J.* **1995**, *68*, 13–24.
333. A. E. Torda; R. M. Scheek; W. F. van Gunsteren, *J. Mol. Biol.* **1990**, *214*, 223–235.
334. D. A. Pearlman, *J. Biomol. NMR* **1996**, *8*, 49–66.
335. A. M. J. J. Bonvin; A. T. Brünger, *J. Mol. Biol.* **1995**, *250*, 80–93.
336. J. Fennen; A. E. Torda; W. F. van Gunsteren, *J. Biomol. NMR* **1995**, *6*, 163–170.
337. J. Kemmink; R. M. Scheek, *J. Biomol. NMR* **1995**, *5*, 33–40.
338. A. Görler; N. B. Ulyanov; T. L. James, *J. Biomol. NMR* **2000**, *16*, 147–164.
339. L. J. Yao; T. L. James; J. T. Kealey; D. V. Santi; U. Schmitz, *J. Biomol. NMR* **1997**, *9*, 229–244.
340. U. Schmitz; A. Donati; T. L. James; N. B. Ulyanov; L. Yao, *Biopolymers* **1998**, *46*, 329–342.
341. R. J. Isaacs; H. P. Spielmann, *J. Am. Chem. Soc.* **2004**, *126*, 583–590.
342. C. D. Schwieters; G. M. Clore, *Biochemistry* **2007**, *46*, 1152–1166.
343. M. Getz; X. Sun; A. Casiano-Negroni; Q. Zhang; H. M. Al-Hashimi, *Biopolymers* **2007**, *86*, 384–402.

Biographical Sketches



Nikolai B. Ulyanov studied mathematics in Moscow State University and worked on computational modeling of DNA bending as part of his Ph.D. project in the group of Dr. Victor Zhurkin in the Engelhardt Institute of Molecular Biology in Moscow. He was a postdoctoral fellow with Prof. Ramaswamy Sarma studying NMR spectroscopy of DNA at the State University of New York at Albany. Currently he is Associate Adjunct Professor at the University of California at San Francisco. His research interests focus on the structure and dynamics of nucleic acids, studied by computational methods and NMR.

Thomas L. James is Professor of Chemistry, Pharmaceutical Chemistry and Radiology at the University of California, San Francisco, where after 13 years he recently stepped down as Chair of the Department of Pharmaceutical Chemistry. He received his Ph.D. from the University of Wisconsin. After 2 years in industry and 2 years of postdoctoral work with Prof. Mildred Cohn in the emerging area of biological NMR, he joined the UCSF faculty.

Most of his research has focused on the development and the use of NMR in biology. Part of which has involved *in vivo* NMR, for example, spectroscopic imaging to investigate stroke, prostatic cancer, and drug toxicology. Other NMR-related research emphasized atomic level understanding with major goals to (1) enhance the accuracy and precision of protein and nucleic acid structures determined, (2) develop the means of describing conformational ensembles, (3) apply those methodologies to study biomolecular structure and dynamics and small molecule–macromolecule interactions, and (4) use 3D nucleic acid structures and computational search algorithms to discover novel ligands to serve as drug leads.

Professor James has authored about 360 publications, including one book written and eight books edited. He has served as Editor or on the Editorial Board of four journals, four years on an NIH study section, and on several advisory boards.

9.09 Derivation of Peptide and Protein Structure using NMR Spectroscopy

Glenn F. King and Mehdi Mobli, The University of Queensland, St. Lucia, QLD, Australia

9.09.1 Introduction

Nuclear magnetic resonance (NMR) spectroscopy is an invaluable tool for determining the three-dimensional (3D) structure of both small and large biomolecules. For very small molecules, determination of high-resolution structures may not be relevant either because the molecules adapt a single rigid structure or because they sample a vast amount of conformational space due to the presence of multiple rotatable bonds with a low energy barrier for interconversion between rotamers. In such cases, structure determination may be simply a matter of determining the local environment about a stereocenter or the orientation of two molecular fragments. However, many peptides and small proteins adopt specific structural forms, with stable secondary and tertiary structures, and in these cases NMR spectroscopy is the most widely used tool for determining their solution conformation (see Chapter 9.06).

The fundamental limitation of any NMR-based protein structure determination endeavor is the ability to unambiguously identify the nuclear resonance frequencies (or chemical shifts) of each NMR-active nucleus in the protein. Inevitably, as the number of nuclei in a protein increases, so does the spectral complexity. A limit is ultimately reached where peak overlap is so severe that resonance frequencies become ambiguous, thus curtailing further analysis. Resonance overlap can be alleviated in several ways. Since resonance dispersion is proportional to magnetic field strength, spectral complexity can be decreased by performing experiments at higher magnetic field strength, which has the added benefit of improving the sensitivity; however, this can be very costly and it generally offers only limited alleviation of the spectral overlap problem. A more efficient means of increasing resonance dispersion is to employ multidimensional NMR experiments in which the chemical shift of one nucleus is related to that of its interacting partner(s), resulting in dispersion of resonances into extra frequency dimensions. The extent to which this can be achieved is related to the sensitivity at which the various nuclei can be detected. Moreover, since NMR spectroscopy is a relatively insensitive spectroscopic method, there is an inherent concentration limit below which NMR studies of peptides and proteins are unlikely to be successful (see Chapter 2.15).

For unlabeled peptides purified from natural sources, one is generally limited to two-dimensional (2D) NMR experiments (although some 3D experiments are possible as outlined in Section 9.09.3.2.5). Therefore, determination of the structure of even small peptides can become problematic due to spectral overlap. The limit at which we are able to conduct NMR structural studies of peptides depends on the chemical shift dispersion of each individual protein, but generally speaking we can expect unlabeled peptides of up to ∼10 kDa (90–95 amino acid residues) to be amenable to structure determination using homonuclear NMR.

Modern molecular biology approaches often allow recombinant peptides and proteins to be produced in various host cells such as bacteria and yeast. In such cases, the protein/peptide of interest can be uniformly labeled with NMR-active ^{13}C and ^{15}N isotopes by growing the host cells in a defined minimal medium. This provides access to a vastly increased range of multidimensional NMR experiments, and it increases spectral dispersion due to the greater chemical shift range of ^{13}C and ^{15}N nuclei compared with ^{1}H, thereby enabling routine structural studies of proteins up to ∼30 kDa (∼270 amino acid residues). Thus, in order to distinguish between the homonuclear and heteronuclear NMR approaches, this chapter has been deliberately partitioned into two sections; one section deals with unlabeled peptides (Section 9.09.3) and the other with isotopically labeled proteins (Section 9.09.4). However, it should be noted that while the types of NMR experiments that can be employed to study these two categories of natural products are dramatically different, both approaches ultimately yield the same types of NMR-derived structural restraints that are ultimately used to reconstruct the 3D conformation of the peptide or protein.

9.09.2 Sample Considerations and Solvent Suppression

9.09.2.1 Heterogeneity of the Nuclear Magnetic Resonance Sample

Two key factors affecting the outcome of protein NMR experiments are the purity and oligomeric nature of the sample. Sample impurities will exacerbate spectral overlap problems and they can lead to incorrect peak assignments, particularly in nuclear Overhauser enhancement spectroscopy (NOESY) spectra; thus, a sample purity of >90% is essential, and >95% is desirable. A generally more difficult problem to overcome is aggregation of the peptide or protein sample at the high concentrations required for NMR analysis (typically $\geq$0.5 mmol l^{-1}). It is absolutely essential that the oligomeric state of the protein sample is known prior to beginning the NMR investigation as this will determine the types of experiments that are feasible. For example, dimerization of a 25 kDa protein would preclude NMR structural studies by taking the protein beyond the readily accessible molecular weight range. Dimerization of a 10 kDa protein would leave it within the accessible molecular weight range but special experimental and/or computational strategies would be required to unravel the intra- and intermolecular nuclear Overhauser enhancements (NOEs).

There are numerous methods available for monitoring protein self-association. These include sedimentation equilibrium and sedimentation velocity experiments performed using an analytical ultracentifuge (AUC)[1] and multiangle laser light scattering (MALLS).[2] Both methods can be used to measure the molecular mass of proteins in solution, with an accuracy of 1–3%, without making any assumptions about the shape of the molecule or its degree of hydration. A variety of NMR-based methods can also be used for monitoring protein self-association; while not as accurate as AUC or MALLS, these methods are very convenient since they can be applied directly to the sample to be used for structure determination and they obviate the need for access to other specialized equipment. One of the simplest and most accurate NMR-based approaches is to derive an estimate of the protein's molecular weight by measuring its translational diffusion coefficient using pulsed-field-gradient spin-echo (PFGSE) NMR.[3]

9.09.2.2 Paramagnetic Ions and Additives

Paramagnetic ions such as Cu^{2+}, Mn^{2+}, high-spin Fe^{3+}, and low-spin Co^{2+} can cause contact broadening of the resonances of nearby nuclei. While paramagnetic ions can be used to probe protein structure,[4] these are special NMR applications and in general these ions should be excluded from samples to be used in high-resolution structure determinations. This can be achieved by treating samples with a metal-chelating agent such as Chelex, or by adding a small amount of ethylenediaminetetraacetic acid (EDTA) (5–50 μmol l^{-1}) to the NMR sample.[5]

Microbial contamination is another potential problem as the highly concentrated protein sample represents an excellent growth medium for algae, bacteria, and fungi during the days to weeks over which NMR experiments will be performed. Algal growth can be eliminated by minimizing exposure of the sample to light. Azide and fluoride can be used to prevent microbial contamination, but azide is volatile below pH 7 and fluoride cannot be used in the presence of metal ions.[5] Broad-spectrum antibiotics such as chloramphenicol are excellent alternatives; chloramphenicol is chemically inert and is effective against both Gram-positive and Gram-negative bacteria at concentrations of 10–50 μmol l^{-1}.[5]

Unlabeled additives such as EDTA are not problematic in heteronuclear NMR studies as they do not yield signals in $^{13}C/^{15}N$-isotope-filtered experiments. For example, it is common practice in heteronuclear NMR studies to add high concentrations (typically 1–15 mmol l^{-1}) of a reducing agent such as dithiothreitol (DTT) or tris(2-carboxy-ethyl)phosphine (TCEP) to avoid oxidation of cysteine residues. High concentrations of nondenaturing detergents have also been used to prevent protein aggregation.[3] However, for homonuclear NMR studies, it might be necessary to use deuterated additives if the additive concentrations are sufficiently high that they would otherwise obscure resonances from the protein.

9.09.2.3 pH

Sample pH is a critical parameter for several reasons. First, it can dramatically affect the solubility of the protein – at high micromolar to millimolar concentrations, many proteins become insoluble when the pH approaches their isoelectric point (pI). Of crucial concern to the NMR experiment is the effect of pH on the rate of exchange of the labile backbone amide protons with solvent protons (most often H_2O). The backbone amide protons are usually the starting point for obtaining resonance assignments in homonuclear scalar correlation experiments such as correlated spectroscopy (COSY) and total correlation spectroscopy (TOCSY) (see Section 9.09.3.2), they provide critical NOE connectivities, and they are often one of the correlated nuclei in heteronuclear triple resonance experiments (see Section 9.09.4.3). Thus, in most NMR experiments, it is desirable to observe as many of the backbone amide-proton resonances as possible.

The exchange of amide protons with solvent water protons is both acid and base catalyzed, with the rate of exchange being lowest at pH values of ~3 and ~5 for the backbone amide and the side chain amide protons of Asn and Gln residues, respectively.[6] Thus, while in theory it might be desirable to work in the pH range 3–5 to maximize the intensity of amide-proton resonances, the protein may be insoluble or may extensively aggregate at these pH values, or its structure may be perturbed. In these cases, the lowest pH consistent with native conformation, tolerable solubility, and negligible aggregation should be chosen for the NMR study. Even a reduction in pH from 7.5 to 6.5 will lead to a 10-fold reduction in the rate of amide-proton exchange. While many earlier homonuclear protein NMR studies were performed at pH <5 in order to limit amide-proton exchange, modern heteronuclear NMR methods and improved water suppression techniques (Section 9.09.2.6) now allow protein structures to be determined at pH values close to neutral. Note that the amide-proton exchange rate can also be reduced by decreasing the temperature, but this has other consequences as outlined in the next section.

9.09.2.4 Temperature

Temperature has an important influence on various aspects of the NMR experiment. Increasing the sample temperature will generally increase signal amplitudes (as long as protein aggregation is not induced) because resonances will become narrower due to the decrease in molecular correlation time (τ_c). This in turn increases the efficiency of coherence transfer through scalar couplings. The rate of exchange of amide protons with solvent water is reduced at lower temperature, making it easier to observe labile amide protons. However, for some experiments, suppression of the water resonance is more efficient at higher temperatures because of its reduced linewidth.

The major factors in deciding which temperature to use are the solubility, state of aggregation, and, most importantly, long-term stability of the peptide or protein sample. Thus, there has been a general trend toward decreased sample temperatures in biomolecular NMR studies as the size of protein being studied has increased (see Chapter 9.07). While many early NMR studies of peptides and small proteins were performed above room temperature (30–50 °C), most recent studies of larger proteins have been carried out at lower temperatures (typically 20–25 °C) due to sample instability at higher temperatures.

9.09.2.5 Ionic Strength

In addition to the sample concentration and specific parameters related to the pulse sequence being used, the spectral signal-to-noise ratio (SNR) depends on various components of the spectrometer hardware, in particular the sensitivity of the probe and preamplifier. The SNR can be related to the temperature of the receiver coil (T_C), its resistance (R_C), the temperature of the sample (T_S), the resistance added to the coil by the sample (R_S), and the noise temperature of the amplifier (T_A) by the following equation:[7]

$$\mathrm{SNR} \sim [T_C R_C + T_A(R_C + R_S) + T_S R_S]^{-0.5} \qquad (1)$$

The sample resistance, R_S, depends on the exact buffer and salt conditions used to solubilize the peptide or protein. Although as a general rule the value of R_S will increase, and the SNR will correspondingly decrease, as the sample ionic strength is increased, R_S is more strictly related to the sample conductivity (σ). Thus, for two different buffers of equivalent ionic strength, the one with lower conductivity will yield the best SNR. In theory, then, dissolving the protein of interest in H_2O would provide the best SNR. However, for proteins and peptides, it is often impractical to use H_2O or a salt/buffer combination with very low conductivity since many proteins are 'salted in' at moderate salt concentrations (see Chapter 9.12). Thus, the choice of buffer will necessarily be a compromise between maximizing protein solubility and minimizing sample conductivity. Nevertheless, for conventional room temperature NMR probes, salt concentrations as high as 1 mol l^{-1} can be used without sacrificing SNR to the point where useful data cannot be collected.[8]

However, high-conductivity buffers are more problematic for modern cryogenically cooled (CC) NMR probes, which have been the most important hardware innovation in biomolecular NMR in the past decade. In CC probes, the receiver coil (T_C) and preamplifier (T_A) are cryogenically cooled to 15–30 K, leading to dramatic improvements in SNR, as predicted by Equation (1). For most multidimensional biomolecular NMR experiments, CC probes provide a 3–4-fold improvement in SNR over conventional probes.[7] However, one consequence of the dramatic decrease in T_C and T_A in CC probes is that the final $T_S R_S$ term in Equation (1) dominates the SNR relationship, and hence the conductivity of the sample can have a dramatic impact on spectral SNR.[7] In general, for CC probes, there will be a very significant deterioration in SNR once the salt concentration exceeds 200 mmol l^{-1}. Thus, if one is using a CC probe, it becomes very important to minimize the conductivity of the salt and buffer without compromising the solubility and stability of the protein sample. The optimization of buffers for use with CC probes is an active area of research and will not be covered in depth here. The reader is referred to tables 1 and 2 in Kelly *et al.*[7] for valuable information regarding the choice of buffers for NMR studies.

An interesting recent development is the use of arginine–glutamate salt (typically 50 mmol l^{-1} L-arginine + 50 mmol l^{-1} L-glutamate) for NMR studies of proteins using CC probes. This zwitterionic salt not only has much lower conductivity than NaCl but has also been shown to help solubilize proteins that are prone to aggregation.[9,10]

9.09.2.6 Solvent Suppression

9.09.2.6.1 Overview

One of the principal advantages of NMR is that molecular information can be probed near physiological conditions. In practice, this equates to dissolving the molecule of interest (such as a peptide or protein) in aqueous solution, often with the addition of appropriate buffers and salts. This unfortunately results in a very strong ^{1}H resonance signal from the protons of the solvent water. This is not surprising considering that the water concentration is near 55 mol l^{-1} whereas that of the solute is generally ~1 mmol l^{-1} or less, resulting in a dynamic range of 10^5:1.

In NMR applications, the solvent signal is most efficiently suppressed by the use of a deuterated solvent (in this case ^2H_2O, often referred to as D_2O); in addition, this provides a deuterium lock signal that is used to correct for magnetic field fluctuations. Unfortunately, many NMR experiments relevant to peptide and protein structure determination require detection of the signals from the backbone amide protons; since these labile protons undergo chemical exchange with the solvent, deuteration renders them invisible in the (^{1}H) NMR spectrum (see also Section 9.09.2.3 on the pH dependence of backbone amide-proton exchange). Consequently, for NMR structure determination studies, the peptide/protein must be dissolved in nondeuterated water (although a small amount of D_2O, typically 5–10%, is still added to the sample for the deuterium lock).

Thus, one of the first practical considerations in setting up an NMR experiment in H_2O is determining the optimal method for suppressing the water signal. Methods developed for this purpose are termed solvent suppression or water suppression techniques. The extent of activity in this field bears witness to its importance and complexity (see Croasmun and Carlson,[11] Price,[12] and Levitt[13] and citations therein). In the following sections, we first discuss the concept of radiation damping before introducing popular methods for solvent suppression in NMR studies of proteins.

9.09.2.6.2 Radiation damping

The signal detected in an NMR experiment is the result of a current generated in the receiver coil due to the bulk magnetization (M_0) of the irradiated nuclei precessing from the transverse (xy) plane toward its equilibrium state along the z-axis due to various relaxation processes.[14] The rate at which this decay into M_{xy} occurs is manifest in an NMR spectrum by the linewidth of the observed signal. Large molecules generally have fast relaxation rates, which results in broad signals, whereas one would intuitively expect the water signal to be much sharper than those of the solute due to its comparatively much slower relaxation rate. In practice, however, the water signal is significantly broader than anticipated on the basis of its intrinsic relaxation rate. This phenomenon, termed radiation damping, is observed when a very strong NMR signal is present (such as when a peptide or protein is dissolved in >90% ^1H_2O) and it is exacerbated when the sensitivity of the receiver coil is very high (as is the case for CC probes). The observed line broadening can be explained by considering that the current induced in the receiver coil due to the strong water signal in turn produces a radiofrequency (rf) magnetic field with the same frequency that rotates the water magnetization back to equilibrium. The extent of this effect can be represented by the time constant T_{RD} (Equation (2)), which describes the rate at which the water magnetization is driven back to equilibrium:[12,15]

$$T_{\mathrm{RD}}^{-1} = 2\pi\gamma\eta Q M_0 \qquad (2)$$

where γ is the gyromagnetic ratio (also referred to as the magnetogyric ratio), η and Q are the filling and quality factors describing the sensitivity of the probe, and M_0 is the equilibrium magnetization per unit volume (assuming the water is irradiated by a 90° pulse). The above relationship shows that an increase in any of the terms on the right-hand side of Equation (2) will lead to increased linewidth. The decay rate of water in highly sensitive CC probes is reduced from seconds to milliseconds due to radiation damping.[16]

Although several techniques involving both pulse sequence[17] and hardware[18,19] modifications have been proposed to specifically reduce radiation damping, the most effective way to achieve this is by suppressing its initiation through solvent suppression. In the following sections, a variety of solvent suppression methods are discussed that also ameliorate the effects of radiation damping by virtue of reducing the intensity of the solvent signal.

9.09.2.6.3 Presaturation

The most direct form of solvent suppression is presaturation of the solvent signal through the application of a frequency-selective, low-power rf pulse at the water resonance over a relatively long period of time (typically seconds) prior to execution of the pulse sequence. This has the effect of equalizing the populations of nuclear spins in the low-energy (α) and high-energy (β) states (i.e., the two available energy states for a spin 1/2 nucleus) with the phases of the individual spins being randomly distributed in the transverse plane. Care should be taken when choosing the resonance frequency of water as errors in tuning can cause a frequency shift.[20]

The presaturation method can in some cases leave an unwanted distortion of the baseline (residual hump) due to magnetic field inhomogeneities across the sample volume. This may be reduced by combining presaturation with the first increment of a NOESY experiment in which the NOESY mixing time is set to zero. The phase cycling of the NOESY experiment reduces signals from parts of the sample where the magnetic field (both external field, B_0, and applied field, B_1) is inhomogeneous. A drawback of this approach is that it restricts the experiment to the use of 90° pulses, which may not be optimal. This restriction is lifted in the FLIPSY[21] approach, making it more generally applicable.

Although commonly used in the past, presaturation is becoming much less popular for biomolecular applications mainly due to the advent of more efficient methods (as discussed below) and also due to a number of shortcomings such as those listed below:

1. Careful minimization of B_0 field inhomogeneity (i.e., shimming) is required.
2. There is bleaching of resonances close to the water irradiation frequency (which is where the H_α resonances of peptides and proteins are often found).
3. There is saturation of amide-proton resonances due to chemical exchange and transfer of this saturation throughout the peptide/protein via dipolar interactions (spin diffusion).
4. Solvent signals originating outside the main sample volume are not efficiently suppressed and can lead to poor solvent suppression.

However, due to its simple setup and generally efficient solvent suppression, presaturation is still a useful method for acquiring an exploratory one-dimensional (1D) experiment in order to gauge properties of the NMR sample.

9.09.2.6.4 Jump–return and binomial sequences

The jump–return sequence differs from presaturation in that instead of saturating the water resonance, it aims to ensure that no net magnetization is produced at this frequency at the end of the pulse sequence. In its simplest form, this sequence consists of two 90° pulses with opposite phases separated by a delay τ. For example, a $90°_x$ pulse rotates the equilibrium (z) magnetization about the x-axis onto the y-axis. Since signals are detected only in the xy-plane, a strong signal would be observed at this point. If the delay τ is ignored, the magnetization is simply rotated back to the z-axis (from the transverse plane) by the second $90°_{-x}$ pulse, due to the opposing phases of the two pulses, thus resulting in no observable signal. However, during the delay, all resonances at frequencies different from that of the carrier frequency (i.e., the frequency at which the rf field is applied) acquire a frequency-dependent component along the x-axis. This frequency dependence is sine-modulated and dependent on the time delay τ according to $\tau = 1/4\Delta\nu$, where $\Delta\nu$ is the difference (in Hz) between the carrier frequency and the frequency at which the maximum intensity will be found along the x-axis. The component of resonances along the x-axis at the end of the τ period will be left completely unperturbed by the $90°_{-x}$ pulse and thus observable.

A very similar sequence is the $1\bar{1}$ sequence, which actually belongs to a family of binomial pulse sequences. Of these sequences, the $1\bar{3}3\bar{1}$ (often referred to as 1–3) is the most popular due to its wider water suppression window (see Figure 1).[22] Due to the frequency-dependent sine modulation of the resonance amplitudes (the so-called 'excitation profile'), the resonances on either side of the water signal (carrier frequency) have opposite signs. In addition, the binomial sequences suffer from baseline distortions due to a strong linear phase gradient (see Figure 1). This baseline distortion can be particularly troublesome in multidimensional experiments and therefore the binomial sequences have not proved popular for multidimensional NMR studies.

9.09.2.6.5 WATERGATE and water-flip-back

The water suppression methods discussed above have lost much of their popularity in biomolecular NMR due to the advent of improved methods that utilize pulsed-field gradients (PFGs). PFGs (often referred to simply as gradients) apply a nonuniform spatial encoding to the NMR sample along the z-axis (only z gradients are discussed as these are available on most modern NMR spectrometers). After the application of a gradient pulse, transverse magnetization (i.e., in the xy-plane) will be dephased. If a spectrum is recorded at this point, no net magnetization will be observed. However, this dephasing is not random (cf. presaturation) and the original signal can be recovered by applying a gradient pulse with the opposite sign. In the meantime, any magnetization along the z-axis will be left unperturbed.

One of the earliest gradient-based solvent suppression techniques, and perhaps still the most widely used, is the WATERGATE sequence.[23] This sequence starts with a hard $90°_x$ pulse, which places the equilibrium magnetization on the $-y$-axis; note that since this is generally the point at which data acquisition commences in most experiments, the WATERGATE sequence can be appended to most sequences in a modular fashion. The initial hard pulse is followed by a gradient pulse (G_z), which dephases both the solute and solvent signals. This is followed by a selective 90° pulse (i.e., one that affects only a narrow frequency range) on the water resonance. At this point, the water magnetization will be dephased and located on the $-z$-axis due to the cumulative effect of

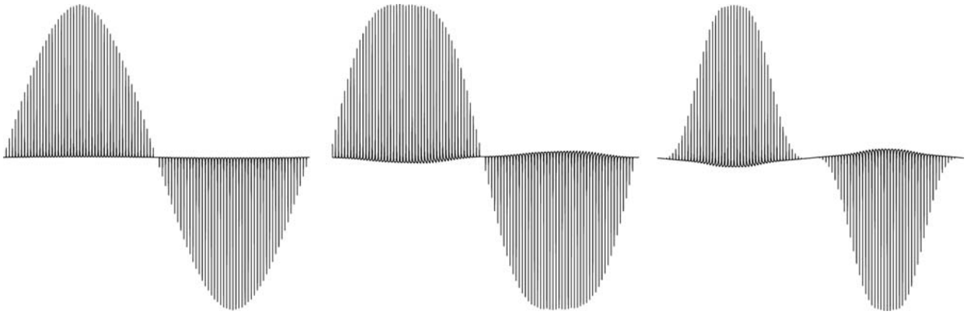


Figure 1 Excitation profiles of the jump–return (left), $1\bar{1}$ (middle), and $1\bar{3}3\bar{1}$ (right) pulse sequences used for suppressing the water resonance.

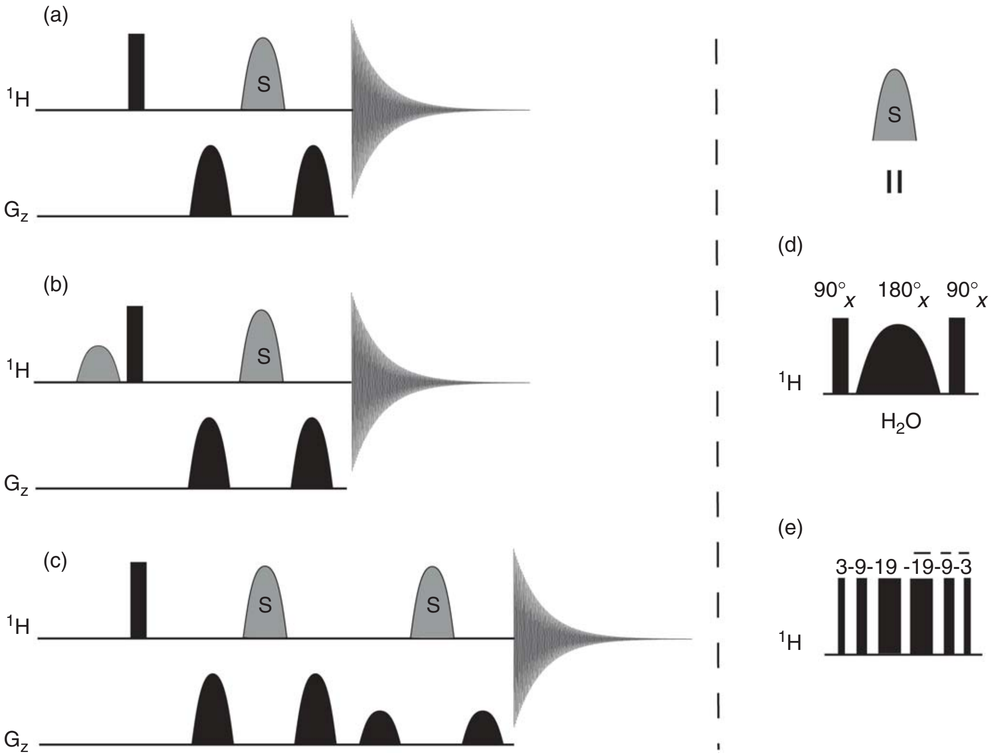


Figure 2 Gradient-echo-based water suppression pulse sequences. (a) WATERGATE; (b) water-flip-back; (c) excitation sculpting; (d and e) examples of the 'S' pulse train that is sandwiched between the gradient echo: (d) water-selective inversion pulse; (e) excitation tailoring using a binomial series.

the hard and soft 90°$_x$ pulses, whereas the solute signals will be unaffected. The next hard 180°$_x$ pulse shifts the water magnetization back onto the $+z$-axis, whereas the solute magnetization is rotated to the $+y$-axis. A final selective pulse rotates the water magnetization back onto the $-y$-axis (where initially it was dephased by G_z). The effect of this pulse train is thus to place the solvent and solute magnetization at opposite ends of the y-axis. When the final G_z pulse is applied, the solute molecules are rephased whereas the water resonances are further dephased (see **Figure 2(a)**).

One of the drawbacks of WATERGATE is that the water molecules have some transverse magnetization at the end of the pulse sequence, albeit dephased. This saturation can be transferred onto labile amide protons through chemical exchange, which, in a similar manner to presaturation, results in reduced sensitivity.

The water-flip-back sequence[24] is different from the WATERGATE sequence in that the initial hard 90°$_x$ pulse is preceded by a selective 90°$_x$ pulse at the water resonance frequency (see **Figure 2(b)**). If we follow through the steps of the WATERGATE sequence as outlined above, we find that, at the end of the pulse sequence, the water magnetization is located along the z-axis and thus is not saturated. This has the advantage of improving the sensitivity of the experiment and further reducing the onset of radiation damping as the gradient pulses remove transverse water magnetization. Since the water is located along the z-axis when the pulse field gradients are applied in the water-flip-back sequence, a lower field strength (compared to WATERGATE) can be used for the gradient pulses.

9.09.2.6.6 Excitation sculpting

In the gradient methods discussed above, a pulse train (S) is placed within two gradients G of equal intensity and duration; such a sequence is a special case of a PFGSE, defined by the sequence G-τ-G. The effect of a PFGSE on a frequency-selective sequence (G-S-G) as in the above examples is to refocus all frequencies experiencing a net inversion (180° rotation) and to dephase stationary components. A drawback of such a PFGSE (apart from those discussed above) is that it will introduce a frequency-dependent phase error determined by the properties of the sandwiched pulse train *S*, leading to undesired baseline distortions. It has been shown that these phase errors can be removed by the addition of a second gradient echo, resulting in a double PFG spin echo (DPFGSE). The second gradient echo 'chips away' at the unwanted magnetization to produce the desired excitation profile, hence the name 'excitation sculpting' (see **Figure 2(c)**).

One of the requirements of this sequence is that the first set of gradients should be different from the second set (i.e., G_1-S-G_1-G_2-S-G_2) so that no dephased magnetization is accidentally refocused by the second set of gradients. Since any sequence *S* can be used regardless of phase properties, it allows for more freedom in designing the sequence *S* to give the desired excitation profile.

It should be noted that when using these sequences there is a dead time between the last pulse and the start of the acquisition (due to instrumental limitations). This can, in some cases, lead to baseline distortions as some of the initial points of the free induction decay (FID) may be lost. Such distortions can effectively be removed by (1) introducing a short delay before the final 180° pulse, (2) incorporating a spin echo at the end of the sequence,[25] or (3) spectral processing methods.[26]

9.09.2.6.7 Coherence pathway selection

The methods outlined above all achieve solvent suppression by manipulating (tailoring) the excitation profile of the observed spectrum. A different approach to achieving this is to take advantage of the fact that the water signal is a singlet that contains no observable homo- or heteronuclear scalar (spin–spin) coupling. Most multidimensional NMR experiments inherently select for a given coherence pathway, generally involving the transfer of magnetization from one nucleus to another through scalar couplings. Thus, the water signal should inherently be suppressed. However, in practice, unwanted coherences are often suppressed by phase cycling and the large (and broad) water signal can still deteriorate the spectral quality because of extreme dynamic range issues (see above). In such cases, PFGs can, as in the above experiments, be used to effectively suppress signals that do not follow the desired coherence pathway. Generally, this is done by applying PFGs at a point during the pulse sequence where the desired magnetization is aligned along the *z*-axis and the water is in the transverse plane. The principles are exactly as discussed above, with the difference being that the water magnetization is manipulated based on its coherence pathway rather than its chemical shift. This principle can be applied to both homonuclear[27] and heteronuclear[28] NMR experiments. In practice, if the coherence selection is not performed at the beginning of the pulse sequence, dynamic range issues and radiation damping may still cause a deterioration in the results. An advantage of this approach is that signals with frequencies similar to the water resonance can be detected since the suppression is not frequency-based, but instead suppresses uncoupled resonances.

9.09.2.6.8 Postprocessing methods

In all of the solvent suppression methods discussed above, there will almost always be a residual water signal remaining at the end of the pulse sequence due to one or a combination of factors such as field inhomogeneities, pulse imperfections, and frequency-dependent delays. Such residual signals can be removed by various postprocessing methods, and these can result in much improved baseline properties that can be important when performing quantitative analyses or when analyzing resonances near the water frequency. It should be noted that to some extent all postprocessing methods are cosmetic and do not address issues such as radiation damping and other experimental issues caused by the solvent.

The most popular postprocessing method for suppression of the solvent signal is the use of a low-pass frequency filter.[29] This method simply filters out frequencies outside a certain bandwidth in the time domain. The filter is usually applied to the central frequency, which generally corresponds to the water resonance. Thus, only the water resonance and nearby resonances remain when the filter is applied. This filtered signal is then subtracted from the original time-domain signal, which results in a visually appealing flat baseline in place of the water signal.

9.09.2.6.9 Summary

The above introduction to solvent suppression serves to elucidate some of the complications imposed by the solvent water signal and common measures for ameliorating them. Note that many modern pulse

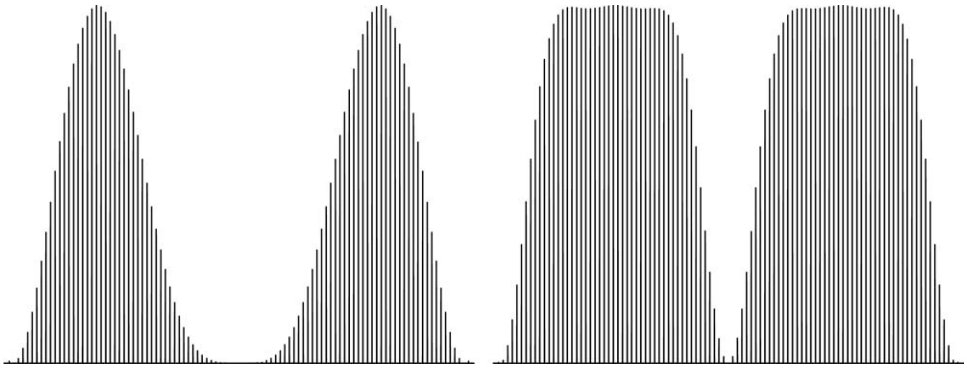


Figure 3 Excitation profiles of the WATERGATE sequence obtained using either a selective 180° pulse sandwiched between two hard 90° pulses (left panel) or a 3-9-19 binomial sequence in place of the 'S' element of the gradient echo (right panel) (see also **Figure 2**).

sequences employ a combination of the above methods. For example, we can conclude that if the pulse angle of the individual pulses in the binomial series is doubled (precluding those that have a τ period when aligned along the z-axis), the net effect of the series is a selective 180° pulse on all resonances except that of the solvent, which is left unperturbed. The same result is achieved by the selective pulse train S in the gradient-echo sequences. Thus, the binomial series can be inserted into any of the PFGSE methods (WATERGATE or excitation sculpting) in place of the S sequence between the gradients. In addition, the effect of the two gradients is to remove the linear phase distortion found when solely using the binomial series for water suppression. This combination of techniques results in an excitation profile consistent with the binomial series but with the advantageous phase properties of the gradient-echo methods, allowing for more elaborate excitation profiles (e.g., the 3-9-19 sequence;[30] see **Figures 2 and 3**). Much recent progress[31–33] in this area has involved combinations of such methods to improve solvent suppression.

9.09.3 Data Acquisition for Nonlabeled Peptides

9.09.3.1 Overview

The general strategy for determination of the 3D structure of proteins and peptides using NMR spectroscopy comprises three distinct stages:

1. The assignment of NMR resonances (^{1}H together with ^{15}N and ^{13}C when possible) to specific atoms or atom groups in the protein.
2. The extraction of experimental constraints from the NMR data, which provide information on the relative spatial positions of these atoms in the protein (see Section 9.09.5).
3. The use of these constraints as input into a computer program that attempts to derive a family of structures for the protein, each of which 'satisfies' the experimental constraints (see Section 9.09.6).

There are two general approaches through which stage (1), resonance assignment, can be accomplished, and the choice between them is determined essentially by the molecular mass of the protein. The first approach, which was pioneered in the laboratory of Kurt Wüthrich[34] and which led to his award of the 2002 Nobel Prize in Chemistry, involves the use of 2D ^{1}H—^{1}H (homonuclear) NMR experiments such as COSY, NOESY, and TOCSY (see below). This approach is still widely used today, but only for proteins smaller than 10 kDa that cannot be isotopically labeled. Nevertheless, the homonuclear NMR strategy is suitable for studying peptides purified from natural sources, and this will be the focus of Section 9.09.3.

9.09.3.2 Homonuclear Resonance Assignment Strategies

9.09.3.2.1 Overview

The advent of 2D NMR techniques in the early 1980s was the key breakthrough that allowed detailed structural information to be extracted from proteins in solution. If information concerning the interactions between spins is to be extracted from 1D experiments, pulses must be selectively applied to particular resonances and their effect(s) on other spins gauged from changes to the 1D spectrum. This method is adequate in the absence of significant spectral overlap, but soon becomes impractical even for molecules of modest size. By contrast, extending such measurements into a second dimension alleviates the overlap problem to a significant degree. These 2D experiments[35] consist of discrete elements – a preparation period; an evolution period (t_1) where spins are 'labeled' as they precess in the *xy*-plane according to their chemical shift; a mixing period, during which correlations are made with other spins; and a detection period (t_2) where an FID is recorded (**Figure 4**).

Note that these elements may be combined to create more complex experiments, which is the basis of the higher dimensionality (3D and 4D) NMR experiments outlined in Section 9.09.4. The FID is signal-averaged as usual (as required for both signal-to-noise and phase cycling considerations)[35] and then the process is repeated a number of times with incremented values of t_1. After Fourier transformation of the series of t_1-incremented experiments, the amplitude of each signal through the series is found to be modulated according to both its intrinsic resonance frequency and the frequency of the proton(s) to which it is correlated during the mixing period (see **Figure 5**).

In addition, transverse relaxation, which occurs during the pulse sequence, will result in smaller peak intensities for increasing values of t_1, so that cross sections of the t_2-transformed data (F_2) have the form of exponentially decreasing sinusoids (cf. FIDs). A Fourier transformation of these cross sections (i.e., with respect to t_1) thus yields a planar spectrum with two frequency dimensions (F_2 and F_1; termed the directly detected and indirectly detected dimensions, respectively) where, in most homonuclear 2D experiments, the 1D spectrum

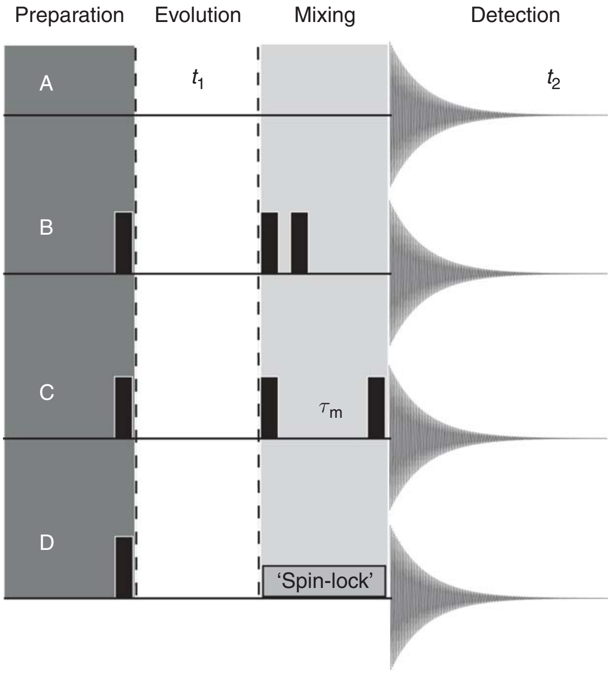


Figure 4 (a) The generic elements of a 2D NMR experiment. The basic pulse sequences shown for (b) DQFCOSY, (c) NOESY, and (d) TOCSY experiments illustrate that the mixing period determines the type of correlation observed in the spectrum. Black rectangles represent 90° pulses. τ_m is the mixing time in the NOESY experiment and the spin-lock time in the TOCSY experiment.

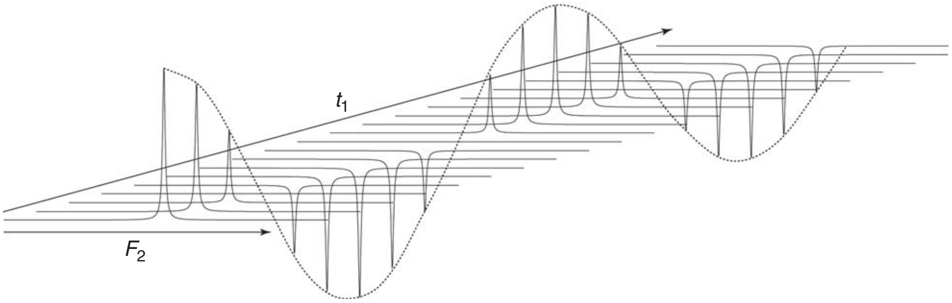


Figure 5 A Fourier-transformed signal in F_2, at frequency ω_2, is modulated in the indirect dimension (t_1) by the incremental delay. This 'interferogram' demonstrates how a sinusoid is created in the indirect dimension when a series of such spectra are collected. Fourier transformation of this sinusoid (along t_1) would thus yield a peak with a frequency ω_1 in F_1.

(sometimes simplified) is present as a diagonal, and correlations between spins are represented by off-diagonal elements known as crosspeaks.

Experiments are distinguished by the nature of the correlations that are probed during the mixing period. Scalar couplings between protons up to three bonds apart are revealed using correlated spectroscopy (COSY)[35] or preferably double-quantum-filtered COSY (DQFCOSY),[36] which has superseded the basic COSY as the experiment of choice for elucidating these couplings due to the narrower lineshapes it produces. NOESY[37,38] connects protons that are close in space (<5.5 Å; see Section 9.09.5.2 for an explanation of the nuclear Overhauser effect). The basic pulse sequences for the COSY and NOESY experiments are given in **Figures 4(b)** and **4(c)**, respectively, and it can be seen that an important difference between them lies in the nature of the mixing period. In DQFCOSY, this consists of two 90° pulses separated by a brief delay (~3 ms), while in NOESY two 90° pulses sandwich an extended mixing time (typically 50–300 ms) during which the NOEs are allowed to build up (there are also phase cycling differences). A further invaluable experiment that also yields scalar connectivities is TOCSY[39] (also called HOHAHA for homonuclear Hartmann–Hahn spectroscopy;[40] **Figure 4(d)**). In TOCSY spectra, correlations are observed between (potentially) all protons within a spin system (i.e., a group of protons that share mutual coupling partners, such as the H_N, H_α, and β-methyl protons of an alanine residue) whether or not they are directly coupled to each other. The coupling is developed during the application of a spin-locking pulse (termed the isotropic mixing period), and the extent to which magnetization is propagated along a spin system depends on the duration of this spin-lock pulse (typically 30–100 ms) and the magnitude of the scalar couplings involved. These three experiments form the basis of the sequential assignment method proposed originally by Wüthrich[34] for the complete assignment of resonances in ^{1}H NMR spectra of polypeptides.

In the first stage of this procedure, the individual spin systems of each amino acid are identified from the scalar-coupled 2D experiments, that is DQFCOSY and TOCSY. Note that all crosspeaks in these experiments correspond to intraresidue connectivities, since there are no interresidue pairs of protons within three bonds of each other. This procedure is greatly aided by inspecting the average for values chemical shift the side chain protons of assigned proteins (**Figure 6**) in the Biological Magnetic Resonance Data Bank (BioMagResBank or BMRB).

This procedure allows residue types to be distinguished, but no information is provided on the positions of these residues within the polypeptide sequence. This information comes from the second stage of the approach, where the characteristic patterns of through-space correlations generated in the NOESY experiment are used to connect sequential pairs of residues and thereby achieve sequence-specific resonance assignment.

9.09.3.2.2 *Suitability of a protein for homonuclear assignment techniques*

It has generally been found that homonuclear resonance assignment cannot be applied successfully to proteins larger than around 10 kDa. There are two reasons for this. First, the complexity (number of crosspeaks) in 2D spectra increases in an approximately linear fashion with the number of chemically inequivalent protons in the molecule. Thus, for proteins larger than around 10 kDa, spectral overlap will generally prevent the exhaustive assignment of resonances necessary for structure determination. The second, more fundamental, reason arises

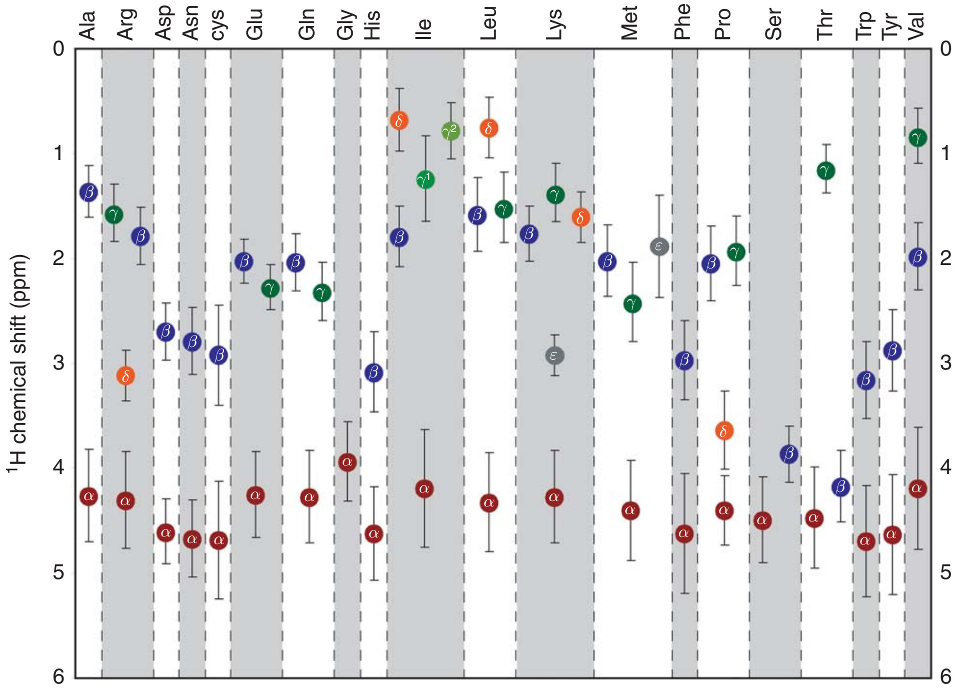


Figure 6 Average chemical shift values for protein backbone and side chain protons extracted from BioMagResBank (http://www.bmrb.wisc.edu). Error bars indicate standard deviations.

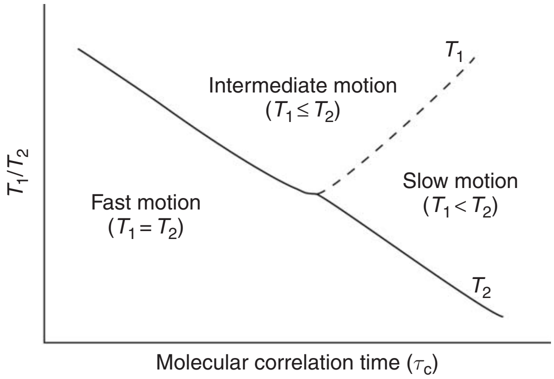


Figure 7 Schematic plot of the relationship between T_1 (longitudinal relaxation time), T_2 (transverse relaxation time), and the molecular correlation time (τ_c). In general, small molecules have short correlation times, whereas large molecules have longer correlation times. T_1 and T_2 are equal in small molecules, whereas T_2 is the dominant relaxation mechanism for large molecules.

from the dependence of the transverse relaxation time, T_2 (and hence the linewidth, $\nu_{1/2}$, which equals $1/\pi T_2$ in the absence of field inhomogeneity), on the molecular correlation time τ_c (see **Figure 7**). τ_c is a measure of how rapidly a molecule tumbles in solution (actually the time taken for a molecule to rotate through one radian), and is given for a spherical molecule by the Stokes–Einstein equation:

$$\tau_c = \frac{4\pi\eta r_h^3}{3kT} \tag{3}$$

where η is the solvent viscosity, r_h is the hydrodynamic radius of the molecule, k is Boltzmann's constant, and T is the temperature. Note that the derived correlation time is an approximate upper limit.

It can be seen from **Figure 7** that, as the molecule tumbles more slowly, T_2 relaxation becomes more efficient compared to T_1 relaxation. (A small T_2 implies efficient relaxation and broad lines, since T_2 is the inverse of the relaxation rate.) This increase in linewidth with molecular size causes two problems: (1) spectral overlap will clearly be worse for broader signals and (2) the efficiency of information transfer between spins in the scalar-coupled experiments (coherence transfer) becomes very poor when the resonance linewidths start to exceed the magnitude of the spin–spin coupling constants. For example, a 7 Hz coupling (an average value for three-bond ^{1}H—^{1}H couplings, $^3\mathcal{J}_{\mathrm{HH}}$) between protons with 20 Hz linewidths gives a COSY-type transfer efficiency of only ~2%.[41] Note, however, that the 10-kDa size limit is only a rough guide, and the exact limit depends on the shape of the protein (which influences the tumbling rate) and the chemical shift dispersion; for example, α-helical domains generally display less dispersion than β-sheets, so that the size limit for predominantly helical proteins will be somewhat lower than that for other proteins.

In order to ascertain whether or not these homonuclear methods will provide complete resonance assignments, a DQFCOSY spectrum of the protein dissolved in H_2O should be recorded. Note that it will be necessary to attenuate the huge signal arising from the solvent (as discussed in Section 9.09.2.6). The majority of the crosspeaks in the so-called fingerprint region of this spectrum ($F_2 \approx 10.0$–6.0 ppm, $F_1 \approx 3.0$–6.5 ppm) will be due to correlations between amide protons and the H_α proton of the same residue. The number of H_N—H_α crosspeaks should be ≥90% of the number of residues in the protein for homonuclear methods to be adequate. However, crosspeaks may be lost for a number of reasons, including transfer of saturation due to the solvent suppression technique used, and the cancellation of antiphase components of the crosspeaks when the linewidths are large compared to the coupling constant. These problems can be circumvented to some degree (see Chapter 4 in Roberts[42]) and attempts should be made to do so before casting the homonuclear assignment strategy aside.

9.09.3.2.3 Spin system identification

Once it has been decided to employ the homonuclear strategy for resonance assignment, good quality DQFCOSY and TOCSY spectra should be acquired in H_2O. The TOCSY spectrum should be acquired with two different mixing times, since the intensity profile for TOCSY crosspeaks is complex, with the intensity of each correlation depending on each of the individual $\mathcal{J}$-couplings if a multistep transfer is involved.[43] Thus, a mixing time that is optimal for long-range transfer from, for example, the amide proton of an Ile residue to its side chain methyl groups may be quite nonoptimal for short-range transfer from H_N to H_α.[43] It may also be useful to record either one or both of these experiments in D_2O, especially if it is suspected that artifacts resulting from incomplete solvent suppression are obscuring crosspeaks involving H_α protons near the water signal. The two spectra complement each other in information content, with TOCSY skewers providing connectivities between most (or often all) protons in the same spin system, and the DQFCOSY distinguishing between direct and indirect connectivities (**Figure 8**).

These spectra are used to identify the type of spin system associated with each H_N—H_α crosspeak, and these spin system types can be matched either to specific amino acids or to groups of amino acids. Note that complete assignment of the longer spin systems such as Lys and Arg is not crucial for sequential assignment, or frequently even for the generation of NOE constraints. Because these residues generally lie on the surface of proteins with their side chains oriented toward the solvent, they are generally very mobile and often exhibit very few structurally useful NOEs

9.09.3.2.4 Sequence-specific resonance assignment

Once all traceable spin systems have been delineated as described above, they need to be matched to specific residues in the sequence. This is achieved using through-space connectivities derived from a NOESY spectrum (or in the case of smaller polypeptides – those with MW ≈ 1000–3000 Da – a ROESY spectrum[45]), which correlates pairs of protons less than ~5.5 Å apart in space, regardless of their relative positions in the primary structure. In general, the shortest mixing time that yields a good quality spectrum is preferable, since indirect effects are observed at longer values. That is, at longer mixing times, magnetization may effectively be transferred between protons that are separated by >5.5 Å (see Section 9.09.5.2).

The sequential assignment procedure relies on the observation of connections between the H_N, H_α, and H_β protons of adjacent residues in the sequence. It has been demonstrated that, irrespective of secondary structure,

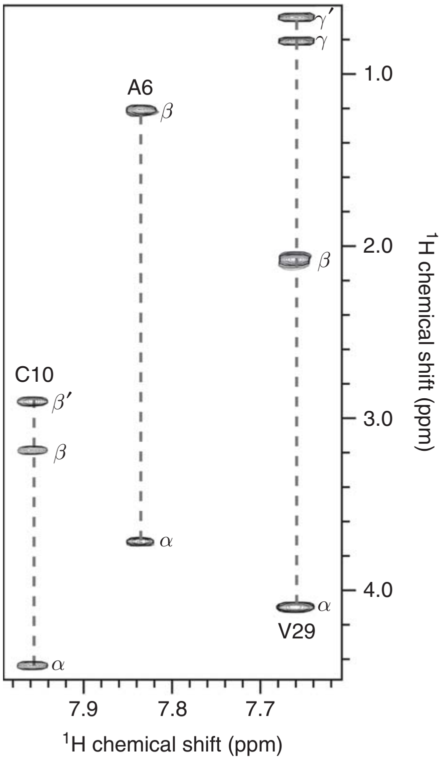


Figure 8 Residue-specific resonance assignment using TOCSY spectra. Portion of the amide region of a 2D ^1H—^1H TOCSY spectrum ($\tau = 80$ ms) of the 37-residue spider toxin κ-atracotoxin-Hv1c.[44] Intraresidue scalar correlations from the backbone amide proton to each of the side chain protons (so-called amide 'skewers') are shown for residues Ala6, Cys10, and Val29. The side chain protons corresponding to each correlation are indicated on the spectrum.

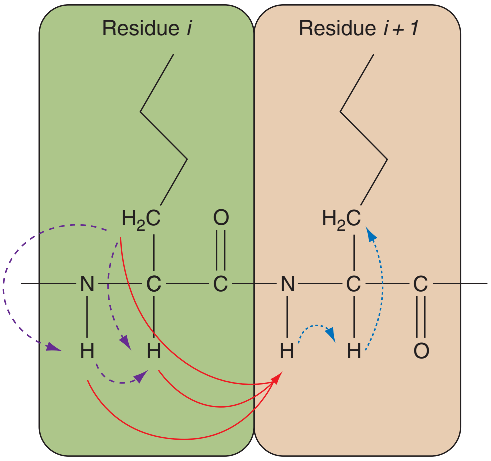


Figure 9 Intra- and interresidue NOE connectivities used to make sequential assignments using the homonuclear strategy. A two-residue protein segment is shown. Blue dotted lines represent intraresidue scalar couplings, which are used to identify the residue type. Purple dashed lines represent intraresidue NOEs, which may assist in this process. Solid red lines show interresidue NOEs, which are used to connect individual spin systems and thereby make sequential assignments.

at least one (generally more) of these pairs of protons will be less than ~3.5 Å apart, and thus should give rise to an NOE with medium-to-strong intensity.[34] The most useful of these are the $d_{\alpha N}(i, i+1)$ (that is, the H_α of a residue i to the H_N of residue $i+1$), $d_{NN}(i, i+1)$, and $d_{\beta N}(i, i+1)$ correlations (see **Figure 9**).

Before commencing this stage of the procedure, it is useful to record a NOESY spectrum in D_2O. In this spectrum, the only signals downfield of ~6 ppm correspond to carbon-bound protons from His, Phe, Trp, and Tyr, so these crosspeaks may be marked as such on the H_2O-NOESY to simplify the assignment task. Note that complete exchange of the backbone amide protons for deuterons may take days, weeks, or even months for some protons, depending on sample temperature and pH (high temperature and pH favor exchange). In fact, partial hydrogen–deuterium exchange may be used to simplify (edit) both scalar-coupled and NOESY spectra if spectral overlap proves to be a problem (as is often the case for predominantly α-helical proteins). DQFCOSY/TOCSY and NOESY spectra collected soon after dissolution of the protein in D_2O will exhibit only a subset of the H_N protons, together with their associated connectivities. Similarly, a sample that has been quantitatively exchanged with D_2O may be freeze-dried and reconstituted in H_2O, and data collected on this so-called reverse-exchanged sample will contain a complementary subset of correlations.

Thus, the presence in the H_2O-NOESY of any of the three classes of NOE listed above is used to infer a sequential juxtaposition of the two amino acids concerned. It should be realized, however, that these types of NOEs can (and often will) be observed between residues that are not neighbors in the sequence, so that caution, as always, should be exercised. Consequently, it is more reliable to base a sequential connection on the evidence of more than one of these three types of NOEs. Breaks in the sequential assignment will inevitably occur – these can be the result of spectral overlap (e.g., two amide protons with identical chemical shifts will prevent the observation of sequential NOEs) or of the structure itself (e.g., the presence of a proline, which lacks an amide proton). Sometimes the H_δ protons of Pro can be used to continue the assignment sequence (i.e., using $d_{\delta N}(i,\ i+1)$ connectivities), although this requires prior assignment of the proline spin system, usually a difficult task in the early stages of the procedure. Once short sequences of spin systems (3–4 residues) have been picked out, these can be mapped onto the polypeptide sequence. In some cases, a unique match may be found, but often there will be several possible assignments, and the segment must be extended in either or both directions as described above, until (hopefully) all but one of the possibilities can be excluded. For a 200-residue protein containing all 20 types of amino acids, there is a 99% probability that a tetrapeptide segment will be unique.[34] This process is repeated until all possible assignments have been made (**Figure 10**).

Although homonuclear resonance assignment is almost exclusively carried out using the sequential assignment method, one other approach has found use in some applications. The main-chain-directed (MCD) method[46] is based on the identification of cyclic patterns of NOEs, which are characteristic of the different types of secondary structure. Because of this, it is less suitable for the assignment of unstructured or irregularly structured sections of a protein.

9.09.3.2.5 Three-dimensional homonuclear NMR

In the late 1980s, a further increase in dimensionality of NMR spectra was proposed and realized: the extension of 2D experiments to a third dimension. A 3D experiment may be considered to be a combination of two 2D experiments in which the detection period (t_2) of the first experiment is replaced by that of a second experiment (of which at least the first 90° pulse of the preparation period is removed). Thus the 3D experiment entails two evolution times (t_1 and t_2), two mixing periods, and a detection period (t_3). The two evolution times are incremented independently, and a 3D Fourier transformation, analogous to the 2D transform described previously, yields three orthogonal frequency axes in a cubic arrangement. Two implementations of this technique involve the combination of the 2D NOESY and TOCSY experiments to give the 3D NOESY–HOHAHA[47] (and the closely related HOHAHA–NOESY[48]) experiments, and of two NOESY sequences, giving a NOESY–NOESY.[49] The extension into the third dimension offers a potential increase in resolution since crosspeaks are now characterized by three frequencies. If, for example, both the H_N and H_α signals for two Ala residues were coincident, their scalar correlations would still be distinguishable in a HOHAHA–NOESY if their H_β protons (to which they would show NOEs) had distinct chemical shifts. Note that this chemical shift difference needs to be larger than the resolution of the spectrum afforded in that dimension – this is less likely than in the corresponding 2D spectra, as the increase in dimensionality effectively results in a decrease in resolution for the same length experiment. Thus, in some cases, this approach can partially alleviate the spectral overlap problem, which hampers the use of homonuclear assignment techniques for larger proteins at a cost of longer experiments.[48,50] However, these experiments have not been used widely as they add little additional information to the

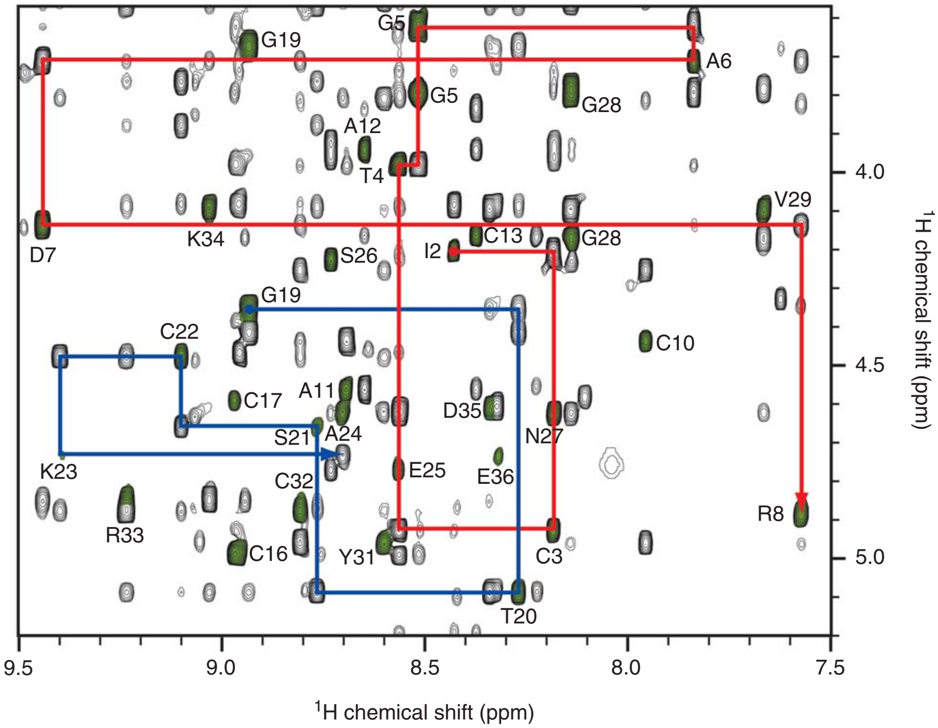


Figure 10 Sequence-specific resonance assignment using interresidue NOEs. A portion of the fingerprint region of a 2D ^{1}H—^{1}H NOESY spectrum ($\tau = 250$ ms) of the 37-residue spider toxin κ-atracotoxin-Hv1c is shown.[44] Intraresidue H_α—H_N NOEs are highlighted in green. Shown are two sets of sequence-specific resonance assignments obtained by connecting adjacent residues (i and $i+1$) via interresidue $H_\alpha(i)$—$H_N(i+1)$ NOEs. The red lines illustrate the backbone 'walk' from Ile2 to Arg8, while the blue lines illustrate the backbone 'walk' from Gly19 to Ala24.

2D experiments. Consequently, it appears that for proteins that are not amenable to a straightforward 2D homonuclear approach, isotopic labeling is probably the method of choice (see Section 9.09.4).

9.09.3.3 Heteronuclear Resonance Assignment Strategies

9.09.3.3.1 Introduction

Traditionally, only homonuclear experiments are used for determination of the structure of peptides when isotopic labels cannot be introduced. The reason for this is primarily the low sensitivity of such experiments, since the natural abundance of the NMR-relevant isotopes of carbon and nitrogen are only 1.1 and 0.4%, respectively (not to mention the unfavorable gyromagnetic ratios of these nuclei compared to ^{1}H). Moreover, although the assignment of these nuclei can aid in obtaining sequence-specific resonance assignments, they do not provide any additional structural restraints. In addition, the low abundance of these isotopes does not allow for heteronuclear-edited experiments (see Section 9.09.4). Due to these drawbacks, heteronuclear experiments were in the past rarely pursued for peptide structure determination. However, with the introduction of residual dipolar couplings (RDCs; see Chapter 9.07) as a source of structural restraints, these experiments may prove useful in structure calculations.[51] The method for acquiring RDC information involves recording a set of two spectra for each experiment, one where the sample is isotropically tumbling (normal solution conditions) and one where the sample is in an aligned medium (generally a phage or liquid crystal solution). The spectral qualities in the second sample are often inferior to those of the first sample, due to the restricted tumbling. In addition, the set of 2D spectra used to extract the coupling constants are recorded without decoupling the heteronucleus, resulting in a splitting, which further reduces sensitivity (at least half the SNR) compared to the traditional decoupled spectrum. These experiments are therefore of interest only when large sample quantities are available (several millimoles per sample). Furthermore, it should be mentioned that recovery of the peptide from the sample containing the alignment media is generally not trivial and therefore often not attempted. Thus, if sample quantities are scarce, this experiment should be left as the final experiment.

The assignment of heteronuclei generally requires acquisition of a heteronuclear single quantum coherence (HSQC) experiment and a heteronuclear multiple bond correlation (HMBC) experiment, and these are described in more detail below.

9.09.3.3.2 Heteronuclear correlation spectroscopy

Although multibond heteronuclear correlation experiments can to some extent aid in the assignment process, these experiments are most easily analyzed once the sequential assignment process is complete. The simplest experiment is the $^{1}H—^{15}N$ HSQC (see also Section 9.09.4), which correlates each backbone amide proton with its directly attached ^{15}N nucleus. For the purpose of assignment, it is best to acquire this experiment under ideal conditions (i.e., not in an aligned medium and with decoupling of the ^{15}N nuclei during acquisition). If assignments have already been obtained for each of the amide protons, then it is simply a matter of collecting an HSQC experiment with enough resolution and sensitivity to assign all of the directly attached nitrogen atoms. Depending on sample conditions and hardware available, this experiment can take from a few hours up to a day to run (**Figure 11**).

The second heteronuclear one-bond experiment to acquire is the $^{1}H—^{13}C$ HSQC experiment. This experiment correlates the chemical shift of each ^{1}H resonance with the ^{13}C chemical shift of the carbon atom to which it is attached. In comparison with the $^{1}H—^{15}N$ HSQC experiment, the $^{1}H—^{13}C$ HSQC spectrum is more crowded, simply due to the large number of one-bond CH correlations; these correlations include $^{1}H—^{13}C_{\alpha}$ and side chain $^{1}H—^{13}C$ methyl/methylene one-bond correlations. The $^{1}H—^{13}C_{\alpha}$ correlations are generally the easiest to assign as they appear in a distinct spectral region ($^{1}H = 3.5–5$ ppm), they have a simple splitting pattern, and they typically have good signal dispersion. In contrast, the crowded and more complex splitting patterns found in the side chain region (0.5–3 ppm) will make the assignment of anything but the methyl groups rather difficult.

One method for improving the level of assignments is to record a ^{13}C HMBC spectrum. This experiment correlates ^{1}H atoms with ^{13}C atoms that share a small coupling constant, thus providing multiple bond correlations. By using the assigned data for the ^{1}H and ^{13}C atoms found by the methods described above, additional assignments may be possible. Apart from aiding in the assignment of side chains, which may be ambiguous from the homonuclear and one-bond heteronuclear assignments, the experiments can provide sequential information as they show correlations to quaternary atoms (such as the C=O). Such assignments

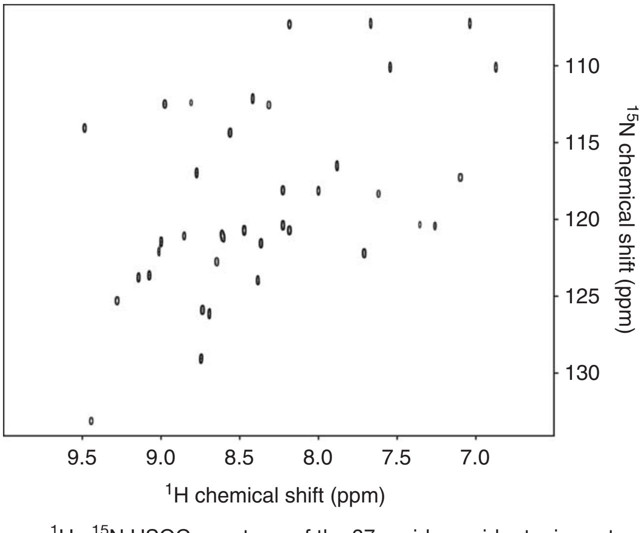


Figure 11 Natural abundance $^{1}H—^{15}N$ HSQC spectrum of the 37-residue spider toxin κ-atracotoxin-Hv1c.[44] The experiment was acquired at 900 MHz using a 0.5 mmol l^{-1} unlabeled peptide sample and an acquisition time of 12 h.

may be particularly useful for prolines (which lack H_N atoms). Although it is theoretically possible to extract coupling information from HMBC spectra, the low sensitivity of this experiment rarely allows for this.

The heteronuclear NMR experiments discussed above highlight how much extra resonance dispersion can be gained via this approach. The power of this added dimension becomes clear if, for example, the ^{1}H—^{15}N HSQC experiment shown above, where each H_N atom is essentially resolved, was to be combined with a TOCSY or NOESY experiment to provide a third frequency dimension. The resulting 3D ^{15}N-HSQC-TOCSY/NOESY spectrum would contain virtually no overlap of interresidue resonances. Such experiments are indeed possible and have been the driving force in producing uniformly ^{15}N- and/or ^{13}C-labeled proteins. This field has been the most intensely researched area of NMR in the past 20 years, and the strategies employed to determine protein and peptide structures using heteronuclear NMR experiments are discussed in the next section (see Chepter 9.19).

9.09.4 Data Acquisition for Isotopically Labeled Proteins

9.09.4.1 Overview

As mentioned in Section 9.09.3, the homonuclear NMR strategy will fail to provide complete and unambiguous assignments for larger proteins – the slower molecular correlation time with its consequent increase in the efficiency of transverse relaxation translates to broader lines and poor coherence transfer via ^{1}H—^{1}H scalar couplings. The increased number of protons in larger proteins also increases the resonance overlap problem. Homonuclear 3D NMR techniques, while providing some relief, still rely heavily on inefficient (for large proteins) ^{1}H—^{1}H scalar couplings. The gain in resolution from the added dimension is also tempered by both the limited frequency range of ^{1}H and the large increase in the number of crosspeaks generated in such experiments (compared to either of the constituent 2D experiments).

The advent of recombinant DNA technology has allowed the relatively facile production of proteins bearing isotopic labels in a variety of arrangements. For example, specific amino acid types may be labeled (e.g., 100% ^{13}C labeling of all carbons in all Leu residues) or the whole protein may be labeled uniformly with ^{13}C and/or ^{15}N. In general, either uniform ^{15}N labeling or ^{15}N/^{13}C labeling is used. The magnetic properties of these nuclei (both with spin quantum number $I = 1/2$) allow them to be utilized in high-resolution NMR, most commonly by exploiting their often large one-bond and two-bond scalar couplings to each other and to directly attached protons (**Figure 12**).

These large couplings constitute a major advantage of heteronuclear over homonuclear multidimensional NMR, as magnetization transfer is very efficient in comparison with the homonuclear case (where $^3\mathcal{J}_{HH} \approx 3$–14 Hz). Thus, the ^{1}H—^{15}N HSQC experiment (discussed further in Section 9.09.3.3.2),[52] which correlates the chemical shifts of ^{15}N nuclei (both backbone and side chain) to their directly attached proton(s), has very high sensitivity because magnetization is transferred via a very large one-bond $\mathcal{J}$-coupling of ~90 Hz (**Figure 12**). The HSQC pulse sequence (see **Figure 13(a)**) involves the initial transfer of ^{1}H magnetization to ^{15}N through the one-bond coupling (using a sequence known as insensitive 'nuclei enhanced by polarization transfer' (INEPT)[53]), an evolution period (t_1) where the magnetization is labeled with the ^{15}N chemical shift, and transfer back to ^{1}H (with reverse-INEPT) for ^{1}H chemical shift detection during t_2. Double Fourier transformation yields a 2D spectrum with no diagonal and a single in-phase crosspeak representing each ^1H_N–^{15}N correlation.

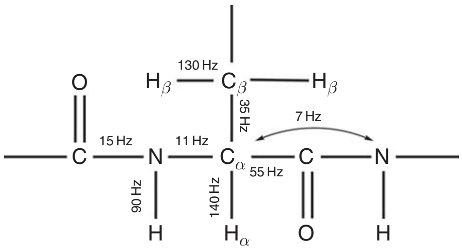


Figure 12 Segment of a polypeptide chain showing the magnitude of the scalar *J*-couplings used in heteronuclear NMR experiments.

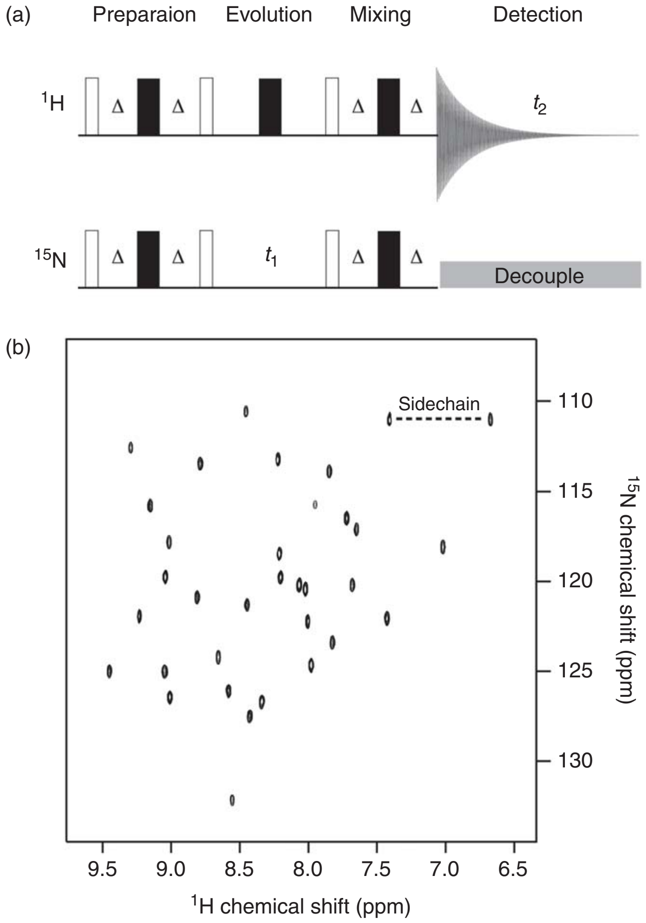


Figure 13 (a) Pulse sequence for the 2D ^{1}H-^{15}N HSQC experiment. Unfilled and filled rectangles represent 90 and 180° pulses. The delays (Δ) are tuned to 1/4*J* to allow magnetization transfer between ^{1}H and ^{15}N. The gray rectangle on the ^{15}N line indicates decoupling for that nucleus during signal acquisition. (b) ^{1}H-^{15}N HSQC spectrum of a 41-residue peptide toxin (0.5 mmol l^{-1}) from the spider *Agelena orientalis* that has been uniformly labeled with ^{15}N. The experiment shows all amide-proton–^{15}N correlations; these arise mainly from the backbone amides and also from side chain amides of Asn and Gln residues. Side chain amide correlations can be readily identified because (1) the ^{15}N nucleus is correlated to two ^{1}H chemical shifts arising from each of the two directly attached protons and (2) each correlation has a weak partner 0.5–0.6 ppm upfield in the ^{15}N dimension that results from the deuterium isotope effect produced by the ~10% semideuterated NHD moieties present in 90% H_2O/10% D_2O solution. One pair of side chain amide correlations is labeled.

Figure 13(b) shows a ^{1}H—^{15}N HSQC spectrum acquired from 0.5 mmol l^{-1} sample of a 41-residue peptide toxin from the spider *Agelena orientalis*. The toxin was produced recombinantly and uniformly labeled with ^{15}N. This HSQC spectrum was collected in 30 min, compared with the 12 h required to acquire a natural abundance spectrum from an unlabeled sample of equivalent concentration (see **Figure 11**). The HSQC, together with the related heteronuclear multiple quantum coherence (HMQC)[54] experiment, forms the cornerstone of a wide range of 2D, 3D, and 4D experiments that are designed to facilitate sequence-specific resonance assignment and determination of protein structure. Note that the HSQC technique is the technique of choice for correlation of ^{1}H and ^{15}N shifts due to generally narrower linewidths in the ^{15}N dimension.[55,56] Furthermore, because these and most of the other heteronuclear experiments described below are designed to observe amide protons, the sample must be in H_2O (rather than D_2O). Consequently, a means of suppressing the H_2O resonance is required (for details see Section 9.09.2.6).

Because of the complex multistep nature of multidimensional experiments, their main adversary is transverse relaxation, T_2. First, T_2 (and hence linewidth) for a particular nucleus determines how efficient coherence transfer via scalar couplings will be for that nucleus. As described above, when linewidths become larger than the magnitude of the scalar couplings concerned, transfer efficiency declines markedly. The linewidths for a 20 kDa globular protein at 25 °C will be ≈12 Hz for H_N, ≈7 Hz for N when proton-coupled (≈4 Hz when decoupled), ≈15 Hz for $^{13}C_\alpha$, and ≈25 Hz for H_α (attached to ^{13}C).[57] These are in general smaller than the couplings used in these experiments, although it is clear that the small $^1\mathcal{J}_{C\alpha N}$ coupling is a primary determinant of which experiments may be carried out with reasonable sensitivity as the protein size increases.

A second problem is that the transverse magnetization associated with a particular nucleus loses phase coherence (and therefore intensity) at a rate characterized by T_2 for that nucleus. Thus it is important to minimize the length of time spent on nuclei such as C_α, which have comparatively short T_2 times (~20 ms for C_α). Thus, experiments that correlate backbone amide nuclei with side chain nuclei of the 'preceding' residue are significantly more sensitive than the corresponding intraresidue experiments, as the former avoids the need for magnetization transfer via the $^1\mathcal{J}_{C\alpha N}$ coupling. For INEPT-type transfer between two nuclei, magnetization must be resident on each nucleus for around $1/2\mathcal{J}$ s. This corresponds to 50 ms spent on C_α for $C_\alpha \rightarrow N$ transfer, but only 9 ms for $C_\alpha \rightarrow C'$ transfer. Clearly the latter pathway allows less transverse relaxation. For similar reasons, most of the experiments described below will decrease rapidly in efficiency for proteins larger than ~20 kDa.

Obviously, it is also important to limit the total length of the pulse sequence so as to minimize T_2 relaxation prior to signal acquisition. Simple concatenation of magnetization transfer and free precession (frequency labeling) periods as described above produced the first triple resonance experiments, but it is possible to overlay such periods so that both free precession and magnetization transfer occur during the same time interval. This gives rise to so-called 'constant time' experiments with significantly shortened pulse sequences.[58,59] This is one of the many 'tricks' employed to increase SNR in heteronuclear multidimensional NMR experiments (including the use of PFGs[60] and sensitivity enhancement[61,62]). The increase in sensitivity gained by application of such tricks can then be used to either improve resolution (shorten experiment time) through the use of nonuniform sampling[63] or study larger systems (or both in favorable cases[64]). In addition, for very large proteins, various sequences incorporating the TROSY principle have been developed, which often also greatly benefit from ^{2}H labeling, since relaxation by ^{2}H is less efficient than by ^{1}H, leading to longer T_2.[65]

These strategies are, however, not discussed in this chapter, as they are methods that only improve the applicability and sensitivity of the experiments covered here. The principles covered here are those that are fundamental to NMR structure determination and valid regardless of pulse sequence elaborations designed to improve sensitivity or resolution. As mentioned above, the most common labeling strategies for NMR studies of proteins are either uniform ^{15}N labeling or uniform double labeling ($^{15}N/^{13}C$). ^{13}C enrichment is generally more expensive, and in some expression systems more difficult than ^{15}N labeling. The experiments that employ single-labeled samples are often referred to as heteronuclear-edited experiments and those that employ double-labeled samples are known as triple resonance experiments. This section is thus split into two segments, which describe the assignment strategy for each of these two cases.

9.09.4.2 Heteronuclear-Edited NMR Experiments

Concatenation of the $^1H—^{15}N$ HSQC (or HMQC) sequence with a $^1H—^1H$ NOESY gives rise to the 3D ^{15}N-edited NOESY–HSQC (or 3D NOESY–HMQC) experiment.[66–68] Here, two of the frequency dimensions represent the amide ^{1}H and ^{15}N chemical shifts, while the third dimension provides information about the chemical shift of protons with which each amide proton is dipolar coupled (i.e., separated by <5.5 Å). The spectrum is routinely viewed as narrow 2D ($^1H—^1H$) strips taken at the ^{15}N chemical shift of each crosspeak in the $^1H—^{15}N$ HSQC spectrum (see **Figure 14**).

As exemplified in **Figure 14**, the increase in resolution compared to a simple 2D NOESY is dramatic, due in part to the lack of a straightforward correlation between ^{15}N chemical shift and the secondary structure in which a residue is located (in contrast to the case of H_N, H_α, and C_α chemical shifts). An analogous combination of TOCSY and HMQC/HSQC yields 3D TOCSY–HMQC/HSQC,[69,70] where the third dimension as described above shows the chemical shifts of protons to which the amide protons would exhibit correlations in a conventional TOCSY (i.e., those protons in the same spin system). Thus, when satisfactory NOESY–HSQC and TOCSY–HSQC spectra are obtained, a semiclassical route to resonance assignment

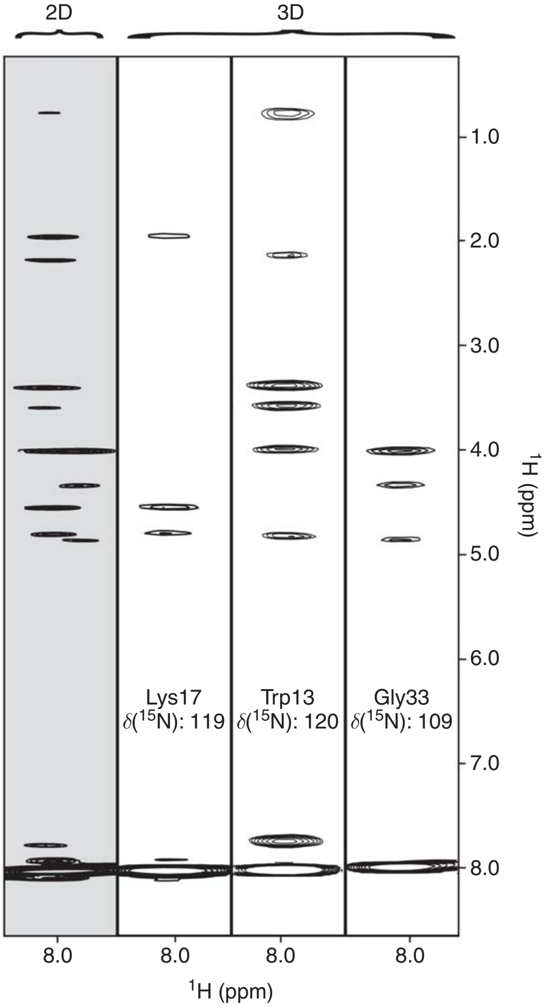


Figure 14 Comparison of 2D NOESY and 3D ^{15}N-edited NOESY–HSQC spectra of a 41-residue peptide toxin from the Australian funnel-web spider *Hadronyche infensa*. A strip from the 2D NOESY spectrum is shown on the far left and it illustrates overlapping NOE correlations from three different amide protons (those of Trp13, Lys17, and Gly33). Fortunately, the ^{15}N nuclei for these three amide groups have unique chemical shifts and hence they appear on different 2D planes in the 3D NOESY–HSQC experiment. Strips from these three planes are shown on the right, and they demonstrate that all of the NOE correlations are perfectly resolved in the 3D experiment.

can be followed. TOCSY skewers from the TOCSY–HSQC are used to identify spin system types and to account for intraresidue NOEs in the NOESY–HSQC. Sequential NOEs can then be identified from the latter spectrum and used to deduce interresidue connectivities as usual. Information from 2D DQFCOSY and 2D NOESY spectra may aid assignment, as many direct scalar correlations and NOEs should be distinguishable, even for large proteins. 2D versions of these two 3D experiments, consisting of ^{15}N shifts in one dimension (F_1) and skewers of NOE/TOCSY correlations to the directly attached amide proton in the other (F_2), have been described.[55,71] These 2D experiments have the advantage of smaller demands on spectrometer time and easier implementation, although their effective resolution compared to the 3D experiments is obviously much poorer.

Note that analogous experiments, such as the ^{13}C-edited HSQC–NOESY,[72] can be performed on ^{13}C-labeled proteins. For labeled proteins, this latter experiment provides the largest number of conformational restraints for protein structure calculations (see Section 9.09.5.2). The ^{15}N-edited NOESY–HSQC only provides distance information for protons that are close in space to amide protons (since magnetization originates and/or terminates on an amide proton). Thus, it provides no information about distances between pairs of carbon-bound protons, which represent by far the largest group of close interproton distances in proteins.

A further implementation of heteronuclear editing of homonuclear spectra can be carried out when specifically labeled samples are available (e.g., ^{15}N labeling of all Leu residues). Normal 2D homonuclear pulse sequences to which a so-called difference echo is appended yield ^{1}H–^{1}H spectra where only the residues carrying the labels appear.[73,74] This can be useful for resolving ambiguities that may be present even in the 3D experiments, or to study interactions of two differently labeled proteins.[75] Some researchers have made full chemical shift assignments by generating many such specifically labeled samples and applying these techniques to each one.[76,77] However, this is a very labor-intensive approach that requires a well-behaved expression system, and it is not expected to be widely applicable.[77]

From a resonance assignment viewpoint, the most significant limitation of the experiments described in this section is that they rely on the transfer of magnetization via small homonuclear couplings ($^3\mathcal{J}_{\alpha\mathrm{N}}$ can be as low as 3 Hz for α-helical regions of a protein), and hence they will fail for larger proteins, as noted earlier. Moreover, the sequential assignment process requires the use of NOEs, which do not provide unambiguous connections as readily as scalar couplings; assignment consequently involves a pattern matching process that is very time consuming and prone to error. The next section describes an approach that largely circumvents these problems and allows routine resonance assignment for proteins up to ~20 kDa (and larger in favorable cases, particularly when deuteration is applied).[78]

9.09.4.3 Triple Resonance Experiments for Protein Backbone Assignment

As previously discussed, the large size of one- and two-bond heteronuclear (and homonuclear $\mathcal{J}_{\mathrm{CC}}$) couplings (**Figure 12**) results in very efficient magnetization transfer (using either HSQC or HMQC sequences) relative to homonuclear scalar transfer through either COSY- or TOCSY-type techniques. Therefore, resonance assignment using NMR spectra that make use of these large couplings represents an appealing alternative. Although two of these couplings ($^1\mathcal{J}_{\mathrm{NC}\alpha}$ and $^1\mathcal{J}_{\mathrm{NC}'}$) are quite small, the heteronuclear experiments have been (and still are being) carefully designed and optimized so as to minimize the problems presented by these couplings (see below). The concept underlying this class of experiments is that magnetization is transferred between nuclei via scalar couplings, such that the frequencies of some or all of the atoms involved in the transfer pathway are sampled. Thus a 3D (or 4D) spectrum is obtained that correlates the chemical shifts of three (or four) nuclei as defined by the coherence pathway chosen. In this way, a number of different interresidue correlations can be made, providing unambiguous sequential assignments (cf. NOE-based connections). The size of these one-bond $\mathcal{J}$-couplings is generally insensitive to conformation, allowing the delays in the pulse sequences to be accurately tuned to the coupling constants.

Although the names of triple resonance NMR experiments appear rather esoteric, they are in fact rationally derived on the basis of the nuclei that are involved in the coherence transfer pathway; nuclei that are used in the transfer pathway, but whose chemical shift is not sampled, appear in parentheses. Thus, the HNCO experiment provides interresidue correlations between the $\mathrm{H_N}$ and N nuclei of residue *i*, and carbonyl carbon of residue $i-1$; in the pulse sequence name, the term 'HN' implies that both the amide proton and its attached nitrogen are frequency-labeled. Note that CO refers to the carbonyl carbon, and that CA, CB, HA, and HB refer to the C_α, C_β, H_α, and H_β nuclei, respectively. The order in which the atoms are listed indicates the direction of the magnetization transfer. In virtually all cases, magnetization starts on ^{1}H and is transferred to a heteronucleus (using either an INEPT or HMQC transfer), so that the sensitivity of the experiment is increased relative to starting with the magnetization on the heteronucleus.[79] In addition, ^{1}H magnetization is always detected in the direct dimension, again for sensitivity reasons, and therefore experiments such as the HNCO are termed 'out-and-back' experiments. That is, after transfer to CO, the reverse pathway is traced, so that the entire experiment is described by $\mathrm{H_N}(i) \rightarrow \mathrm{N}(i) \rightarrow \mathrm{CO}(i-1) \rightarrow \mathrm{N}(i) \rightarrow \mathrm{H_N}(i)$. The HN(CA)CO,[80] also an out-and-back experiment, selects a symmetrically related intraresidue pathway: $\mathrm{H_N}(i) \rightarrow \mathrm{N}(i) \rightarrow \mathrm{C}_\alpha(i) \rightarrow \mathrm{CO}(i) \rightarrow \mathrm{C}_\alpha(i) \rightarrow \mathrm{N}(i) \rightarrow \mathrm{H_N}(i)$. The C_α is not frequency labeled and hence it appears in parentheses. Note that the C_α nuclei (together with C_β nuclei in other experiments) are treated separately from the carbonyl carbons in these sequences. To make this possible, specialized pulses must often be generated that excite specific spectral regions; for example, a pulse may be required that excites the CO region of the carbon spectrum but not the

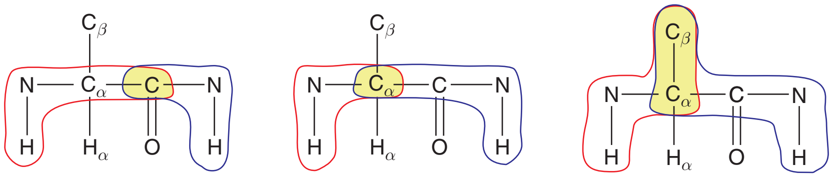


Figure 15 Schematic illustration of three different pairs of triple resonance NMR experiments that can be used for making sequence-specific resonance assignments. Left panel: HNCACO and HNCO; middle panel: HNCA and HN(CO)CA; right panel: HNCACB and CBCA(CO)NH. In each case, the experiment listed first, which is shown in red, provides intraresidue correlations (and sometimes also interresidue correlations), whereas the experiment listed second, shown in blue, provides only interresidue correlations.

aliphatic region. This requires either a spectrometer with four separate amplifiers or, if this is not available, the use of off-resonance frequency-selective pulses (for a review of selective pulses, see Kessler *et al.*[81]).

A formidable array of triple resonance experiments has been developed since the concept was first introduced.[82–84] Consequently, a number of different triple resonance strategies are available for making sequence-specific resonance assignments, and the strategy that works best will depend on the size, and hence the relaxation properties, of the protein concerned (in particular C_α and H_α atoms, see below). These experiments mostly consist of complementary pairs (see **Figure 15**), with one providing exclusively interresidue connectivities and the other providing both intra- and interresidue connections. This situation arises because $C_\alpha(i)$ is coupled to both $N(i)$ ($^1\mathcal{J}_{C\alpha N}$) and $N(i-1)$ ($^2\mathcal{J}_{C\alpha N}$), so that any sequence that transfers magnetization between C_α and N will branch off in two directions. Crosspeaks resulting from these two pathways are often distinguishable in the final spectrum due to the lower intensity of the interresidue signals (see **Figure 16**), which arise from the smaller $^2\mathcal{J}_{C\alpha N}$ coupling constant. Conversely, the interresidue pathway can be selected exclusively by routing magnetization (originating from either $N(i)$ or $C_\alpha(i-1)$) through $C'(i-1)$. Thus, the basic strategy is to create clusters of nuclei and then to link these clusters (preferably in two or more independent ways) to generate fragments of sequentially linked amino acid residues. These fragments can then be matched with the known amino acid sequence of the protein using spin system information, generally with the help of chemical shift information.

The HNCA[58,83] and HN(CO)CA[58,85] are one pair of experiments that can be used for sequence-specific resonance assignment (see **Figures 15 and 16**). The HNCA correlates the $N(i)H(i)$ unit with both $C_\alpha(i)$ and $C_\alpha(i-1)$, while the HN(CO)CA, by virtue of the transfer through $CO(i)$, provides only the interresidue $N_H(i) \rightarrow C_\alpha(i-1)$ correlation. In theory, these two experiments should suffice to elucidate sequential assignments for all H_N, N, and C_α nuclei. However, in practice, overlap and/or missing signals preclude this, and a second linkage between residues, involving an atom other than C_α, is required. One possibility is to use H_α; the HN(CA)HA[86] and HN(COCA)HA[80] experiments achieve this connection. Another method uses CO as the linking nucleus. The HNCO[58,87] is one of the most sensitive triple resonance experiments (see below), correlating the $H(i)N(i)$ unit with $CO(i-1)$, while the complementary experiment HN(CA)CO gives the intraresidue $H(i)N(i) \rightarrow CO(i)$ correlation (and often a weaker interresidue crosspeak; see **Figure 16**). Unfortunately, the latter experiment suffers from lower sensitivity because of a long residence time on C_α, which has a fast transverse relaxation rate, and hence it is not suitable for large proteins.

Thus, two of these sets of two 3D spectra can provide a pair of interresidue links; the same result can be arrived at by recording two complementary 4D experiments, which sample an extra chemical shift each during the magnetization transfer pathway. For example, the HNCAHA[88–91] and HN(CO)CAHA[88,91] combine the first four 3D triple resonance experiments described above, resulting in two connections between residues (H_α and C_α chemical shifts) in each experiment. An alternative and powerful method for providing two interresidue links is achieved using two 3D experiments, the CBCANH[90] (or the very similar, but more sensitive, HNCACB)[92] and the CBCA(CO)NH (see **Figures 15–17**).[93,94] These experiments connect $H(i)N(i)$ units of one residue with the C_α and C_β atoms of both the same and the preceding residues, in a manner similar to the HNCA/HN(CO)CA pair (i.e., up to four crosspeaks are seen at each combination of N/H_N frequencies). Thus, the HNCACB yields the frequencies of up to six nuclei from a single 3D data set ($H_N(i)$, $N(i)$, $C_\alpha(i)$, $C_\beta(i)$, $C_\alpha(i-1)$, and $C_\beta(i-1)$). Interpretation of the spectra is simplified, however, by the opposite signs of the

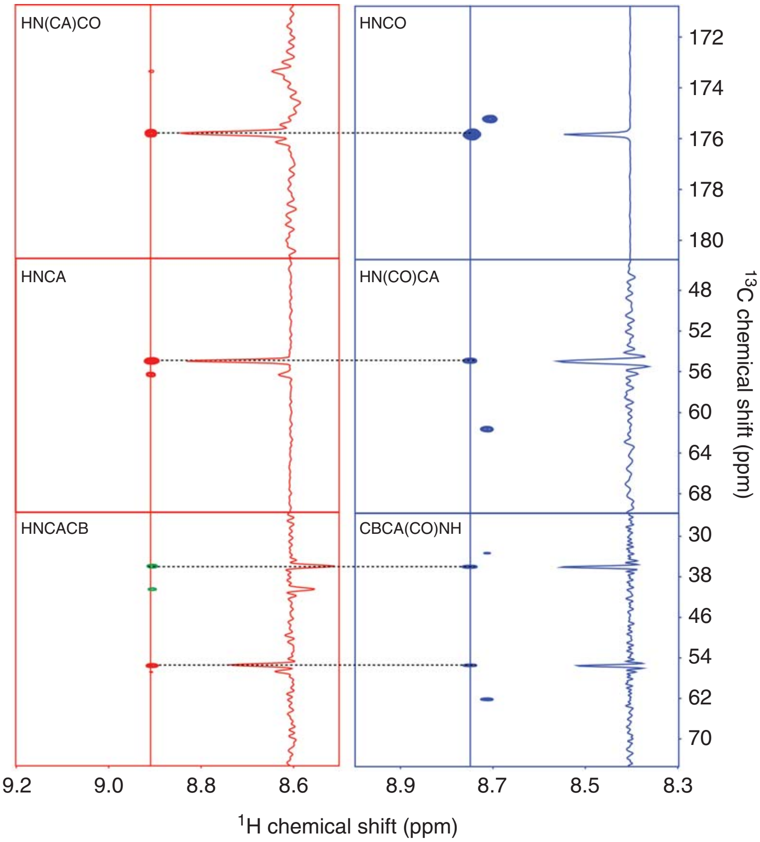


Figure 16 2D planes taken from pairs of 3D triple resonance NMR experiments designed to obtain sequence-specific resonance assignments for a 41-residue peptide. The 2D planes were extracted along the ^{15}N dimension. The planes in red show intraresidue correlations, while the spectra in blue show interresidue correlations. The vertical lines show the frequency positions at which a 1D trace was extracted along the ^{13}C dimension. These 1D traces are shown on the right-hand side of each 2D plane, and they provide an indication of the relative sensitivity of each of the experiments. Note that the HNCACB spectrum contains peaks of opposite sign for the C_α correlations (shown in red) and the C_β correlations (shown in green). The horizontal dotted lines highlight the correlations that are obtained in both spectra, which enable sequence-specific assignment (also see **Figure 17**).

C_α and C_β correlations. Two additional features add to the utility of these experiments. First, the chemical shifts of C_α and C_β enable facile identification of the spin systems of several residues (Ala, Thr, Ser, and also Gly due to the absence of a C_β correlation), providing entry points for sequence-specific assignment. Second, the measured C_α and C_β chemical shifts overlap with correlations obtained in the experiments used for obtaining side chain assignments (Section 9.09.4.4). A closely related experiment that correlates the amide unit of residue *i* with the side chain protons (rather than the carbons) of residue $i-1$ is the HBHA(CBCACO)NH, which forms an assignment pair with the HNHAHB.[93,94]

The 3D HCACO[59,83,95,96] and 3D HCA(CO)N[59,83,95,96] are also very useful experiments. The HCACO has high sensitivity (because the small $\mathcal{J}_{C\alpha N}$ couplings are avoided) and correlates three intraresidue atoms; it is therefore used widely to complement the other triple resonance experiments. The HCA(CO)N is one of the less sensitive experiments, but its particular benefit lies in its ability to provide correlations to amide nitrogens that are connected to broadened amide protons,[97] and/or assign prolines (which do not have a H_N atom). Such protons may occur in flexible regions (such as loops and the N- and C-termini) and may fail to provide crosspeaks with sufficient intensity in many of the other triple resonance experiments, which both begin with and detect H_N magnetization. Because both of these experiments detect H_α in the direct dimension, the spectra must be recorded in D_2O. This may give rise to a small isotope shift of the ^{13}C resonances (C′ and C_α), although

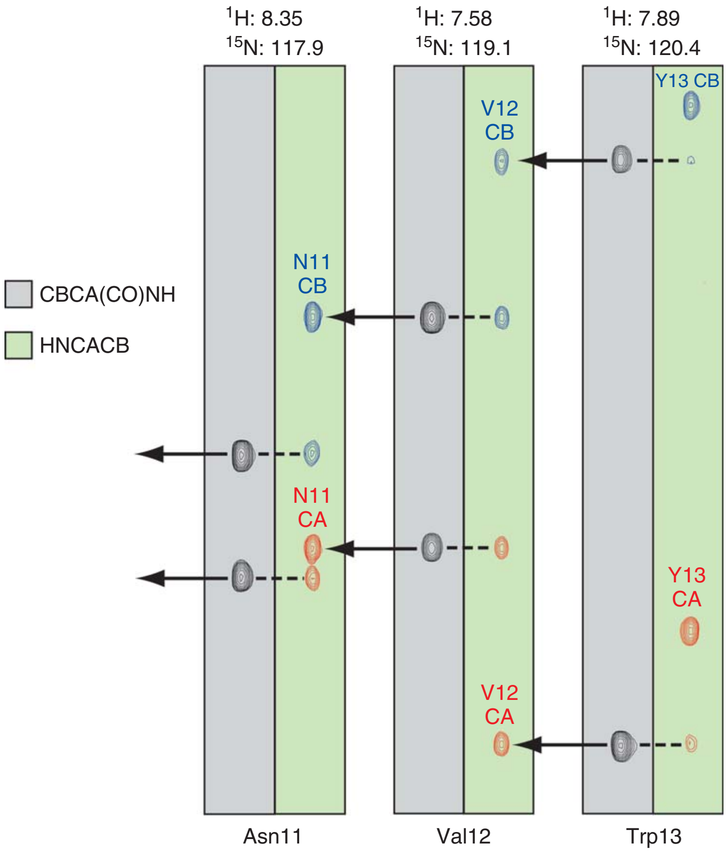


Figure 17 An example of the pattern matching process used to obtain sequence-specific resonance assignments from pairs of triple resonance NMR spectra. The spectra were acquired from a 41-residue peptidic spider toxin (0.5 mmol l^{-1}) at 900 MHz. Shown are pairs of strips from the CBCA(CO)NH (gray) and HNCACB (green) spectra taken at the same ^{1}H and ^{15}N frequencies. Note that the interresidue $H_N(i)N(i) \rightarrow C_\alpha C_\beta(i-1)$ correlations appear in both strips (as indicated by the horizontal dotted lines). These correlations need to be matched with intraresidue correlations in another pair of strips to provide a set of sequence-specific resonance assignments. For example, note how the two 'weak' interresidue correlations in the HNCACB strip for Val12 can be matched with the two 'strong' intraresidue correlations in the HNCAB strip for Asn11, indicating that these residues are adjacent.

such shifts are readily accounted for. The suppression of residual water can also saturate H_α protons that lie directly underneath the water resonance, thus preventing the observation of correlations for these residues.[98]

9.09.4.4 Triple Resonance Experiments for Protein Side Chain Assignment

There are two further experiments that together provide multiple interresidue connectivities; however, they are primarily used to complete the side chain assignment process. In the 3D H(CCO)NH–TOCSY[99,100] and H(C)NH–TOCSY,[99,101] magnetization begins on the side chain protons of residue $i-1$ and is first transferred to the attached carbon using INEPT. It is then propagated along the side chain via an isotropic (TOCSY-like) mixing sequence. In the former experiment, the magnetization that arrives at $C_\alpha(i-1)$ at the end of the mixing period is transferred to $CO(i-1)$ and then to $N(i)$ and finally $H_N(i)$ for detection. In the latter sequence, direct transfer from $C_\alpha(i-1)$ to $N(i-1)$ and $N(i)$ yields, as for the HNCA, both intra- and interresidue connectivities (although the latter are rather weak). During these experiments, the chemical shifts of the side chain protons $H_x(i-1)$, $N(i)$, and $H_N(i)$ are sampled (and also $N(i-1)$ and $H_N(i-1)$ in the H(C)NH–TOCSY). In principle, the information provided by the H(C)NH–TOCSY is available in a ^{15}N-edited TOCSY–HSQC spectrum.

In practice, this is often not the case since magnetization is transferred along the side chain via small and inefficient homonuclear couplings in the case of the TOCSY–HSQC compared with the large $^1\mathcal{J}_{CH}$ and $^1\mathcal{J}_{CC}$ couplings in the H(C)NH–TOCSY.

These experiments can be recorded in four dimensions with the side chain carbons comprising the fourth dimension (i.e., HCNH–TOCSY and HC(CO)NH–TOCSY),[99] in order to provide ^{13}C chemical shifts and a potential increase in resolution if ambiguities still remain. Note also that the H(CCO)NH–TOCSY is very similar in concept to the HBHA(CBCACO)NH described above. The main difference is that transfer from C_β to C_α in the latter case is through a COSY-type step. This excludes magnetization that may have originated on C_γ from being observed, in contrast to the isotropic mixing sequence used in the two TOCSY-type experiments.

3D HCCH experiments,[102–106] that is HC(C)H–COSY and HC(C)H–TOCSY, provide the ability to obtain assignments for side chain protons and carbons in larger proteins (up to ~30 kDa) where the 3D TOCSY–HSQC fails because of the large ^{1}H linewidths. In these experiments, proton magnetization is first transferred to carbon ($^1\mathcal{J}_{CH} \approx 125$–$150$ Hz) from where it is propagated along the carbon skeleton of the residue via the one-bond ^{13}C—^{13}C couplings ($^1\mathcal{J}_{CC} \approx 30$–$55$ Hz) using either HOHAHA- or COSY-type methods. Finally, ^{1}H magnetization is detected during the acquisition period following transfer of magnetization from ^{13}C to directly attached protons via the large $^1\mathcal{J}_{CH}$ coupling. In this way, indirect proton correlations (such as are normally found in TOCSY) are observed using only larger heteronuclear couplings (>30 Hz). Each crosspeak is characterized by the frequencies of two (directly or indirectly coupled) protons and the carbon to which the magnetization was first transferred. Consequently, there is much redundant information in these spectra, since the transfer will proceed in both directions. The HCCH–COSY works in an analogous manner, except that a single step $^{13}C \rightarrow {}^{13}C$ transfer is used, such that correlations are seen only between directly coupled protons (i.e., protons separated by ≤ 3 bonds).

The HCCH experiments avoid the pitfalls caused by proton linewidths being greater than $^3\mathcal{J}_{HH}$ couplings by transferring magnetization exclusively via large one-bond couplings. These experiments have proven extremely useful in the assignment of both proton and carbon side chain resonances in medium- to large-sized proteins. Note also that the HCCH experiments are best carried out in D_2O, in order to avoid the large water signal that could otherwise obscure a number of H_α resonances and their associated correlations.

Finally, a number of more specialized experiments have been proposed that deal with specific problems in the side chain assignment process. For example, although the aromatic protons of the Phe and Tyr residues have traditionally been assigned using NOE connectivities between H_β and H_δ protons, heteronuclear NMR experiments have been developed that allow these aromatic protons to be unambiguously assigned using scalar rather than dipolar correlations. For example, 2D (HB)CB(CGCD)HD and (HB)CB(CGCDCE)HE experiments can be used to correlate the $^{13}C_\beta$ chemical shift of Phe and Tyr residues with the ^{1}H chemical shift of the H_δ and H_ε protons, respectively, using magnetization transfer solely via scalar couplings.[107]

9.09.4.5 Summary of Resonance Assignment Strategies

With such an array of possible NMR methods for resonance assignment, it must be decided which is the most suitable approach for the protein of interest. Although peptides and many smaller proteins may be amenable to the homonuclear approach described in Section 9.09.3, there are a number of advantages in using a heteronuclear approach employing ^{15}N- or ^{15}N/^{13}C-labeled protein. As long as a suitable expression system is available, labeling is relatively straightforward and (at least for ^{15}N) not excessively expensive. For smaller proteins, ^{15}N labeling will significantly simplify the resonance assignment process by allowing the use of ^{15}N-edited NOESY–HSQC and TOCSY–HSQC experiments, and a number of medium-sized proteins have been assigned using this ^{15}N-directed strategy.[69,76,108,109] Note, however, that the ^{15}N-edited NOESY–HSQC only yields NOEs involving at least one amide proton, and NOEs between carbon-bound protons will have to be obtained from analysis of 2D NOESY spectra. However, the additional advantages of ^{15}N labeling, such as the ability to measure amide-proton exchange rates conveniently and probe backbone dynamics, make it an attractive strategy even for proteins smaller than 10 kDa.

For proteins smaller than ~10 kDa, the ^{15}N-only approach may prove adequate. If not, double labeling with both ^{15}N and ^{13}C will be necessary, and assignment will be most readily achieved using triple resonance experiments. A key feature of the triple resonance strategy is that nearly all resonance assignments can be made on the basis of scalar couplings. A further advantage is that they provide ^{15}N and ^{13}C chemical shifts; the former are used in amide exchange and backbone dynamics studies, while the latter contain information on secondary structure (see below) and together these shifts can be used to derive estimates of backbone ϕ and ψ dihedral angles (see Section 9.09.5.3). A decision on which of the armory of triple resonance experiments to use will be partly based on the relaxation properties (i.e., size) of the protein. This is essentially due to the rapid transverse relaxation of C_α and H_α and its steep dependence on molecular correlation time (see **Figure 7**). Experiments that involve long residence times on these nuclei (e.g., HNCACB and HN(CA)CO) will often be of limited use for proteins larger than ~20 kDa.

Although several discrete triple resonance strategies that in theory can yield complete assignments have been outlined above, in practice a whole battery of these experiments are often applied to a protein.[110] This is generally the result of problems that arise during the assignment process and which could not be predicted *ab initio*, such as the chemical shift coincidence of several atoms in a cluster with those from another cluster. Alternatively, some crosspeaks may be weak or absent in a given spectrum, for one of a number of reasons. For example, amide protons in less ordered regions may exchange rapidly with solvent protons; this will broaden their NMR signal, reducing the obtainable SNR for the corresponding crosspeak(s) in multidimensional spectra. Conversely, the carbon atoms of such mobile regions will be much sharper than those of the remainder of the protein (since they are effectively rotating independently of the bulk of the protein, and therefore exhibit longer T_2 values). The observation of weaker correlations in the presence of such narrow, intense ones can be exacerbated by artifacts such as t_1 noise associated with the stronger crosspeaks.[97] The decision as to which triple resonance experiments to use therefore remains to some degree empirical, and often needs to be determined separately for individual cases and for individual spectrometers. As a starting point however, the more sensitive 'out-and-back' experiments should be attempted first (e.g., HNCO, HNCA, HN(CO)CA, and CBCA(CO)NH).

9.09.5 Extraction of Structural Constraints

9.09.5.1 Overview

Once resonance assignment is complete, the next step in the structure determination process can be tackled, namely the extraction of structural restraints from the NMR data. These restraints can then be used as input to a computer algorithm that attempts to calculate a 3D structure of the protein (or generally a family of structures) that is consistent with these restraints. The most important source of conformational information is homonuclear ^{1}H–^{1}H NOEs, which are observed between protons that are spatially separated by ≤ 5.5 Å. Scalar coupling constants (especially $^3\mathcal{J}$-couplings) provide information about torsion (dihedral) angles that can be used to derive angle restraints for structure calculations. Hydrogen bonds can be inferred or observed directly using NMR (see Section 9.09.5.4) and distance restraints defining these hydrogen bonds can also be used in structure calculations. Finally, the chemical shifts of various nuclei have been shown in many cases to be reliable indicators of protein secondary structure. In the sections below, we consider the information content of each of these classes of NMR data. Residual dipolar couplings (RDCs) have been briefly mentioned in Section 9.09.3.3.1 and they are discussed in more detail in Chapter 9.07.

9.09.5.2 Interproton Distances

When two protons are close in space, they are said to be dipolar-coupled (as opposed to scalar-coupled, which is a through-bond coupling mechanism). The modulation of this dipolar coupling as a result of molecular tumbling allows relaxation of the protons, and this relaxation may be manifested as a NOE, which can be observed as a crosspeak between the two protons in a two- or higher dimension NOESY experiment (NOESY). For large molecules such as proteins, this relaxation occurs predominantly through coupling of the dipole modulation to a simultaneous mutual spin flipping of the protons, which is essentially a zero frequency process. As the protein gets larger, its tumbling (and hence the dipole modulation) occurs at a slower frequency (Equation (3)). The prevalence (spectral density) of low-frequency processes[35] therefore increases, and coupling to the zero frequency proton–proton cross-relaxation process becomes more efficient. As relaxation becomes more efficient, the rate of buildup of an NOE during the mixing period of a NOESY increases. The buildup rate (ρ) also depends on the strength of the dipolar interaction between the two protons, so that

$$\rho = \frac{1}{r^6}\tau_c \tag{4}$$

The $1/r^6$ dependence of ρ causes the buildup rate to fall off very rapidly with internuclear distance, with the result that NOEs are short-range interactions that are typically not observed between protons separated by more than ~5.5 Å (but see below). Nevertheless, this provides extremely valuable structural information since spatially proximal protons will yield a crosspeak in NOESY spectra regardless of how distal they are in the amino acid sequence. Since several thousand NOEs will be observed for even a protein of modest size, NOEs provide the most important structural restraints for structure calculations. But first they need to be assigned to specific proton pairs, quantified, and converted into distance information.

For small proteins ($\leq$5 kDa), complete or nearly complete assignment of NOE connectivities can usually be accomplished by visual inspection of 2D NOESY spectra (see Clore and Gronenborn,[111] and references therein). Many of these NOEs would have been identified as a matter of course if the homonuclear assignment procedure described in Section 9.09.3 was employed. Any ambiguities in NOE assignments may be resolved by an iterative back-calculation procedure,[112–114] whereby structure calculations are first carried out using only unambiguously assigned NOEs. The resulting preliminary structures are used to resolve multiple assignment possibilities for unassigned NOEs by excluding those possibilities that are grossly inconsistent with the calculated structures. The newly assigned NOEs are included in a second round of calculations, and the new, better defined structures used to assign NOEs that remained ambiguous after the first iteration. This procedure can be repeated until no further ambiguities are resolved.

For large proteins, severe resonance overlap precludes a simple 2D approach, and resolution enhancement by incorporation of an extra dimension (or two) into the experiment is required. This can potentially be achieved in a 3D homonuclear experiment (3D NOESY–HOHAHA or NOESY–NOESY), but far more preferable is editing of the NOESY according to the frequencies of attached heteronuclei using an isotopically labeled sample. Thus a range of experiments are available, depending on the labeling pattern present in the protein: 3D ^{15}N-edited NOESY, 3D ^{13}C-edited NOESY, 4D ^{13}C,^{13}C-edited NOESY, and 4D ^{13}C,^{15}N-edited NOESY. These experiments, in particular the 4D versions, provide dramatic increases in resolution (see, for example, Clore *et al.*[115]), despite restricted digital resolution, because so few crosspeaks appear in each plane. A disadvantage of the ^{15}N-edited experiments is that they only yield NOEs involving at least one amide proton. Thus for a high-resolution structure, ^{13}C-edited NOESY experiments are essential so that NOEs between pairs of carbon-bound protons can be detected.[116,117]

Once the identity of all or most of the observable NOEs is established, the proximity of each proton pair must be gauged. As noted above, the rate of buildup of an NOE is proportional to the distance between the two protons. However, because all other proton pairs in a molecule give rise to oscillating fields at similar frequencies (as a result of molecular tumbling), these pairs can contribute to the cross-relaxation of a proton. This phenomenon is termed spin diffusion (SD) since it results from a stepwise (diffusive) transfer of magnetization away from a given proton pair via other neighboring protons. The observable results of spin diffusion are (1) a change in the shape of the buildup curve for direct NOEs (see below) and (2) the appearance of crosspeaks between spatially distal protons. Obviously, the latter effect reduces the useful information content of an NOE experiment and SD should therefore be minimized as much as possible. Since SD is an indirect phenomenon, its buildup has an initial lag phase in comparison to direct NOEs, and thus the use of short mixing times allows its effects to be discarded to a first approximation. Note that since cross-relaxation is more efficient for larger values of τ_c, SD becomes more of a problem the larger a protein is, necessitating the use of shorter and shorter mixing times.

The buildup of direct NOEs is approximately linear at short mixing times. In this regime (the isolated spin-pair approximation (ISPA)),[118] it is assumed that the intensity of the observed crosspeak is directly proportional to r^{-6}. The proportionality can be estimated by measuring the intensities (I_{ref}) of NOEs between protons that are separated by a fixed, conformation-independent distance (d_{ref}), such as geminal methylene (1.7 Å) or orthoaromatic (2.45 Å) protons. Unknown distances, d_{ij}, can therefore be calculated as

$$d_{ij} = d_{\mathrm{ref}} \left(\frac{I_{\mathrm{ref}}}{I_{ij}} \right)^{1/6} \tag{5}$$

where I_{ij} is the intensity of the crosspeak of interest. These derived distances represent 'upper limits' for the interproton distance, since a number of mechanisms may operate to reduce the observed NOE intensity and lead to the estimation of an artificially longer distance. In addition, the use of a single reference distance can introduce systematic errors.[119] Consequently, a better way to derive less biased distance estimates is to use two different types of reference distances in combination.[119]

For structure calculations, error ranges need to be placed on the derived distances. The most conservative method is to simply assign an upper distance bound of 5–6 Å to all proton pairs that yield an NOE, irrespective of NOE intensity. This can be useful for rapidly determining the overall fold of a protein, but it clearly discards useful information. An alternative approach is to partition the distances into broad categories: for example, 1.8–2.8, 1.8–3.5, and 1.8–5.0 Å for strong, medium, and weak NOEs, respectively, where 1.8 Å is the van der Waals contact distance between two hydrogen atoms.[120] If stereospecific assignments are not available for pairs of methylene protons or Leu/Val methyl groups, a so-called pseudoatom is created midway between each pair in the structure calculations, and the upper distance bound is relaxed for those NOEs (by 1 Å for a methylene pair).[121] This categorization procedure is still rather conservative, but it is better to underinterpret than overinterpret the data. In any case, the most important factor in determining the final quality of an NMR structure is the total 'number' of distance restraints, not their 'precision'.[34,122]

Because ISPA is an approximation, care should be taken not to overinterpret the interproton distances derived from Equation (5). The isolated spin-pair approximation is really only applicable to backbone protons such as H_α and H_N for which the effective correlation time for modulation of the dipolar coupling is equivalent to, or very close to, the molecular correlation time τ_c. This will not be the case for protons found in more dynamic regions of the protein, such as flexible loops, or protons at the tip of long side chains such as those of Arg and Lys residues. A 'uniform averaging model' developed to account for this flexibility[123] shows that the relationship between NOE intensity and internuclear distance for protons in more dynamic regions of the protein is closer to r^{-4} than the r^{-6} in the ISPA model. Thus, structure calculation programs such as CYANA[114] use a more complicated NOE calibration model that is based on an r^{-6} dependence for NOEs involving only backbone protons and an r^{-4} dependence for backbone–side chain and side chain–side chain NOEs. Regardless of whether automatic or manual calibration is used, care should be taken to ensure that an appropriate calibration model is used for converting NOE intensities into interproton distances.

9.09.5.3 Backbone Dihedral Angles

The backbone conformation of a peptide or protein can be completely defined by specifying the value of the ϕ, ψ, and ω dihedral angles for each amino acid residue. Since peptide bonds invariably assume the *trans* conformation ($\omega = 180°$), except in rare instances, experimental determination of ϕ and ψ would obviously be extremely useful for defining a protein's 3D structure. NMR can be used to obtain estimates of the value of dihedral angles (although not with a high degree of precision) by taking advantage of the fact that the magnitude of three-bond coupling constants has a characteristic dependence on the dihedral angle θ between the two coupled atoms. This dependence is described by a Karplus equation[124] of the type

$$\mathcal{J}(\theta) = A\cos^2\theta - B\cos\theta + C \tag{6}$$

where the constants *A, B*, and *C* have been determined empirically for various types of dihedral angle.[125] The Karplus relationship has the form shown in **Figure 18**, which shows the dependence of the $^3\mathcal{J}(H_NH_\alpha)$ coupling constant on the protein backbone dihedral ϕ. Although there are potentially multiple solutions for ϕ at a given value of $^3\mathcal{J}(H_NH_\alpha)$, in practice ϕ in proteins is mostly restricted to the range $\phi = -30$ to $-180°$,[126] such that unique solutions are possible for many values of $^3\mathcal{J}(H_NH_\alpha)$. In particular, regular secondary structural elements

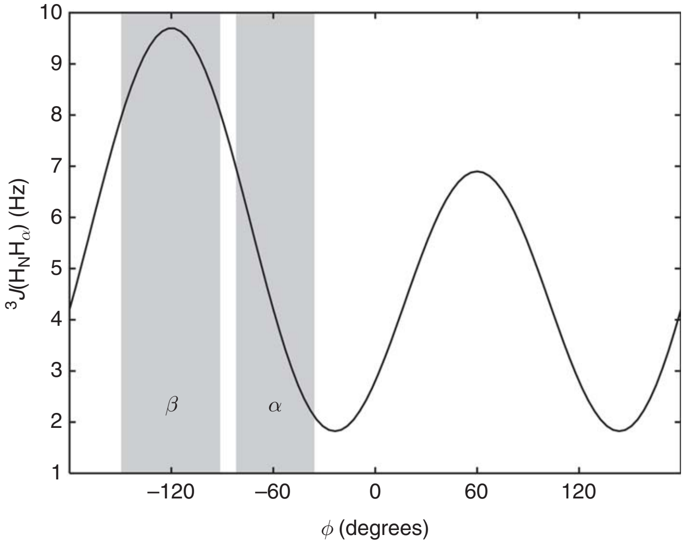


Figure 18 Plot of the coupling constant $^3J(H_NH_\alpha)$ as a function of the associated backbone dihedral angle ϕ (based on the Karplus parameterization described by Billeter *et al.*[130]). The approximate regions of ϕ space associated with α-helix and β-sheet are indicated and it can be seen that these two secondary structural elements give rise to distinct $^3J(H_NH_\alpha)$ values of <6 and >8 Hz, respectively.

have characteristic $^3\mathcal{J}(H_NH_\alpha)$ values of ~3–6 Hz for α-helices and >8 Hz for β-strands. Intermediate values between these two ranges may represent rigid structures with a well-defined angle, but more often indicate averaging of the torsion angle through internal motion (e.g., in flexible loops or flexible N- and C-terminal regions). Thus, although values of $^3\mathcal{J}(H_NH_\alpha)$ can often be measured for most or all residues in a peptide or protein, typically only those values that correspond to these secondary structure elements are converted to angle restraints for use in structure calculations. For example, it is common in structure calculations to restrain ϕ to -120 ± 30 and $-60 \pm 20°$ for $^3\mathcal{J}(H_NH_\alpha) > 8$ Hz and $^3\mathcal{J}(H_NH_\alpha) < 6$ Hz, respectively.[127] These dihedral angle restraints are extremely useful in structure calculations since a single dihedral angle restraint typically constrains the solution confirmation of a protein much more than a single NOE-derived interproton distance restraint.

The ϕ dihedral angle is associated with as many as six different coupling constants: $^3\mathcal{J}(H_NH_\alpha)$, $^3\mathcal{J}(H_NC_\beta)$, $^3\mathcal{J}(H_NC')$, $^3\mathcal{J}(C'_{i-1}H_\alpha)$, $^3\mathcal{J}(C'_{i-1}C_\beta)$, and $^3\mathcal{J}(C'_{i-1}C')$. However, $^3\mathcal{J}(H_NH_\alpha)$ is the most experimentally accessible coupling and the most widely used for estimation of ϕ, and hence we will focus our discussion on methods for estimating this coupling constant. For small unlabeled peptides, $^3\mathcal{J}(H_NH_\alpha)$ can often be measured directly from the separation of the two components of the amide-proton doublet in a high-resolution 1D spectrum (or from the three components of the amide-proton triplet in the case of glycine residues, where H_N is coupled to two H_α protons). Alternatively, $^3\mathcal{J}(H_NH_\alpha)$ can be measured from the magnitude of the F_2 antiphase splitting of the H_N–H_α crosspeaks in a high-resolution DQFCOSY spectrum.[128] However, this method is generally not suitable for large proteins,[129] since, irrespective of the coupling constant, the minimum separation of the antiphase components in a DQFCOSY crosspeak is 0.576 times the linewidth.[119] Thus, for proteins with linewidths ≥10 Hz, $^3\mathcal{J}(H_NH_\alpha)$ couplings smaller than ~5 Hz will be overestimated. It might still be possible to broadly group the couplings into those that are <6 Hz and those that are >8 Hz, but even these estimates become unreliable for proteins larger than ~15 kDa.

A number of heteronuclear NMR methods have been developed for measuring $^3\mathcal{J}(H_NH_\alpha)$ in proteins that can be isotopically labeled with ^{15}N. These methods utilize the large one-bond ^{15}N–^1H coupling constant, rather than relying on the measurement of antiphase splittings, and they allow measurement of $^3\mathcal{J}(H_NH_\alpha)$ in proteins as large as 20 kDa. The $\mathcal{J}$-modulated ^1H–^{15}N HSQC[129–131] consists essentially of a normal HSQC with an extra delay period τ_2 appended prior to signal acquisition. During this delay, the amide-proton magnetization, which has been labeled during τ_1 with the attached ^{15}N frequency, evolves according to its coupling to H_α.

Consequently, the intensity of the observed crosspeaks is modulated according to both ${}^3\mathcal{J}(\mathrm{H_N H_\alpha})$ and τ_2, according to the equation

$$\overline{V}(\tau_2) = A[\cos(\pi\mathcal{J}\tau_1)\cos(\pi\mathcal{J}\tau_2) - 0.5\sin(\pi\mathcal{J}\tau_1)\sin(\pi\mathcal{J}\tau_2)]\mathrm{e}^{-\tau_2/T'_2} \tag{7}$$

where $\overline{V}(\tau_2)$ is the crosspeak volume as a function of the delay time τ_2, A is the crosspeak volume at $\tau_2 = 0$, $\mathcal{J}$ is the ${}^3\mathcal{J}(\mathrm{H_N H_\alpha})$ coupling constant, and T_2' is the apparent ^{1}H transverse relaxation time. A number of spectra with incremented values of τ_2 are recorded, and the change in crosspeak intensity with τ_2 can be fitted with Equation (7). A modification of the basic pulse sequence for this experiment has also been proposed,[130] which is optimized for larger proteins where fast transverse relaxation is especially problematic. ${}^3\mathcal{J}(\mathrm{H_N H_\alpha})$ can also be estimated from the in-phase splitting of the HMQC-type crosspeaks in the ^{15}N dimension in a ^{1}H–^{15}N HMQC-$\mathcal{J}$ experiment.[132] Because the linewidths in this dimension are significantly narrower than in the homonuclear DQFCOSY, small splittings are resolvable even for relatively large proteins.[132]

The most popular heteronuclear method for measuring ${}^3\mathcal{J}(\mathrm{H_N H_\alpha})$ is the 3D HNHA experiment.[133] In comparison with the $\mathcal{J}$-modulated HSQC, the HNHA further alleviates spectral overlap by dispersing signals into a third frequency dimension that reports the chemical shift of $\mathrm{H_\alpha}$. In the 3D HNHA spectrum, ${}^1\mathrm{H_N}$–${}^1\mathrm{H_\alpha}$ correlations are observed as crosspeaks with opposite phase to the diagonal ${}^1\mathrm{H_N}$–${}^1\mathrm{H_N}$ peaks. The intensity ratio of the crosspeaks and diagonal peaks yields an accurate estimate of the ${}^3\mathcal{J}(\mathrm{H_N H_\alpha})$ coupling constant, namely,

$$I_{\mathrm{cross}}/I_{\mathrm{diagonal}} = -\tan^2\left({}^3\mathcal{J}(\mathrm{H_N H_\alpha})2\zeta\pi\right) \tag{8}$$

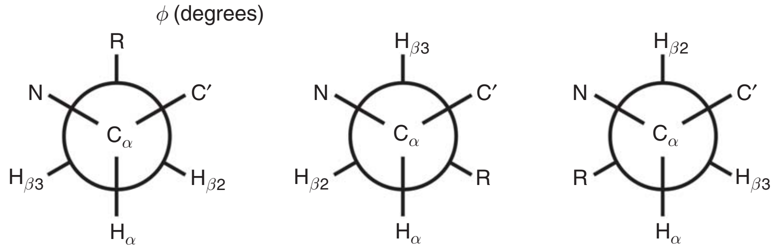
where 2ζ is the length of the transfer period in the pulse sequence. This method is very efficient for small proteins and peptides but begins to fail for proteins larger than 10 kDa due to the unfavorable relaxation properties of $\mathrm{H_\alpha}$.

The ψ backbone dihedral angle is much less experimentally accessible than ϕ since the coupling constants related to this angle are typically small (almost zero in the case of ${}^3\mathcal{J}(\mathrm{N}_i\mathrm{N}_{i+1})$). Thus, although a number of experiments such as the HCACO[N]–ECOSY[134] have been designed to measure values of ${}^3\mathcal{J}(\mathrm{H_\alpha N}_{i+1})$, they are rarely used. Instead, for isotopically labeled proteins, an alternative approach is available for estimation of both ϕ and ψ backbone dihedral angles that completely obviates the need to measure coupling constants. This approach relies on the fact that the secondary chemical shifts of the ${}^1\mathrm{H_\alpha}$, ${}^{13}\mathrm{C}'$, ${}^{13}\mathrm{C_\alpha}$, ${}^{13}\mathrm{C_\beta}$, and amide nitrogen nuclei are conformation dependent[135–137] and that these shifts have been determined as a matter of course during the sequence-specific assignment process. The TALOS program takes advantage of this relationship by taking chemical shift information for triplets of residues from the protein being studied and searching for the best match in a database of proteins for which both a high-resolution X-ray crystal structure and complete NMR chemical shift assignments are available.[138] On average, TALOS provides confident estimates of the ϕ and ψ dihedral angles for ~70% of residues in a protein, and less than 2% of these predictions are likely to be incorrect. Hence, the TALOS-derived angle estimates are essentially 'free' structural information and they should be used whenever possible. The PREDITOR web-server (http://wishart.biology.ualberta.ca/preditor) provides a similar function; however, it is also capable of providing estimates of side chain χ_1 and backbone ω angles in favorable circumstances, and its accuracy can be improved by reference to the structures of homologous proteins if they are available in the Protein Data Bank (PDB).[139]

9.09.5.4 Side chain Dihedral Angles

The side chain dihedral angle χ_1 is important for high-resolution definition of protein structures as it determines the angle at which each amino acid side chain branches out from the protein backbone. Moreover, in combination with certain types of NOEs, it can allow stereospecific assignment of prochiral β-methylene protons, which improves the precision of NMR structures by obviating the need to include pseudoatom distance corrections.[140–142]

χ_1 can be inferred from the measurement of ^{1}H-^{1}H couplings, and it generally corresponds to one of the three possible staggered rotamers, where each atom attached to $\mathrm{C_\beta}$ is either in one of the two *gauche* positions relative to $\mathrm{H_\alpha}$ or in the *trans* position (**Figure 19**). For unlabeled proteins, χ_1 determination relies on the measurement of ${}^3\mathcal{J}(\mathrm{H_\alpha H_\beta})$, whereas for ^{15}N-labeled proteins the measurement of ${}^3\mathcal{J}(\mathrm{NH_\beta})$ is often more useful.



Rotamer	g^2g^3	g^2t^3	t^2g^3
χ_1	60°	180°	–60°
$^3J(H_\alpha H_{\beta2})$	<5 Hz	<5 Hz	>10 Hz
$^3J(H_\alpha H_{\beta3})$	<5 Hz	>10 Hz	<5 Hz
$^3J(NH_{\beta2})$	~5 Hz	~1 Hz	~1 Hz
$^3J(NH_{\beta3})$	~1 Hz	~1 Hz	~5 Hz
NOE($H_\alpha H_{\beta2}$)	Strong	Strong	Weak
NOE($H_\alpha H_{\beta3}$)	Strong	Weak	Strong
NOE($H_N H_{\beta2}$)	Weak	Medium/strong	Strong
NOE($H_N H_{\beta3}$)	Medium/strong	Strong	Weak

Figure 19 Newman projections of the three possible staggered conformers about the χ_1 dihedral angle. The combination of $^3J(NH_\beta)$ and $^3J(H_\alpha H_\beta)$ coupling constants can be used to define χ_1 and obtain stereospecific assignment of the β-methylene protons. Alternatively, either the $^3J(NH_\beta)$ or $^3J(H_\alpha H_\beta)$ coupling constants can be used in combination with the intensities of the H_N–H_β and H_α–H_β NOEs to obtain this information.

For amino acids with a β-methine proton (Val, Thr, Ile), the magnitude of $^3\mathcal{J}(H_\alpha H_\beta)$ determines whether H_β is *trans* or *gauche*.[143] For amino acids with a β-methylene group, there are two relevant homonuclear couplings, $^3\mathcal{J}(H_\alpha H_{\beta2})$ and $^3\mathcal{J}(H_\alpha H_{\beta3})$. **Figure 19** shows their values for each of the three staggered rotamers. For amino acids with a single H_β proton, a DQFCOSY spectrum recorded in D_2O (to eliminate any multiplet structure arising from the $^3\mathcal{J}(H_N H_\alpha)$ coupling and to reduce loss of signals under the solvent peak) may be used to measure $^3\mathcal{J}(H_\alpha H_\beta)$. However, for amino acids with a β-methylene group, modified COSY experiments such as E.COSY[144] or P.E.COSY,[145] which simplify the multiplet structure of the H_α–H_β crosspeaks, are better suited to this task. These experiments rely on a splitting of the crosspeak of interest by a further large passive coupling; for the H_α-$H_{\beta2/\beta3}$ system, the splitting is by the geminal $^3\mathcal{J}(H_\alpha H_\beta)$ coupling ($\sim$14 Hz). As a consequence, couplings can be measured between signals with substantially broader lines than is possible in a simple DQFCOSY.[119] However, spectral overlap, exacerbated by broader lines, still thwarts these experiments for large proteins, and heteronuclear experiments are an attractive alternative.

A modified 3D HCCH–TOCSY experiment, where proton decoupling is not applied during the ^{13}C chemical shift evolution time, has been used for this purpose.[146] 2D planes showing ^{13}C and ^1H are taken through the indirect proton dimension, and the crosspeaks in the resulting ^1H-^{13}C correlation spectra have an E.COSY format. That is, the crosspeaks consist of two in-phase signals, separated in the ^{13}C dimension by the passive $^1\mathcal{J}(C_\alpha H_\alpha)$ coupling and in the ^1H dimension by $^3\mathcal{J}(H_\alpha H_\beta)$ coupling, which can be measured directly. Alternatively, Clore *et al.*[140] have shown that the size of the $^3\mathcal{J}(H_\alpha H_\beta)$ coupling for each β-proton is directly reflected in the intensity of $H_N \rightarrow H_\beta$ correlations in a 3D HOHAHA–HMQC experiment. That is, for the g^2t^3 and t^2g^3 conformations, only one $H_N \rightarrow H_\beta$ crosspeak is generally visible (or else one is much more intense), corresponding to the H_β with the large $^3\mathcal{J}(H_\alpha H_\beta)$ coupling. For g^2g^3, both correlations are absent, while disordered side chains sampling some or all of the possible rotamers display two crosspeaks of similar intensity.

For labeled proteins, the 3D HNHB experiment[147] has become the most popular approach for determining χ_1 and for providing stereospecific assignment of prochiral β-methylene protons. In contrast with the experiments described above, it allows measurement of the three-bond heteronuclear coupling, $^3\mathcal{J}(\mathrm{NH}_\beta)$, between the amide nitrogen and the H_β protons rather than the homonuclear $^3\mathcal{J}(\mathrm{H}_\alpha\mathrm{H}_\beta)$ coupling; however, as for $^3\mathcal{J}(\mathrm{H}_\alpha\mathrm{H}_\beta)$, the magnitude of this coupling is related to χ_1. The intensity of the H_β crosspeak in the HNHB spectrum is a reflection of the magnitude of $^3\mathcal{J}(\mathrm{NH}_\beta)$; however, the original version of this experiment required an additional 2D reference spectrum to be acquired for proper quantification.[147,148] Thus, it is more common to run a modified version of this experiment that allows the magnitude of $^3\mathcal{J}(\mathrm{NH}_\beta)$ to be extracted directly from the ratio of the diagonal and crosspeak intensities in a single 3D HNHB spectrum according to the equation[149]

$$I_{\mathrm{cross}}/I_{\mathrm{diag}} = -\tan^2\left({}^3\mathcal{J}\left(\mathrm{NH}_\beta\right)\pi T\right) \tag{9}$$

where T is a fixed delay in the pulse sequence. More often than not, however, the HNHB is analyzed qualitatively, much as described above for the 3D HOHAHA–HMQC experiment. This is possible because $^3\mathcal{J}(\mathrm{NH}_\beta)$ is only ~1 Hz or less when H_β is proximal to H_N, leading to a weak or unobservable crosspeak. Thus, for the g^2g^3 and t^2g^3 conformations, one often observes only a single $\mathrm{H}_\mathrm{N}\mathrm{H}_\beta$ crosspeak that corresponds to the H_β with the large $^3\mathcal{J}(\mathrm{NH}_\beta)$ *trans* coupling (~5 Hz). For the g^2t^3 rotamer, both crosspeaks are absent, while disordered side chains that sample some or all possible rotamers display two crosspeaks of similar intensity. Thus, for amino acids with prochiral β-methylene protons, χ_1 must be either 60 or –60° if only one H_β crosspeak is visible, whereas both crosspeaks will be absent if $\chi_1 = 180°$.

It is not possible to distinguish between the g^2g^3 and t^2g^3 rotamers on the basis of the $^3\mathcal{J}(\mathrm{NH}_\beta)$ coupling alone (unless the identities of the pro-R and pro-S β-methylene protons are already known). The $^3\mathcal{J}(\mathrm{H}_\alpha\mathrm{H}_\beta)$ coupling has a similar 'blindspot' with respect to discriminating between the g^2t^3 and t^2g^3 rotamers. As outlined in **Figure 19**, a *combination* of $^3\mathcal{J}(\mathrm{NH}_\beta)$ and $^3\mathcal{J}(\mathrm{H}_\alpha\mathrm{H}_\beta)$ would allow resolution of the three possible rotamers, but these two couplings are rarely both available. One usually measures $^3\mathcal{J}(\mathrm{H}_\alpha\mathrm{H}_\beta)$ from an ECOSY experiment if the protein is unlabeled, whereas $^3\mathcal{J}(\mathrm{H}_\alpha\mathrm{H}_\beta)$ is typically determined using an HNHB experiment if ^{15}N-labeled protein is available. Fortunately, by combining measurement of either the $^3\mathcal{J}(\mathrm{H}_\alpha\mathrm{H}_\beta)$ or $^3\mathcal{J}(\mathrm{NH}_\beta)$ coupling constant with knowledge of the relative intensities of the H_α-H_β and H_N-H_β crosspeaks in the NOESY spectra, it is possible to determine χ_1 and stereospecifically assign the β-methylene protons (see **Figure 19**).

The strategy for making stereospecific assignments is best explained with an example. Let us imagine a ^{15}N/^{13}C-labeled peptide containing a single Asn residue with magnetically inequivalent β-methylene protons at 2.72 and 2.83 ppm. The HNHB reveals an intense H_β crosspeak at 2.83 ppm ($^3\mathcal{J}(\mathrm{NH}_\beta) \sim 5$ Hz) but no crosspeak at 2.72 ppm ($^3\mathcal{J}(\mathrm{NH}_\beta) < 1$ Hz). Thus, χ_1 must be 60 or −60° but at this stage we cannot tell which. Now let us imagine that the ^{15}N-edited NOESY–HSQC spectrum reveals a very strong H_N–H_β crosspeak for the H_β proton at 2.72 ppm but only a weak H_N–H_β crosspeak for the H_β proton at 2.83 ppm. These NOE intensities are consistent with the assignment of χ_1 to 60 or −60°, but we still cannot distinguish between them. Finally, the ^{13}C-edited NOESY–HSQC spectrum reveals a weak H_α–H_β crosspeak for the H_β proton at 2.72 ppm but an intense H_α–H_β crosspeak for the H_β proton at 2.83 ppm. Thus, χ_1 must be −60°, as both H_β protons would yield intense H_α–H_β crosspeaks if χ_1 was 60° (compare the Newman projections for the g^2g^3 and t^2g^3 rotamers in **Figure 19**). Moreover, if $\chi_1 = -60°$, then the H_β proton at 2.72 ppm with a small $^3\mathcal{J}(\mathrm{NH}_\beta)$ coupling constant must be $\mathrm{H}_{\beta2}$, while the H_β proton at 2.83 ppm with a large $^3\mathcal{J}(\mathrm{NH}_\beta)$ coupling constant must be $\mathrm{H}_{\beta3}$. Using this approach, one can typically obtain stereospecific assignments (and associated χ_1 values) for 50% or more of the pairs of β-methylene protons in a peptide or small protein.

9.09.5.5 Hydrogen Bonds

Linus Pauling and Robert Corey surmised in the early 1950s (i.e., well before any protein structures had been experimentally determined) that the most stable protein folds would be those that maximized hydrogen bond formation, while still maintaining normal bond lengths and bond angles, and avoiding unfavorable steric overlap. They showed that there are only two polypeptide folds that adhere to this rule, and they christened them α-helices and β-sheets. The subsequent determination of over 56 000 protein structures using NMR spectroscopy and X-ray crystallography has confirmed that α-helices and β-sheets are indeed the major secondary structure elements in folded proteins. Thus, the experimental identification of hydrogen bonds can be extremely helpful both for defining elements of secondary structure in folded proteins and for use as conformational restraints in protein structure calculations (see Chapter 9.03).

Due to their small atomic mass, hydrogen atoms diffract X-rays poorly and hence they can only be resolved in crystal structures solved at extremely high resolution (<1.2 Å); thus, hydrogen bonds are 'inferred' in most protein crystal structures. Similarly, in NMR studies, hydrogen bonds have historically been inferred from the presence of 'slowly exchanging amide protons'. The rate of exchange of amide protons with solvent can be slowed by many orders of magnitude in folded proteins compared with unstructured peptides,[150,151] and this slow exchange is largely due to the existence of hydrogen bonds involving the amide proton, often in regular elements of secondary structure. Amide-proton exchange rates are typically measured by monitoring the change in intensity with time of amide-proton crosspeaks in a 2D spectrum following dissolution of the protein in 100% D_2O. Several types of 2D spectra can be used for this purpose, including TOSCY, COSY, or best of all, a ^{1}H–^{15}N HSQC spectrum if ^{15}N-labeled protein is available. The advantages of the HSQC spectrum are that dispersion is generally better and good signal-to-noise can be achieved much faster than with homonuclear 2D experiments. These experiments are usually analyzed qualitatively, with an amide proton being declared as 'slowly exchanging' if its corresponding crosspeak is still apparent in the spectrum after a certain period of time following dissolution of the protein in D_2O. However, this must be done with caution since, even in an unstructured peptide, there are intrinsic differences in the exchange rates for different types of amino acid residues.

A more quantitative approach involves acquisition of a series of spectra following dissolution of the protein in D_2O, after which a single exponential function can be fitted to the change in peak intensity with time in order to derive a pseudo-first-order rate constant for the exchange process. The rate constants for the exchange of each residue in an unstructured peptide[152–154] can then be divided by the observed rate constants for that residue in the protein under investigation to give a so-called protection factor. Large protection factors (>1000), when observed for amide protons that exhibit NOEs and coupling constants characteristic of regular secondary structure, can be used to generate restraints for structure calculations.

More recently, it has been realized that it is possible to use NMR to measure scalar couplings 'across' hydrogen bonds in both proteins[155–159] and nucleic acids.[160,161] (see Chapter 9.08). This has the dual advantage of providing direct proof for existence of the hydrogen bond while simultaneously revealing the identity of the donor and acceptor atoms. The $^{h3}\mathcal{J}_{NN}$ couplings in nucleic acids are relatively easy to access experimentally as they range from 2.5 to 11 Hz. (The notation $^{hn}\mathcal{J}_{AB}$ indicates a *trans* scalar coupling between nuclei A and B in which one of the *n* bonds is actually a hydrogen bond.)[161] However, the through-hydrogen-bond scalar couplings in proteins are much smaller and therefore more difficult to measure: the typical ranges for $^{h3}\mathcal{J}_{NC'}$ and $^{h2}\mathcal{J}_{HC'}$ in proteins are -0.2 to -0.9 and -0.6 to -1.3 Hz, respectively.[162,163]

For peptides and small proteins (<10 kDa) that can be isotopically labeled, $^{h3}\mathcal{J}_{NC'}$ couplings can be visualized using a 'long-range HNCO' experiment. This is essentially a conventional HNCO experiment (discussed in Section 9.09.4.3) in which the time for magnetization transfer from N → C′ via INEPT is substantially increased in order to favor transfer via the small three-bond $^{h3}\mathcal{J}_{NC'}$ coupling as opposed to the larger one-bond $^1\mathcal{J}_{N(i)C'(i-1)}$ coupling.[162–164] The experiment can be acquired as a 3D HNCO or, more commonly, as a 2D H(N)CO, in which the chemical shift of the ^{15}N nucleus is not recorded (see **Figure 20**). The large INEPT delays required for magnetization transfer via $^{h3}\mathcal{J}_{NC'}$ means that this experiment becomes very inefficient for larger proteins with short T_2 values. However, in these cases, a TROSY version of the experiment can be used with perdeuterated protein in which the amide deuterons have been converted to protons by exchange in H_2O buffer.[162–164] In favorable cases, this can extend the size range for this experiment to 30 kDa.[165] Note that, as shown in **Figure 20**, the long INEPT transfer time in the long-range HNCO also allows observation of intraresidue N → C′ correlations via $^2\mathcal{J}_{N(i)C'(i)}$, which is of similar magnitude to $^{h3}\mathcal{J}_{NC'}$ (i.e., ~ -1 Hz).

9.09.6 Calculation of Structures from NMR Data

9.09.6.1 Overview

The final step in protein structure determination using NMR is to use a computer program that combines the NMR-derived conformational restraints with additional restraints resulting from the covalent structure of the protein (i.e., bond lengths and bond angles) in order to calculate a 3D structure that is consistent with all of these

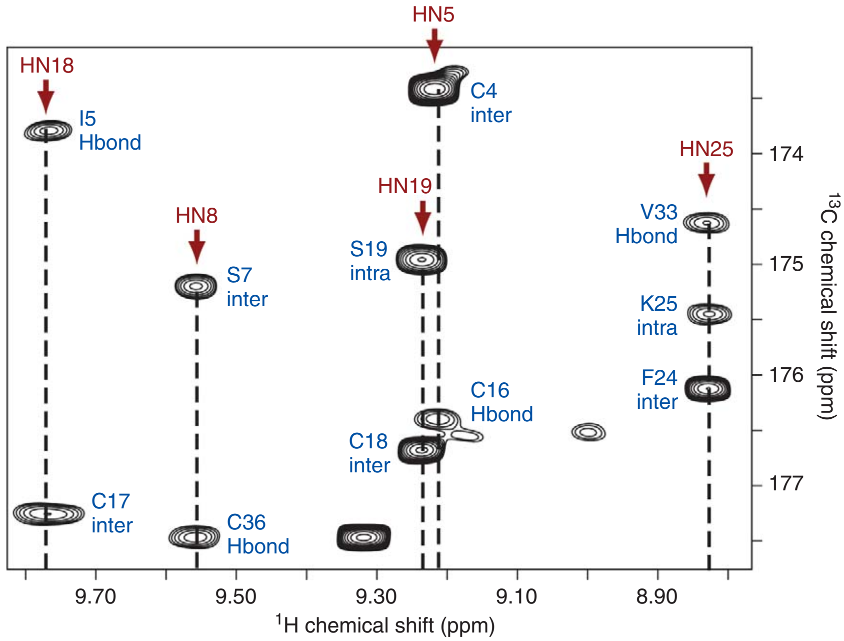


Figure 20 Long-range 2D H(N)CO experiment acquired at 600 MHz using a 1 mmol l^{-1} sample of the 37-residue spider toxin ω-atracotoxin-Hv1a.[127] The long INEPT transfer time allows observation of sequential interresidue correlations (inter) via $^1J_{N(i)C(i-1)}$ (~-15 Hz), intraresidue correlations (intra) via $^2J_{N(i)C'(i)}$ (~-1 Hz), and through-hydrogen-bond correlations (H-bond) via $^{3h}J_{NC'}$ (~-1 Hz). For example, the ^1H_N nucleus of Lys25 shows a correlation to its own C′ as well as the C′ of its neighbor Phe24 and its hydrogen bond partner Val33.

restraints. The primary experimental restraints are interproton distances derived from NOESY crosspeak intensities (Section 9.09.5.2), dihedral angle restraints derived from either *J* coupling constants or database searches based on chemical shift information (Sections 9.09.5.3 and 9.09.5.4), and hydrogen bond restraints based on either measurement of amide-proton exchange rates or a long-range HNCO experiment (Section 9.09.5.5). Residual dipolar couplings can also be used to provide orientational restraints (see Chapter 9.07) but these are rarely used for peptides and small proteins that are the focus of this chapter. The direct use of chemical shifts in structure calculations has not yet become 'mainstream' but nevertheless this is a promising area of investigation that we discuss in Section 9.09.6.3.3. We shall first briefly consider how to parameterize NMR-derived conformational restraints and then examine the various types of computational approaches that can be used to derive protein structures based on NMR data.

9.09.6.2 Parameterization of NMR-Derived Conformational Restraints

The primary experimental aim in any protein structure determination via NMR is to collect enough conformational restraints so that the 3D structure can be uniquely reconstructed using a computer algorithm. In trying to parameterize the NMR-derived information about interproton distances, dihedral angles, and hydrogen bonds, it is important to remember that it is the *quantity* rather than the *precision* of the restraints that is important.[34,122] Hence the parameterization should be conservative; overrestraining the distances and angle estimates is more likely to lead to errors than conservatively applied restraints. Our intention here is to provide a rough guide to such parameterization.

Hydrogen bonds are usually parameterized using a pair of distance restraints, one between the amide proton and its acceptor carbonyl oxygen (typically 1.8–2.0 Å), and the other between the amide nitrogen and the carbonyl oxygen (typically 2.7–3.0 Å).[166] However, hydrogen bonds in proteins can be considerably longer, as well as shorter,[167–169] than implied by these restraints and hence they are overly restrictive. Thus, we recommend setting hydrogen bond restraints of 1.7–2.2 and 2.7–3.2 Å for the H_N—O and N—O distances, respectively. (If one is using the program CYANA[114] to automatically generate hydrogen bond restraints, then these distances will need to be edited in the *H-bond.cya* macro.)

The parameterization of dihedral angle restraints will depend on the source of the angle estimates. As discussed in Section 9.09.5.3, it is common practice in homonuclear NMR studies to restrain the backbone ϕ dihedral angle to -120 ± 30 and $-60 \pm 20°$ for $^3\mathcal{J}(\mathrm{H_N H_\alpha}) > 8$ Hz and $^3\mathcal{J}(\mathrm{H_N H_\alpha}) < 6$ Hz, respectively.[127] However, more fine-grained angle estimates can be obtained in heteronuclear NMR studies when programs such as TALOS[138] are used to match chemical shifts against a database. In these cases, each estimate of the ϕ and ψ dihedral angles has an associated error, which is simply the standard deviation of the set of dihedral angles derived from the database matches. This error can range from as little as a few degrees to 50° or more. Our experience has shown that doubling the error estimates from TALOS (i.e., using an error of 2 rather than 1 standard deviation) produces reliable results and avoids over-restraining the structures.[2,170] In both homonuclear and heteronuclear NMR studies, it is usual to restrain the side chain χ_1 dihedral angle to ± 20 or $\pm 30°$ around the preferred rotamer determined from analysis of coupling constants and NOEs as outlined in Section 9.09.5.4.

Interproton distances are the dominant conformational restraints derived from NMR experiments, and hence their parameterization is important. As discussed in Section 9.09.5.2, it is still relatively common practice to 'manually' partition interproton distance restraints into broad categories such as 1.8–2.8, 1.8–3.5, and 1.8–5.0 Å for strong, medium, and weak NOEs. However, this approach is not recommended since it provides only a very coarse-grained set of restraints and it is both time consuming and somewhat arbitrary. Regardless of whether restraints are to be derived from 2D or 3D NOESY spectra, it is better to integrate the crosspeaks and use a structure calculation program such as CYANA[114] to automatically derive interproton distance estimates from these crosspeak intensities using a uniform-averaging-type model that uses an internal calibration and takes account of the type of NOE (i.e., backbone–backbone, backbone–side chain, or side chain–side chain). Adjustments to the derived parameterization can then be easily made if it is believed that the derived distances are too tight or too loose (although the latter is much less of a problem than the former).

Finally, it is important to add pseudoatom corrections[121] for pairs of methylene protons and Leu/Val methyl groups that have not been stereospecifically assigned. This can be done automatically by programs such as CYANA[114] during the process of converting NOE intensities into interproton distance restraints.

9.09.6.3 Structure Calculation Methods

Although the first protein structure determined using NMR was reported in 1985,[171] there is still no consensus method for deriving a 3D structure from NMR-derived conformational restraints. Indeed, a comprehensive overview of the variety of approaches and computer programs available for calculating protein structures using NMR-derived restraints is beyond the scope of this chapter. Rather, we will provide a brief overview of the two most commonly used structure calculation methods, torsion angle dynamics and simulated annealing.

9.09.6.3.1 Torsion angle dynamics

Torsion angle dynamics (TAD) programs, such as DISMAN[172] and its refined descendants DIANA,[173] DYANA,[174] and CYANA,[114,175] operate by minimizing a variable target function in torsion angle space. The programs begin with a random 3D structure generated on the basis of the known amino acid sequence of the protein and standard bond lengths and angles. The starting structure is then refined by varying the torsion (dihedral) angles in order to minimize a variable target function that includes terms for the various types of experimental restraints. For example, the part of the target function (T) dealing with violations of upper distance bounds in DIANA is[173]

$$T = w_u \sum \left[\Theta_u \left(\frac{d_{ij}^2 - u_{ij}^2}{2u_{ij}} \right) \right]^2 \tag{10}$$

where d_{ij} is the distance between atoms i and j in the current structure, u_{ij} is the upper bound on this distance, w_u is a weighting factor for upper-bound violations, and Θ_u is the Heaviside (step) function, which equals 0 if $d_{ij} \leq u_{ij}$ or 1 if $d_{ij} > u_{ij}$. The target function contains similar terms for experimentally derived lower distance bounds, a term for dihedral angle restraints, and a van der Waals' repulsion term that places a lower limit on interatomic distances in order to avoid unfavorable steric clashes; the latter term is a 'soft' model for the more computationally expensive repulsive term in the Lennard-Jones 6–12 potential (see Equation (11)).

The problem with trying to minimize the target function by introducing all restraints simultaneously is that the function will have many local minima. The variable target function approach was introduced by Braun and Gō[172] in an attempt to alleviate this problem. Instead of introducing all restraints simultaneously, one first optimizes using only local restraints (such as intraresidue and sequential restraints), and then introduces sequentially more long-range restraints until they have all been added into the calculation. This has the effect of optimizing the local conformation prior to determining the overall fold of the protein. Since torsion angles are the only independent variables in these calculations, TAD is less computationally intensive than the metric–matrix distance geometry approach[176] that was popular in the early days of protein NMR but which has now fallen out of favor.

9.09.6.3.2 Dynamical simulated annealing

Dynamical simulated annealing (DSA)[177] is a variant of restrained molecular dynamics (RMD).[178] There are numerous programs available for performing molecular dynamics (MD) simulations, including GROMOS,[178] AMBER,[179] CHARMM,[180] X-PLOR/CNS,[181] and OPLS.[182] In MD simulations, Newton's equations of motion are solved for all atoms under the influence of a physical force field (V_{physical}), which for a protein has the form[183]

$$\begin{aligned} V_{\text{physical}} = {} & \sum_{\text{bonds}} \frac{1}{2} K_b (b - b_0)^2 + \sum_{\text{angles}} \frac{1}{2} K_\theta (\theta - \theta_0)^2 \\ & + \sum_{\substack{\text{improper}\\ \text{dihedrals}}} \frac{1}{2} K_\zeta (\zeta - \zeta_0)^2 + \sum_{\text{dihedrals}} K_\varphi [1 + \cos(n\varphi - \delta)]^2 \\ & + \sum_{\text{pairs}(i,j)} C_{12}(i,f)/r_{if}^{12} - C_6(i,j)/r_{ij}^{6} + q_i q_j / 4\pi\varepsilon_0 \varepsilon_r r_{ij} \end{aligned} \tag{11}$$

where the K terms are force constants. The first term is a harmonic potential representing covalent bond stretching along bond b; the force constant K_b and minimum-energy bond length b_0 vary with the type of covalent bond. A similar term is used to describe bending of bond angles (θ). Two forms are used to describe distortions of dihedral angles: a harmonic term is used for dihedral angles ζ that are not allowed to make transitions (e.g., dihedral angles within aromatic rings), whereas a cosinusoidal term is used for dihedral angles φ that may make 360° turns. The final term is a sum over all pairs of nonbonded interatomic interactions: the first part sums the van der Waals interactions (a typical Lennard-Jones 6–12 potential) and the second part sums all electrostatic (Coulombic) interactions. There are numerous variants of this physical force field, including explicit inclusion of terms for hydrogen bonds.

The general RMD strategy for refining protein structures based on NMR-derived restraints is to add restraining potentials to the force field so that the structure can be refined against both the covalent geometry and nonbonded interactions (i.e., V_{physical}) as well as terms representing the experimentally derived distance ($V_{\text{distances}}$) and dihedral angle (V_{dihedral}) restraints, namely,

$$V_{\text{total}} = V_{\text{physical}} + V_{\text{distances}} + V_{\text{dihedral}} \tag{12}$$

A restrained molecular dynamics simulation using this expanded force field is performed using random or TAD-generated structures as the starting point for the simulation. The motion of the molecule is simulated for sufficient time to enable it to sample large regions of conformational space with a view toward converging on the structure with the global energy minimum or somewhere close to it by the end of the simulation. In the final stage of the RMD simulation, the structures are energy-minimized.

DSA is similar to high-temperature RMD except that the nonbonded van der Waals' and Coulombic terms in the force field (i.e., the last term in Equation (11)) are replaced with a simple quadratic van der Waals' repulsive term (V_{repel}) with repulsive force constant k_{repel}. The first stage of the simulation is performed at very high temperature (1000 K is typical) with k_{repel} set to a very low value so that atoms can move freely under the influence of the experimental terms, even being allowed to pass ‘through’, or very near to, each other. The value of k_{repel} is gradually incremented to reduce nonbonded contacts and then the temperature is lowered once k_{repel} reaches its maximum value. Finally, the structures are energy-minimized in a full RMD-type force field (i.e., the V_{repel} term is replaced by the full Lennard-Jones and Coulombic terms given in Equation (11)). The advantage of DSA over RMD is that the initial high temperature, combined with weak constraints on nonbonded interactions, enables the molecule to sample regions of conformational space that would be energetically inaccessible in classical room temperature RMD. Thus, the molecules are more likely to reach the global energy minimum corresponding to a structure with good covalent geometry, favorable nonbonded interactions, and minimal violations of the experimental constraints.

Of course, one could use randomly generated initial conformations to calculate structures using RMD or DSA. However, the disadvantage of this approach is that much computational time is wasted in calculating atomic trajectories for structures that are far removed from those that will ultimately satisfy the experimental constraints. The structures derived via the comparatively rapid TAD approach, on the other hand, usually satisfy the large majority of experimental restraints, thus minimizing the amount of computationally intensive RMD/DSA that is necessary to produce final refined structures. Hence, it is fairly common nowadays to generate initial structures via TAD and then to refine these using DSA. Note, however, that the computational time for the DSA approach is reduced relative to RMD due to the simplified force field, and consequently many groups have found it profitable to generate final conformations from random starting structures using only DSA.

9.09.6.3.3 Chemical shifts as structural restraints

The resonance frequencies, or chemical shifts, of various atoms are typically used only for assignment of individual atoms so that structural parameters specific to those atoms can be extracted from the NMR data. The chemical shift itself however is a measure of the local magnetic field experienced by the nucleus, which is dependent on

1. the electronic configuration of the atom itself, which is influenced by neighboring atoms (e.g., partial atomic charges, steric interactions, hydrogen bonding),
2. local magnetic fields due to anisotropic fields generated by nearby atoms (e.g., currents from aromatic rings and carbonyl groups), and
3. bulk solvent properties and direct intermolecular interactions.

Thus, the chemical shift itself is an incredibly rich and precise source of information. Indeed, these shifts can be thought of as a unique fingerprint of the molecule under the conditions of the measurement. The obvious question that arises is why a 3D structure cannot be derived from chemical shifts. In theory, there is no reason why this should not be possible; however, in practice, the *ab initio* models required for accurately calculating the above contributions in a dynamic and complex system such as a protein, in particular when solvent effects are considered, are beyond our current computational capabilities. All is not lost however, as the fact that proteins consist of a limited number of residue types that are connected through the same repetitive bonding structure allows for derivation of vastly simplified empirical models for approximating the abovementioned complex contributions.

The earliest models simply correlated the H_N and H_α chemical shift to hydrogen bonding and secondary structure of the protein.[136,137,184–189] For example, the H_α protons in α-helices display a marked upfield shift relative to random coil peptides,[34] while the opposite is true for residues in β-sheets. Subsequent studies correlated the secondary chemical shift of $^{13}C'$, $^{13}C_\alpha$, $^{13}C_\beta$, and amide nitrogen nuclei to various backbone torsion angles.[135,136,190–198] Recently, by coupling this approach with some form of empirical Monte Carlo structure calculation protocol (force field or *de novo* structure determination), several investigators have been able to determine protein structures from chemical shift data alone.[199–202] These approaches require only assignment of the backbone nuclei plus C_β in order to predict the 3D fold of the protein. Side chain assignments, which are generally more difficult to obtain, are not required. However, at present, these approaches cannot produce high-resolution protein structures that would be useful for applications such as structure-based drug design.

9.09.6.4 Assessing the Quality of Structures Derived from NMR Data

Since NMR-based structure calculation methods generate a family of structures that 'satisfy' the NMR-derived structural restraints, it has become common practice to assess the quality of structures by measuring the root mean squared deviation (RMSD) of individual structures from the mean structure. However, this is a measure of the precision of the structures rather than their accuracy. Moreover, the global RMSD is not a particularly good measure of precision as it does not discriminate between a structure that is poorly reproduced on average and one that is accurately reproduced except for a single ill-defined segment. Residue-by-residue, segment-by-segment, or domain-by-domain RMSD comparisons are often better indicators of precision, although they lack the visual impact of an overlay of an ensemble of structures based on minimization of the global RMSD.

Measurement of the accuracy of NMR-derived structures is a much more difficult task than estimating their precision. An absolute measure of the accuracy of an NMR-derived structure is not possible in the absence of any knowledge about the 'true' structure and therefore it has to be measured by some statistic.[203] One advantage of iterative relaxation matrix analysis (IRMA),[204,205] in which the structure is iteratively refined by comparison of the experimental NOESY spectrum with a synthetic spectrum back-calculated from the coordinates of the current structural model, is that it enables an NMR '*R* factor' to be calculated,[203,205] which is analogous to the *R* factor (or reliability index) used in crystallography. However, IRMA is not widely used for structure calculations and hence NMR *R* factors are rarely reported.

The most reliable indicator of the quality of an NMR-derived structure is its stereochemical merit as judged by programs such as PROCHECK-NMR,[206,207] WHAT IF,[208] and MolProbity.[209] PROCHECK reports numerous measures of stereochemical merit, including Ramachandran plot quality, deviations of bond lengths, bond angles, and dihedral angles from ideality, unfavorable side chain rotamers, and bad nonbonded interactions. An added bonus is that the program cleans up the coordinate files for submission to the PDB by ensuring that the atom labels conform to IUPAC-IUB nomenclature and by performing some basic stereochemical checks on the file. MolProbity, which can be accessed online at http://molprobity.biochem.duke.edu, additionally offers all-atom contact analysis and more detailed Ramachandran and side chain rotamer analysis. It also provides an overall 'MolProbity score' that allows the structure to be ranked on a percentile basis against other structures in the PDB. A MolProbity score that caused a structure to be ranked in the 25th percentile or lower would be a cause for concern, and it should provoke a detailed analysis of the MolProbity output, including any bad steric clashes, poor rotamer distributions, or less than 80% of residues in the most favored region of the Ramachandran plot.

A word of caution, however, is warranted when using these programs. In contrast with X-ray crystallography, where highly dynamic regions of the protein do not appear in the electron density maps and thus are omitted from the final coordinate file, all regions of the protein are modeled in NMR structure calculations. Highly dynamic regions of the protein, in which multiple conformations are accessed during the timescale of the NMR experiment, will have either a completely ill-defined conformation due to the lack of NOE information or an unrealistic one due to time- and population-weighted averaging of the NOEs and coupling constants. These regions of the protein are likely to have poor Ramachandran plot quality and bad side chain rotamer distributions, but these analyses are meaningless when applied to such mobile regions. Thus, if these regions are included in the PROCHECK or MolProbity analysis, they will reduce the overall stereochemical quality of the structural ensemble and may give a false indication of the quality of the well-structured region of the protein or peptide. Thus, these regions should be 'omitted' from the stereochemical analysis, just as they effectively are in the analysis of X-ray crystal structures.

9.09.7 Conclusions and Future Prospects

NMR is unrivaled in its ability to provide structural information on peptides and small proteins, as evidenced by the fact that it accounts for ~75% of the structures with mass less than 5 kDa deposited in the PDB. Moreover, it has the distinct advantage of also being able to provide information about protein dynamics and intermolecular interactions, as detailed elsewhere in this volume. Perhaps its major disadvantage relative to X-ray crystallographic approaches is that it is relatively slow and involves a great deal of user intervention.

Thus, much effort is currently being devoted to speeding up data acquisition and automating the process of spectral assignment and structure determination.[210]

There are numerous methods that have been developed for expediting data acquisition, including projection reconstruction,[211,212] multiway decomposition,[213] GFT NMR,[214] and nonuniform sampling (NUS).[63,215] We routinely use the last approach in our laboratory and have found that it can reduce the time for acquiring 3D triple resonance experiments such as the CBCA(CO)HN from a few days to a few hours. Although the raw data must be processed via maximum entropy reconstruction (MaxEnt)[63,216,217] rather than a conventional Fourier transform, it yields conventional spectra that are amenable to either classical manual analysis or automated assignment approaches; for example, the spectra shown in **Figure 16** were collected using NUS and processed using MaxEnt. A distinct advantage of these rapid data acquisition approaches is that they often allow data to be collected with higher digital resolution in the indirect dimensions, which facilitates automated spectral assignment. For example, using the NUS/MaxEnt approach, it takes less than 20 h to acquire a set of 2D HNCO, 3D CBCA(CO)HN, and 3D HNCACNB spectra that are of sufficient quality for the online PINE server (http://pine.nmrfam.wisc.edu)[218] to routinely achieves 100% sequence-specific backbone assignment.

Programs such as CYANA[114,175,219] and ARIA[220,221] also now provide the ability to automatically assign NOESY spectra and calculate structures, which dramatically improves the speed of the NMR structure determination process since, particularly for homonuclear NMR, much more time is usually spent analyzing the data than collecting the data. In many ways, the CANDID module[222] employed in CYANA mimics the iterative manual approach: a set of initial NOESY assignments are made based on various criteria (including the possibility of a crosspeak being assigned to multiple NOEs), a structure is calculated based on these assignments plus any dihedral angle and hydrogen bond information input to the program, the NOEs are adjusted based on the initial set of structures, and then the process is repeated through a cycle of seven iterations to produce the final ensemble of structures. In contrast with the manual approach, which can take weeks or even months, the automated process performed by CYANA takes $\sim$40 min on a laptop for a peptide of $\sim$5 kDa, or just a few minutes on even a modest server. We strongly recommend this approach if the speed of structure determination is a concern.

In conclusion, while NMR remains the dominant technique for determining the structures of peptides and small proteins, there are numerous developments that promise to dramatically improve the rate at which a protein structure can be determined using NMR. We predict that it will not be long before peptide/protein structure determination within 1 week becomes relatively routine.

Acknowledgments

Peptide studies reported from this laboratory were supported by Australian Research Council Discovery Grant DP0774245 to G.F.K. We thank Dr. Scott Robson for proofreading an earlier version of this chapter and for making numerous helpful suggestions.

Abbreviations

2D	two-dimensional
3D	three-dimensional
AUC	analytical ultracentrifugate
BMRB	Biological Magnetic Resonance Data Bank
CC	cryogenically cooled
COSY	correlated spectroscopy
DQFCOSY	double-quantum-filtered correlated spectroscopy
DSA	dynamical simulated annealing
DTT	dithiothreitol
HMBC	heteronuclear multiple bond correlation
HMQC	heteronuclear multiple quantum coherence
HOHAHA	homonuclear Hartmann–Hahn spectroscopy

HSQC	heteronuclear single quantum coherence
INEPT	insensitive nuclei enhanced by polarization transfer
IRMA	iterative relaxation matrix analysis
ISPA	isolated spin-pair approximation
MALLS	multiangle laser light scattering
MaxEnt	maximum entropy reconstruction
MD	molecular dynamics
NMR	nuclear magnetic resonance
NOE	nuclear Overhauser enhancement
NOESY	nuclear Overhauser enhancement spectroscopy
NUS	nonuniform sampling
PFG	pulsed-field gradient
PFGSE	pulsed-field-gradient spin-echo
RDC	residual dipolar coupling
RMD	restrained molecular dynamics
RMSD	root mean squared deviation
SNR	signal-to-noise ratio
TAD	torsion angle dynamics
TCEP	tris(2-carboxy-ethyl)phosphine
TOCSY	total correlation spectroscopy

Nomenclature

T_1	longitudinal relaxation time
T_2	transverse relaxation time
γ	gyromagnetic ratio
τ_c	molecular correlation time

References

1. G. J. Howlett; A. P. Minton; G. Rivas, *Curr. Opin. Chem. Biol.* **2006**, *10*, 430–436.
2. S. L. Rowland; W. F. Burkholder; K. A. Cunningham; M. W. Maciejewski; A. D. Grossman; G. F. King, *Mol. Cell* **2004**, *13*, 689–701.
3. A. J. Dingley; J. P. Mackay; B. E. Chapman; M. B. Morris; P. W. Kuchel; B. D. Hambly; G. F. King, *J. Biomol. NMR* **1995**, *6*, 321–328.
4. G. Otting, *J. Biomol. NMR* **2008**, *42*, 1–9.
5. W. U. Primrose, Sample Preparation. In *NMR of Macromolecules*; G. C. K. Roberts, Ed.; IRL Press: Oxford, 1993.
6. Y. Bai; J. S. Milne; L. Mayne; S. W. Englander, *Proteins Struct. Funct. Genet.* **1993**, *17*, 75–86.
7. A. E. Kelly; H. D. Ou; R. Withers; V. Dotsch, *J. Am. Chem. Soc.* **2002**, *124*, 12013–12019.
8. B. Pan; Z. Deng; D. Liu; S. Ghosh; G. P. Mullen, *Protein Sci.* **1997**, *6*, 1237–1247.
9. A. P. Golovanov; G. M. Hautbergue; S. A. Wilson; L. Y. Lian, *J. Am. Chem. Soc.* **2004**, *126*, 8933–8939.
10. G. M. Hautbergue; A. P. Golovanov, *J. Magn. Reson.* **2008**, *191*, 335–339.
11. W. R. Croasmun; R. M. K. Carlson, *Two-Dimensional NMR Spectroscopy: Applications for Chemists and Biochemists*, 2nd ed.; VCH Publishers: New York, 1994.
12. W. S. Price, *Annu. Rep. NMR Spectrosc.* **1999**, *38*, 289–354.
13. M. H. Levitt, *Concept Magn. Reson. A* **1996**, *8*, 77–103.
14. P. Luginbuhl; K. Wüthrich, *Prog. Nucl. Magn. Reson. Spectrosc.* **2002**, *40*, 199–247.
15. N. Bloembergen; R. V. Pound, *Phys. Rev.* **1954**, *95*, 8–12.
16. X. Mao; C. Ye, *Sci. China C Life Sci.* **1997**, *40*, 345–350.
17. V. Sklenář, *J. Magn. Reson. A* **1995**, *114*, 132–135.

18. D. Abergel; C. Carlotti; *J. Magn. Reson. B* **1995**, *109*, 218–222.
19. C. Anklin; M. Rindlisbacher; G. Otting; F. H. Laukien, *J. Magn. Reson. B* **1995**, *106*, 199–201.
20. S. Y. Huang; C. Anklin; J. D. Walls; Y. Y. Lin, *J. Am. Chem. Soc.* **2004**, *126*, 15936–15937.
21. D. Neuhaus; I. M. Ismail; C. W. A. Chung, *J. Magn. Reson. A* **1996**, *118*, 256–263.
22. P. J. Hore, *J. Magn. Reson.* **1983**, *55*, 283–300.
23. M. Piotto; V. Saudek; V. Sklenář, *J. Biomol. NMR* **1992**, *2*, 661–665.
24. S. Grzesiek; A. Bax, *J. Am. Chem. Soc.* **1993**, *115*, 12593–12594.
25. J. Cavanagh; W. J. Fairbrother; A. G. Palmer, III; N. J. M. Skelton, *Protein NMR Spectroscopy: Principles and Practice*; Academic Press: San Diego, CA, 1996.
26. H. Barkhuijsen; R. de Beer; W. M. M. J. Bovée; D. van Ormondt, *J. Magn. Reson.* **1985**, *61*, 465–481.
27. R. E. Hurd, *J. Magn. Reson.* **1990**, *87*, 422–428.
28. G. Wider; K. Wüthrich, *J. Magn. Reson. B* **1993**, *102*, 239–241.
29. D. Marion; M. Ikura; A. Bax, *J. Magn. Reson.* **1989**, *84*, 425–430.
30. V. Sklenář; M. Piotto; R. Leppik; V. Saudek, *J. Magn. Reson. A* **1993**, *102*, 241–245.
31. E. Prost; P. Sizun; M. Piotto; J.-M. Nuzillard, *J. Magn. Reson.* **2002**, *159*, 76–81.
32. A. J. Simpson; S. A. Brown, *J. Magn. Reson.* **2005**, *175*, 340–346.
33. B. D. Nguyen; X. Meng; K. J. Donovan; A. J. Shaka, *J. Magn. Reson.* **2007**, *184*, 263–274.
34. K. Wüthrich, *NMR of Proteins and Nucleic Acids*; John Wiley & Sons: New York, 1986.
35. R. R. Ernst; G. Bodenhausen; A. Wokaun, *Principles of Nuclear Magnetic Resonance in One and Two Dimensions*; Clarendon Press: Oxford, 1987.
36. U. Piantini; O. W. Sorensen; R. R. Ernst, *J. Am. Chem. Soc.* **1982**, *104*, 6800–6801.
37. J. Jeener; B. H. Meier; P. Bachmann; R. R. Ernst, *J. Chem. Phys.* **1979**, *71*, 4546–4553.
38. A. Kumar; R. R. Ernst; K. Wüthrich, *Biochem. Biophys. Res. Commun.* **1980**, *95*, 1–6.
39. L. Braunschweiler; R. R. Ernst, *J. Magn. Reson.* **1983**, *53*, 521–528.
40. D. G. Davis; A. Bax, *J. Am. Chem. Soc.* **1985**, *107*, 2820–2821.
41. E. R. P. Zuiderweg; S. R. Van Doren, *Trends Analyt. Chem.* **1994**, *13*, 24–36.
42. G. C. K. Roberts, *NMR of Macromolecules: A Practical Approach*; Oxford University Press: New York, 1993.
43. J. Cavanagh; W. J. Chazin; M. Rance, *J. Magn. Reson.* **1990**, *87*, 110–131.
44. X.-H. Wang; M. Connor; R. Smith; M. W. Maciejewski; M. E. H. Howden; G. M. Nicholson; M. J. Christie; G. F. King, *Nat. Struct. Biol.* **2000**, *7*, 505–513.
45. A. Bax; D. G. Davis, *J. Magn. Reson.* **1985**, *63*, 207–213.
46. S. W. Englander; A. J. Wand, *Biochemistry* **1987**, *26*, 5953–5958.
47. G. W. Vuister; R. Boelens; R. Kaptein, *J. Magn. Reson.* **1988**, *80*, 176–185.
48. S. S. Wijmenga; C. P. M. Mierlo, *Eur. J. Biochem.* **1991**, *195*, 807–822.
49. R. Boelens; G. W. Vuister; T. M. G. Koning; R. Kaptein, *J. Am. Chem. Soc.* **1989**, *111*, 8525–8526.
50. G. W. Vuister; R. Boelens; A. Padilla; G. J. Kleywegt; R. Kaptein, *Biochemistry* **1990**, *29*, 1829–1839.
51. S. Meier; D. Haussinger; E. Pokidysheva; H. P. Bachinger; S. Grzesiek, *FEBS Lett.* **2004**, *569*, 112–116.
52. G. Bodenhausen; D. J. Ruben, *Chem. Phys. Lett.* **1980**, *69*, 185–189.
53. G. A. Morris; R. Freeman, *J. Am. Chem. Soc.* **1979**, *101*, 760–762.
54. A. Bax; R. H. Griffey; B. L. Hawkins, *J. Am. Chem. Soc.* **1983**, *105*, 7188–7190.
55. T. J. Norwood; J. Boyd; J. E. Heritage; N. Soffe; I. D. Campbell, *J. Magn. Reson.* **1990**, *87*, 488–501.
56. A. Bax; M. Ikura; L. E. Kay; D. A. Torchia; R. Tschudin, *J. Magn. Reson.* **1990**, *86*, 304–318.
57. A. D. Bax; S. Grzesiek, *Acc. Chem. Res.* **1993**, *26*, 1–138.
58. S. Grzesiek; A. Bax, *J. Magn. Reson.* **1992**, *96*, 432–440.
59. R. Powers; A. M. Gronenborn; G. M. Clore; A. Bax, *J. Magn. Reson.* **1991**, *94*, 209–213.
60. L. E. Kay, *Curr. Opin. Struct. Biol.* **1995**, *5*, 674–681.
61. A. G. Palmer, III; J. Cavanagh; P. E. Wright; M. Rance, *J. Magn. Reson.* **1991**, *93*, 151–170.
62. J. Cavanagh; M. Rance, *J. Magn. Reson.* **1990**, *88*, 72–85.
63. M. Mobli; J. C. Hoch, *Concept Magn. Reson. A* **2008**, *32*, 436–448.
64. T. Luan; V. Jaravine; A. Yee; C. Arrowsmith; V. Orekhov, *J. Biomol. NMR* **2005**, *33*, 1–14.
65. K. Pervushin, *Q. Rev. Biophys.* **2000**, *33*, 161–197.
66. E. R. P. Zuiderweg; S. W. Fesik, *Biochemistry* **1989**, *28*, 2387–2391.
67. D. Marion; L. E. Kay; S. W. Sparks; D. A. Torchia; A. Bax, *J. Am. Chem. Soc.* **1989**, *111*, 1515–1517.
68. S. W. Fesik; E. R. P. Zuiderweg, *J. Magn. Reson.* **1988**, *78*, 588–593.
69. D. Marion; P. C. Driscoll; L. E. Kay; P. T. Wingfield; A. Bax; A. M. Gronenborn; G. M. Clore, *Biochemistry* **1989**, *28*, 6150–6156.
70. S. W. Fesik; E. R. P. Zuiderweg, *Q. Rev. Biophys.* **1990**, *23*, 97–131.
71. A. M. Gronenborn; A. Bax; P. T. Wingfield; G. M. Clore, *FEBS Lett.* **1989**, *243*, 93–98.
72. M. Ikura; L. E. Kay; R. Tschudin; A. Bax, *J. Magn. Reson.* **1990**, *86*, 204–209.
73. M. R. Bendall; D. T. Pegg; D. M. Doddrell; J. Field, *J. Am. Chem. Soc.* **1981**, *103*, 934–936.
74. R. Freeman; T. H. Mareci; G. A. Morris, *J. Magn. Reson.* **1981**, *42*, 341–345.
75. A. P. Golovanov; R. T. Blankley; J. M. Avis; W. Bermel, *J. Am. Chem. Soc.* **2007**, *129*, 6528–6535.
76. Y. Muto; K. Yamasaki; Y. Ito; S. Yajima; H. Masaki; T. Uozumi; M. Wälchli; S. Nishimura; T. Miyazawa; S. Yokoyama, *J. Biomol. NMR* **1993**, *3*, 165–184.
77. L. P. McIntosh; A. J. Wand; D. F. Lowry; A. G. Redfield; F. W. Dahlquist, *Biochemistry* **1990**, *29*, 6341–6362.
78. V. Kanelis; J. D. Forman-Kay; L. E. Kay, *IUBMB Life* **2001**, *52*, 291–302.
79. A. E. Derome, *Modern NMR Techniques for Chemistry Research*; Pergamon Press: Oxford, 1987.
80. R. T. Clubb; V. Thanabal; G. Wagner, *J. Magn. Reson.* **1992**, *97*, 213–217.
81. H. Kessler; S. Mronga; G. Gemmecker, *Magn. Reson. Chem.* **1991**, *29*, 527–557.
82. L. E. Kay; G. M. Clore; A. Bax; A. M. Gronenborn, *Science* **1990**, *249*, 411–414.

83. M. Ikura; L. E. Kay; A. Bax, *Biochemistry* **1990**, *29*, 4659–4667.
84. M. Sattler; J. Schleucher; C. Griesinger, *Prog. Nucl. Magn. Reson. Spectrosc.* **1999**, *34*, 93–158.
85. A. Bax; M. Ikura, *J. Biomol. NMR* **1991**, *1*, 99–104.
86. R. T. Clubb; V. Thanabal; G. Wagner, *J. Biomol. NMR* **1992**, *2*, 203–210.
87. L. E. Kay; M. Ikura; A. Bax, *J. Magn. Reson.* **1991**, *91*, 84–92.
88. L. E. Kay; M. Wittekind; M. A. McCoy; M. S. Friedrichs; L. Mueller, *J. Magn. Reson.* **1992**, *98*, 443–450.
89. W. Boucher; E. D. Laue; S. L. Campbell-Burk; P. J. Domaille, *J. Biomol. NMR* **1992**, *2*, 631–637.
90. S. Grzesiek; A. Bax, *J. Magn. Reson.* **1992**, *99*, 201–207.
91. E. T. Olejniczak; R. X. Xu; A. M. Petros; S. W. Fesik, *J. Magn. Reson.* **1992**, *100*, 444–450.
92. M. Wittekind; L. Mueller, *J. Magn. Reson. B* **1993**, *101*, 201–205.
93. S. Grzesiek; A. Bax, *J. Am. Chem. Soc.* **1992**, *114*, 6291–6293.
94. S. Grzesiek; A. Bax, *J. Biomol. NMR* **1993**, *3*, 185–204.
95. S. Grzesiek; A. Bax, *J. Magn. Reson. B* **1993**, *102*, 103–106.
96. A. Palmer; W. Fairbrother; J. Cavanagh; P. E. Wright; M. Rance, *J. Biomol. NMR* **1992**, *2*, 103–108.
97. S. Seip; J. Balbach; S. Behrens; H. Kessler; K. Flukiger; De R. Meyer; B. Erni, *Biochemistry* **1994**, *33*, 7174–7183.
98. M. L. Remerowski; T. Domke; A. Groenewegen; H. A. M. Pepermans; C. W. Hilbers; F. J. M. Ven, *J. Biomol. NMR* **1994**, *4*, 257–278.
99. T. M. Logan; E. T. Olejniczak; R. X. Xu; S. W. Fesik, *FEBS Lett.* **1992**, *314*, 413–418.
100. G. T. Montelione; B. A. Lyons; S. D. Emerson; M. Tashiro, *J. Am. Chem. Soc.* **1992**, *114*, 10974–10975.
101. B. A. Lyons; G. T. Montelione, *J. Magn. Reson. B* **1993**, *101*, 206–209.
102. S. W. Fesik; H. L. Eaton; E. T. Olejniczak; E. R. P. Zuiderweg; L. P. McIntosh; F. W. Dahlquist, *J. Am. Chem. Soc.* **1990**, *112*, 886–888.
103. A. Bax; G. M. Clore; A. M. Gronenborn, *J. Magn. Reson.* **1990**, *88*, 425–431.
104. L. E. Kay; M. Ikura; A. Bax, *J. Am. Chem. Soc.* **1990**, *112*, 888–889.
105. A. Majumdar; H. Wang; R. C. Morshauser; E. R. P. Zuiderweg, *J. Biomol. NMR* **1993**, *3*, 387–397.
106. E. T. Olejniczak; R. X. Xu; S. W. Fesik, *J. Biomol. NMR* **1992**, *2*, 655–659.
107. T. Yamazaki; J. D. Forman-Kay; L. E. Kay, *J. Am. Chem. Soc.* **1993**, *115*, 11054–11055.
108. B. J. Stockman; N. R. Nirmala; G. Wagner; T. J. Delcamp; M. T. DeYarman; J. H. Freisheim, *Biochemistry* **1992**, *31*, 218–229.
109. C. Redfield; L. J. Smith; J. Boyd; G. M. P. Lawrence; R. G. Edwards; R. A. G. Smith; C. M. Dobson, *Biochemistry* **1991**, *30*, 11029–11035.
110. J. Anglister; S. Grzesiek; H. Ren; C. B. Klee; A. Bax, *J. Biomol. NMR* **1993**, *3*, 121–126.
111. G. M. Clore; A. M. Gronenborn, *Crit. Rev. Biochem. Mol. Biol.* **1989**, *24*, 479–564.
112. C. Eccles; P. Güntert; M. Billeter; K. Wüthrich, *J. Biomol. NMR* **1991**, *1*, 111–130.
113. P. Guntert; K. D. Berndt; K. Wüthrich, *J. Biomol. NMR* **1993**, *3*, 601–606.
114. P. Güntert, *Methods Mol. Biol.* **2004**, *278*, 353–378.
115. G. M. Clore; L. E. Kay; A. Bax; A. M. Gronenborn, *Biochemistry* **1991**, *30*, 12–18.
116. P. J. Kraulis; P. J. Domaille; S. L. Campbell-Burk; T. Van Aken; E. D. Laue, *Biochemistry* **1994**, *33*, 3515–3531.
117. M. J. M. Burgering; R. Boelens; D. E. Gilbert; J. N. Breg; K. L. Knight; R. T. Sauer; R. Kaptein, *Biochemistry* **1994**, *33*, 15036–15045.
118. A. M. Gronenborn; G. M. Clore, *Prog. Nucl. Magn. Reson. Spectrosc.* **1985**, *17*, 1–32.
119. I. L. Barsukov; L.-Y. Lian, Structure Determination from NMR Data I. In *NMR of Macromolecules*; G. C. K. Roberts, Ed.; Oxford University Press: New York, 1993; pp 315–357.
120. I. D. Kuntz; J. F. Thomason; C. M. Oshiro, *Methods Enzymol.* **1989**, *177*, 159–204.
121. K. Wüthrich; M. Billeter; W. Braun, *J. Mol. Biol.* **1983**, *169*, 949–961.
122. G. M. Clore; M. A. Robien; A. M. Gronenborn, *J. Mol. Biol.* **1993**, *231*, 82–102.
123. W. Braun; C. Bösch; L. R. Brown; N. Gō; K. Wüthrich, *Biochim. Biophys. Acta* **1981**, *667*, 377–396.
124. M. Karplus, *J. Am. Chem. Soc.* **1963**, *85*, 2870–2871.
125. A. Pardi; M. Billeter; K. Wüthrich, *J. Mol. Biol.* **1984**, *180*, 741–751.
126. J. S. Richardson, *Adv. Protein Chem.* **1981**, *34*, 167–339.
127. J. I. Fletcher; R. Smith; S. I. O'Donoghue; M. Nilges; M. Connor; M. E. H. Howden; M. J. Christie; G. F. King, *Nat. Struct. Biol.* **1997**, *4*, 559–566.
128. D. Marion; K. Wüthrich, *Biochem. Biophys. Res. Commun.* **1983**, *113*, 967–974.
129. D. Neuhaus; G. Wagner; M. Vasak; J. Kägi; K. Wüthrich, *Eur. J. Biochem.* **1985**, *151*, 257–273.
130. M. Billeter; D. Neri; G. Otting; Y. Q. Qian; K. Wüthrich, *J. Biomol. NMR* **1992**, *2*, 257–274.
131. D. Neri; G. Otting; K. Wüthrich, *J. Am. Chem. Soc.* **1990**, *112*, 3663–3665.
132. L. E. Kay; A. Bax, *J. Magn. Reson.* **1990**, *86*, 110–126.
133. G. W. Vuister; A. Bax, *J. Am. Chem. Soc.* **1993**, *115*, 7772–7777.
134. A. C. Wang; A. Bax, *J. Am. Chem. Soc.* **1995**, *117*, 1810–1813.
135. S. Spera; A. Bax, *J. Am. Chem. Soc.* **1991**, *113*, 5490–5492.
136. D. S. Wishart; B. D. Sykes; F. M. Richards, *J. Mol. Biol.* **1991**, *222*, 311–333.
137. D. S. Wishart; B. D. Sykes, *J. Biomol. NMR* **1994**, *4*, 171–180.
138. G. Cornilescu; F. Delaglio; A. Bax, *J. Biomol. NMR* **1999**, *13*, 289–302.
139. M. V. Berjanskii; S. Neal; D. S. Wishart, *Nucleic Acids Res.* **2006**, *34*, 63–69.
140. G. M. Clore; A. Bax; A. M. Gronenborn, *J. Biomol. NMR* **1991**, *1*, 13–22.
141. J. D. Forman-Kay; G. M. Clore; P. T. Wingfield; A. M. Gronenborn, *Biochemistry* **1991**, *30*, 2685–2698.
142. H. J. Dyson; G. P. Gippert; D. A. Case; A. Holmgren; P. E. Wright, *Biochemistry* **1990**, *29*, 4129–4136.
143. A. Demarco; M. Llinas; K. Wüthrich, *Biopolymers* **1978**, *17*, 637–650.
144. C. Griesinger; O. W. Soerensen; R. R. Ernst, *J. Am. Chem. Soc.* **1985**, *107*, 6394–6396.
145. L. Mueller, *J. Magn. Reson.* **1987**, *72*, 191–196.

146. A. Bax; D. Max; D. Zax, *J. Am. Chem. Soc.* **1992**, *114*, 6923–6925.
147. S. J. Archer; M. Ikura; D. A. Torchia; A. Bax, *J. Magn. Reson.* **1991**, *95*, 636–641.
148. A. Bax; G. W. Vuister; S. Grzesiek; F. Delaglio; A. C. Wang; R. Tschudin; G. Zhu, *Methods Enzymol.* **1994**, *239*, 79–105.
149. P. Düx; B. Whitehead; R. Boelens; R. Kaptein; G. W. Vuister, *J. Biomol. NMR* **1997**, *10*, 301–306.
150. A. Hvidt; S. O. Nielsen, *Adv. Protein Chem.* **1966**, *21*, 287–386.
151. S. W. Englander; N. R. Kallenbach, *Q. Rev. Biophys.* **1983**, *16*, 521–655.
152. F. K. Junius; J. P. Mackay; W. A. Bubb; S. A. Jensen; A. S. Weiss; G. F. King, *Biochemistry* **1995**, *34*, 6164–6174.
153. S. Grzesiek; H. Doebeli; R. Gentz; G. Garotta; A. M. Labhardt; A. Bax, *Biochemistry* **1992**, *31*, 8180–8190.
154. R. S. Molday; S. W. Englander; R. G. Kallen, *Biochemistry* **1972**, *11*, 150–158.
155. P. R. Blake; B. Lee; M. F. Summers; M. W. Adams; J. B. Park; Z. H. Zhou; A. Bax, *J. Biomol. NMR* **1992**, *2*, 527–533.
156. P. R. Blake; J. B. Park; M. W. W. Adams; M. F. Summers, *J. Am. Chem. Soc.* **1992**, *114*, 4931–4933.
157. F. Cordier; S. Grzesiek, *J. Am. Chem. Soc.* **1999**, *121*, 1601–1602.
158. F. Cordier; M. Rogowski; S. Grzesiek; A. Bax, *J. Magn. Reson.* **1999**, *140*, 510–512.
159. G. Cornilescu; J.-S. Hu; A. Bax, *J. Am. Chem. Soc.* **1999**, *121*, 2949–2950.
160. A. J. Dingley; S. Grzesiek, *J. Am. Chem. Soc.* **1998**, *120*, 8293–8297.
161. K. Pervushin; A. Ono; C. Fernández; T. Szyperski; M. Kainosho; K. Wüthrich, *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 14147–14151.
162. S. Grzesiek; F. Cordier; A. J. Dingley, *Methods Enzymol.* **2001**, *338*, 111–133.
163. A. J. Dingley; F. Cordier; V. A. Jaravine; S. Grzesiek, Scalar Couplings Across Hydrogen Bonds. In *BioNMR in Drug Research*; O. Zerbe, Ed.; Wiley-VCH: Weinheim, 2003; pp 207–226.
164. F. Cordier; L. Nisius; A. J. Dingley; S. Grzesiek, *Nat. Protoc.* **2008**, *3*, 235–241.
165. Y. X. Wang; J. Jacob; F. Cordier; P. Wingfield; S. J. Stahl; S. Lee-Huang; D. Torchia; S. Grzesiek; A. Bax, *J. Biomol. NMR* **1999**, *14*, 181–184.
166. G. Wagner; W. Braun; T. F. Havel; T. Schaumann; N. Gõ; K. Wüthrich, *J. Mol. Biol.* **1987**, *196*, 611–639.
167. E. N. Baker; R. E. Hubbard, *Prog. Biophys. Mol. Biol.* **1984**, *44*, 97–179.
168. I. K. McDonald; J. M. Thornton, *J. Mol. Biol.* **1994**, *238*, 777–793.
169. R. S. Lipsitz; Y. Sharma; B. R. Brooks; N. Tjandra, *J. Am. Chem. Soc.* **2002**, *124*, 10621–10626.
170. V. Y. Gorbatyuk; N. J. Nosworthy; S. A. Robson; N. P. S. Bains; M. W. Maciejewski; C. G. dos Remedios; G. F. King, *Mol. Cell* **2006**, *24*, 511–522.
171. M. P. Williamson; T. F. Havel; K. Wüthrich, *J. Mol. Biol.* **1985**, *182*, 295–315.
172. W. Braun; N. Gō, *J. Mol. Biol.* **1985**, *186*, 611–626.
173. P. Güntert; W. Braun; K. Wüthrich, *J. Mol. Biol.* **1991**, *217*, 517–530.
174. P. Güntert; C. Mumenthaler; K. Wüthrich, *J. Mol. Biol.* **1997**, *273*, 283–298.
175. P. Güntert, *Prog. Nucl. Magn. Reson. Spectrosc.* **2003**, *43*, 105–125.
176. T. F. Havel; K. Wüthrich, *Bull. Math. Biol.* **1984**, *46*, 673–698.
177. M. Nilges; G. M. Clore; A. M. Gronenborn, *FEBS Lett.* **1988**, *229*, 317–324.
178. R. M. Scheek; W. F. van Gunsteren; R. Kaptein, *Methods Enzymol.* **1989**, *177*, 204–218.
179. W. D. Cornell; P. Cieplak; C. I. Bayly; I. R. Gould; K. M. J. Merz; D. M. Ferguson; D. C. Spellmeyer; T. Fox; J. W. Caldwell; P. A. Kollman, *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.
180. B. R. Brooks; R. E. Bruccoleri; B. D. Olafson; D. J. States; S. Swaminathan; M. Karplus, *J. Comput. Chem.* **1983**, *4*, 187–217.
181. A. T. Brunger, *Nat. Protoc.* **2007**, *2*, 2728–2733.
182. W. L. Jorgensen; J. Tirado-Rives, *J. Am. Chem. Soc.* **1988**, *110*, 1657–1666.
183. W. F. van Gunsteren; H. J. C. Berendsen, *Angew. Chem. Int. Ed. Engl.* **1990**, *29*, 992–1023.
184. A. Pardi; G. Wagner; K. Wüthrich, *Eur. J. Biochem.* **1983**, *137*, 445–454.
185. G. Wagner; A. Pardi; K. Wüthrich, *J. Am. Chem. Soc.* **1983**, *105*, 5948–5949.
186. A. Pastore; V. Saudek, *J. Magn. Reson.* **1990**, *90*, 165–176.
187. M. P. Williamson, *Biopolymers* **1990**, *29*, 1423–1431.
188. K. Ösapay; D. A. Case, *J. Am. Chem. Soc.* **1991**, *113*, 9436–9444.
189. L. Szilagyi, *Prog. Nucl. Magn. Reson. Spectrosc.* **1995**, *27*, 325–443.
190. I. Ando; H. Saito; R. Tabeta; A. Shoji; T. Ozaki, *Macromolecules* **1984**, *17*, 457–461.
191. H. Saito, *Magn. Reson. Chem.* **1986**, *24*, 835–852.
192. J. Glushka; M. Lee; S. Coffin; D. Cowburn, *J. Am. Chem. Soc.* **1989**, *111*, 7716–7722.
193. A. C. De Dios; J. G. Pearson; E. Oldfield, *Science* **1993**, *260*, 1491–1496.
194. H. Le; E. Oldfield, *J. Biomol. NMR* **1994**, *4*, 341–348.
195. M. Iwadate; T. Asakura; M. P. Williamson, *J. Biomol. NMR* **1999**, *13*, 199–211.
196. D. S. Wishart; D. A. Case, *Methods Enzymol.* **2001**, *338*, 3–34.
197. S. Neal; A. M. Nip; H. Zhang; D. S. Wishart, *J. Biomol. NMR* **2003**, *26*, 215–240.
198. Y. Wang; O. Jardetzky, *J. Biomol. NMR* **2004**, *28*, 327–340.
199. A. Cavalli; X. Salvatella; C. M. Dobson; M. Vendruscolo, *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 9615–9620.
200. Y. Shen; O. Lange; F. Delaglio; P. Rossi; J. M. Aramini; G. Liu; A. Eletsky; Y. Wu; K. K. Singarapu; A. Lemak; A. Ignatchenko; C. H. Arrowsmith; T. Szyperski; G. T. Montelione; D. Baker; A. Bax, *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 4685–4690.
201. D. S. Wishart; D. Arndt; M. Berjanskii; P. Tang; J. Zhou; G. Lin, *Nucleic Acids Res.* **2008**, *36*, W496–W502.
202. J. A. Vila; J. M. Aramini; P. Rossi; A. Kuzin; M. Su; J. Seetharaman; R. Xiao; L. Tong; G. T. Montelione; H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 14389–14394.
203. A. T. Brunger; G. M. Clore; A. M. Gronenborn; R. Saffrich; M. Nilges, *Science* **1993**, *261*, 328–331.
204. B. A. Borgias; T. L. James, *Methods Enzymol.* **1989**, *176*, 169–183.
205. P. D. Thomas; V. J. Basus; T. L. James, *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 1237–1241.
206. R. A. Laskowski; M. W. MacArthur; D. S. Moss; J. M. Thornton, *J. Appl. Crystallogr.* **1993**, *26*, 283–291.
207. R. A. Laskowski; J. A. C. Rullmann; M. W. MacArthur; R. Kaptein; J. M. Thornton, *J. Biomol. NMR* **1996**, *8*, 477–486.

208. G. Vriend, *J. Mol. Graph.* **1990**, *8*, 52–56.
209. I. W. Davis; A. Leaver-Fay; V. B. Chen; J. N. Block; G. J. Kapral; X. Wang; L. W. Murray; W. B. Arendall, III; J. Snoeyink; J. S. Richardson; D. C. Richardson, *Nucleic Acids Res.* **2007**, *35*, 375–383.
210. A. S. Altieri; R. A. Byrd, *Curr. Opin. Struct. Biol.* **2004**, *14*, 547–553.
211. E. Kupče; R. Freeman, *J. Am. Chem. Soc.* **2006**, *128*, 6020–6021.
212. S. Hiller; F. Fiorito; K. Wüthrich; G. Wider, *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 10876–10881.
213. D. Malmodin; M. Billeter, *J. Am. Chem. Soc.* **2005**, *127*, 13486–13487.
214. S. Kim; T. Szyperski, *J. Am. Chem. Soc.* **2003**, *125*, 1385–1393.
215. M. W. Maciejewski; A. S. Stern; G. F. King; J. C. Hoch, Nonuniform Sampling in Biomolecular NMR. In *Handbook of Modern Magnetic Resonance, Part II*; G. A. Webb, Ed.; Springer: Dordrecht, 2006; pp 1287–1293.
216. M. Mobli; M. W. Maciejewski; M. R. Gryk; J. C. Hoch, *J. Biomol. NMR* **2007**, *39*, 133–139.
217. M. Mobli; M. W. Maciejewski; M. R. Gryk; J. C. Hoch, *Nat. Methods* **2007**, *4*, 467–468.
218. A. Bahrami; A. Assadi; J. L. Markley; H. Eghbalnia, *PLoS Comp. Biol.* **2009**, e1000307.
219. P. Guntert, *Eur. Biophys. J.* **2009**, *38*, 129–143.
220. J. P. Linge; M. Habeck; W. Rieping; M. Nilges, *Bioinformatics* **2003**, *19*, 315–316.
221. M. Habeck; W. Rieping; J. P. Linge; M. Nilges, *Methods Mol. Biol.* **2004**, *278*, 379–402.
222. T. Herrmann; P. Güntert; K. Wüthrich, *J. Mol. Biol.* **2002**, *319*, 209–227.

Biographical Sketches

Glenn King graduated BSc and Ph.D. from the University of Sydney before undertaking postdoctoral studies with Professor Iain Campbell, FRS, at the University of Oxford. Glenn was a faculty member in the Department of Biochemistry at the University of Sydney from 1989 to 1998 before joining the University of Connecticut as Professor of Biochemistry and Microbiology in 1999. He returned to Australia in 2007 to take up a position as Professorial Research Fellow at the Institute for Molecular Bioscience at the University of Queensland. One of his major interests over the past 10 years has been the structure, function, and potential applications of peptide toxins expressed in spider venoms. Glenn recently founded Vestaron, an agricultural biotechnology company based in the United States, which aims to develop environmentally friendly insecticides based on natural insecticidal peptides. Glenn serves on the Scientific Advisory Boards of several companies and he is a Fellow of the American Academy of Microbiology.

Mehdi Mobli is an ARC Senior Research Associate at the Institute for Molecular Bioscience (IMB) at The University of Queensland, where he is currently managing an NMR structural genomics project. Mehdi received his undergraduate degree in Chemical Engineering from Chalmers University of Technology in Gothenburg, Sweden, and did his graduate work on calculation of chemical shifts in organic molecules in the laboratory of Professor Raymond Abraham at The University of Liverpool, UK. After a brief stint with Professor Jeffrey Hoch at the University of Connecticut, USA, working on methods for processing nonuniformly sampled multidimensional NMR data, Mehdi returned to the University of Manchester, UK, to work on the structure and dynamics of heparan sulfate derived from the capsular polysaccharides of pathogenic *Escherichia coli* strains. Mehdi is the coauthor (along with Professor Abraham) of the recently published monograph *Modelling ^{1}H NMR spectra of Organic Compounds*.

9.10 Mass Spectrometry: An Essential Tool for Trace Identification and Quantification

Charles H. Hocart, Australian National University, Canberra, ACT, Australia

9.10.1 Introduction

9.10.1.1 Overview

Mass spectrometry (MS) is an essential tool in the identification and quantification of natural products, primarily because of its speed, sensitivity, selectivity, and its versatility in analyzing solids, liquids, and gases. Indeed there are reports of viable viruses being collected after passage through a mass spectrometer.[1] MS has become an interdisciplinary methodology, impacting very many areas of science from physics, through chemistry, to biology.

The first mass spectrometer was constructed in the 1890s and was critical to the discovery of the electron by Sir Joseph John Thompson (winner of the Nobel Prize for Chemistry in 1906). Since then, MS has proven to be a technique of immense importance to scientific endeavors in a variety of fields, initially physics with the discovery of the electron and then stable isotopes and later, biology where it has been an essential tool for the high-throughput identification of proteins and their posttranslational modifications (PTMs). It is interesting to note that Thompson[2] observed in his book *Rays of Positive Electricity and Their Application to Chemical Analysis* that the new technique could be profitably used for chemical analysis. However, this potential was largely ignored until World War II when MS came to be used to monitor the cracking process in oil refineries and to separate ^{235}U and ^{238}U for use in the atomic bomb.

The last century has also seen considerable innovation and development of the technique and three further Nobel Prizes have been awarded for the discovery of isotopes of nonradioactive elements (1922, Francis Aston), development of new analyzers (1989, Wolfgang Paul – quadrupoles (Q's) and ion trap), and soft desorption ionization methods (2002, Koichi Tanaka and John B. Fenn – laser desorption ionization and electrospray ionization (ESI), respectively). MS continues to evolve and innovations in hardware and software are being driven by demands from medicine and biology for instruments with better mass accuracy, better mass resolution, increased dynamic range, faster data acquisition, and enhanced tandem MS capabilities. Samples of increasing complexity and diminishing size are being presented for analysis and entirely new fields of endeavor, such as proteomics and metabolomics, have been established, based on modern and continuing developments in MS. Those interested in the history of MS are referred to Grayson,[3] Griffiths,[4] and to Watson and Sparkman.[5]

Today, MS instruments are used in identifying and quantifying, for example, drugs, pollutants, products of chemical syntheses, planetary atmospheric components, biopolymers, and metabolites from microorganisms, plants, and animals. These analytes range in size from a few mass units (e.g., elemental gases) to hundreds of kilodaltons (kDa) (e.g., proteins and protein complexes) and cover a large range of polarities (e.g., hydrocarbons to sulfated carbohydrates). In addition to the wide applicability, another attractive feature of mass spectrometric analyses is that they can potentially be performed with a large degree of specificity and sensitivity (e.g., zeptomolar concentrations – 10^{-21} mol l^{-1}). Thus these instruments are used by a multitude of research disciplines and regulatory authorities (e.g., drug testing in sport,[6,7] Olympic Games,[8] space exploration,[9] geological dating,[10] biological tissue imaging,[11] wine industry,[12] metabolomics,[13,14] proteomics,[15–17]).

Although mass spectrometers are of widespread utility, it is also important to understand their limitations. Particular instruments are usually designed and dedicated to a narrow range of tasks dictated by their linkage to specific modes of sample presentation (e.g., solids probe, liquid chromatograph, gas chromatograph, or a proton transfer reaction drift tube) and methods of ionization (e.g., electrospray or electron impact). A well-equipped MS laboratory will therefore contain a variety of instruments with different capabilities.

9.10.1.2 Scope of the Present Work

MS is most commonly applied to problems of identification and quantification, particularly in the area of natural products chemistry. I hope in this brief chapter to give the nonspecialist chemist or biologist some basic background in MS and its capabilities so that they can sensibly engage with the MS specialist or MS literature in seeking solutions to their particular analytical problems. To this end, we will look specifically at the components of a mass spectrometer, the presentation of samples, the ionization processes available, and how the data generated from an analysis can be used for identification and quantification. Readers should also refer to complementary chapters in this volume on chromatography (chromatographically separated components of mixtures may be fed directly into the MS source for analysis) and proteomics (high-throughput technique for identification and quantification of large sets of proteins by MS (see Chapters 9.11–9.13).

In keeping with the philosophy of this series, only selective references to the literature have been made and wherever possible these have been review and tutorial style articles.

9.10.2 Components of a Mass Spectrometer

9.10.2.1 The Mass Spectrometer – Overview

The mass spectrometer may be divided into a number of discrete components: a sample inlet, an ion source, one or more analyzers, a detector, and finally, a computer to both collect data and control the operational parameters of the instrument (**Figure 1**). In principle, gas-phase neutral molecules are ionized so that they may be separated by the electric and/or magnetic fields of the analyzer according to their mass (m) to charge (z) ratios (m/z). The ions are then detected and recorded as a mass spectrum, graphing the ion abundance against the m/z ratio of the individual ions (**Figure 2**).

To enhance the passage of the ion stream, the ion source, analyzer region, and detector are held under vacuum. At atmospheric pressure (760 torr), there is a density of some 10^{19} molecules ml^{-1}, yielding a mean free path of 10^{-4} cm. However, in an evacuated region at 10^{-6} torr, the density drops to 10^{10} molecules ml^{-1} and the mean free path is extended out to 10^{3} cm, increasing the probability that an ion will be able to physically traverse the instrument without collision with a residual gas molecule. This requirement for a maximal mean free path is particularly important for the beam-type instruments (magnetic sectors and multiple analyzer instruments) and for ion cyclotron resonance (ICR) cells. In the latter, ions may literally travel many kilometers over an observation period of 1 s.[18] In some respects, the linear and Q ion traps are the exception, in that although the analyzers are held within a vacuum system, the traps themselves contain a helium buffer gas, which is required to collisionally cool the trapped ions.

9.10.2.2 Ion Source and Ionization Methods

Prior to analysis, the sample must be volatilized and ionized.[19] These processes can be separate or linked, depending on the nature of the sample and the ionization process being used. Samples may be presented for MS analysis in solid, liquid, or gaseous form and, furthermore, they may be a mixture of components. In the case of mixtures, separation is usually necessary for unambiguous identification or quantification because the

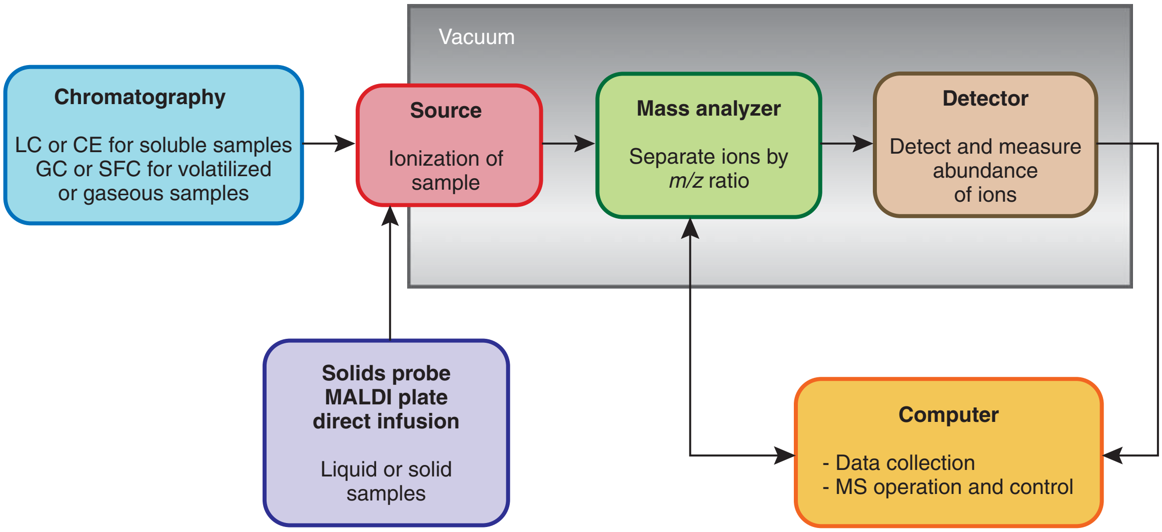


Figure 1 Principle components of a mass spectrometer. For mass spectrometry (MS) analysis of a sample, the neutral analyte(s) must first be ionized, positively or negatively, to allow manipulation by the magnetic and/or electric fields in the MS analyzer. Ions are sorted according to their mass to charge ratio (*m/z*), which is then plotted against their intensity to generate a mass spectrum. The flight path of the ions is evacuated to maximize the mean free path of the ions and to reduce the possibility of unfavorable interactions with residual air molecules.

simultaneous presence of two or more components in the source region will result in an overlapping or mixed spectrum. Mixtures are therefore often separated by gas chromatography (GC) or capillary electrophoresis (CE) or supercritical fluid chromatography (SFC) or liquid chromatography (LC), with the eluted and separated components being supplied directly into the MS source.[20] These hyphenated approaches are known as GC/MS, CE/MS, SFC/MS and LC/MS, respectively.

If chromatography is not required, samples may be introduced directly into the ion source. In the case of gaseous samples, volatilization is of course unnecessary, and the sample can be introduced into the source using appropriate gas handling techniques. Nonpolar, thermally stable, low-molecular-weight solids and liquids can be placed in metal or glass crucibles (solids probe or direct insertion probe) or may be directly applied to a wire loop (direct exposure probe). The crucible or wire loop is then heated to thermally desorb or volatilize the sample. Some polar low-molecular-weight compounds may also be directly analyzed after being chemically derivatized[21–27] to mask the polar functional groups and thereby increase volatility (e.g., by alkylation or silylation) and improve thermal stability (Section 9.10.4.3.2). Otherwise, samples may be dissolved in an appropriate solvent and subject to either an atmospheric pressure ionization process or be laser desorbed from a solid matrix as described below.

9.10.2.2.1 Electron ionization

Electron ionization (EI), originally developed by Dempster,[28] is widely used in MS for relatively volatile samples that are thermally stable and have relatively low molecular weight. Samples are typically presented in the effluent from a GC or are volatilized from a solids probe inserted into the high vacuum source. Ionization is effected by interaction between the gas-phase analyte molecules and a stream of high-energy electrons (typically 70 eV) drawn from a filament. Ionization occurs by removal of an electron to form an odd-electron ion, $M^{+\cdot}$ (Equation (1)). EI generally creates a singly charged positive ion, and any doubly or triply charged ions are of very low abundance. EI is also a high-energy process and excess energy remaining after ionization can be dissipated by fragmentation (possibly with rearrangement) of covalent bonds in the molecular ion, to lose either a radical (e.g., $CH_3^{\cdot}$) (Equation (2)) or a neutral species (e.g., H_2O or CH_3OH) (Equation (3)).

$$\text{Ionization}: \quad M + e^{-} \rightarrow M^{+\cdot} + 2e^{-} \qquad (1)$$

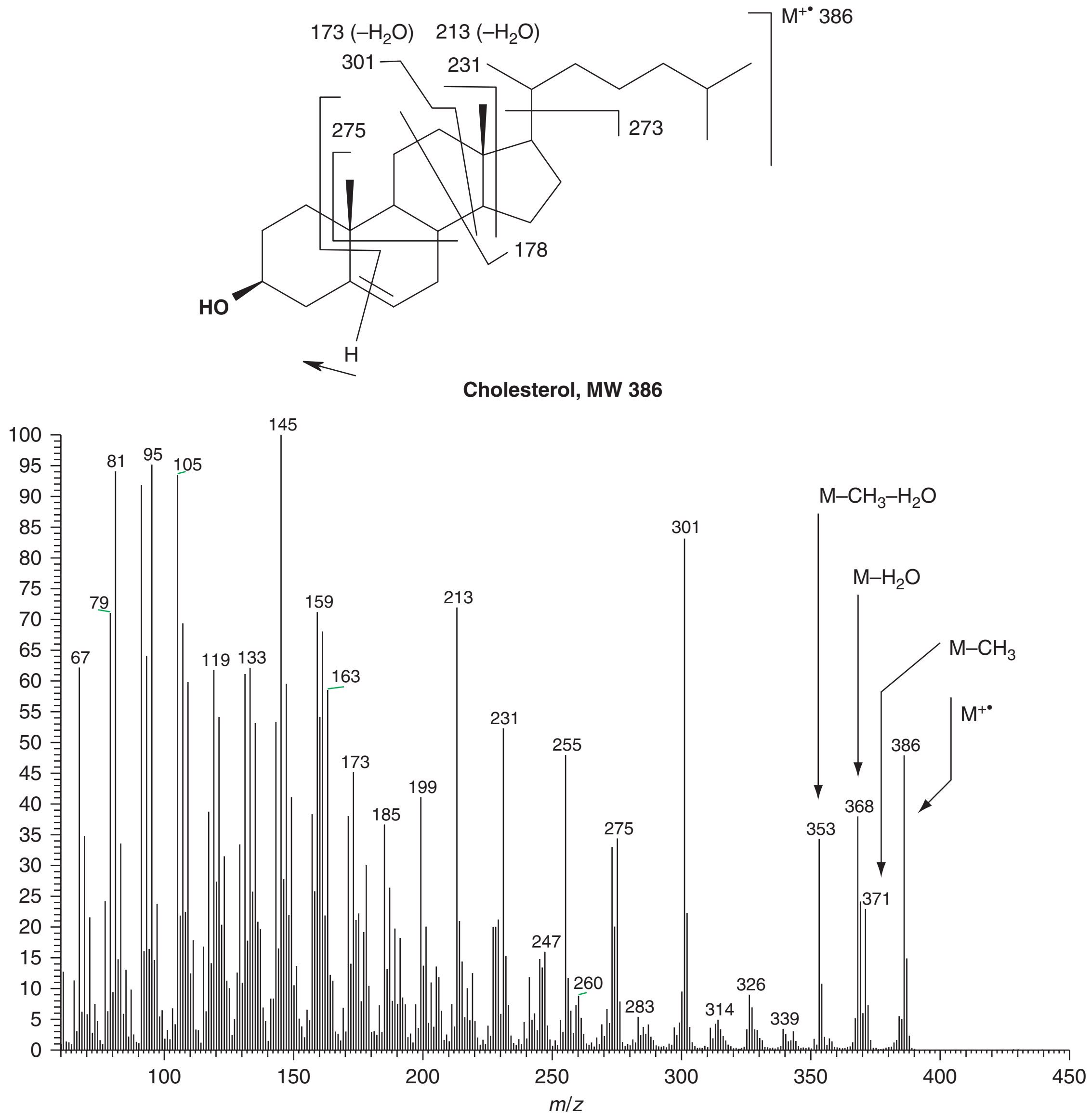


Figure 2 Mass spectrum of cholesterol generated by electron ionization (EI). The EI mass spectrum of cholesterol is characterized by the presence of a molecular ion at *m/z* 386 and by extensive fragmentation and contains information on the steroid nucleus and side chain. There is no indication as to the position of the double bond in this spectrum but the 3-hydroxy-Δ^5 structure can be identified after conversion to an ester.[29,30]

$$\text{Fragmentation}: \quad M^{+\cdot} \rightarrow [M-R]^{+} + R^{\cdot} \quad \text{radical loss} \tag{2}$$

$$M^{+\cdot} \rightarrow [M-R]^{+\cdot} + R \quad \text{neutral loss} \tag{3}$$

The fragmentation observed during EI is defined by the chemical structure of the analyte and the resulting highly reproducible pattern of fragmentation may be used for structural elucidation and identification of unknowns[5,31–34] (Figures 2 and 3(a)). This reproducibility has been exploited to develop user-generated and commercial libraries of spectra (some containing several hundred thousand spectra), which can be rapidly searched for comparable spectra. For some compounds, fragmentation may be so extensive that the molecular ion does not appear in the EI spectrum (e.g., Figure 3(a)). If this molecular mass information is required, then the analyst will have to resort to one of the softer or less energetic ionization processes such as chemical ionization (CI, **Figures 3(b)**, **3(c)**, and **3(e)**) or ESI (**Figure 3(g)**) as outlined below.

The displaced electron is generally assumed to be the electron with the lowest ionization energy. In order of probability, this will be a nonbonding electron followed by a π bond electron and then a σ bond electron. Thus EI yields, in the first instance, a molecular ion which is a radical cation with an unpaired electron. In principle, any remaining energy will then be dissipated by bond cleavages that result in the formation of the most stable cation with a paired electron (even-electron ion). These even-electron ions may be formed by homolytic or heterolytic cleavages. This whole process happens very rapidly ($<10^{-8}$ s) and is the reason for the close similarity of EI spectra produced across all different instruments. It is important to remember that mass spectral reactions in the EI source are unimolecular. This is because the pressure in the EI source is too low for bimolecular (ion–molecule) reactions to occur.

9.10.2.2.2 Chemical ionization (positive and negative) and electron capture ionization

Like EI, CI is also typically applied to samples presented via a GC interface or volatilized from a solids probe. It is a less energetic or soft form of ionization and is designed to minimize fragmentation.[35–38] CI is usually carried out in a source similar to that used for EI except that a reagent gas, commonly methane, isobutane, or ammonia, is added at a pressure of 0.3–1 torr. The electron beam then interacts with the reagent gas to produce reagent ions (**Table 1**) and thermal electrons. The neutral analyte molecules are then ionized by ion–molecule reactions to produce positive and negative analyte ions (**Figures 3(b), 3(c), and 3(e)**). The thermal electrons are also available for electron capture by electrophilic analytes, yielding negative analyte ions.

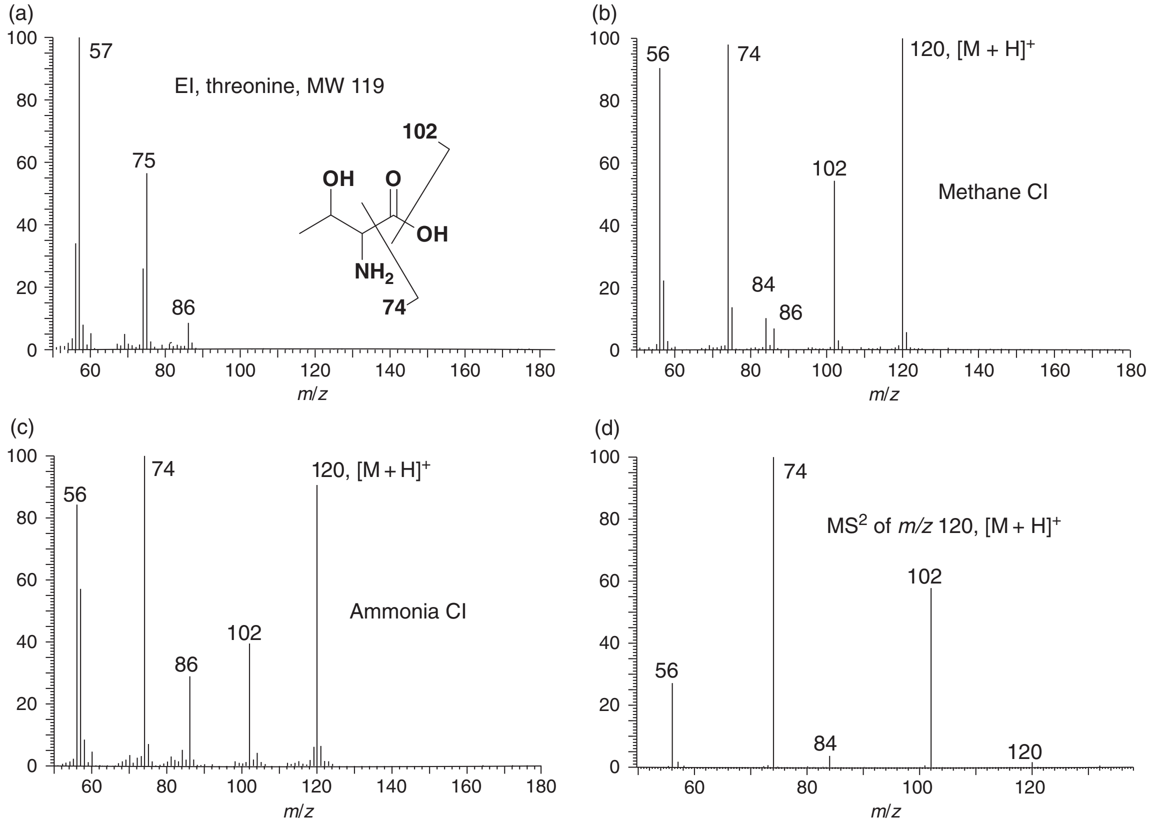


Figure 3 (Continued)

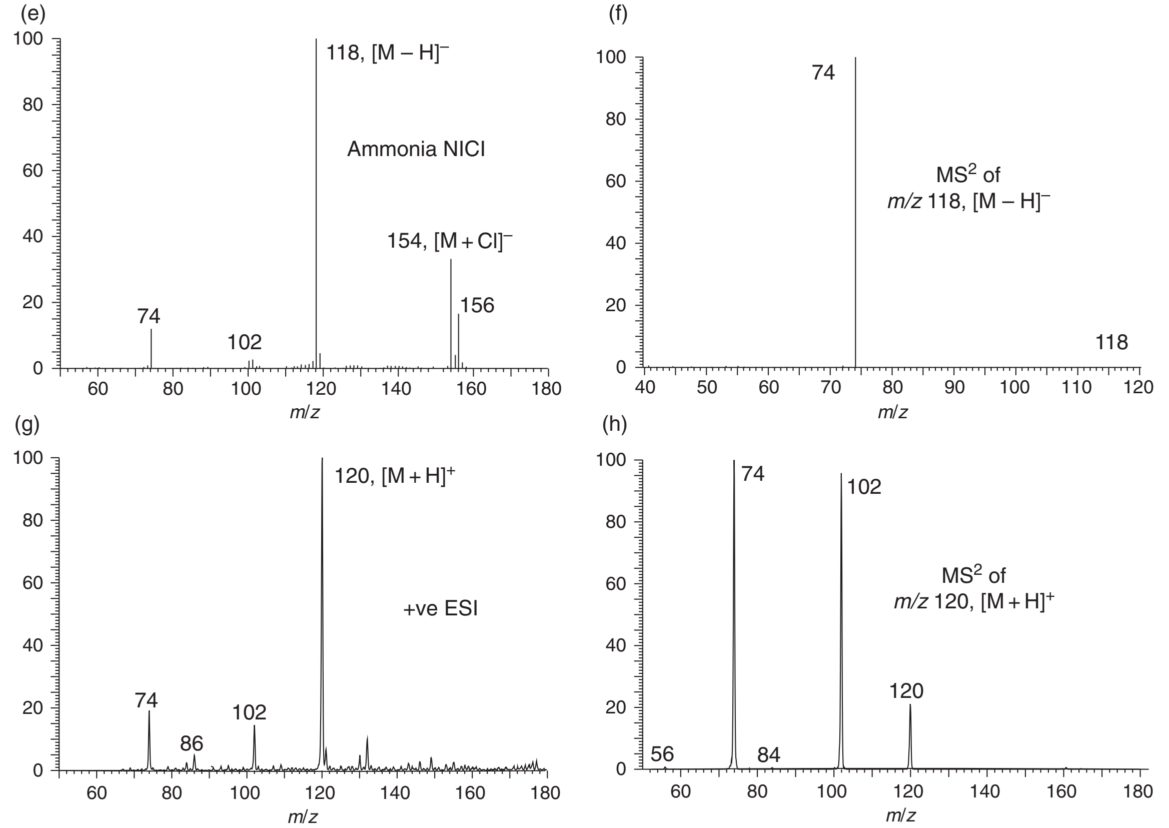


Figure 3 Mass spectra of the amino acid threonine. (a) Electron ionization (EI) mass spectra generated at an ionization energy of 70 eV. No molecular ion is observed. For an interpretation of the EI fragmentation, see Bieman and McCloskey[39] and Junk and Svec.[40] (b) Chemical ionization (CI) spectra generated using methane as the reagent gas. A prominent $[M+H]^+$ ion is observed at *m/z* 120. A discussion of the CI fragmentation may be found in Milne *et al*.[41] and Solovev *et al*.[35] (c) CI spectra generated using ammonia as the reagent gas. The spectrum is very similar to that in (b). (d) Tandem MS^2 experiment selecting *m/z* 120 from the ammonia CI spectra. (e) Negative ion chemical ionization (NICI) spectra demonstrating proton abstraction $[M-H]^-$ *m/z* 118 and adduct formation with chlorine $[M+Cl]^-$ *m/z* 154 and 156. Threonine HCl was dissolved in 50% EtOH/water. (f) Tandem MS^2 of *m/z* 118, in the negative mode. (g) Electrospray ionization (ESI) spectra, again featuring a prominent $[M+H]^+$ ion at *m/z* 120. (h) Tandem MS^2 experiment of *m/z* 120 from the ESI spectra.

Table 1 Common reagent and analyte ions in chemical ionization (CI)

Reagent gas	*Major reagent ions*	*Product ions*
Positive CI		
Methane, CH_4	$CH_5{}^+$, $C_2H_5{}^+$, $C_3H_5{}^+$	$[M+H]^+$, $[M+C_2H_5]^+$
Isobutane, C_4H_{10}	$C_4H_9{}^+$	$[M+H]^+$, $[M+C_4H_9]^+$
Ammonia, NH_4	$NH_4{}^+$	$[M+H]^+$, $[M+NH_4]^+$
Negative CI		
Chloroform	Cl^-	$[M-H]^-$, $[M+Cl]^-$
Ammonia	$NH_2{}^-$	$[M-H]^-$, $[M+NH_2]^-$
N_2O/CH_4(1:1)	OH^-	$[M-H]^-$

It is important to remember that these reactions are all occurring simultaneously in the source and that either the positive or negative ions can be selectively extracted from the source into the mass analyzer by placing the appropriate voltages on the extracting and focusing lenses. In the case of Q analyzers (Section 9.10.2.3.3), the switching between positive and negative polarity can be accomplished very rapidly so that positive and negative ions may be analyzed from a single GC peak. This technique is known as pulsed positive ion/negative ion CI (PPINICI).

In positive ion chemical ionization (PICI), the neutral analyte is most commonly ionized by proton transfer (Equation (4)) or adduct formation (Equations (5) and (6)).

When, for example, methane is used as a reagent gas, the $[M+1]^+$, $[M+29]^+$, and $[M+41]^+$ series of ions (Equations (4)–(6)) is good confirmation of the analyte molecular mass.

$$M + CH_5^+ \rightarrow [M+H]^+ + CH_4 \quad \text{proton transfer} \tag{4}$$

$$M + C_2H_5^+ \rightarrow [M+C_2H_5]^+ \quad \text{adduct formation} \tag{5}$$

$$M + C_3H_5^+ \rightarrow [M+C_3H_5]^+ \quad \text{adduct formation} \tag{6}$$

Less commonly, charge transfer (Equation (7)) and hydride abstraction (Equation (8)) may be observed.

$$M + CH_4^{+\cdot} \rightarrow M^{+\cdot} + CH_4 \quad \text{charge transfer} \tag{7}$$

$$M + C_2H_5^+ \rightarrow [M-H]^+ + C_2H_6 \quad \text{hydride abstraction} \tag{8}$$

Under CI conditions, negative reagent ions are also formed (**Table 1**) and these can effect analyte ionization by hydride abstraction (Equation (9)) or anion attachment (Equation (10)) (**Figure 3(e)**).

$$M + NH_2^- \rightarrow [M-H]^- + NH_3 \quad \text{hydride abstraction} \tag{9}$$

$$M + NH_2^- \rightarrow [M-NH_2]^- + NH_3 \quad \text{anion attachment} \tag{10}$$

As mentioned above, thermal electrons are also generated in the CI source, along with the reagent ions. These can be exploited for electron capture ionization (ECI), particularly in the case of molecules containing electrophilic moieties such as F, Cl, NO_2, and CN and this may confer advantages of increased sensitivity and selectivity for a particular analyte. These electron-capturing groups can of course be added into the target analyte by appropriate derivatization prior to analysis, to selectively enhance the possibility of electron capture.[21–27] It should be noted that this electron capture process is not, strictly speaking, negative CI as the analyte molecules are interacting with the thermal electrons and not the reagent ions derived from the CI gas.

There are three different mechanisms for ECI:

$$M + e^-(\sim 0.1\,\text{eV}) \rightarrow M^{-\cdot} \quad \text{resonance electron capture} \tag{11}$$

$$M + e^-(0-15\,\text{eV}) \rightarrow [M-A]^- + A^{\cdot} \quad \text{dissociative electron capture} \tag{12}$$

$$M + e^-(>10\,\text{eV}) \rightarrow [M-B]^- + B^+ + e^- \quad \text{ion pair formation} \tag{13}$$

The sensitivity of ECI analysis is generally two to three orders of magnitude greater than that of CI or EI analysis. Little fragmentation occurs during ECI, and this mode of ionization is generally employed for quantification of trace amounts of known compounds.

9.10.2.2.3 Ionization by proton transfer reaction

Recently, a variant of CI has been specifically developed to monitor in real time low concentrations of volatile organic compounds (VOCs).[42] VOCs are normally present in complex mixtures that could be separated by GC; however, these separations are relatively slow (15–60 min) and are not suitable for real-time monitoring. Proton transfer reaction mass spectrometry (PTR-MS) uses CI based on proton transfer from hydroxonium ions (H_3O^+). These hydroxonium ions are produced in an external glow discharge ion source operating in pure water vapor. The reagent ions are then passed into a drift tube that is continuously flushed with the ambient air containing the VOCs of interest. The H_3O^+ ion does not react with any of the common constituents of the atmosphere (N_2, O_2, Ar, or CO_2) as their proton affinities are lower than those of water. However, most VOCs have proton affinities higher than water ($>166.5\,\text{kcal mol}^{-1}$), and so proton transfers to the VOCs occur exothermically as a consequence of ion–molecule reactions in the drift tube. For the most part, these proton transfers are nondissociative and the mass analyzer can monitor a single ion species for each individual VOC (Equation (14)).

$$H_3O^+ + M_{VOC} \rightarrow MH^+_{VOC} + H_2O \tag{14}$$

However, some dissociation to $[M-OH]^+$ (Equation (15)) or $[M-OR]^+$ (Equation (16)), depending on the chemical class of the analyte, can occur.

$$MH^+ \rightarrow [M-OH]^+ + H_2O \tag{15}$$

$$MH^+ \rightarrow [M-OR]^+ + ROH \tag{16}$$

Some further selectivity can be introduced into the process by using ammonia as the reagent gas (NH_4^+ reagent ions) so that proton transfers occur only with compounds with a proton affinity >204 kcal mol^{-1}.

A more sophisticated, though less common version of PTR-MS is selected ion flow tube mass spectrometry (SIFT-MS). In this technique, a mixture of reagent ions is generated in a gas discharge ion source and then a Q mass filter is used to select one reagent ion, which is then injected into an inert carrier gas (usually He), for reaction with the gaseous sample, which is also injected into the carrier gas. The products of the ion–molecule reactions are then analyzed by a second mass analyzer. Recently, a triple cell PTR Fourier transform ion cyclotron resonance MS (FTICR MS) has been built to encompass the whole process with the advantage of high mass resolution and accuracy to characterize the ion–molecule reaction products[43] (see also Section 9.10.2.3.8). This, however, is achieved at the cost of sensitivity (1 ppm compared with 0.1 ppb).

The most common reagent ions used in SIFT-MS are H_3O^+, NO^+, and $O_2{}^{+\cdot}$, and their reactions with many different classes of volatile organics have been well documented.[44] The NO^+ reagent ion can react with the VOCs, depending on their chemistry, in one or two of several different ways – charge transfer (Equation (17)), hydride ion transfer (Equation (18)), hydroxide ion transfer (Equation (19)), alkoxide ion transfer, and ion–molecule association (Equation (20)).

$$M + NO^+ \rightarrow M^{+\cdot} + NO^{\cdot} \tag{17}$$

$$M + NO^+ \rightarrow [M-H]^+ + HNO \tag{18}$$

$$M + NO^+ \rightarrow [M-OH]^+ + HNO_2 \tag{19}$$

$$M + NO^+ \rightarrow [M+NO]^+ \tag{20}$$

VOCs mostly react with $O_2{}^{+\cdot}$ via charge transfer (Equation (21)) or dissociative charge transfer (Equation (22)); however, this reagent ion has found most use in monitoring NO, NO_2, and CS_2 as NO^+, $NO_2{}^+$, and $CS_2{}^{+\cdot}$, respectively.

$$M + O_2^{+\cdot} \rightarrow M^{+\cdot} + O_2 \tag{21}$$

$$M + O_2^{+\cdot} \rightarrow [M-R]^+ + R^{\cdot} + O_2 \tag{22}$$

Reactions of different chemical classes with the H_3O^+, NO^+, and $O_2{}^{+\cdot}$ reagent ions may be found in Smith and Španěl.[44]

9.10.2.2.4 Electrospray ionization

Electrospray is a process of transferring solution ions, typically large, nonvolatile polar molecules such as proteins, peptides, and carbohydrates, into the gas phase by ion desorption or ion evaporation.[45] Samples are supplied to the source directly via a syringe or, most commonly, as the eluent from an LC column. The liquid is passed through a metal needle held at high voltage (1–3 kV with respect to the sample cone or MS inlet) and sprayed into the ionization chamber at atmospheric pressure. A coaxial nebulizer gas may assist spray formation in the case of high solvent flow rates. As the charged droplets evaporate and shrink in size, the charge concentration in the droplets increases to the point where like-charge repulsion overcomes surface tension and the droplets explode to form microdroplets. The process is repeated and ultimately ions are ejected (desorbed) into the gas phase. These ions are then attracted into the off-axis or orthogonal sample inlet (counterelectrode) of the mass spectrometer. This off-axis geometry has the advantage of excluding neutral molecules and solvent clusters from the mass spectrometer.

ESI is a very 'soft' process, inducing little fragmentation, but in the case of molecules with a number of chargeable sites, a distribution of charge states is generated (Figures 3(g) and 6). The distribution and nature of the charges is very much a function of the sample solvent. In protic solvents such as water or mixtures of water and methanol or acetonitrile, sample ions will form a protonated, $[\mathrm{M}+n\mathrm{H}]^{n+}$, or deprotonated, $[\mathrm{M}-n\mathrm{H}]^{n-}$, series of multicharged ions. If alkali metals or ammonia is present in solution, then cationization will also be observed.

The number of charges that can be carried on an electrosprayed molecule depends on a number of factors including the size of the molecule, the tertiary structure (e.g., some charge-carrying sites – basic amino acids in +ve ESI – may be physically removed from exposure to the solvent at the center of a folded protein), the number of sites on which a charge may be localized (acidic and basic sites), and the nature of the solvent (pH and presence of salts). The effect of solvent pH on the abundance and distribution of the charges on myoglobin is illustrated in Figure 4.

The multicharging phenomenon means that ions of very large mass can be detected with conventional analyzers with mass ranges up to 3000 u. As a general rule, there will be one charge for every 8–10 amino acids (~1000 mass units). Thus, for example, a protein or protein complex of 200 000 Da can be readily analyzed if it can accommodate 100 charges. Hence,

$$\frac{200\,000\ \mathrm{Da}}{100z} = 2000\ m/z$$

This distribution of charges, especially when there may be more than one molecular species, can represent a very confusing picture. This situation may be further compounded by the presence of additional ion series that can occur when protonation competes with cations such as sodium and potassium. However, it is possible to deconvolute the multiple charge states and to calculate the mass of the molecule in question, by application of simple algebra.

First, it is reasonable to assume when looking at the multicharged envelope of an unknown that adjacent peaks differ by one charge. For the most part, this will represent a proton, as the multicharged envelopes due to sodium and potassium tend to be much less abundant. In the myoglobin spectrum (Figure 4), two adjacent ions have been labeled M_1 (higher value) and M_2 (lower value) and these will carry n_1 and n_2 charges (protons), respectively.

Thus

$$n_2 = n_1 + 1 \tag{23}$$

Second, the observed m/z values of each of the peaks can be written as

$$M_1 = \frac{M_\mathrm{r} + n_1 H}{n_1} \tag{24}$$

where M_r is the mass of the unknown, n is the number of charges, H is the mass of a proton, M_1 is the m/z experimental value, and

$$M_2 = \frac{M_\mathrm{r} + n_2 H}{n_2} = \frac{M_\mathrm{r} + (n_1+1)H}{n_1+1} \tag{25}$$

The charge state, n_1, can then be calculated from

$$n_1 = \frac{M_2 - H}{M_1 - M_2} \tag{26}$$

The mass of the unknown, M_r, can then be determined from

$$M_\mathrm{r} = n_1(M_1 - H) \tag{27}$$

Where the multicharged series is due to cationization, the mass of H should be replaced by that of the cation (e.g., Na^+ or K^+). Fortunately, most modern ESI-MS data systems have computer-based deconvolution algorithms to automate this process (Figure 4(c)).

ESI is most commonly associated with the analysis of large biomolecules of medium to high polarity, and it is a major tool for proteomic analyses,[17] but it can also be used for the MS analysis of small molecules provided they contain basic groups (e.g., amino, amide) for positive ESI or acidic groups (e.g., carboxylic acid, hydroxyl) for negative ESI.

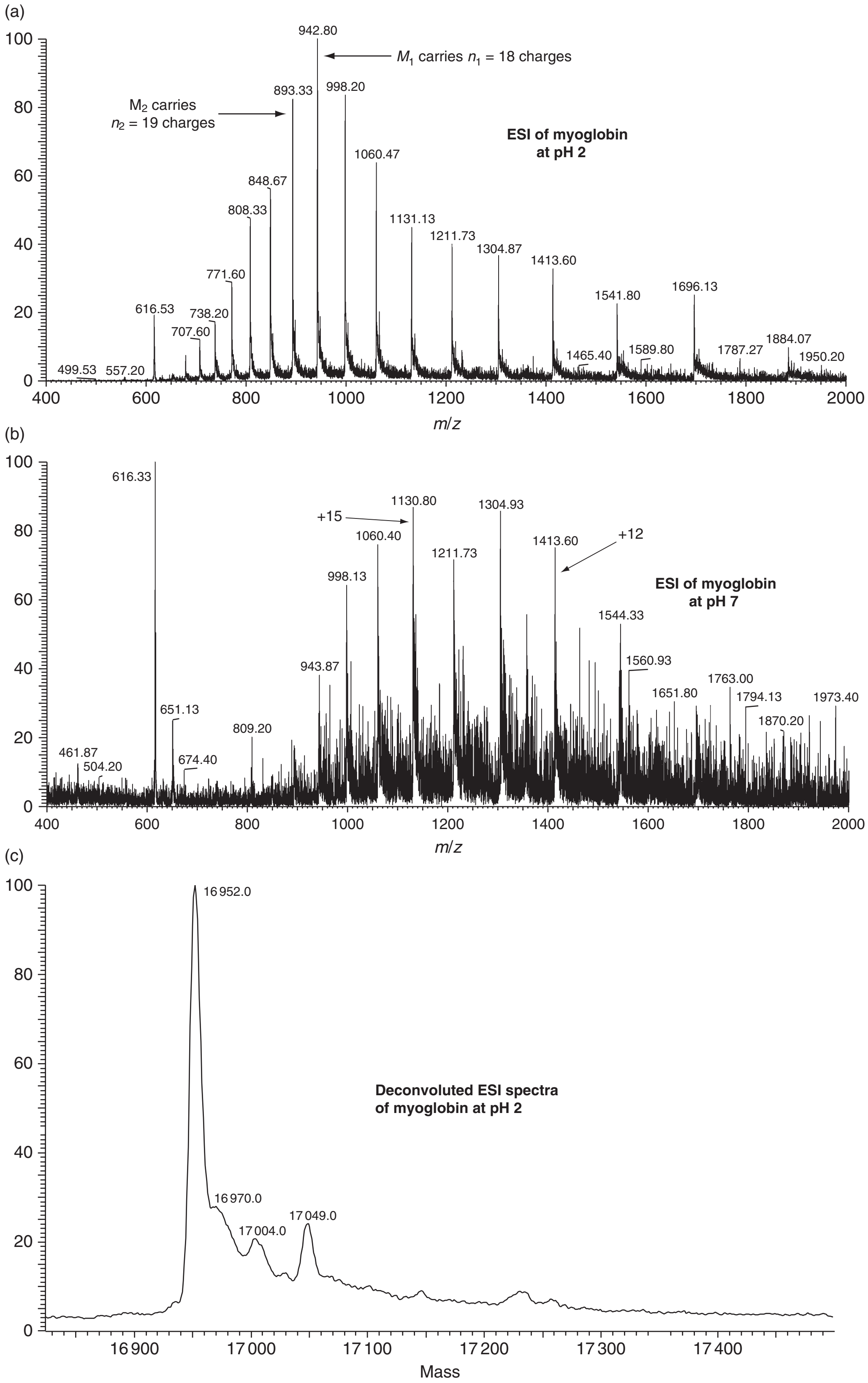


Figure 4 ESI mass spectrum of horse heart myoglobin (M_r, 16 951.49 Da) illustrating the multiple charge phenomena. Note that protonation is most effective at acid pH (a) rather than neutral pH (b) significantly altering the abundance and distribution of charge on myoglobin. Determination of the charge state can be made from first principles, using adjacent pairs of ions, from Equation (26), and the mass of the protein from Equation (27). The average mass determined from using all the data is 16 952.0 Da (c).

One of the great advantages of ESI is that generally it is very successful without the added complications of derivatization. Derivatization is often carried out under harsh conditions and the risk of sample degradation or the formation of multiple derivatives is very real. Nevertheless, derivatization can be a useful adjunct to ESI and there are many reports of derivatization being used to improve the ionization efficiency (and hence the sensitivity of an assay) by increasing the hydrophobicity or adding a group with a fixed charge to the analyte (see the review by Zaikin and Halket[46]).

Although ESI can be performed at quite high flow rates (up to 1–2 ml min^{-1}), the trend has been to run at lower and lower flow rates. Low flow rates mean that the coaxial nebulizer gas and the heated drying gases are no longer required, simplifying the construction and operation of the source. However, the most attractive feature of low flow rates is the dramatic improvement in the ESI efficiency with nano-ESI (20–50 nl min^{-1}) producing smaller initial droplets (200 nm diameter compared with 1–2 μm, a 100–1000-fold reduction in volume) allowing a much greater proportion of the sample to pass into the gas phase and then into the MS analyzer. Consequently, smaller amounts of sample are required, allowing more sophisticated biological experiments to be attempted on smaller samples. The second advantage of using low flow rates in ESI is that the problem of ion suppression is reduced. Analytes and other components in the spray compete for charge so that analytes with the lowest ionization energy will be preferentially ionized at the expense, for example, of more abundant analytes with higher ionization energy. Therefore, when using ESI, caution should be exercised in extrapolating from the observed spectrum ion abundance to the concentration of the neutral analyte in solution.

9.10.2.2.5 Atmospheric pressure chemical ionization

The atmospheric pressure chemical ionization (APCI) source is similar in design to the ESI source but the process of ionization is quite different.[47] The liquid sample solution is sprayed through a heated nebulizer into the source at atmospheric pressure. A corona discharge acts to ionize the atmospheric gases and solvent molecules to generate a series of reagent ions, in a manner similar to CI. Ionization of the analyte molecules then occurs by ion–molecule reactions, with minimal fragmentation. In most cases, only singly charged ions are generated and these are then extracted out of the source into the MS analyzer.

Unlike ESI, APCI actively generates ions from neutrals, making small (up to 1000–2000 Da), low to medium polarity analytes amenable to MS analysis. However, APCI is not as readily adaptable to low flow conditions as ESI because it is reliant on a concentrated cloud of solvent molecules to generate the necessary reagent ions.

9.10.2.2.6 Atmospheric pressure photoionization

Atmospheric pressure photoionization (APPI) is a relatively new technique[48–51] but the source design is almost identical to that used for APCI except that the corona discharge needle is replaced by a krypton discharge lamp, which irradiates the hot vaporized plume from the heated nebulizer with photons (10 and 10.6 eV). The mechanism of direct photoionization is quite simple. Where the ionization energy of the molecule is less than the energy of the photon, absorption of a photon is followed by ejection of an electron to form the molecular radical ion $M^{+\cdot}$ (Equation (28)).

$$M + h\nu \rightarrow M^{+\cdot} + e^{-} \quad \text{direct APPI} \tag{28}$$

However, in an atmospheric pressure environment, the major ion observed is $[M+H]^{+}$, the result of ion–molecule reactions abstracting a proton from protic solvents to yield $[M+H]^{+\cdot}$ (Equation (29)).[50] Charge may also be lost by proton transfer or electron attachment.

$$M^{+\cdot} + S \rightarrow [M+H]^{+} + [S-H]^{\cdot} \tag{29}$$

It should be noted that direct photoionization is not a very efficient process due to the strong absorption by the nebulizing gases and the solvent. Ionization efficiencies may be significantly enhanced by the use of a dopant such as toluene or acetone or anisole, which is added in excess to the vaporized solvent plume.[50,51] These dopants can be photoionized (Equation (30)) and the resultant reagent ions are then available to ionize the analyte by ion–molecule reactions, resulting in proton transfer (Equation (31)) and charge exchange (Equation (32)).

$$\mathrm{D} + h\nu \rightarrow \mathrm{D}^{+\cdot} + \mathrm{e}^- \quad \text{dopant APPI} \tag{30}$$

$$\mathrm{D}^{+\cdot} + \mathrm{M} \rightarrow [\mathrm{M}+\mathrm{H}]^+ + [\mathrm{D}-\mathrm{H}]^{\cdot} \quad \text{proton transfer} \tag{31}$$

$$\mathrm{D}^{+\cdot} + \mathrm{M} \rightarrow \mathrm{M}^{+\cdot} + \mathrm{D} \quad \text{charge exchange} \tag{32}$$

All the reactions are dependent on the ionization energies and proton affinities of the analyte, solvent, and dopant. Thus there are three possibilities for ionization in the positive mode, direct photoionization, proton transfer, and charge exchange.

The APPI source is also an effective generator of thermal electrons and is thus well suited to the generation of negatively charged analyte ions by ECI. Thermal electrons are readily generated by the 10 eV photons striking a metal surface (3–5 eV electron binding energy) and as can be seen from Equations (28) and (30), a thermal electron is generated for every photoionization event.

The great advantage of APPI is that it can be used to ionize nonpolar classes of compounds such as alkanes, alkenes, and aromatics that are not ionized by ESI or APCI and it can be interfaced with normal-phase chromatography,[49,51] where the corona discharge (APCI) and the high-voltage discharge (ESI) present a potential explosion hazard.

The full potential of APPI, particularly in the context of combined APCI/APPI or ESI/APPI sources, has yet to be explored. The photoionization and fragmentation of peptides/proteins is not well characterized and may represent another method, along with electron caphere dissociation (ECD) (Section 9.10.3.2.3) and electron transfer dissociation (ETD) (Section 9.10.3.2.5), of generating sequence information.[51,52] Also, unlike APCI, photoionization can be applied to very low solvent flow rates (less than 5 μl min^{-1}) relying on the analyte interacting with a photon of sufficient energy, and not on the solvent as a charge carrier. This alleviates the ESI and APCI problem of ion suppression where some analytes are unable to compete for charge from the charge carriers.

9.10.2.2.7 Matrix-assisted laser desorption ionization

Matrix-assisted laser desorption ionization (MALDI), like ESI, is capable of ionizing and launching very large molecules (e.g., polysaccharides, synthetic polymers, peptides, and proteins) into the gas phase and is a major analytical tool for high-throughput proteomic studies.[17,53] In many respects, MALDI is a complementary technique to ESI and both techniques are often applied to the same sample when determining protein identity. ESI produces macromolecular ions from solution, whereas MALDI produces them from the solid state.

In principle, the sample is cocrystallized with a matrix onto a stainless-steel target or a target with a hydrophilic spot surrounded by a hydrophobic surface designed to concentrate the sample into a small area.[54–57] The dried sample is then illuminated with a pulse of laser light (usually UV but also infrared (IR)) that is absorbed by the matrix chromophore. The photon energy is then transferred from the matrix to the embedded analyte which in turn is ionized and desorbed from the target. Singly charged ions, $[\mathrm{M}+\mathrm{H}]^+$, are typically produced and because this is another 'soft' ionization process, little fragmentation occurs. This makes for a relatively simple interpretation of the spectra; however, it must be noted that the lower end of the mass scale ($<\sim$800 m/z) is dominated by a plethora of intense matrix-derived ions. The lack of multiple charging of large analytes means that MS analyzers with an extended m/z range, such as time-of-flight (ToF) (see below), must be used.

Successful MALDI analysis is dependent on a number of factors, not the least of which is selection of an appropriate matrix. The matrix must be soluble in solvents compatible with the analyte (usually an aqueous/organic solvent mixture) and it must be possible to cocrystallize the analyte and matrix onto the target. The matrix must also be vacuum stable and be able to absorb at the emission wavelength of the laser. In addition, it must be able to cause codesorption of the analyte and promote analyte ionization. See **Table 2** for a list of commonly used MALDI matrices.

Other important factors that need to be optimized for MALDI analysis include the molar ratio of analyte to matrix ($\sim$1:10^4 is a good starting value) and the power or fluence (energy per unit area) of each laser shot. The best spectra, in terms of minimizing fragmentation and achieving the best resolution, are acquired at just above the laser fluence for ion formation. However, at low laser power, few ions are generated by a single laser pulse, so MALDI spectra are typically accumulated over tens or even hundreds of laser pulses. One of the drawbacks with MALDI is that the quality of the spectra generated is very dependent on good sample preparation and

Table 2 Common UV absorbing MALDI matrices (nitrogen laser, $\lambda = 337$ nm) and their area of application

Matrix	*Analyte*
Picolinic acid (PA)	DNA, RNA
3-Hydroxypicolinic acid (HPA)	DNA, RNA
3-Aminopicolinic acid (APA)	DNA, RNA
Dihydroxybenzoic acid (DHB)	Oligosaccharides
α-Cyano-4-hydroxycinnamic acid (αCHCA)	Peptides, lipids, oligonucleotides
Sinapinic acid (SA)	Proteins
2-(4-hydroxyphenylazobenzoic acid (HABA)	Polymers
2,4,6-Trihydroxyacetophenone (THAP)	Polymers, glycopeptides, oligonucleotides
6,7-Dihydroxycoumarin	Lipids

even then some parts of the sample surface, the so-called 'sweet spots', will generate better quality spectra than others. Practice and automated sample preparation, however, go some way in reducing this problem.

The sensitivity of the MALDI technique is generally comparable with that achieved by ESI but any advantage is offset, where automation is not available, by the work required in sample preparation and the difficulty of reproducibility. However, MALDI has a clear advantage over ESI in that targets holding a successful sample preparation can be stored and exploited repeatedly, at leisure. By comparison, ESI samples are nebulized and the sample consumed.

While MALDI is reputed to be relatively insensitive to contaminants (e.g., buffers, detergents, and salts), it must be said that the cleaner the sampler, the better the sensitivity and the better the coverage of analytes because ionization suppression is reduced.

A recent and exciting development of the MALDI technique has seen it adapted to molecular imaging of biological tissue sections (see discussion in Section 9.10.4.5).

9.10.2.2.8 Secondary-ion mass spectrometry

Secondary-ion mass spectrometry (SIMS) is an ionization technique that with the advent of ESI and MALDI had largely fallen out of favor with chemists and biologists. However, it has undergone something of a revival as its ability to chemically characterize a surface is now being applied to MS imaging of biological tissues (see Section 9.10.4.5). In this technique, a solid surface is bombarded with a continuous beam of highly focused, high-energy ions such as gold ($Au_3{}^+$), cesium (Cs^+), or bismuth ($Bi_3{}^+$) from a liquid metal ions gun (LMIG) or ions of Buckminster fullerene ($C_{60}{}^+$).[11,58–60] These ions penetrate the sample surface to a certain depth, depositing their energy through nuclear collisions and generating secondary ions (protonated or cationized) along the way. These secondary ions ($<\sim 500\ m/z$) are sputtered or emitted from the surface and are then directed to the entrance of the mass spectrometer for analysis.

9.10.2.2.9 Ambient ionization methods

Recently, a new family of ionization techniques that are distinguished by their ability to ionize analytes from surfaces under ambient conditions have been developed.[61] These methods are also characterized by the fact that no prior separation or extraction of the sample is required. Of these methods two have so far been well characterized, desorption electrospray ionization (DESI)[62] and direct analysis in real time (DART).[63]

DESI is closely related to ESI, with surface samples being ionized by a stream of charged solvent droplets to produce low-energy intact molecular ions. This technique has been successfully applied to a wide range of analytes (e.g., proteins, peptides, oligosaccharides, amino acids, terpenes, steroids, and lipids) that have been desorbed from a variety of surfaces, including paper, fabric, plastic, skin, and plant tissues. Ionization has also been demonstrated at up to 3 m away from the MS analyzer using an extended heated ion transfer capillary[64] and sensitivities down to attomole levels have also been reported.[62]

DART uses a glow discharge plasma to excite a heated stream of inert gas, usually nitrogen or helium, which is directed onto the surface to be analyzed. These excited state atoms and molecules have been shown to effect, like DESI, low-energy ionization of a variety of analytes (e.g., chemical warfare agents, pharmaceuticals, explosives, peptides) from a range of different surfaces (currency, concrete, skin, plant tissue, fabric, and glass). Again, like DESI, excellent sensitivities have been reported.

9.10.2.3 Mass Analyzers

After sample ionization, the ions are passed to the mass analyzer(s) where they are separated according to their mass to charge ratio (m/z). This separation can be based on a number of different ion properties, including momentum (magnetic sectors), kinetic energy (electrostatic analyzer), path stability (linear Q's), resonance frequencies (Q ion traps, linear ion traps), orbital frequencies (ion cyclotrons), velocity (ToF), or axial frequency (Orbitrap), as ions transit or are contained by combinations of electric and/or magnetic fields. The principle of operation, compatibility with different ionization sources, mass accuracy, mass resolution, and utility for tandem MS experiments of these different analyzers will be briefly discussed. Other factors that can be used to compare the performance of mass analyzers include the mass range limit, scan speed, efficiency of ion transmission, mass accuracy, and mass resolution (**Table 3**). More prosaic considerations include, of course, cost and vendor support. A more in-depth discussion of this subject matter will be found in Gross,[33] McLuckey and Wells,[65] Tarantin,[66] and Wollnik.[67]

Finally, it is important to realize that there is no one analyzer that is superior to all others. The choice of analyzer, therefore, must be based on the information required from the particular type of sample, remembering that analyses based on different mass analyzers can provide complementary information.

9.10.2.3.1 Resolution and accuracy

No discussion of MS data or comparison of mass analyzers would be complete without including some definition of the data quality, particularly, the accuracy of the data and the resolving power at which they were obtained.

Table 3 Common mass analyzers: their attributes and typical specifications

Mass analyzer	*Measures*	*Upper mass*	*Resolving power*	*Accuracy (ppm)*	*Dynamic range*[a]	*Cost*[b]
E[c]	Kinetic energy					
B[c]	Momentum					
EB or BE[d]		10^4	10^2–10^5	1–5	10^9	++++
Q[c]	Path stability	10^4	10^2–10^4	100	10^7	+
ToF[e]	Flight time	>10^4	>10^4	5–50[f]	10^2–10^4	+++
QIT[g]	Resonance frequency	>10^3	10^3–10^4	50–100	10^2–10^3	++
LIT[h]	Resonance frequency	>10^3	10^3–10^4	100	10^2–10^4	+++
FTICR[i]	Orbital frequency	>10^4	>10^6 at m/z 100	<1	10^2–10^5	+++++
Orbitrap[j]	Axial frequency	>10^4	6×10^4 at m/z 400	2–5[k]	10^3–10^4	++++

[a] Linear dynamic range.

[b] + = low cost; +++++ = high cost.

[c] May be configured with other analyzers for tandem-in-space experiments (e.g., QqQ, QqLIT, QqToF, and QqFTICR).

[d] Double-focussing BE or EB analyzer may be configured with other analyzers for tandem-in-space experiments (e.g., EBE).

[e] ToF combined with reflectron may be configured with other analyzers for tandem-in-space experiments (e.g., ToF–ToF, QIT-ToF, and QqToF).

[f] 1–5 ppm with a lock mass.

[g] Trapping-type instrument capable of tandem-in-time experiments and can be linked to ToF analyzer (QIT-ToF).

[h] Trapping-type instrument capable of tandem-in-time experiments and can be linked to Q, FTICR, or Orbitrap analyzers (e.g., QqLIT-FTICR, and LIT-Orbitrap).

[i] Trapping-type instrument capable of tandem-in-time experiments and can be configured to analyze fragments generated externally (e.g., QqFTICR or LIT-FTICR).

[j] Trapping-type instrument configured to analyze fragments generated externally (e.g., LIT-Orbitrap).

[k] < 1 ppm with a lock mass.

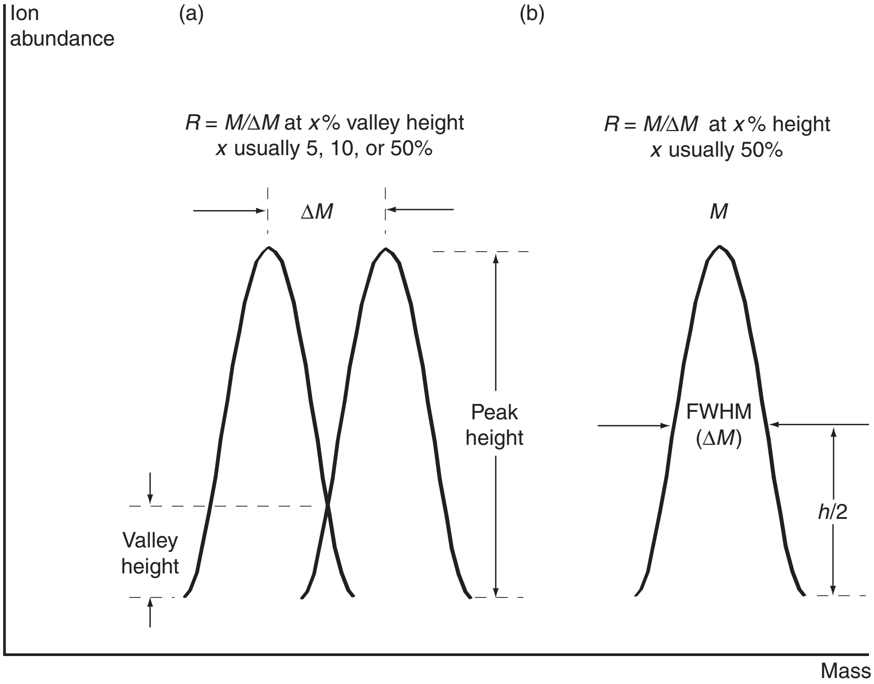


Figure 5 There are two definitions of mass resolution. These are based on either two overlapping peaks of equal intensity separated by ΔM (a) or a single well-defined peak with ΔM defined as the full-width at half-maximum height (FWHM) (b).

There are two commonly used definitions for mass resolution (R). The first, used with magnetic sector instruments, is defined as the ability to separate two neighboring ions in a mass spectrum where ΔM_x is the difference in m/z between the two peaks. The two peaks should be of equal size and similar shape and the degree of overlap (x) should be specified (**Figure 5(a)**). The latter is often specified as 10 or 50% of the valley height. M is the average of the two masses.

$$R = \frac{M}{\Delta M_x} \tag{33}$$

A more convenient definition, commonly used with trapping and ToF analyzers, pertains to a single well-resolved peak where ΔM_x is the peak width at a specified height x, usually half-maximum height (full-width at half-maximum height, FWHM) (**Figure 5(b)**). It should be noted that this FWHM definition of resolution equates to about twice that calculated from the 10% valley definition.

Resolution can vary over the mass range and this should also be specified. For example, Q mass filters and ion traps are usually operated at 'unit mass resolution' ($\Delta M_x = 1$) constant over the whole mass range. Thus the peaks at 100 m/z and 101 m/z will be separated at a resolution of 100 and the peaks at 1000 m/z and 1001 m/z will be separated with a resolution of 1000.

Mass accuracy is the difference (ΔM) between the measured accurate mass M and the calculated exact mass. It can be stated as absolute units of mass (differences of so many millimass units, mmu, 10^{-3} u) or as a relative mass accuracy in parts per million.

$$\text{Relative mass accuracy} = \frac{\Delta M}{M} \times 10^{6}\ \text{ppm} \tag{34}$$

Mass accuracy is also closely bound up with mass resolution as failure to achieve sufficient resolution of the ion of interest, away from interfering isobaric ions, will seriously impinge on the attainable mass accuracy. An appropriate level of mass accuracy and mass resolution in a mass spectrum can enable the determination of the elemental composition of the ions and can allow the analyst to distinguish, for example, glutamine from lysine ($\Delta M = 0.036$ u) and phenylalanine from oxidized methionine ($\Delta M = 0.033$ u) (see Section 9.10.4.3.3).

9.10.2.3.2 Magnetic and electric

The use of magnetic and electric fields to separate ions was introduced by Thompson[2] in his parabola mass spectrometer. The many developments that followed on from this culminated in the modern 'double-focusing' mass spectrometer that is available today. In principle, ions may be deflected by magnetic (B, momentum analyzer) or electric fields (E, kinetic energy analyzer). An ion, extracted with accelerating voltage (V) from the ion source and introduced orthogonally into a magnetic field, will follow a circular trajectory the radius (r) of which will be dependent on the ion's m/z value, its velocity v, and the magnetic field strength B.

The magnetic force zvB will be balanced by the centrifugal force mv^2/r.

Thus

$$zvB = \frac{mv^2}{r} \quad \text{or} \quad \frac{mv}{z} = Br \tag{35}$$

Hence it can be seen that the magnetic sector separates ions according to their momentum to charge ratio.

If the velocity (v) of the ion as calculated from the kinetic energy (E_k) of the ion emerging from the source

$$E_\mathrm{k} = zV = \frac{mv^2}{2} \tag{36}$$

is substituted into Equation (35), we derive

$$\frac{m}{z} = \frac{B^2 r^2}{2V} \tag{37}$$

from which it can be seen that changing the magnetic field (B) as a function of time will allow the successive passage of ions with varying m/z values. Ions with the same m/z value and the same kinetic energy will follow the same trajectory through the magnetic field. However, the process of ionization in the source results in ions being created with a small spread of kinetic energy. This energy dispersion then acts to limit the resolution achievable by the magnetic analyzer. This limitation can be countered by the addition of an electrostatic analyzer set to pass ions of a defined kinetic energy.

An ion entering an electrostatic field travels in a circular path of radius r such that the centrifugal force is balanced by the electrostatic field strength (E).

For ions carrying z charges

$$\frac{mv^2}{r} = zE \tag{38}$$

Substituting for the ion's kinetic energy (Equation (36))

$$r = \frac{2E_\mathrm{k}}{zE} \tag{39}$$

It can be seen from Equation (39) that the ion path is independent of the mass and that the electric field is a kinetic energy analyzer. The combination of the magnetic sector's directional focusing and the electrostatic analyzer's energy focusing results in a dramatic increase in the overall mass resolution and accuracy of the instrument. However, high resolving power is achieved at the cost of sensitivity because ions are selected within an increasingly narrow spread of energy and direction, with the rest being discarded. This double focusing characteristic can be obtained with the magnetic and electric analyzers arranged in either of the so-called forward (EB) or reverse (BE) geometries.

These types of mass spectrometers are today rarely used for biological applications, primarily because of their expense, size, and the relatively slow scan speed, which is incompatible with the trend toward fast, high-resolution LC and GC. The relatively low transmission efficiency also serves to limit the sensitivity of these instruments. The polarity of the magnetic field cannot be rapidly changed to perform, for example, PPNICI and rapid switching to selectively monitor a discrete number of ions (selected ion monitoring, SIM) is possible only over a narrow mass range. In addition, the high-voltage sources lend themselves to discharges when interfaced to liquid chromatographs or the relative high pressures in CI sources. The most important attribute of the double-focusing BE and EB instruments has been the acquisition of high mass accuracy and high mass resolution measurements; however, much of this demand is increasingly being met by ToF analyzers and by FTICR and Orbitrap instruments. Samples are usually introduced into these types of mass spectrometers by either a solids probe or GC.

9.10.2.3.3 Quadrupole

The Q mass filter consists of four parallel rods of circular or hyperbolic cross section (~10 cm long), extending in the *z* direction (direction of the ion beam). A high-frequency oscillating electric field is created in the space between the rods by rapidly switching the voltages applied to the rods, with adjacent rods having opposite polarity. The voltages are made up of a DC component (U) and a radio frequency (RF) component ($V\cos\omega t$). The forces acting on ions within the central volume (radius r) of the rods are given by

$$F_x = ma_x = z(U + V\cos t)\frac{2x}{r^2} \tag{40}$$

$$F_y = ma_x = -z(U + V\cos t)\frac{2y}{r^2} \tag{41}$$

Ions are thus alternately attracted and repelled by the rod voltages as they pass through these quadrupolar fields along the central axis of the rods. The equations of motion are complex (Mathieu equations[33]), but in principle, only ions with a narrow range of m/z values will be able to traverse the field for particular values of U and V. Other ions will undergo unstable oscillations and be ejected. From these equations, it can be seen that mass and charge are the only factors describing the ion trajectories. Scanning of the mass spectrum is achieved by varying U and V while maintaining the ratio of U/V constant. Q performance is dependent on the number of RF cycles experienced by the ion as it traverses the rods, so the accelerating voltage (and thus ion velocity) applied to ions entering the rods is limited to approximately 10–20 eV. These low accelerating voltages mean that Q analyzers can tolerate higher pressures than the high-voltage sources of magnetic analyzers and are more suited to interfacing with atmospheric pressure sources (e.g., ESI and APCI).

Q's are compact, robust, and inexpensive. They have high ion transmission properties and because scanning is achieved by sweeping electric potentials, the mass range can be rapidly scanned, so they are readily adapted to interfacing with fast chromatography. The ability to rapidly change electric potentials in the source means that it is possible to rapidly switch between analyzing positive and negative ions in alternate scans, something that is impossible with BE-or EB-type instruments, which would require a change in the direction of the magnetic field. The potentials on the Q rods can also be rapidly switched to allow the selective monitoring of a discrete number of ions (SIM). Most importantly, Q's are readily interfaced to a variety of ion sources and methods of ionization. However, the mass range is limited (2000–4000) and they are not generally capable of high mass resolution. The circular cross-section rods only approximate the required quadrupolar trapping fields and higher mass resolution can be achieved by the use of the more expensive hyperbolic rods.

Q's are also used in the so-called 'RF-only' mode (DC voltage set to zero) allowing transmission of ions with a wide range of m/z values and characteristically focusing them into the central region between the rods. This latter property means that RF-only Q's have found wide use as ion guides or collision cells, to focus an ion beam or to improve the transmission of collision products. The amplitude of the RF voltage determines the low mass cutoff and, theoretically, all ions of m/z greater than the low cutoff value are transmitted. However, there is some discrimination against ions of high mass. Hexapoles and octapoles are used in a similar manner but have better wide band pass characteristics. All these RF-only multipole devices are designated 'q' in the shorthand used to describe instrumental configurations. These RF-only multipoles are commonly found in hybrid mass spectrometers used for tandem MS (Section 9.10.3) serving as collision cells and to efficiently transport ions between differentially pumped regions of the instrument.

Q mass spectrometers may be found interfaced with most of the sample introduction and ionization methods described above with the exception of MALDI.

9.10.2.3.4 Quadrupole 3D-ion trap

The quadrupole ion trap (QIT) is about the size of a small fist and consists of a ring electrode and two hyperbolic end electrodes (see March and Todd[68] for a detailed theory of operation and history of development). Like the linear ion trap (LIT, see below), the QIT operates at relatively high pressure (10^{-3} torr) with a helium buffer gas that assists the ions to maintain a stable orbital frequency. The buffer gas also serves as the collision gas for collision-induced dissociation (CID) during MS/MS experiments.

Ions may be created inside the QIT or, more commonly, externally. An oscillating saddle field inside the trapping volume contains and focuses the ions into the center of the trap. From here the operator can scan the ions out of the trap to create a classic full mass spectral scan of the ions in the trap. Alternatively, a particular ion can be selected (isolated), collisionally fragmented and a scan of all the product ions generated (MS^2 scan). This whole process can be repeated with any one of these fragment ions (MS^3 scan) and as long as there are sufficient ions remaining in the trap to provide an adequate signal-to-noise ratio (S/N), the process can be repeated (**Figure 6**).

The number of ions that can be retained in the QIT, or indeed in any trapping-type instrument, is limited by space charging effects. Space charging occurs when the cloud of ions becomes sufficiently dense that coulombic repulsion between the like-charged ions starts to overcome the trapping potential, resulting in degraded mass resolution and accuracy. Limiting the number of ions in the trap at any one time normally controls this effect.

The QIT is compatible for use with the full range of methods for introducing solids, liquids, and gases – solids probe, GC, and LC – and with all the ionization methods described above including MALDI.

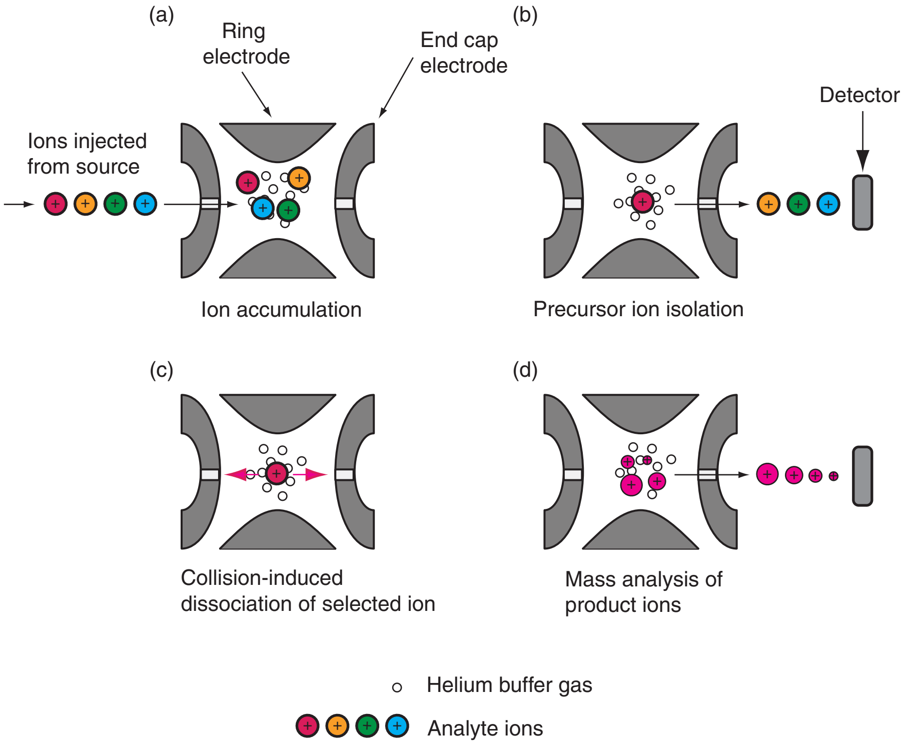


Figure 6 Schematic of collision-induced dissociation (CID) in the quadrupole ion trap (QIT) (MS^2 experiment). In separate events, ions from the source are accumulated and trapped in the space at the center of the electrodes (a). Ions with a specified *m/z* value are retained in the trap and all others ejected (b). The specified ions are then collisionally fragmented by axial excitation between the two end caps (c). The resulting product ions are then sequentially ejected to generate the product ion spectrum (d). In an MS^3 experiment, one of these product ions may be selectively retained in the trap, excited, and fragmented.

9.10.2.3.5 Linear 2D-ion trap

The two-dimensional linear ion trap (2D-LIT) is a logical development of the Q mass filter, described above, in that by the imposition of appropriate potentials at the entrance and exit of the Q's, ions with a range of *m*/*z* values can be trapped within the axial quadrupolar field (see March and Todd[69] for a detailed theory of operation and history of development). In common with the QIT, the LIT operates at relatively high pressure (10^{-3} torr) with a helium buffer gas. The buffer gas collisionally cools the ions and also acts as a collision gas for MS/MS experiments.[70,71]

The LIT has several advantages over the QIT. The larger volume means that more ions can be contained within the LIT before space charging becomes evident. This results in a greater dynamic range and improved sensitivity that can translate into lower detection limits for MS/MS analysis. Trapping efficiencies are also enhanced, as ions entering the trap have to overcome the trapping potential only on the front section. Once in the trap, the ions are collisionally cooled by interaction with the helium buffer gas and thereafter lack the energy to escape the trapping potential on the front section. Once in the trap, the ions are collisionally cooled by interaction with the helium buffer gas and thereafter lack the energy to escape the trapping potential on the front and back sections. This is in contrast to the QIT where there is only a narrow time window in which the amplitude and phase of the RF voltage are such that ions can pass through the end cap to enter the trap. This limits the trapping efficiency for the QIT to <5% compared to 29% for the LIT. At other phases and amplitudes, ions will have either too little or too much momentum so that the ions do not experience a sufficient number of collisions with the QIT buffer gas to be cooled and trapped.[70] In summary, the LIT has a significant sensitivity advantage over the QIT.

Using mass selective instability with resonance ejection, ions are scanned out of the trap through slits in the center of two opposite center section rods and focused onto two separate conversion dynodes. In the case of the QIT, where ions are scanned out of both end cap electrodes, the only place for a detector is behind the end cap opposite the ion entrance, so that only half of the ions scanned out of the trap are detected. Both the QIT and LIT operate at unit mass resolution with similar scan rates and both have the capacity to generate higher resolution spectra at slower scan rates.

In theory, the LIT should have the same universal utility as the QIT in terms of the types of samples and in being interfaced with GC or LC but to date only the LC interface is commercially available.

9.10.2.3.6 Orbitrap

A new mass analyzer, the Orbitrap, is a modified development of the 'Knight-style' Kingdon trap.[68,72–73] The Orbitrap radially traps ions about a central spindle electrode that is contained by an outer barrel-like electrode maintained at a vacuum of more than 3×10^{-10} torr. The *m*/*z* values of the ions are then measured from the frequency of the ion's harmonic oscillations along the axis of the central electrode. These axial frequencies are independent of the energy and spatial spread of the ions and they are detected as a broadband image current of a time-domain signal that is converted to a mass spectrum by fast Fourier transform algorithms.[74]

The Orbitrap is available as a stand-alone instrument and as a hybrid consisting of a linear ion trap coupled to the Orbitrap via a C-trap, which is responsible for focusing and injecting ions tangentially into the Orbitrap (LTQ-Orbitrap).

The performance characteristics of this analyzer are quite remarkable with mass accuracies of <2 ppm at a resolving power of 60 000 using an external calibration[75] and of <1 ppm with internal calibration.[76] As such, it has attracted the attention of analysts and instrument developers alike. New features have included ETD (see Section 9.10.3.2.5) and options for higher energy collisions in the C-trap or in an additional octapole collision cell.[77]

9.10.2.3.7 Time-of-flight

Conceptually, the ToF analyzer is very simple, in that ions of the same kinetic energy, E_k (extracted from the ion source with accelerating voltage *V*), but differing *m*/*z* values take different times *t* to traverse a fixed distance *d*. Thus lighter ions travel the fastest and are detected before the heavier ones. For an ion of mass *m*, the electric charge *q* is equal to the number *z* of electron charges *e* (*ez*).

$$E_{\mathrm{k}} = ezV = \frac{mv^2}{2} \tag{42}$$

$$t = \frac{d}{v} \tag{43}$$

Substituting for velocity into Equation (43) yields

$$t^2 = \frac{m}{z} \quad \frac{d^2}{2Ve} \tag{44}$$

To measure the flight time, the ions must be accelerated from the source in discrete packets. The resolving power of this simple experimental setup, linear ToF, was not good and after some initial popularity, the technique languished. The resolution was limited by the fact that at the time when ions are accelerated out of the source, they are not neatly lined up at the starting line. Rather they are positioned throughout the source and have a range of different kinetic energies. For the ultimate resolution, ions of the same mass (isobaric ions) positioned anywhere within the ion source need to arrive simultaneously at the detector.[78,79]

When the laser-induced ion plume is formed, there is no immediate application of the source extraction field and the plume is allowed to expand as if in a field-free region. If we consider just a group of isobaric ions, the more energetic ions fly faster and reach further into the source region than less energetic ones. Then at a chosen time, one of the electrodes of the extraction region is appropriately pulsed with high voltage to create the extraction potential. The ions in the tailing end of the plume (the originally less energetic) find themselves in a higher potential than the rest, and eventually acquire slightly higher velocity, enough to catch up with the leading-end ions by the time they reach the detector position.

Variations in the longitudinal velocity of isobaric ions can also be corrected by the use of a reflectron. This is basically an electric field that initially slows the ions and then accelerates, or reflects, them back out toward the detector. The more energetic ions will penetrate deeper into the decelerating field than less energetic ions of the same m/z value and experience a longer flight path and a longer flight time. The end result is that ions of a given m/z value will arrive at the detector in a much narrower time span (time focusing). The combination of delayed extraction, to compensate for positional differences of the ions, the addition of one or more reflectrons in the flight path, to compensate for different ion kinetic energies, and fast digital electronics, can boost the mass resolution of the ToF analyzer to better than 10^4 (FWHM).

At the start of the ToF renaissance, these analyzers were associated with MALDI sources as the discontinuous laser pulses are ideally suited to the pulsed nature of the ToF analyzer. However, continuous ion beams (e.g., EI and ESI) have also been coupled with ToF analyzers.[78] This has been achieved by locating the ToF analyzer orthogonal to the continuous ion beam axis. An orthogonal accelerating voltage is then applied to the beam and a discrete linear ion packet can then be pulsed into the ToF. During the time that the ions are moving in the drift region, and in the reflectron, the orthogonal acceleration volume is refilled by the continuous beam, hence the high, mass analyzer efficiency that is characteristic of ToF analyzers. For illustrative purposes, Guilhaus *et al.*[78] compared the approximate mass analyzer efficiency of a Q scanning over a 1000 u mass range and a ToF analyzer, with calculations of 0.025 and 25% maximum efficiencies, respectively. As Guilhaus *et al.*[78] have stated,

> In scanning instruments some of the mass range is detected all of the time while in TOF instruments all of the mass range is detected some of the time.

ToF analyzers are relatively small and of medium expense and so represent a good alternative to magnetic sector and Q analyzers, especially when their speed and sensitivity advantages are considered. Their mass accuracy and ease of calibration are also well established. ToF analyzers also have the highest practical mass range of all mass analyzers. However, the digitizer speed may place limitations on the instrumental dynamic range. The very fast acquisition rates that are achieved in ToF analyzers mean that they are also ideally suited to analyze fast GC separations with the added benefit that the high acquisition rates mean that coeluting components are much more readily deconvoluted than when a slower analyzer such as a Q is used.

9.10.2.3.8 Fourier transform ion cyclotron resonance

The FTICR is a trapping-type instrument with the ICR cell being held within the field of a superconducting magnet.[18,33,80] The cell itself consists of three pairs of opposing plates in the form of a cube or a cylinder. Ions are injected into the cell along the axis of the magnetic field and are then electrostatically trapped within the cell by the trapping potential placed on the two trapping plates that are orthogonal to the direction of travel. These ions are then subjected to an excitation pulse from the excitation plates and they will then, under the direction of the Lorentz force, spiral out from the center of the cell into a circular orbit. As noted above (Section 9.10.2.3.2), ions introduced orthogonally into a magnetic field will, under the direction of the Lorentz force, follow a circular trajectory, the radius (r) of which will be dependent on the ion's m/z value, its velocity v, and the magnetic field strength B.

The Lorentz force, qvB (q, charge; v, velocity), can be equated to the centripetal force

$$\frac{mv^2}{r} = qvB \tag{45}$$

and the angular frequency (ω) of the ions trapped in these circular orbits (cyclotron motion) is given by

$$\omega = \frac{v}{r} \tag{46}$$

so that substituting for v from Equation (46) into Equation (45) yields

$$\begin{gathered} m\omega^2 r = q\omega rB \\ \omega = \frac{qB}{m} \end{gathered} \tag{47}$$

the cyclotron equation.

From Equation (47) it can be seen that while the ion cyclotron frequency (ω) of an ion is a function of its mass, charge, and the magnetic field, it is independent of the ion's initial velocity.

The cyclotron orbits of thermal energy ions when they first enter the ICR cell are both too small and incoherent to be detected. However, if an excitation pulse is applied at the cyclotron frequency, the resonant ions will absorb energy and be brought into phase with the excitation pulse. They will have a larger orbital radius and the ion packets will orbit coherently. The ions may then be detected as an image current induced in the receiver plates. Additionally, this excitation pulse increases the kinetic energy of the trapped ions to the extent that fragmentation can be collisionally induced by ion–molecule reactions. Alternatively, the excitation pulse may be used to increase the cyclotron radius so that ions are ejected from the ICR cell.

Normally, many different ions will be present within the cell but they may all be excited by a rapid frequency sweep. The m/z values of the ions present in the ICR cell, and their abundance, may then be extracted mathematically from the resultant complex image current using a Fourier transformation to generate the mass spectrum of the ions. An important feature of FTICR is that the ions are detected nondestructively and that longer acquisition times over a narrower m/z range may be used to increase the measured mass resolution and the S/N.

FTICR instruments have a stringent requirement for a very low background pressure (10^{-10} torr) to minimize ion–molecule reactions and for this reason most analytical experiments are accessed through a variety of external ion sources that are separated from the cell by several stages of differential pumping. This vacuum requirement and the cryogenic cooling needed to run the superconducting magnet make this form of MS very capital intensive and expensive to run. However, this is offset by the extraordinary mass accuracy (sub-ppm with internal calibration), mass resolution (>10^6 at 100 u), and sensitivity (able to detect a few hundred ions at a time) that may be achieved.[18] FTICR instruments can also serve as platforms for a variety of unique dissociation techniques (e.g., IRMPD, infrared multiphoton dissociation; ECD; EDD, electron detachment dissociation; see Section 9.10.3.2) and this combined with their high mass accuracy and high mass resolution means they are ideally suited to identify and characterize large intact biomolecules – the 'top-down' approach.[81]

FTICR instruments that are designed to analyze low-molecular-weight molecules, such as VOCs, do not require superconducting cryogenic magnets and can be built using structured permanent magnets. Dehon *et al.*[43] built a dedicated PTR-FTICR (proton-transfer reaction Fourier transform ion cyclotron resonance) containing a cascade of three differentially pumped cells within the same magnetic field. The first cell is used as an ion source (10^{-5} torr), from which the selected H_3O^+ ions are drifted via the second cell into the third cell where they react with the sample (10^{-7}–10^{-5} torr). After the reaction, the ions are drifted back to the second cell for FTICR analysis. Although this instrumental approach is not as sensitive as in PTR-MS instruments (1 ppm compared with 0.1 ppb), the mass resolution and mass accuracy of the FTICR means that molecular formulas may be readily determined for the VOCs.

9.10.2.3.9 Ion mobility spectrometry

In drift tube ion mobility spectrometry (IMS), a packet of ions is drawn through an inert gas under the influence of a weak electric field. The extent of interaction with the inert gas and the rate of progress through the drift tube are dependent on the collisional cross section (shape and size) of the ion and on the number of charges carried by the ion. The requirement to gate the packets of ions entering the IMS and the need to wait for the ions to clear the drift tube result in a low duty cycle. If the sample is being supplied in a continuous flow, as in, for example, an ESI source, then much of the sample will be lost to the analysis. When this is combined with losses through radial diffusion, the overall sensitivity of the technique is poor. Nevertheless, the prospect of a technique to preprocess ions prior to MS analysis has proved attractive and in recent times two variations of this technique, circumventing these disadvantages, have been successfully developed for combination with MS.

In high-field asymmetric waveform ion mobility spectrometry (FAIMS), a continuous stream of ions is fed into the device inlet in a stream of dry carrier or bath gas.[82,83] The ions are then exposed to alternating strong and weak electric fields of opposite polarity across the carrier gas flow. The differential collisional interaction of ions with the carrier gas in the oscillating asymmetric fields results in different ions experiencing a net movement to one or the other wall electrode. If no other voltage is applied, the ions will eventually collide with one of the wall electrodes and be lost. However, if a low compensation voltage (CV) of correct magnitude and polarity is applied, then selected subsets of ions will be passed to the mass analyzer with a concomitant increase in their S/N and improved detection limits. For mixtures of ions, the CV can also be scanned. The ion separation achieved in the FAIMS device can also be refined by the use of different carrier gases.[82]

In the traveling wave IMS (TWIMS),[84] ions are initially accumulated in a trap ion guide and then released as an ion packet into the ion mobility ion guide. Here axial motion through the stack is generated by a repeating sequence of transient DC voltages providing a continuous series of 'traveling waves'. Ions are then separated as they are driven ahead of these potential hills through the stacked ring ion guides before transfer to the MS analyzer.

Although a relatively new adjunct to MS, IMS, whether the FAIMS or the TWIMS variety, has demonstrated a wide-ranging usefulness, particularly with respect to analyzing complex mixtures. It has been used, for example, to separate positional isomers of small molecules, to remove chemical noise and thereby improve detection limits and sensitivities of assays, and to examine conformational forms of multicharged protein ions. In proteomic experiments, IMS can be used to select out triply charged ions for ETD and doubly charged ions for CID, ignoring the single-charged peptides, solvent ion clusters, and other chemical noise (e.g., phthalate ions). Both FAIMS and TWIMS can be used with existing LC techniques to separate ions in a continuous stream and are in principle compatible with all types of ion sources and analyzers.

9.10.3 Tandem Mass Spectrometry

As you will see in the following section, it is quite common for an instrument to contain more than one analyzer. A shorthand nomenclature has been adopted to describe such instrumental configurations, using the analyzer abbreviations outlined above (Section 9.10.2.3), in which the order of the abbreviations represents the order of the analyzers traversed by the ion beam. For example, QqQ designates the very common triple Q instrument with the two scanning Q's separated by an RF-only Q that acts as the collision chamber. Other examples will be discussed below.

The 'soft' ionization processes described above (Section 9.10.2.2) typically generate single- or multicharged molecular ions with little accompanying fragmentation. To obtain structurally informative fragments, these ions must be subject to a second round of mass spectral analysis. This is known as MS/MS or tandem MS. In the first MS stage, an ion is selected or isolated in the mass spectrometer, activated and fragmented, most commonly by CID, and the product ions mass analyzed in the second MS stage. Depending on the instrument being used, it is possible to perform multistage mass spectrometry (MS^n) and to construct ion fragmentation pathways as part of an exercise in structural elucidation. It is also possible to use tandem MS to add a large degree of selectivity and to improve sensitivity in an assay by removing background chemical noise (see discussion below, Sections 9.10.4.2.2 and 9.10.4.5.7). With the demand for the analysis of increasingly complex samples, often coupled with a 'soft' ionization process, tandem MS along with mass determination with high accuracy and resolution has become an essential feature of modern biological mass spectrometers.

9.10.3.1 Analyzers

9.10.3.1.1 Tandem-in-space

For the beam-type mass analyzers (sector, ToF, and Q), each stage of mass analysis is performed in discrete mass analyzers usually separated by a collision cell. This arrangement is called tandem-in-space. The use of multiple analyzers means that analyzers can be independently selected for the different stages of analysis based on the desired performance characteristics.

Two common instrumental configurations for tandem-in-space experiments are the so-called QqQs, which consist of two Q mass filters, Q_1 and Q_3, separated by an RF-only Q collision cell (q) (**Figure 7**), and the QqToF class of instruments, which use a ToF analyzer in place of the third Q. The QqQ analyzer arrangement has the advantages of cost and ease of operation associated with Qs but leaves the analyst with limited mass resolution and mass accuracy with which to select and analyze ions. The replacement of the third Q by a ToF analyzer, although representing an increase in cost, gives the operator access to high-resolution/high mass accuracy data, in addition to greatly improved full scan sensitivity.[85] Today's generation of collision cells use RF-only multipoles (hexapoles and octapoles) or ring guides, which have improved transmission characteristics over the RF-only Q, but these are still commonly denoted as 'q' in instrumental shorthand.

Recent developments in instrumentation have seen the commercial release of traps combined with ToF analyzers (QIT-ToF), quadrupoles with traps (QqLIT), and traps with traps (LIT-FTICR, LIT-ToF, and LIT-Orbitrap), all taking advantage of the MS^n capabilities of the ion trap mass analyzers. The ability to select ions in a separate analyzer, prior to the final stage of MS analysis, serves to enhance the dynamic range and sensitivity of the final MS analysis. The development and characteristics of these hybrid combinations have been reviewed by Glish and Burinski[86] and Hagar.[87] The ToF–ToF combination with high mass accuracy and high mass resolution in both MS stages is also commercially available.[88]

9.10.3.1.2 Tandem-in-time

Tandem MS may also be performed intime using a trapping-type analyzer (e.g., LIT, QIT, and FT-ICR) (**Figure 6**). The experimental efficiency of this arrangement is usually higher than that of tandem-in-space instruments as ions do not have to be transferred between analyzers; however, experiments take longer to complete and sample presented to the mass analyzer from a continuous source while the trap is in the analysis mode will be lost. The different stages of the tandem-in-time experiment all take place in a temporal sequence within the same physical space. In these experiments, the selected precursor ion is retained in the trap and all other ions expelled. The selected ion is then activated and fragmented and the fragments analyzed to generate the MS/MS (MS^2) spectrum of the precursor ion. As long as there are sufficient ions still available in the trap, this process may be extended by selectively retaining one of the fragment ions and repeating the fragmentation process to generate the MS/MS/MS or MS^3 spectrum.

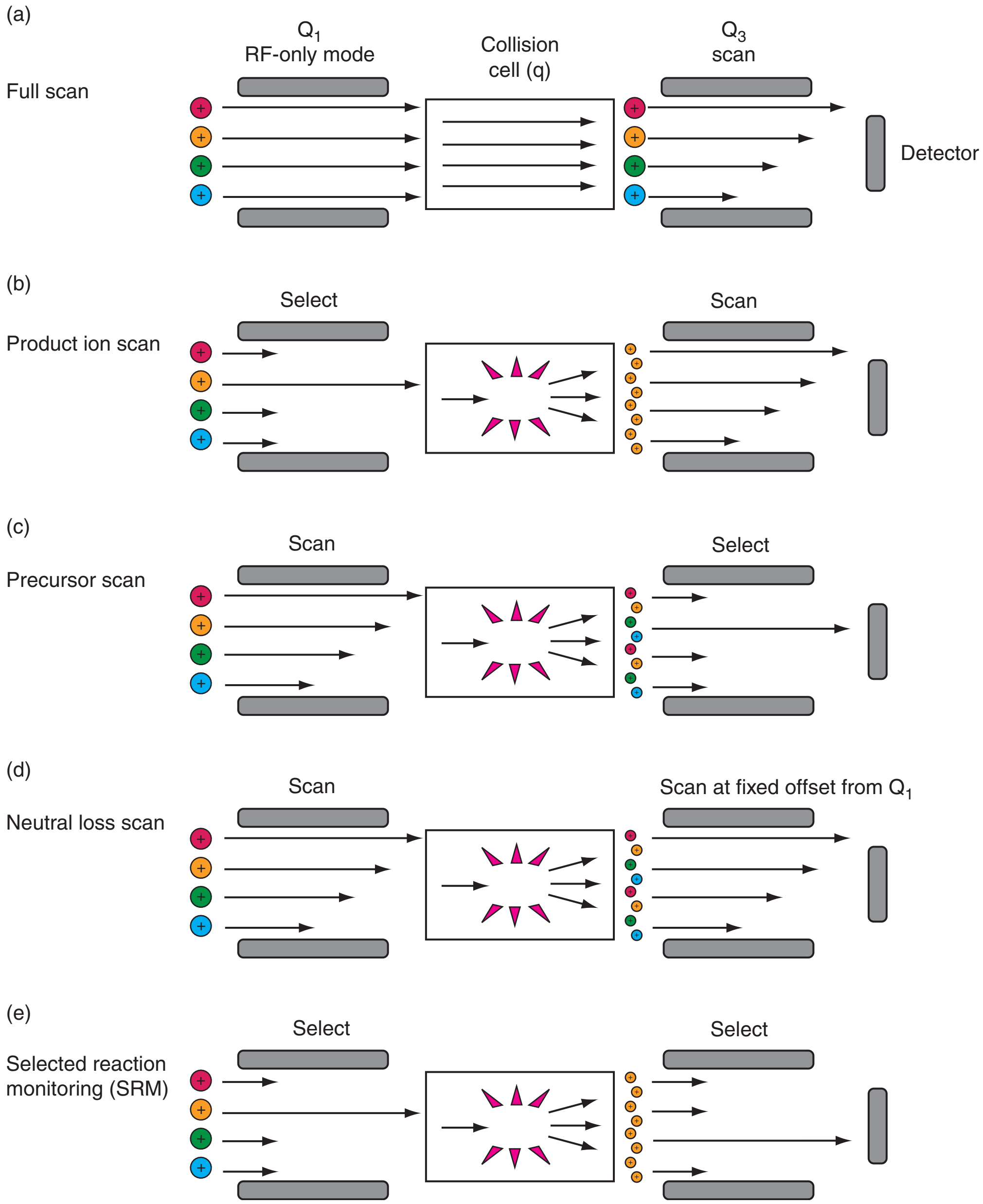


Figure 7 Scan modes for a tandem-in-space instrument, the triple quadruple (QqQ). (a) Full scan: all source ions are passed through to Q_3 while Q_1 and q (collision cell) are set to the RF-only mode. (b) Production scan: Q_1 is set to pass a selected ion (precursor ion). This is fragmented in the collision cell and products are analyzed by scanning Q_3. (c) Precursor scan: Q_1 scans all the source ions into the collision cell for collision-induced dissociation (CID). Q_3 is set to pass a selected product ion. A signal recorded at Q_3 is correlated with the corresponding precursor ion passing through Q_1. (d) Neutral loss scan: Q_1 is set to scan ions into the collision cell for CID. The Q_3 scan is offset by a specified mass, equal to the mass of the neutral, relative to Q_1. (e) Selected reaction monitoring (SRM): an ion selected in Q_1 is fragmented and a specific fragment is then recorded after selection by Q_3. SRM is commonly used in quantitative work to improve assay selectivity and sensitivity.

9.10.3.2 Fragmentation

9.10.3.2.1 Collision-induced dissociation

Fragmentation of ions in tandem experiments requires an input of energy to break internal covalent bonds. This is most commonly achieved by converting the kinetic energy of a collision, between the selected ion and an inert collision gas such as helium or argon, into vibrational energy. Fragmentation then occurs when the internal energy exceeds the activation energy required to cleave a particular bond. It is also important to note that bond cleavage may be preceded by an internal rearrangement, such as, for example, hydrogen scrambling or the McLafferty rearrangement.

For tandem-in-space experiments, CID occurs in a collision cell, physically located in the field-free region between the mass analyzers (**Figure 7**). The cell is differentially pumped and the flow of gas into the cell is carefully controlled. Increasing the collision gas pressure attenuates the main beam and, at the same time, the probability of ions undergoing single, double, triple, etc. collisions will increase, as will the scattering of the ion beam. Modern gas cells are usually either an RF-only multipole or a set of ring guides that are designed to contain and refocus, as much as possible, ions scattered from the direction of travel of the main beam. In high-energy collisions (KeV), the collision gas is usually helium as its high ionization energy reduces the risk of charge exchange. In low collision energy systems (1–200 eV), heavier gases such as argon or xenon have been used to improve the effectiveness of the CID process. The extent of fragmentation obtained from CID in tandem-in-space configurations is dependent on both the energy of the ions entering the collision cell and the pressure of the collision gas. Higher energy ions will be able to access fragmentations with higher activation energies and higher pressure of the collision gas will result in multiple collisions producing more extensive fragmentation, fragmenting fragments.

For small ions, a single collision may be sufficient to induce the dissociation of a covalent bond; however, as the size of an ion increases, so does the number of vibrational degrees of freedom over which the collisional energy may be distributed. Thus the effectiveness of CID decreases with mass. Nevertheless, CID is quite effective for relatively large, multicharged polypeptides (up to 5 kDa), where the bond cleavage may be assisted by the coulombic repulsion of the multiple charges.

The CID spectra generated in traps are qualitatively different from those generated by tandem-in-space experiments. In traps, only the selected ion is activated and once fragmentation of this ion has occurred, no further collisions can take place. Thus the fragment ion spectrum generated in a trap will be simpler and less informative than that from a tandem-in-space experiment. Where the precursor ion undergoes, for example, the simple loss of a neutral such as water, the product ion spectrum will be relatively uninformative, consisting mainly of $[\mathrm{MH}-\mathrm{H_2O}]^+$. However, this may be readily overcome by a broadband activation, which is applied to all ions in a range 20 m/z below the precursor ion.

FTICR cells are typically operated under very high vacuum (10^{-10} torr) and CID within the cell must be initiated by injecting a collision gas after an ion packet of a designated m/z value has been selected by expulsion of all other ions from the cell. The selected ions are then activated by sustained off-resonance irradiation (SORI). This results in the ion orbit expanding and contracting with time and in the process the ions undergo multiple, low-energy collisions with the injected collision gas. The collision gas is then pumped away and the fragment ion spectrum measured. Alternatively, if the FTICR is part of a tandem-in-space instrument (e.g., QqFTICR or LIT-FTICR), then ions can be fragmented by CID outside of the cell, with the product ions presented to the FTICR for mass analysis. Fragmentation in the FTICR cell may also be accomplished by photon-induced dissociation (PID, Section 9.10.3.2.2), ECD (Section 9.10.3.2.3), and EDD (Section 9.10.3.2.4) (see also discussion in Section 9.10.3.2.6 on the use of CID/IRMPD and ECD/ETD for protein/peptide sequencing, and **Table 4**).

Finally, when using ESI or APCI, it is also possible to perform in-source CID at atmospheric pressure. This is sometimes referred to as 'pseudo-MS/MS'. By increasing the entrance cone voltage, newly formed ions can be accelerated toward the entrance cone, colliding with other molecules, mostly atmospheric nitrogen, and fragmenting. It is important to note that there is no mass selection for a precursor ion and that selection is entirely based on chromatographic separation.

9.10.3.2.2 Photon-induced dissociation

Gas-phase ions may be fragmented by photoexcitation (PID), particularly by IR photons tuned to the vibrational frequency of covalent bonds. The cross section of an ion for photon absorption is low compared with its collisional cross section, so PID is most commonly associated with the use of intense light sources (lasers) and FTICR where the period the ion is exposed to the photons can be lengthened to increase the

Table 4 Comparison of CID/IRMPD with ECD/ETD for peptide sequencing

CID/IRMPD	ECD/ETD
• Molecules are vibrationally excited by either physical collision with a neutral gas (CID) or by absorption of an IR photon (IRMPD). • Vibrational energy can be distributed over whole molecule. • Larger analytes need more energy and efficiency drops off with increasing size. • The weakest, most labile bonds break first (loss of water, PTMs, etc). • In peptides, CID/IRMPD generates *y*- and *b*-series ions. • Low-energy interactions do not allow isomeric Leu/Ile to be distinguished.	• An electron is directly (ECD) or indirectly, via an anion ($M^{-\cdot}$ from fluoranthrene) (ETD), transferred to a cation (positively charged peptide). • Not applicable to singly charged ions. • The electron is accepted by an amide-associated proton on the peptide backbone. This very unstable radical reacts very quickly to cleave the peptide bond at the site of reaction. • Rapid process does not allow time to distribute energy over whole molecule. • Can be applied to whole proteins in 'top-down' proteomic analysis. • Bonds cleaved are those that accept the electron, not simply the weakest bonds. • PTMs are preserved. • Can distinguish isomeric Leu/Ile via secondary fragmentation of radical $z^{\cdot}$ ions. • In peptides and proteins, ETD generates *c*- and *z*-series ions but some *y* ions may also be observed. • Masses of *c* and *z* ions may be 1 Da lighter or heavier, respectively, because of extensive hydrogen rearrangement.

In modern instrumentation, these techniques may both be applied to the same peptide to generate complementary sequence data and PTM data (IRMPD and ECD in FTICR MS,[89] CID and ETD in QIT and LIT,[90] and CID and ETD in QToF[91]).

chance of absorption and subsequent dissociation. For IRMPD, these IR photons may be supplied by a laser (10.6 μm from a CO_2 laser) or they may be radiated from a heated blackbody (blackbody infrared dissociation, BIRD). IRMPD, BIRD, and CID all induce fragmentation by the addition of excess vibrational energy to covalent bonds and consequently yield very similar patterns of fragmentation. For example, when applied to protonated peptides and proteins, CID and IRMPD cleave the weakest bonds, the PTMs (e.g., phosphorylation, sulfation, γ-carboxylation, and N-and O-glycosylation) and the backbone peptide amide C–N bonds, to yield the characteristic series of N-terminal *b* ions and C-terminal *y* ions (**Figure 8**). The advantage of IRMPD and BIRD over CID is that no pump-down time is required to remove the collision gas and thus high-resolution detection can be effected immediately.

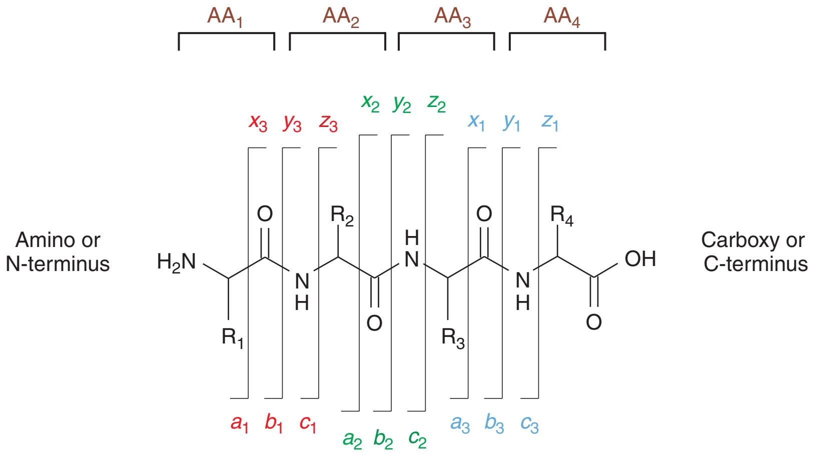


Figure 8 Roepstorff and Fohlman[92] notation for peptide fragmentation. For the *x*-, *y*-, and *z*-ion series, charge is retained on the C-terminus fragment and for the *a*-, *b*-, and *c*-ion series, charge is retained on the N-terminus fragment. Cleavage of the C_α–C bond gives rise to the *a* and *x* ions (e.g., by EDD); cleavage of the C–N amide bond, the *b* and *y* ions (e.g., CID); and cleavage of the N–C_α amine bond (e.g., by ECD or ETD), the *c* and *z* ions.

9.10.3.2.3 Electron capture dissociation

An alternative method for achieving covalent bond cleavage in the FTICR cell, and one that has been mostly applied to sequencing peptides and proteins, is ECD.[93,94] The multicharged peptide and protein ions from ESI are an ideal target for ECD as the cross section for electron capture increases by approximately the square of the ionic charge. The capture of a thermal electron (~0 eV) is an exothermic reaction and in protonated peptides and proteins, results in cleavage of disulfide (S–S) bonds along with cleavage of the backbone N–C$_\alpha$ amine bond, yielding the characteristic complementary pairs of *c* and $z^{\cdot}$ (~90%) or $a^{\cdot}$ and *y* (~10%) fragment ions used for sequencing.

$$[\mathrm{M}+n\mathrm{H}]^{n+}+\mathrm{e}^{-}\rightarrow([\mathrm{M}+n\mathrm{H}]^{(n-1)+\cdot})_{\text{transient}}\rightarrow \text{fragments} \tag{48}$$

$$\mathrm{R}_1-\mathrm{S}-\mathrm{S}-\mathrm{R}_2+\mathrm{e}^{-}\rightarrow \mathrm{R}_1-\mathrm{SH}+{}^{\cdot}\,\mathrm{S}-\mathrm{R}_2 \tag{49}$$

A unique feature of ECD is that the N-terminal fragment ions, the *c* ions, contain an extra hydrogen atom from the proton neutralized by the electron capture. The complementarity of the $c/z^{\cdot}$ pair can thus be confirmed by the fact that their mass sum is 1 u greater than the M_r of the protein.

The ECD process, by its nature, is a very rapid process and bond dissociation occurs faster than the redistribution of intramolecular vibrational energy that occurs with CID. This explains the dissociation of the strong N–C$_\alpha$ amine bonds in the presence of the weaker C–N amide bonds in peptides and proteins.[93,94] Consequently, any labile PTMs (e.g., phosphorylation, sulfation, γ-carboxylation, N- and O-glycosylation) are preserved and may be unequivocally located in the peptide/protein sequence. See also discussion in Section 9.10.3.2.6 on the use of ECD/ETD and CID/IRMPD for protein/peptide sequencing, and **Table 4**.

Recently, it has been observed that ECD can also occur for electrons with energies in the range of 3–13 eV,[95] the so-called hot ECD (HECD), with the excess energy going into secondary fragmentation, including cleavage of the C–N amide bonds (*b*- and *y*-ion series) in multicharged peptides. Significantly, the isobaric isoleucine and leucine residues were reported as losing $^{\cdot}C_2H_5$ and $^{\cdot}C_3H_7$, respectively, allowing these isomeric amino acids to be distinguished.[96]

Unlike CID, where the applied intramolecular vibrational energy can be redistributed and dissipated across the whole molecule, so that the efficiency of CID is diminished with increasing analyte size, ECD can be used to sequence large, undigested proteins. This has enabled the development of the 'top-down' approach to proteomics, which has the advantage of directly sequencing a protein, along with its PTMs, rather than having to infer the sequence following an enzymic digestion (e.g., trypsin or Lys-C) and an *in silico* reassembly from the MS/MS data on the enzymic peptides. It has also been noted that there are a number of side chain losses in ECD that can aid, for example, in distinguishing the isobaric amino acid residues, leucine and isoleucine. However, the ECD technique is accessible only in the expensive FTICR instruments (see also discussion in Section 9.10.3.2.6 on the use of CID/IRMPD and ECD/ETD for protein/peptide sequencing, and **Table 4**).

9.10.3.2.4 Electron detachment dissociation

EDD is a promising new FTICR technique, and is the negative ion complement to ECD. Both these electron-mediated techniques involve a radical ion intermediate, produced by either electron attachment to multiply charged cations (ECD) (Equations (48) and (49)) or electron removal from multiply charged anions (EDD) (Equation (50)).

$$[\mathrm{M}-n\mathrm{H}]^{n-}+\mathrm{e}^{-}_{\sim 20\ \mathrm{eV}}\rightarrow([\mathrm{M}-n\mathrm{H}]^{(n-1)-\cdot})_{\text{transient}}+2\mathrm{e}^{-}\rightarrow \text{fragments} \tag{50}$$

Many compounds such as glycosaminoglycans (GAGs), nucleic acids, acidic peptides, or peptides with acidic PTMs such as phosphorylation or sulfation do not readily form positive ions, especially in mixtures where ion formation is favored by the more basic mixture components. When positive ions can be formed, then the spectra are usually characterized by the abundant loss of the PTM (e.g., sulfate from GAG) or, in the case of oligonucleotides, a proton from the sugar–phosphate backbone.

However, these acidic compounds do readily form negative ions. Exposure of these anions to energetic electrons (~20 eV) is reported to produce a ‘positive radical charge (hole)’ which is exothermically neutralized by an electron.[97] For peptides, this results in C_α–C bond cleavage, to form complementary $a^{\cdot}$ and x fragment ions, with retention of the acidic PTM.[94,98] Thus like ECD, EDD can also cleave covalent bonds without affecting weaker noncovalent interactions. EDD has also been shown to preferentially cleave S–S and C–S bonds.[99] For GAGs, structurally informative glycosidic bond cleavages and cross ring cleavages can be generated without loss of the labile sulfate group.[100,101] and the glucuronic acid and iduronic acid epimers in heparan sulfate tetrasaccharides can also be distinguished.[102] EDD has also been used to partly characterize synthetic polyamidoamine dendrimers[103] and fragmentation was found to complement that obtained with CID. Complete sequences of short oligonucleotides of both DNA and RNA have been determined with EDD[104,105] and EDD may also be used to probe the tertiary structure of nucleic acid.[106]

9.10.3.2.5 Electron transfer dissociation

With respect to peptide/protein sequencing, ECD has some very attractive features, producing random cleavages along the peptide backbone (*c*- and *z*-type ions) and at the same time preserving labile PTMs such as phosphorylation. Unfortunately, this technique is not readily transferable from FTICR instruments to the relatively low-cost and more common instruments that trap ions by RF electrostatic fields (QIT and LIT), where the bulk of this work is performed, as these analyzers are unable to trap the required dense cloud of thermal electrons. However, the Hunt group[107,108] have developed an alternative method of delivering electrons to multiply charged cations, using anion–cation interactions to effect ETD.

Radical cations ($M^{-\cdot}$) of a polyaromatic hydrocarbon, usually fluoranthrene, are generated by methane CI, externally to the trap.

$$C_{16}H_{10} + e^{-}{}_{\text{thermal}} \rightarrow C_{16}H_{10}{}^{-\cdot}\ (m/z\ 202) \tag{51}$$

These radical cations are then injected into the trap where they are mixed with the multiply charged peptide cations to which an electron is then transferred, leading to their direct dissociation into *c*- and *z*-type ions by the same mechanism responsible for ECD. The process is rapid (milliseconds) and quite compatible with the chromatographic timescale of LC–MS.

$$[M + 3H]^{3+} + C_{16}H_{10}{}^{-\cdot} \rightarrow [M + 3H]^{2+\cdot} + C_{16}H_{10} \tag{52}$$

$$[M + 3H]^{2+\cdot} \rightarrow [c + 2H]^{+} + [z + H]^{+\cdot} \tag{53}$$

ETD is also applicable to large intact proteins but the multiply charged fragments are difficult to interpret because of the limited resolving power of the Q traps. However, it is possible to deprotonate the multiply charged fragment ions and to reduce their charge by a further round of cation–anion interactions with even-electron benzoate anions.[108]

$$[M + 7H]^{7+} + 6C_6H_5COO^{-} \rightarrow [M + H]^{+} + 6C_6H_5COOH \tag{54}$$

More recently, it has been demonstrated that the reagent anions used for either ETD or proton transfer can be derived from the same neutral compound. The radical anions used for ETD, $[M]^{-\cdot}$, are converted into even-electron proton transfer reagent anions. $[M + H]^{-}$, by changing the potential on the methane CI source.[109] This voltage switch can be acheived in milliseconds allowing for rapid sequential ion–ion reactions and opens up the possibility of top-down sequencing of intact proteins in RF ion traps.

It is unusual for either CID or ETD to provide complete sequence information from any one peptide but the use of both techniques provides complementary information, which can greatly extend the sequence coverage (**Table 4**). In addition, because the energy from the ETD process is directed into cleaving the C_α–N bond, the labile PTMs are preserved and their location in the peptide sequence can then be determined[90] (see also discussion in Section 9.10.3.2.6 on the use of CID/IRMPD and ECD/ETD for protein/peptide sequencing, and **Table 4**).

It should be noted that ETD is a relatively inefficient process for doubly protonated peptide precursors $[M+2H]^{2+}$, which are the ions most commonly found in 'bottom-up' proteomics experiments. This situation may be retrieved, however, by using a supplemental low-energy CID method (ETciD) to target the nondissociated electron transfer (ET) product, $[M+2H]^{2+\cdot}$. CID of the ET product then yields *c*- and *z*-type fragment ions. Swaney *et al.*[110] have reported that in a large-scale analysis of doubly charged tryptic peptides, the use of ETciD resulted in a median sequence coverage of 89% compared to 63 and 77% for ETD and CID, respectively.

9.10.3.2.6 Combined use of dissociation techniques

In proteomics experiments, neither CID/IRMPD (*b*- and *y*-series ions) nor ECD/ETD (*c*- and *z*-series ions) fragmentation, when used on its own, is capable of generating a complete set of sequence ions from which an unambiguous primary structure of a peptide may be derived (IRMPD and ECD in FTICR MS,[86] CID and ETD in QIT and LIT,[90] CID and ETD in QToF[91]). Thus database searching, which is at the heart of MS/MS-based proteomics, is vulnerable to misidentification of peptides (false positives).[111] The respective characteristics of both these processes are summarized in **Table 4**.

In CID/IRMPD, the C–N bond is cleaved to generate the well-defined *b*- and *y*-series ions and, in ECD/ETD, the N–C_α bond is cleaved to generate the *c*- and *z*-series ions. However, the latter series is not so well defined as the masses of the *c* and *z* ions may be 1 Da lighter or heavier, respectively, due to hydrogen rearrangements.[112] The other important difference between these two processes is the retention of PTMs (phosphorylation of serine, threonine, and histidine, along with the N- and O-glycosylations) in ECD/ETD and the ability to distinguish the isomeric amino acids, leucine, and isoleucine.

On the basis of the complementarity of CID/IRMPD and ECD/ETD, Zubarev *et al.*[111] have concluded that *de novo* sequencing of peptides using these two fragmentation techniques in conjunction with high mass accuracy[113,114] can be achieved with >95% reliability. They have furthermore stated that it is

> . . . only de novo sequencing which can guarantee error-free sequence identification.

The complementarity of ECD/ETD has also been confirmed in a comprehensive comparison by Molina *et al.*[115] on some 19 000 peptides. They found that by combining the respective peptide fragmentation data they could achieve a 92% sequence coverage for an average tryptic peptide.

ECD may also be used in a complementary manner with the newly developed EDD technique. Although the fragmentation efficiency of EDD (average 3.6%) was low compared with ECD (average 15.7%), Kjeldsen *et al.*[116] have recently demonstrated that the combination of the two techniques could increase the overall amino acid sequence coverage of proteins and PTM characterization.

Further developments in the use of these complementary combinations of dissociation techniques will aid in generating a more comprehensive and reliable system of identifying and characterizing proteins and their PTMs. Such progress is likely to be based on a comprehensive understanding of gas-phase peptide chemistry and fragmentation.[117]

9.10.4 Experimental Use of Mass Spectrometry

Whether you are using MS for identification or quantification, it is important to first have as much information about the sample as possible, particularly the matrix, and whether the sample is a mixture or a pure compound, and to clearly identify the data that need to be obtained. This will influence decisions on

- type of instrument required (e.g., do you require exact mass data – high accuracy and resolution, or selected reaction monitoring (SRM) for a complex sample),
- method of presentation to the mass spectrometer (e.g., solids probe, GC, or LC),
- most appropriate method of ionization, and finally
- the scan mode to be used.

In this section, we will look briefly at factors to be considered in selecting an ionization method, the choice of a scan mode, how mass spectral data can be used to identify an unknown compound or a known compound, and the factors to be considered in setting up a quantitative mass spectral assay.

9.10.4.1 Spoilt for Choice – Which Ionization Method to Choose?

The general range of application for each of the ionization methods described above (Section 9.10.2.2) is illustrated in **Figure 9**. For low-molecular-weight samples (<~1000 mass units) of moderate polarity, there will be several options. If analyte identification is required, then GC/MS will be a good option as the GC will permit high-resolution separation of mixture components and the use of EI will generate spectra that can be searched against large databases of EI spectra (e.g., NIST-Wiley). Also molecular weight can be confirmed by the use of CI. However, samples for GC/MS analysis will generally need to have polar functional groups (e.g., carboxylic acids, hydroxyls, amines) derivatized prior to analysis to improve volatility and thermal stability. There are a very large number of possibilities for derivatization and the reader is best referred to the very comprehensive literature that is available (e.g., Knapp,[21] Blau and Halket,[22] Halket and Zaikin,[23–27,118] Zaikin and Halket,[46,119]). Identification may be further aided by acquiring high-resolution, high mass accuracy data from which elemental formulas may be derived, particularly if constraints can be introduced with respect to the presence and number of particular elements (see Section 9.10.4.3.3).

If samples are to run without derivatization, then recourse may be had to the 'soft' liquid spray ionization processes, APCI, APPI, and ESI. ESI is the best choice for polar molecules (**Figure 9**) such as drugs and their metabolites and is by far the best choice of these three for peptides and proteins. At the relatively nonpolar end of the spectrum, APCI and APPI will be the preferred choice. However, APPI will have an advantage in terms of operability at low flow rates and its ready application to normal-phase chromatography and to lower polarity compounds than APCI. All three of these ionization methods produce molecular ions, and perhaps some adduct ions, yielding molecular weight information, but little fragmentation. In these cases, CID must be used in a tandem MS experiment to generate structurally informative fragmentation. Unfortunately, there is little in the

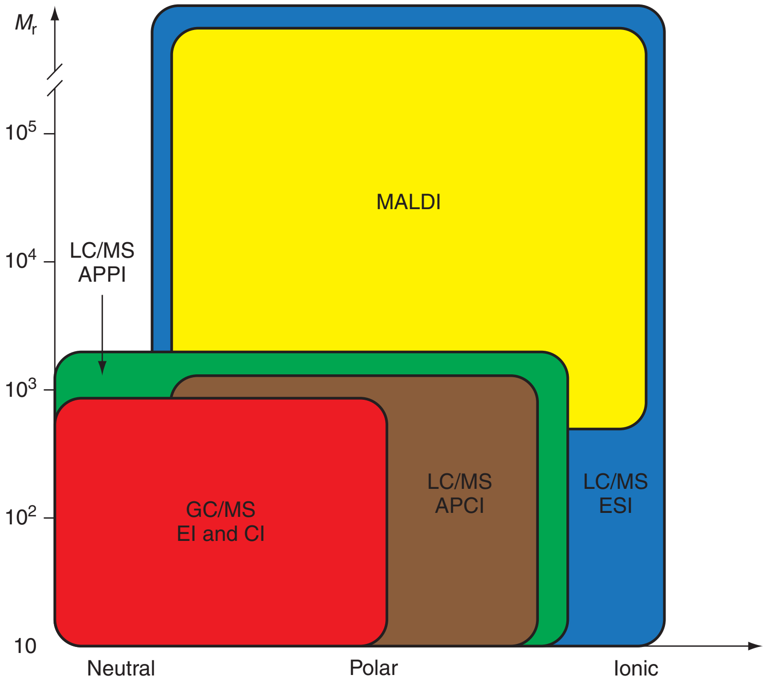


Figure 9 Approximate ranges of analyte polarity and size that may be suited to different ionization techniques. With respect to the surface desorption techniques, DESI and DART, they are comparable in their range of application to ESI and APCI, respectively.

way of MS/MS libraries available for searching, and structural elucidation will have to rely on user-generated libraries and on an interpretation of the spectra from first principles.[5,31–33] Interpretation will be greatly assisted by the acquisition of high mass accuracy, high mass resolution data to generate elemental formulas (Section 9.10.4.3.3). It is worth noting that despite the very widespread use of tandem mass spectrometers, there are no standard conditions for the acquisition of MS/MS data (see Hopley *et al.*,[120] for a discussion on the attempts to develop tandem MS/MS libraries). This is because the amount of energy that can be put into the fragmentation process is dependent on both instrument design and the experimental conditions. For beam instruments (tandem-in-space), the variable experimental conditions will include the energy of the ions entering the collision cell, the collision gas (e.g., Ar, He, N_2), the pressure of the collision gas, and the dimensions of the collision cell. Similarly, for trap-type instruments, the extent of fragmentation will be dependent on the collision gas, energy of the ions, type of fragmentation (CID, PID, ECD, EDD, ETD), and duration of the process. This is in contrast to EI-generated spectra that are normally acquired with electrons of 70 eV energy. The induced EI fragmentation is readily reproduced across all brands of MS instruments with some variation in the intensity of fragment ions.

In summary, where samples require chromatography prior to MS, the separation technique as well as the size and polarity of the analyte will influence which ionization technique will be most appropriate. Thus GC/MS will use EI or CI and LC/MS will use ESI, APCI, APPI, or a combined source (e.g., APCI/APPI or ESI/APPI) to ionize the chromatographic eluant. In the case of LC/MS, the most effective form of ionization may not be easily predicted and some experimentation may be required.

For high-molecular-weight samples, most commonly proteins and peptides, but also polysaccharides and synthetic polymers, the choice of an ionization method will be limited to ESI (Section 9.10.2.2.4) and/or MALDI (Section 9.10.2.2.7) (see also **Table 5** for a comparison of ESI and MALDI). As mention above, MALDI is a solid-phase-based ionization technique and ESI is a flow-based liquid technique. Both readily generate

Table 5 Comparison of ESI-MS with MALDI-MS

Advantages of ESI-MS	Advantages of MALDI-MS
• Typical sensitivity in the range of femtomole to low picomole or attomole concentration. Best sensitivity combining capillary LC and nano-ESI. • Can be readily interfaced to LC outlet, permitting multidimensional chromatography (MuDPIT analysis of whole proteomes). • Desalting and sample cleanup can be performed online. • Soft ionization with little fragmentation; however, labile posttranslational modifications such as phosphates are often lost. • Soft ionization permits observation of noncovalently bound protein complexes. • Multiple charging of proteins and peptides permits analysis of high-mass ions on low-mass range analyzers. • Multicharged ions fragment more efficiently in CID than singly charged ions	• Typical sensitivity in the range of femtomole to low picomole or attomole concentration. • Sample analysis is very rapid – a few seconds to analyze mixture of peptides. • Sample can be stored and reanalyzed at leisure. • Sample may be further purified *in situ* if required. • Soft ionization with little fragmentation. Phosphate groups may be retained at very low laser power. • Proteins and peptides usually ionized with a single charge simplifying interpretation. • High practical mass limit but at low-resolution.
Disadvantages of ESI-MS	Disadvantages of MALDI-MS
• Sample injected onto LC column, or directly injected into source, is totally consumed. • Multiple charging of proteins and peptides complicates interpretation of MS and MS/MS data. • Mixture analysis requires use of LC interface to minimize problems of ion suppression. • Gradient chromatography of a single sample can require 10 s or minutes to hours to complete.	• Intense matrix background below ~800 Da. • Can only be interfaced with LC in an off-line mode. • Cannot analyze noncovalently bound protein complexes. • High-throughput productivity requires automation of sample preparation. • Reflectron required for good mass resolution.

These are complementary techniques and while many analytes, including proteins and peptides, may be equally well ionized by either method, some will only be ionizable by ESI and not MALDI and *vice versa*.

gas-phase ions, with MALDI-generated ions being singly charged and ESI-generated ions carrying multiple charges. Samples for ESI are readily, and conveniently, analyzed in an online mode accepting the separated analytes from an interfaced LC column. However, this means that the MS analysis must be completed in the time it takes to elute each individual component. In very complex samples, where separation is incomplete, analysis must be completed very quickly and it is possible that minor components, or poorly ionized components, may not be analyzed at all. It should be noted that capillary LC while yielding enhanced chromatographic separation also results in narrower chromatographic peaks and a requirement for even faster MS analysis particularly where quantification is desired. MALDI, on the other hand, is not readily interfaced with chromatography and is usually performed off-line in conjunction with an automated plate spotter, yielding highly reproducible mixtures of sample and matrix and consequently very consistent high-quality MALDI spectra. However, the size of the fractions collected compromises, to some extent, the chromatographic separation but this disadvantage is offset by the ability to archive the samples stored on the MALDI target so that the MS analysis can be repeated at leisure. When coupled with an automated off-line plate spotter, MALDI is faster than ESI and is often used for high-throughout proteomic analysis.

Spectra generated by ESI are more complex than MALDI because of the multicharging phenomena. However, multiply charged ions are more amenable to CID because the coulombic repulsion of like charges aids the fragmentation process. CID of MALDI-generated singly charged ions lack this advantage. As the size of the analyte ions increases, the efficiency of CID diminishes, as there is a corresponding greater capacity for the ion to absorb and redistribute the impact energy. Nevertheless, this has proven to be a very effective method of amino acid sequencing for 'bottom-up' proteomics on peptides (typically 800–5000 Da) from trypsin digests.

In general, MALDI and ESI are of comparable sensitivity (femtomolar to low picomolar levels of peptides/proteins); however, it is impossible to make definitive comparisons for two reasons. First, there have been, and continue to be, technological advances improving the sensitivity of both techniques whereby, for example, the use of special hydrophobic surfaces on MALDI targets has been matched by the development of nano-ESI. Second, it has been observed that in complex proteomic analyses, perhaps only 30–50% of all proteins are adequately ionized by both ESI and MALDI, with the remainder being best ionized by either ESI or MALDI alone. Thus there is a good case to be made for the use of both techniques in a comprehensive proteomic analysis (**Table 5**).

9.10.4.2 MS Scan Modes

9.10.4.2.1 Single MS analyzer (nontrapping) scan modes

In the case of beam-type mass analyzers (B, BE, Q, ToF, and hybrids), the analyzer is commonly operated to scan over a defined mass range, generating a mass spectrum of all ions generated in the ion source. This is often referred to as a full scan (**Figure 10(a)**). Alternatively, the Q, B, or BE analyzers can be set to pass a selected ion or a selected series of ions (**Figure 10(b)**). This is known as SIM. This scan mode is commonly used for single Q analyzers (Section 9.10.2.3.3) because they can be rapidly switched between different ions over a large mass range.

The SIM mode is designed to enhance the sensitivity of an assay by concentrating the analyzer time onto only the ions of interest. For example, if instead of scanning a mass range of 500, the analyzer is set to monitor just 5 m/z values, then the number of ions counted in each of those channels will be 100 times that observed in the scanning mode. This improvement in ion statistics translates directly into improved sensitivity; however, this must be offset against the loss of a great deal of analytical information. It should also be noted that SIM is inappropriate for ToF analyzers because instead of scanning, they sample the entire mass range at any one time (Section 9.10.2.3.7). However, postacquisition processing of ToF data can be used to extract the time-based intensity trace for any of the ions in the mass range monitored. SIM is also problematic for B analyzers except over a narrow mass range; however, a double-focusing instrument (BE or EB) (Section 9.10.2.3.2) does have the possibility to further enhance selectivity by performing SIM with high mass resolution, as does the more common ToF analyzer.

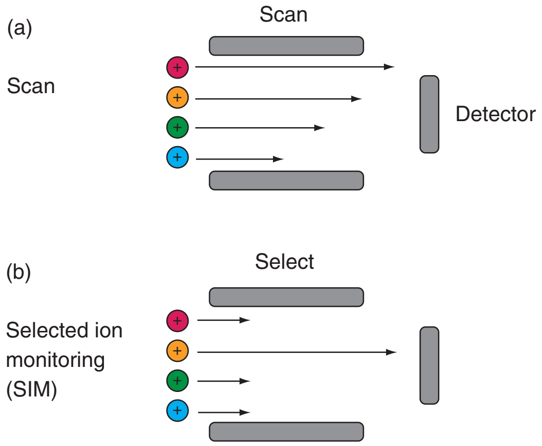


Figure 10 Scan modes for a single beam-type analyzer (e.g., Q, B, E). (a) Full scan. (b) Selected ion monitoring scan, commonly used in quantitative work to improve assay sensitivity.

9.10.4.2.2 Tandem MS scan modes

There are five main scan modes possible using MS/MS and these will be described and illustrated using the QqQ as an example (**Figures** 7(a)–7(e)). Double-focusing BE or EB instruments are also capable of tandem MS using linked scans to monitor ion transitions; however, they suffer from the disadvantage of having to either select ions with low mass resolution or detect ions with low mass resolution. Consequently, these instruments are rarely used today for tandem MS (see Gross,[33] for further discussion).

The first scan type to consider is the full scan of all the ions generated in the ion source using no mass selection. This is done by setting both Q_1 and Q_2 to pass all ions (RF-only mode) through to Q_3, which is then scanned in the normal way (**Figure** 7(a)) to generate a spectrum of ions present in the source.

The product ion scan entails the mass selection of a precursor ion in the first stage (Q_1), fragmentation (CID or ETD) in the collision cell, and then mass analysis of all resultant fragment masses in the second stage of mass analysis (Q_3) (**Figure** 7(b)). This experiment can be performed by beam (tandem-in-space) or trap (tandem-in-time) instruments. It is commonly performed to identify transitions used for quantification by tandem MS or as part of an exercise in structural elucidation.

In the precursor ion scan, the first mass analyzer (Q_1) sequentially scans all precursor ions into the collision cell (**Figure** 7(c)) for fragmentation. The second analyzer (Q_3) is then set to transmit a single specified ion product. The resulting mass spectrum is then a record of all the precursor ions that give rise to the specified common product ion, such as, for example, the metabolites of a particular drug, or class of compounds, which can be fragmented to a common structural moiety. The precursor ion scan can be carried out only with tandem-in-space instruments.

For the neutral loss scan, the first mass analyzer (Q_1) scans all the masses (**Figure** 7(d)). The second mass analyzer (Q_3) also scans, but at a fixed offset from the first mass analyzer. This offset corresponds to a neutral loss that is commonly observed for a particular class of compounds; for example, the loss of 44 u (CO_2) from $[M-H]^-$ ions will be indicative of carboxylic acids. Alkyl loss (C_nH_{2n+1}) will be seen in the loss of 15, 29, or 43, etc. and the loss of 18 u (H_2O) will be indicative of a primary alcohol. A comprehensive table of common neutral fragments may be found in McLafferty and Tureček.[32] The mass spectrum is then a record of all precursor ions that lose the specified neutral fragment. Again, neutral loss scans cannot be performed with trap-type MS instruments or with ToF analyzers. However, postacquisition analysis software can be used to search for the specified neutral loss.

SRM is a version of the product ion scan and is used in experiments designed to identify and quantify targeted analytes (**Figure** 7(e)). Both mass analyzers, Q_1 and Q_3, are set to pass predetermined masses. These correspond, first, to a specific precursor ion (Q_1) and, second, to a fragmentation or transition (Q_3) that is characteristic of the selected analyte. Typically, the MS will be rapidly switched between several sets of such transitions representing different analytes, internal standards (ISs), or possibly an alternative confirmatory transition. Thus SRM adds a considerable degree of selectivity to an assay in that the fragmentation monitored is specific to the target analyte and is unlikely to also occur with any background chemical noise that is also selected by Q_1. If the first mass analyzer can be operated with high mass accuracy and high mass resolution, this will further enhance selectivity. Sensitivity in SRM is also concomitantly improved, because by removing all the background chemical noise, the S/N of the monitored ion is improved. SRM may be performed by both tandem-in-space and trap-type instruments.

9.10.4.3 Identification – Unknown Small Molecules

Identification or structural elucidation of an unknown compound is one of the most challenging tasks that can be undertaken by an analytical chemist. Where the analyst has milligram or more amounts of the unknown, MS is often used in conjunction with other techniques such as nuclear magnetic resonance (NMR) and IR spectroscopy. However, when there are only limited quantities of sample, the sensitivity of MS makes this the technique of choice for assembling structural information and, where no definitive conclusion can be reached on the mass spectral data alone, serves to limit the search to a particular class of chemicals or set of isomers.

9.10.4.3.1 An LC/MS approach

A useful first step is to analyze the sample using LC interfaced to ESI or APCI on a tandem mass spectrometer capable of accurate mass measurement at high resolution. The LC will serve to separate the unknown away from the matrix components and will reduce the potential for ion suppression. A short linear gradient of acetonitrile against 0.05 mol l^{-1} ammonium acetate on a reverse-phase C-18 column represents a good starting point (see, e.g., Eckers *et al.*,[121] Arthur *et al.*,[122] Tozuka *et al.*,[123] and Wolff *et al.*[124]).

Spectra obtained from ESI or APCI will generally yield molecular ions of the type $[M+H]^+$ and little, if any, fragmentation. However, fragmentation can be induced by CID (Section 9.10.3.2.1) followed by an MS/MS analysis of the product ions (Section 9.10.4.2.2). High mass accuracy/high mass resolution measurement of the molecular ion can, with appropriate constraints, generate an elemental formula (see discussion in Section 9.10.4.3.3).

The MS/MS analysis of the molecular ion will also yield a structurally informative set of fragment ions. Again, if exact mass data can be obtained for these, a further set of elemental formulas may be obtained. Unfortunately, because there are only limited collections of library spectra generated by MS/MS (see Section 9.10.3.2), the experimenter will generally have to resort directly to a first-principles interpretation.

Considerable structural information is to be found in the fragmentation patterns of gas-phase ions. Although this discussion is often held in the context of EI-generated spectra, it is important to remember that the reactions of gas-phase ions are not dependent on their method of formation but rather on their intrinsic structural properties and their internal energy. Thus structural information can be obtained from fragmentation that is induced by a high-energy ionization process, such as EI, as well as from collisionally induced fragmentation of a $[M+H]^+$ ion that may have been generated by a ‘soft’ ionization process. Instructive examples of structural elucidation of drug metabolites using MS^n fragmentation trees and exact mass data are described by Eckers *et al.*,[121] Arthur *et al.*,[122] Tozuka *et al.*,[123] and Wolff *et al.*[124] and of complex lipids are described by Hsu and Turk.[125,126] In brief, the drugs and their metabolites are exhaustively fragmented and the fragments compared to locate the biologically induced structural changes – for example, oxidations, cleavages, alkylations, and conjugations. The relationship between the product ions and the precursor ions, and their elemental composition are key elements in assembling the structural features of the unknowns. One must also be aware of the possibility of isomers and the use of an appropriate separation technique may be required in addition to the MS and MS^n data.

9.10.4.3.2 GC/MS approach

If the unknown is sufficiently volatile or can be made volatile by derivatization, then the LC/MS approach can be complemented by the use of GC/MS. A useful starting point in this regard is to make a trimethylsilyl derivative. Silylation is applicable to a wide range of nonsterically hindered functional groups, including alcohols, phenols, thiols, amines, oximes, and carboxylic acids (**Table 6**).[22,127] The dried sample (1 μg

Table 6 Trimethylsilylation reagents

Silylation reagent	*Abbreviation*
Strong silyl donors	
N,*O*-Bis(trimethylsilyl)acetamide	BSA
N,*O*-Bis(trimethylsilyl)trifluoroacetamide	BSTFA
N-Methyl-*N*-trimethylsilyltrifluoroacetamide	MSTFA
Moderate strength silyl donors	
Trimethylsilyldiethylamine	TMSDEA
Weak selective donors	
Trimethylimidazole (hydroxyl groups)	TMSIM
Hexamethyldisilazane (hydroxyl groups)	HMDS
Trimethylchlorosilane (hydroxyl groups)	TMCS

A range of reagents, with differing degrees of reactivity, are commonly available to make trimethylsilyl (TMS) derivatives. The silylating potential can be increased by the choice of an appropriate solvent (e.g., pyridine, DMF, acetonitrile) or by the addition of a catalyst (e.g., 1–20% TMCS).[22,127]

maximum) is dissolved in pyridine (10 μl) to which is added an equal volume of a strong silylation reagent such as *N*,*O*-bis(trimethylsilyl)trifluoroacetamide (BSTFA) plus 1% trimethylchlorosilane (TMCS) or *N*-methyl-*N*-trimethylsilyltrifluoracetamide (MSTFA). It is important that the silylation reagent is present in an excess of at least 2:1 molar ratio to active hydrogens and that the sample is dry. Unhindered moieties will be quickly silylated but derivatization times and the need for heat vary widely depending on the degree of steric hindrance but unless determined otherwise, heating at 70 °C for 20–30 min will ensure the reaction is driven to completion for most active hydrogens. A variation on this approach, widely used in metabolomics experiments where a large range of chemical classes need to be derivatized, is to initially protect any carbonyl moieties by methyloximation (10 μl of a 20 mg ml^{-1} solution of methoxyamine HCl in pyridine at 40 °C for 90 min) (see, e.g., Fiehn[128]). The silylation reagent can then be directly added at the end of the methoximation reaction. After silylation, the sample may be injected onto a general purpose capillary column such as one coated with 5% phenyl-95% methylpolysiloxane. The column temperature is then increased to drive off the less volatile components. Although silylation is a good general derivatization reaction, it should be remembered that there are many other possibilities available, especially if a particular class of analyte is being assayed (e.g., Knapp,[21] Blau and Halket,[22] Halket and Zaikin,[23–37] Zaikin and Halket[46,119]). If assay sensitivity is important then derivatization with electron-capturing groups for ECI should be considered (Section 9.10.2.2.2).

The advantage of GC/MS over LC/MS is that extensive libraries of EI data are available for searching (see Section 9.10.4.3.4). Where necessary, this can be complemented by molecular weight data from CI. Library identification then requires confirmation by comparing the column retention time and MS and MS^n data with that of a standard. If no library match is found, then a similar process of determining elemental formulas (see Section 9.10.4.3.3) and interpretation of the fragmentation data from first principles must be followed (see Section 9.10.4.3.4).

A more sophisticated method of comparing retention times, and one that is applicable across different column phases and temperature programs, is by way of retention time indices or Kovát's indices (KIs).[129] The KI for a particular analyte is calculated against a homologous series of *n*-alkanes, coinjected with the sample.

$$\mathrm{KI} = 100n + \frac{100(t_x - t_n)}{(t_{n+1} - t_n)} \tag{55}$$

where n is the number of carbon atoms in the *n*-alkane standard that elutes immediately prior to the analyte of interest, t_x is the retention time of the analyte, t_n is the retention time of the *n*-alkane standard that elutes immediately prior to the analyte of interest, and t_{n+1} is the retention time of the *n*-alkane standard that elutes immediately after the analyte of interest.

Thus in metabolomic experiments, the KI for individual metabolites is an important piece of confirmatory information where the mass spectral differences between isomers are minimal or nonexistent (**Figure 11**). Unfortunately, no such comparative set of indices are available for LC.

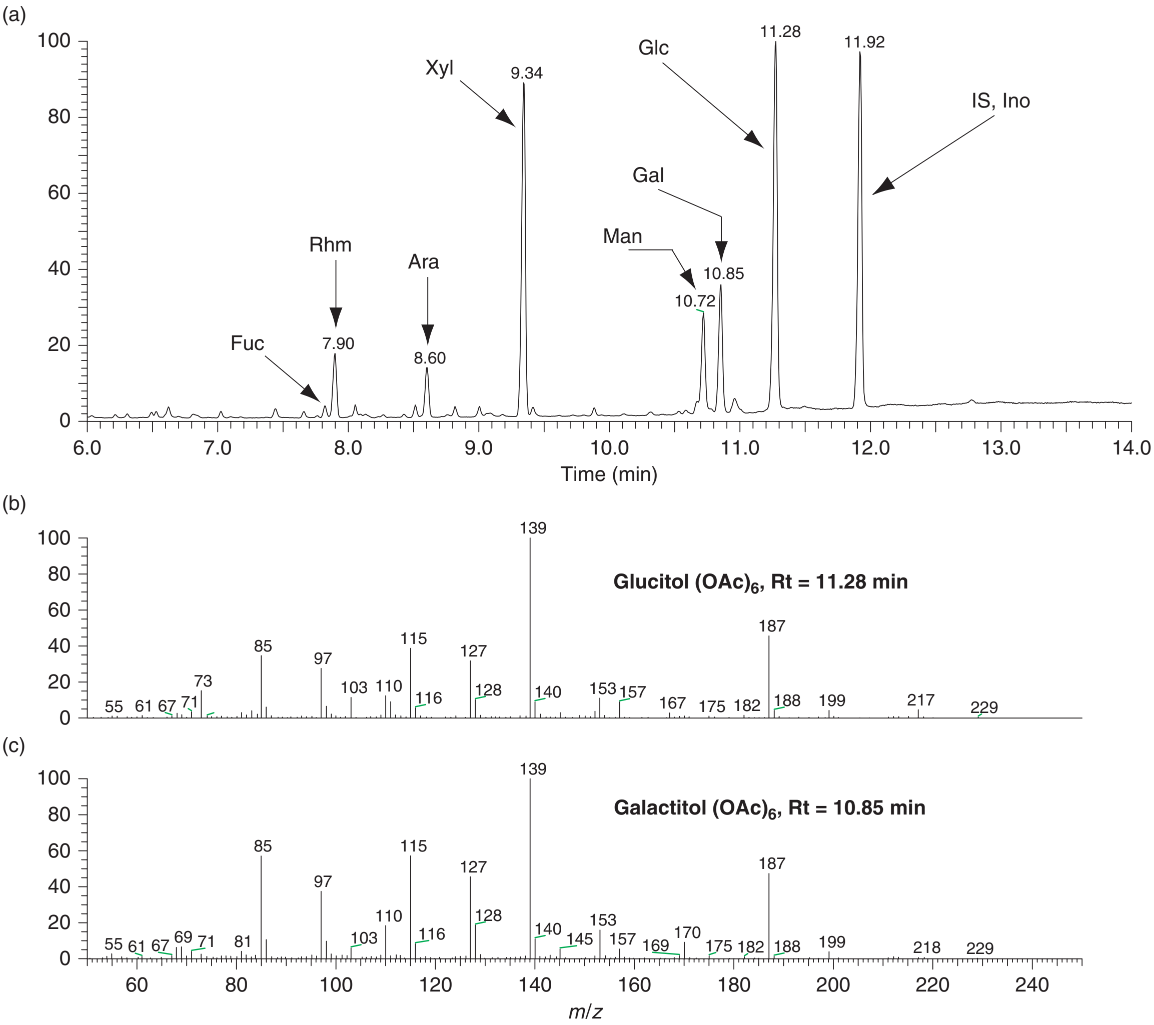


Figure 11 GC/MS assay of alditol hexa-acetates quantified against inositol internal standard (IS). (a) In the chromatogram shown here the monosaccharides making up a plant cell wall are being quantified as their alditol acetates, using inositol (Ino) as the (IS). The GC separation of these reduced sugars is essential for their identification. The mass spectra of the alditol acetates of the hexoses, glucose (Glc) (b), galactose (Gal) (c), and mannose (Man), are essentially identical, as are the mass spectra of the alditol acetates of the pentoses, xylose (Xyl) and arabinose (Ara), and the deoxysugars, rhamnose (Rhm) and fucose (Fuc).

9.10.4.3.3 Determination of elemental formula

Accurate mass data can be a significant aid in identifying compounds as it can yield the elemental composition of the molecular ion and the associated fragment ions.[130] The theoretical mass of a compound can be readily calculated from tables of elemental masses (**Table 7**) and there are software packages to automate this procedure. Although it is possible to accurately measure mass at low-resolution, the analyst runs the risk of including extraneous isobaric ions in the measurement. This will result in the mass measurement being skewed or shifted by the interfering ion(s) producing an erroneous elemental formula.[131] Note that as higher mass ions are analyzed, the number of possible elemental formula consistent with the measured mass also increases along with the requirement to resolve away isobaric ions. This situation is nicely summarized by the editor of the *JASMS*[132] in the journal's guidance on the use of accurate mass data:

> When valence rules and candidate compositions encompassing $C_{0\text{-}100}$, $H_{3\text{-}74}$, $O_{0\text{-}4}$, and $N_{0\text{-}4}$ are considered at nominal parent mass 118, there are no candidate formulae closer together than 34 ppm. At nominal parent mass 500, there are

Table 7 Stable isotopic masses and abundances[a]

Isotope	*Mass (u)*	*Natural abundance (%)*
^{1}H	1.007 825 031 9(6)[b]	99.988 5(70)[b]
^{2}H	2.014 101 777 9(6)	0.011 5(70)
^{12}C	12 (exactly, by definition)[c]	98.93(8)
^{13}C	13.003 354 838(5)	1.07(8)
^{14}N	14.003 074 007 4(18)	99.636(20)
^{15}N	15.000 108 973(12)	0.364(20)
^{16}O	15.994 914 622 3(25)	99.757(16)
^{17}O	16.999 131 50(22)	0.038(1)
^{18}O	17.999 160 4(9)	0.205(14)
^{19}F	18.998 403 20(7)	100
^{23}Na	22.989 769 66(26)	100
^{28}Si	27.976 926 49(22)	92.223(19)
^{29}Si	28.976 494 68(22)	4.685(8)
^{30}Si	29.973 770 18(22)	3.092(11)
^{31}P	30.973 761 49(27)	100
^{32}S	31.972 070 73(15)	94.99(26)
^{33}S	32.971 458 54(15)	0.75(2)
^{34}S	33.967 866 87(14)	4.25(24)
^{36}S	35.967 080 88(25)	0.01(1)
^{35}Cl	34.968 852 71(4)	75.76(10)
^{37}Cl	36.965 902 60(5)	24.24(10)
^{39}K	38.963 7069(3)	93.258 1(44)
^{40}K	39.963 998 67(29)	0.011 7(1)
^{41}K	40.961 825 97(28)	6.730 2(44)
^{79}Br	78.918 337 9(20)	50.69(7)
^{81}Br	80.916 291(3)	49.31(7)
Electron (e^-)[d]	5.485 799 09(27) × 10^{-4}	–
Proton (H^+)[d]	1.007 276 452	–

[a] Data are derived from an IUPAC Technical Report by deLaeter *et al.*[133]

[b] The (±) uncertainty of the measurement is indicated in parentheseis.

[c] In mass spectrometry, the unit of measurement is the unified atomic mass unit (u), which is defined as 1/12 the mass of a ^{12}C atom ($1\,u = 1.660\,540\,29 \times 10^{-27}$ kg).

[d] The very high mass resolution and accuracy that are available from BE, EB, FTICR, Orbitraps, and ToF analyzers mean that calculations of exact masses need to also account for electrons in the analyte ion.[134,135] Thus, for example, an electron is lost in the formation of a radical cation ($M^{+\cdot}$) and a protonated molecule ($[M+H]^+$) gains a proton not a hydrogen atom ($\Delta m = 1\,e^-$). An electron is gained in the formation of a radical anion ($M^{-\cdot}$) and a deprotonated molecule ($[M-H]^-$) loses a proton not a hydrogen atom. While this error will be insignificant for large molecules such as proteins, for small molecules, the error can be as large as several ppm. For example, ignoring the mass of three electrons in triply charged GluFib, $[M+3H]^{3+}$, leads to an error of 1 ppm in the exact mass calculation.[134]

These are the isotopes most commonly encountered in natural product chemistry. Silica is encountered as a generic derivative for GC (e.g., trimethylsilyl and *tert*-butyldimethylsilyl derivatives).

five compositions that have a neighbouring candidate less than 5 ppm away. Using $C_{0\text{-}100}$, $H_{25\text{-}110}$, $O_{0\text{-}15}$, and $N_{0\text{-}15}$ at mass 750.4, there are 626 candidate formulae that have a neighbouring possibility less than 5 ppm away. Thus, for a measurement at m/z 118, an error of only 34 ppm uniquely defines a particular formula. At m/z 750.4 an error of 0.018 ppm would be required to eliminate all extraneous possibilities.

In practice, it is important to be able to restrict the type and number of elements in any possible formula so as to improve the degree of confidence in selecting the most appropriate formula and to eliminate impossible or unlikely combinations of elements. Further information to help with this may be gleaned from an examination of the isotope pattern of the molecular ion. Most of the elements present in organic compounds (C, H, N, O, P, and S; Si must obviously be included if silyl derivatives were used for GC/MS) have two or more stable isotopes (**Table 7**). This information can be used, for example, to estimate the number of carbons present in an unknown from knowing that the ^{13}C isotope has an abundance equal to 1.1% that of the ^{12}C isotope. Thus an ion containing 10 carbons will have a ^{13}C abundance ratio of 11% and by extension an ion containing 20 carbons will have a ^{13}C abundance ratio of 22%. A number of other elements such as chlorine (^{35}Cl:^{37}Cl, 1:3), bromine (^{79}Br:^{81}Br, 1:1), sulfur (^{32}S:^{33}S:^{34}S: 100:1:5), and silicon (^{28}Si:^{29}Si:^{30}Si, 100:5:3) also have distinctive isotope patterns, recognition of which will aid in restricting the possible elemental formula of the unknown (for further discussion, see McLafferty and Tureček,[32] Gross,[33] and Watson and Sparkman[5]). Algorithms such as that described by Pickup and McPherson[136] and Hsu[137] can be used to model isotope distributions in elemental formulas and the comparison between the experimental and theoretical isotopic distribution can be assigned a goodness-of-fit score.

Further constraints may be identified by application of the nitrogen rule,[5,32,33] which states that a compound containing the common elements, C, H, O, S, Si, P, and the halogens, will have an odd nominal molecular weight if it contains an odd number of nitrogens. A compound with zero or an even number of nitrogen atoms will have an even nominal molecular weight. This is because every element with an odd mass has an odd valence and every element with an even mass has an even valence, with nitrogen being an exception, having an odd valence and an even mass.

In addition, a consideration of the valency of the constituent elements leads to the derivation of a general algorithm for the number of rings and double bonds (R + DB) present in an ion.[32,33,138] Thus,
for the elemental formula $C_cH_bN_nO_o$

$$(\mathrm{R} + \mathrm{DB}) = c - 0.5b + 0.5n + 1 \tag{56}$$

Other monovalent elements (F, Cl, Br, and I) are counted as hydrogens, trivalent elements (P) are counted as nitrogen, and tetravalent elements (Si) are included with carbon. For chemically possible formulae, $r + db >$ -1.5. Odd-electron ions ($M^{+\cdot}$) will have an integer value and even-electron ions will have 0.5 $r + db$ more than expected, so round up to next lowest integer.[32,33] By way of example, Kind and Fiehn[139] have described an integrated application of accurate mass data to metabolite identification, constrained by isotope abundance information and valence rules, in addition to the KI (Section 9.10.4.3.2).

In the ideal case, the high mass accuracy and high mass resolution determination of the molecular ion will yield an unambiguous formula but this says nothing about the connectivity of the constituent atoms. For the trivial case of C_2H_6O, the exact mass is 46.041 864 8 but this does not distinguish ethanol (CH_3CH_2OH) from dimethyl ether (CH_3OCH_3). However, fragmentation occurs in a mostly predictable fashion and an examination of the molecular ion fragments will often reveal a distinctive 'fingerprint' including structurally diagnostic ions (m/z 31 for ethanol and m/z 29 for dimethyl ether).

9.10.4.3.4 Database searching and interpretation of fragmentation from first principles

For the EI spectra of unknowns, a very valuable first step toward identification is to perform a simple spectral comparison with an EI library. As noted above, EI spectra (Section 9.10.2.2.1) are highly reproducible and are not instrument dependent. The widely available NIST-Wiley Library, for example, contains several hundred thousand spectra. A satisfactory match of the unknown and reference spectra can be confirmed experimentally against a reference standard. However, it should be noted that the EI mass spectra of stereoisomers and geometric isomers are often very similar, exhibiting the same fragmentation pattern and similar abundances of fragments. This can be seen, for example, in the spectra of glucitol hexaacetate and galactitol hexaacetate (**Figures 11(b)** and (**c**)). In these cases, the ambiguity of the MS identification can be overcome by a comparison of the GC retention times (Rt) (**Figure 11(a)**) (the use of KIs is described in Section 9.10.4.3.2). Judicious selection of the phase coating the inside of the capillary column (a wide range of different chemistries and polarities are available) that is interfaced with the MS instrument will permit the separation of many of these stereoisomers and geometric isomers. An alternative approach is to make a chemical derivative that the MS can distinguish. For example, mass spectra of fatty acid methyl esters (FAMEs) containing two double bonds are essentially identical regardless of the location of the double bond; however, if instead a dimethyloxazoline derivative is made, the location of the double bond can be readily determined (**Figure 12**).

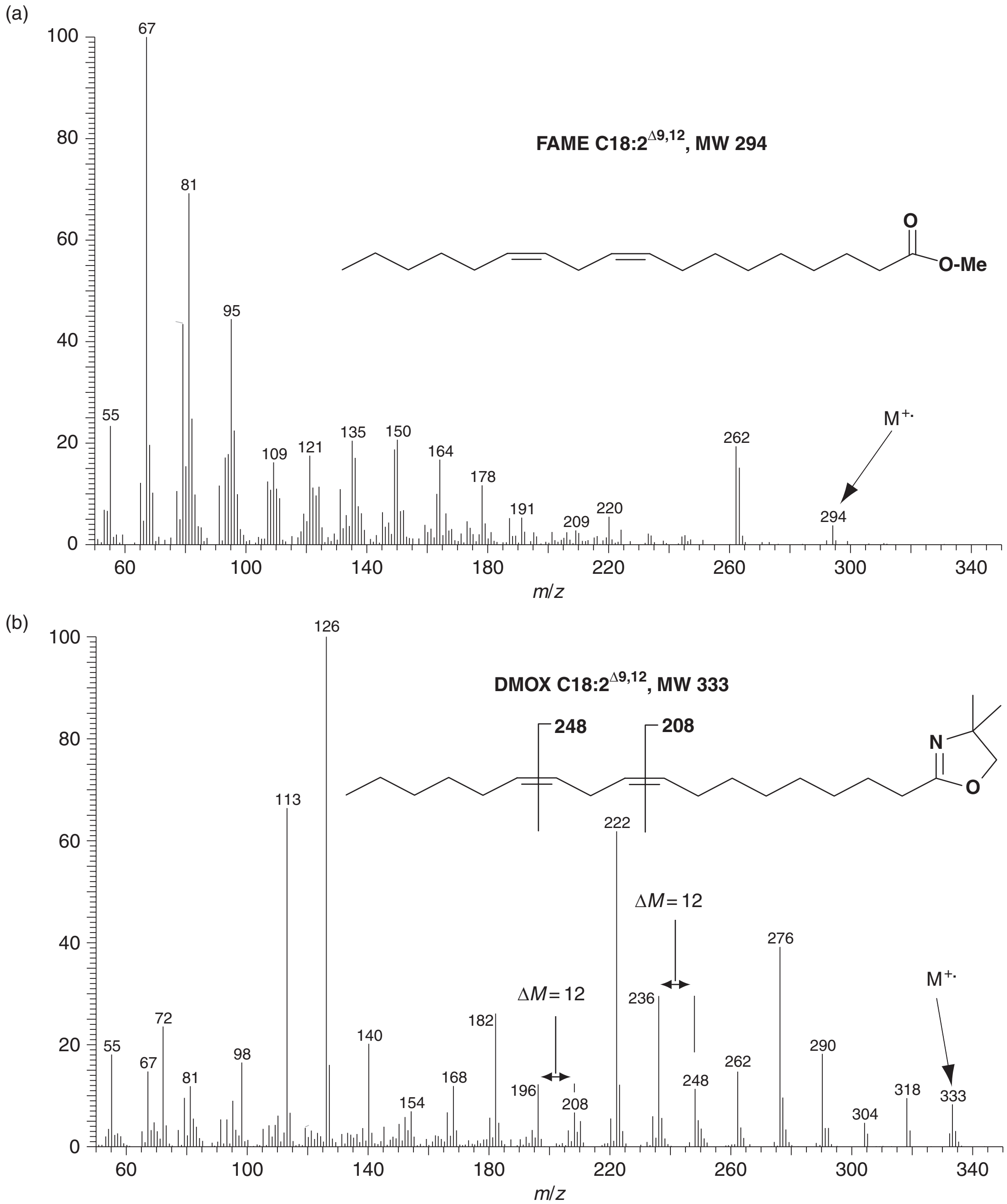


Figure 12 The mass spectrum of the fatty acid methyl ester (FAME) of linolenic acid (C18:2$^{\Delta 9,12}$) contains no readily discernable structural information beyond the molecular ion (a). However, the dimethyloxazoline (DMOX) derivative, in which the charge is retained by the heterocyclic ring, can undergo charge remote fragmentation yielding a mass spectrum from which the location of the double bonds, but not their geometry (*cis* versus *trans*), can be readily determined (b). The latter stereochemistry can usually be distinguished by the GC retention time on an appropriate column.

Where no satisfactory comparable spectra can be found by a database search, the more laborious process of interpreting the spectra from first principles must be attempted. As mention before, this process will be considerably aided if elemental compositions of the molecular ion and the EI fragments are available.

Much effort has been expended on providing rational mechanisms for fragmentation and these are well summarized by, for example, Budzikiewicz *et al.*,[31] McLafferty and Tureček,[32] Gross,[33] de Hoffmann and Stroobant[34], and Watson and Sparkman.[5] In EI, an odd-electron ion ($M^{+\cdot}$) is generated and the subsequent bond cleavages that follow result in the formation of the most stable cation with paired electrons (even-electron ion). The soft ionization techniques such as CI, ESI, APCI, and MALDI produce molecular species by the addition or abstraction of a proton, yielding an ion with an even number of electrons (e.g., $[M+H]^+$). These ions are more stable than radical cations and their fragmentation is more likely to reflect steric effects, so isomers with essentially identical EI spectra often give rise to different soft ionization spectra and may fragment differently following CID.[32]

A comprehensive description of these processes is beyond the scope of this chapter and the reader is referred to one of many texts on the interpretation of fragmentation and to tables of common neutral losses and of common ion series for particular classes of compounds (see, e.g., Budzikiewicz *et al.*,[31] McLafferty and Tureček,[32] Gross,[33] de Hoffmann and Stroobant,[34] Dass,[140] and Watson and Sparkman[5]).

As mentioned above, mass spectral interpretation will be greatly aided if high mass accuracy data at high mass resolution are available to determine the elemental formula of the unknown and its fragments. Also there is increasing use of gas-phase ion/molecule reactions that can be exploited for class and functional group identification.[141]

9.10.4.4 Criteria for Identification of a Known Compound

Forensic laboratories and regulatory authorities responsible for the quality of food, drugs, and environmental pollution are major users, directly and indirectly, of MS for the purpose of identification and quantification. The major question they face is: How much information is required to support their claim, within the specified confidence limits, for the presence of known specified compounds in their samples? This is not an easy question to answer and is usually dealt with by defining the core analytical technology and a set of minimal performance criteria for acceptable identification and by reserving the right to assess methods on a case-by-case basis. While there may be a general consensus on the broad issues of what is required for confirmation of identity, there is no general agreement on specifics and a number of different approaches and specific requirements are used around the world. By way of example, we will look at the requirements of two such regulatory authorities, the US Food and Drug Administration (FDA) and the European Union (EU), with respect to residues of banned substances, mostly veterinary drugs, in animal products. The identification criteria set by both these regulatory authorities are quite stringent and similar types of criteria are also required for other regulatory authorities and also by editors of research journals.

In addition to the mass spectral aspects of these assays, which are outlined below, there may also be extensive requirements to be met by the analyst with respect to compliance with good laboratory practice, which governs the operations of analytical laboratories and includes sampling regimes, assay validation procedures (e.g., limits of detection, limits of quantification, accuracy, reproducibility, and ruggedness), and laboratory accreditation (e.g., staff training, laboratory equipment, documentation, quality assurance, and quality control).[142–145]

9.10.4.4.1 FDA Guidance for Industry 118

In its Guidance for Industry 118,[144] the FDA requires that methods for confirmation of identity include the use of a

- comparison standard,
- chromatography interfaced to MS (GC/MS or LC/MS), and
- mass spectral matching.

The use of a standard is a fairly obvious requirement but where matrix effects alter either the chromatography or the spectrum, the authority will allow the use of a control extract spiked with the standard instead of using a pure standard. However, the analyst must then be able to demonstrate the absence of interference in a control extract containing no standard.

The FDA asks that the use of MS be combined with chromatography but specifications are only listed for GC/MS and LC/MS. The omission of interfaces such as CE/MS SFC/MS is a reflection of the conservative

Table 8 FDA criteria for mass spectral matching

MS scan	*Requirements*
Full scan	• At least three structurally specific ions that completely define the molecule are present above a specified level. • General correspondence between relative abundance of sample and standard ions (within the range of ±20%). • Prominent ions, not from analyte, can be explained.
SIM	• Relative abundance of three structurally specific ions of sample and standard should be within ±10%. • Relative abundance of four or more structurally specific ions of sample and standard should be within ±15%.
MS^n full scan	• All structurally specific ions present in standard spectra should be present in sample spectra. • General correspondence between relative abundance of sample and standard ions (within the range of ±20%). • Prominent ions, not from analyte, can be explained.
MS^n SRM	• If precursor ion is completely dissociated and only two structurally specific ions are monitored, the relative abundance of sample and standard ions should match within ±10%). • If three or more structurally specific ions are monitored, the relative abundance of sample and standard ions should match within ±20%).

Summary of FDA requirements for identifying animal drug residues.[144]

nature of regulatory authorities with respect to the unproven reliability of these techniques to robustly deliver reproducible chromatograms, not only on a day-to-day basis but also over an extended period of time. There is, however, flexibility in the type of chromatogram that may be used: total ion currents (TICs), reconstructed ion currents (RICs), SIM, and SRM are all acceptable with the provision that the retention times for the standard and the analyte should be within 2% for GC/MS and 5% for LC/MS.

With respect to mass spectral matching, the criteria for identification vary depending on the technique used for mass spectral data acquisition (see summary of requirements in Table 8). It is interesting to note that while the FDA does not rule out the use of exact mass measurements, it views these data as problematical as there are no generally accepted specific standards for their use. The problem here is that it is difficult to be definitive about the resolving power required, particularly, when analytes have masses greater than m/z 500. Clearly the resolving power and accuracy must be sufficient to exclude all reasonable alternative elemental compositions and they recommend that if exact mass measurements are to be used then multiple structurally specific ions should be measured.

9.10.4.4.2 EU performance of analytical methods

The EU takes a slightly different approach to the FDA in setting the criteria for identification (Tables 9–11) but agrees with the FDA in accepting only GC/MS and LC/MS methods and in their requirement for an analyte standard.[145] The EU tolerance for chromatographic performance is more stringent than the FDA, requiring GC and LC retention times for standards and samples to be within ±0.05 and ±2.5%, respectively. In addition to outlining a set of performance criteria for the different types of MS data (Tables 9 and 10), the EU uses a system of identification points to score the MS data (Table 11). Identification under this system is acceptable only if a certain number of identification points have been accumulated. So, for example, identification using GC/MS^2 for one precursor ion and two product ions will earn four identity points, and identification using GC/MS and LC/MS, monitoring two ions with each technique, will also accrue four identity points. This level of identification is deemed sufficient for identification of their Group A banned substances (veterinary drug residues in meat for human consumption). It is interesting to note that the EU has set no qualifications around the acceptability of exact mass data, save that resolution should be greater than 10 000 (10% valley) for the entire mass

Table 9 EU criteria for mass spectral matching

MS scan	*Requirements*
Full scan	
	• A minimum of four diagnostic ions (molecular ion, adducts, fragments, and isotope ions) with an intensity >10% in the standard must be observed in the sample. • The molecular ion must be included if the relative intensity is ≥10% of the base peak. • The relative intensities of the sample diagnostic ions are required to match those of the standard, within specified tolerances (**Table 10**).
SIM	
	• The molecular ion shall be one of the selected diagnostic ions. • The S/N for each diagnostic ion shall be ≥3:1. • A minimum of four identity points (Group A, banned substances) (**Table 11**) must be accumulated and these must be derived from at least one ion ratio measurement, meet the specified intensity tolerances (**Table 10**), and no more than three techniques can be used to achieve the minimum number of identity points.

Summary of EU requirements for identifying animal drug residues.[145]

Table 10 EU maximum permitted tolerances for relative ion intensities

Relative intensity (% base peak)	*GC–EI–MS (%)*	*GC/CI-MS, GC/MS^n, LC/MS, LC/MS^n (%)*
>50	±10	±20
>20–50	±15	±25
>10–20	±20	±30
≤10	±50	±50

Table 11 EU identification points earned

MS technique	*Identity points/ion*
Low-resolution (LR)-MS	1.0
LR-MS^{n}precursor ion	1.0
LR-MS^{n}product ions	1.5
High-resolution (HR)-MS	2.0
HR-MS^{n}precursor ion	2.0
HR-MS^{n}product ions	2.5

A minimum of four identity points are required to confirm the presence of a Group A substance (banned veterinary products).

range and indeed it assigns two identity points for each measured ion (**Table 11**). This lack of qualification ignores the fact that the number of candidate elemental compositions increases markedly with mass (Section 9.10.4.3.3). This point is well illustrated by Nielen *et al.*,[146] who using the anabolic steroid, stanozolol, and the β-agonist Clenbuterol-R, as models, demonstrate the current EU mass accuracy criteria can yield false negative results.

9.10.4.5 Quantification

As noted by the Reverend Stephen Hales as long ago as 1727, scientific insight into the processes of nature can be obtained only through the discipline of measurement.

> Since we are assured that the all-wise Creator has observed the most exact proportions, of number, weight and measure, in the make of all things, the most likely way therefore, to get any insight into the nature of those parts of the creation, which come within our observation, must in all reason be to number, weigh and measure.
>
> *Vegetable Staticks*, Stephen Hales 1977–1761

It should be no surprise therefore that mass spectrometers are most commonly used for quantification. In addition to quantitative applications by regulatory authorities and industry (e.g., petrochemical, pharmaceutical, food, forensic, and environmental areas), the postgenomic era has witnessed an explosion in the use of mass spectrometers to determine and quantify gene function as exhibited in the gene products – proteins and metabolites. This has given rise to two new and unique areas of endeavor, proteomics and metabolomics. Both of these aim to analyze the complete respective sets of proteins or metabolites, present in a cell, tissue or organism at any one time point. The importance of these analyses lies in the fact that they provide information that is not directly attainable from the genomic sequence, including, for example, insight into developmental processes and responses to environmental stimuli and pathogens, at the cellular level. These data can then be linked to genomic and transcriptomic data to present the scientist with a holistic or systems biology view of an organism (see, e.g., Weckwerth *et al.*[147] and Trauger *et al.*[148]).

The challenge of proteiomic and metabolomic analysis lies in the complexity (e.g., PTMs of proteins and the array of different chemical classes of metabolites), and the large range of concentrations, of the components present in the sample and in the need for high-throughput and reproducible methodologies for their identification and quantification. A detailed discussion of protein and peptide analysis by MS may be found elsewhere in this volume (see Chapter 9.12).

9.10.4.5.1 Components of an MS-based metabolite assay

Although techniques such as NMR and IR spectroscopy have found some utility in metabolite analysis, the most common approach has been to draw upon the versatility, speed, and high degree of specificity and sensitivity inherent in tandem MS.[149] In the case of complex samples, this specificity and sensitivity can be enhanced by interfacing the mass spectrometer to some form of high-resolution chromatography such as GC, nano-LC, or CE.

Using MS for quantification is no different in principle to using any other detector, and generally encompasses sample quenching, homogenization to break down tissue and cell structure, extraction, separation, sample analysis, calibration standard analysis, and finally data processing (**Figure 13**).[143] Also included will be assay validation–determining the limits of quantitation, selectivity, accuracy, precision, and linear dynamic range of the assay (see FDA Guidance for Industry,[144] Bioanalytical Method Validation,[150] Pritchard and Barwick,[151] and Boyd *et al.*,[143] for a detailed discussion on validation and quality assurance in analytical chemistry). In addition, a mass spectral assay will include specific consideration of the following items:

- Optimizing the quenching, extraction, and purification processes, being cognisant of reagents that may be incompatible with MS (e.g., nonvolatile salts and detergents; see also discussion on chemical noise and contamination in Section 9.10.4.5.8).[152]
- Selecting the method of sample introduction to the MS, for example, GC/MS (need to consider analyte derivatization; Section 9.10.4.3.2) or LC/MS.[152]
- Choosing the best ionization method (Section 9.10.4.1). For a metabolomics experiment, the most complete metabolite coverage may require analysis with multiple ionization methods.[153]
- Choice of quantitative standard – IS (isotopically labeled of not), external standard, or standard addition.[154]
- Selecting the most appropriate method of ion analysis – full scan, SIR, SRM.

It is important to remember, however, that in any quantitative assay, MS is just one part of a closely integrated overall procedure and that failure and compromise in any one step will invalidate the entire procedure. Some of the above points are illustrated in the development of a mass spectral assay for salicylic acid in tomato (**Figures 14–17**).

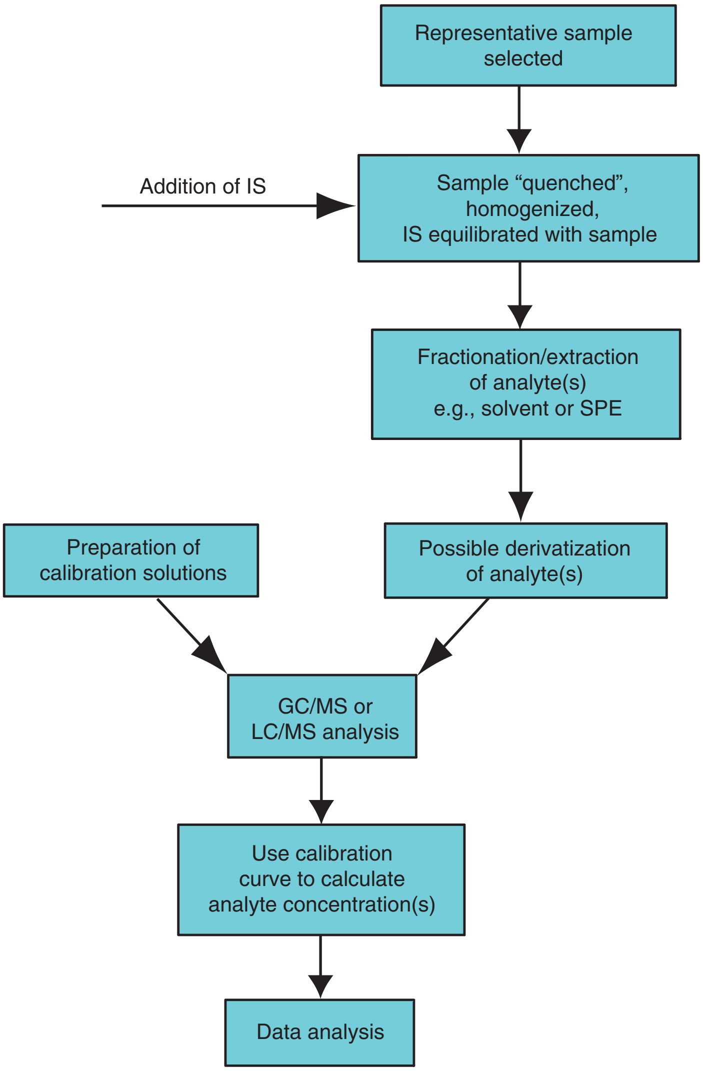

Representative sample selected
Addition of IS
Sample "quenched", homogenized, IS equilibrated with sample
Fractionation/extraction of analyte(s) e.g., solvent or SPE
Preparation of calibration solutions
Possible derivatization of analyte(s)
GC/MS or LC/MS analysis
Use calibration curve to calculate analyte concentration(s)
Data analysis

Figure 13 Flow chart for a quantitative assay using an internal standard. The most critical steps are the selection of a representative sample, the accurate preparation of the standards, and finally the addition of the standard to the sample – planning, weighing, making up to volume, and pipetting. It is sobering to remember that failure in any one of these will invalidate the entire assay no matter how sophisticated the instrumentation or how powerful the statistics applied to data analysis.

9.10.4.5.2 Sample preparation

Cellular processes are dynamic and the level of a particular metabolite at any one time will represent the balance of biosynthesis, biochemical transformation into other metabolites, degradation, transportation into and out of the cell, and sequestration into and out of storage forms. Depending on the rates of these respective processes, the level of a metabolite can be subject to large and rapid change during quenching. Similarly, subtle changes introduced by developmental processes or genetic manipulation can also induce large changes in the level of metabolites (see, e.g., Schwab[155]). In any metabolite analysis, it is important that the analytical sample accurately represents the cellular or tissue status at the time the sample is taken. This means that quenching must very rapidly terminate all biological processes and that chemical degradation is minimized.[13,152,156] Typically, quenching is achieved by extremes of temperature (<20 or >80 °C) or acidity ($\mathrm{pH} < 2$ or >10), possibly in the presence of an organic solvent, and/or an antioxidant, and in conjunction with homogenization. If the analysis is directed at a particular metabolite or class of metabolites (targeted analysis), the optimization

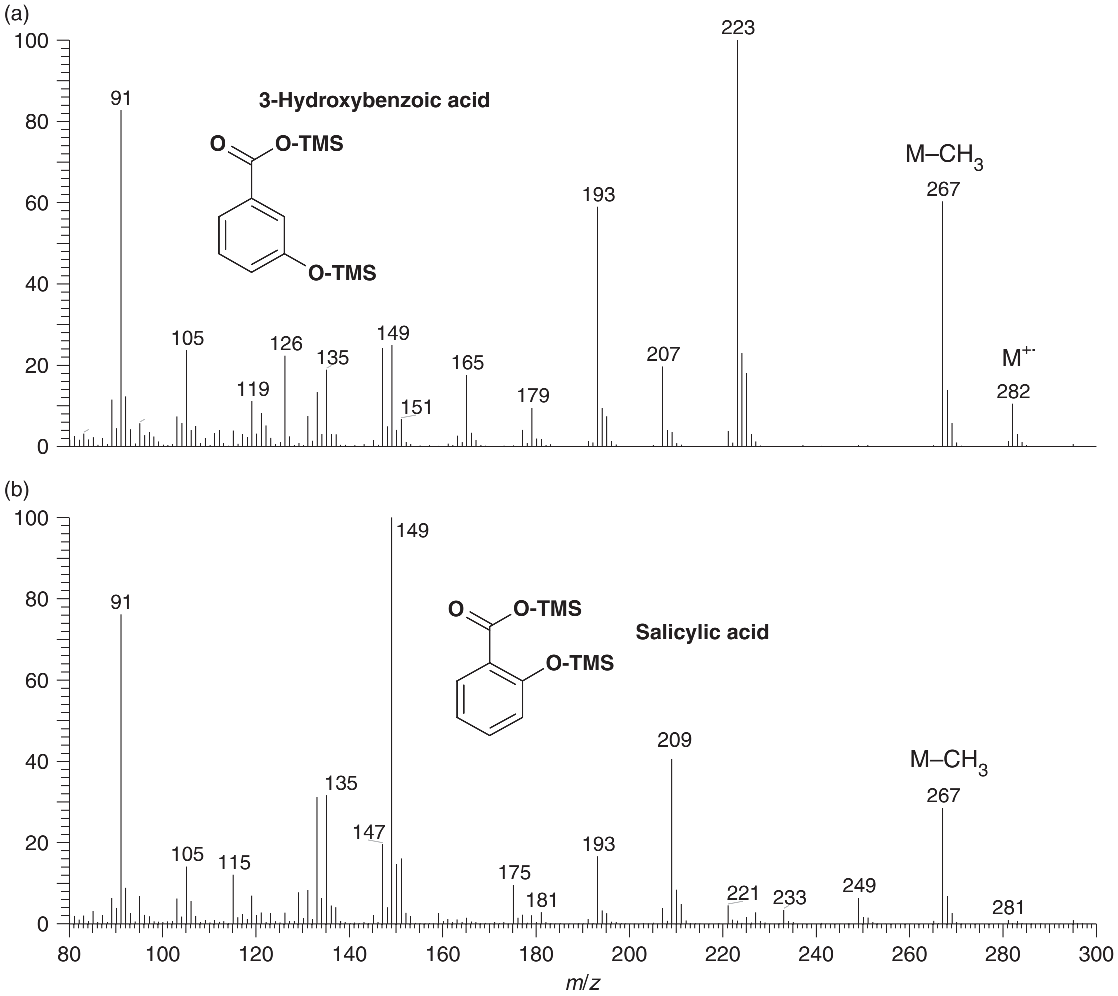


Figure 14 Quantifying salicylic acid in tomato. Full-scan mass spectra of the per-trimethylsilyl derivatives of 3-hydroxybenzoic acid internal standard (IS) (a) and salicylic acid (b). In both cases, the $[M - CH_3]^+$ ion (*m/z* 267), a structurally significant ion, was chosen for selected reaction monitoring.

of quenching is readily monitored. However, if, as in a metabolomic analysis, the objective is to analyze as many metabolites as possible, then it is inevitable that some compromise will have to be made, in which case reproducibility of the process becomes very important (see the review by Villas-Bôas,[157] and two case studies of optimizing metabolomic assays in blood plasma,[158] and plant tissue.[159]

9.10.4.5.3 Fractionation and extraction of sample

For a targeted assay, considerable effort is typically devoted to extracting the analyte or analyte class away from the sample matrix. Depending on the matrix, this may be as simple as a liquid–liquid extraction, selecting appropriate solid-phase extraction (SPE) chemistry (e.g., C-18 or ion exchange), or using affinity chromatography (specific lectins or antibodies bound to an inert matrix).[13,152,156] This step is designed to, first, reduce the possibility of interference by isobaric ions and, second, to reduce the possibility of ion suppression in the ionization process (ESI and APCI). This sample purification step may then be complemented by a high-resolution chromatographic separation interfaced to the MS source (e.g., nano-LC or GC).

ESI suppression has been correlated with high concentrations of nonvolatile matrix materials present in the spray and it is thought that this acts by inhibiting the formation of smaller droplets. Salts (e.g., phosphates and

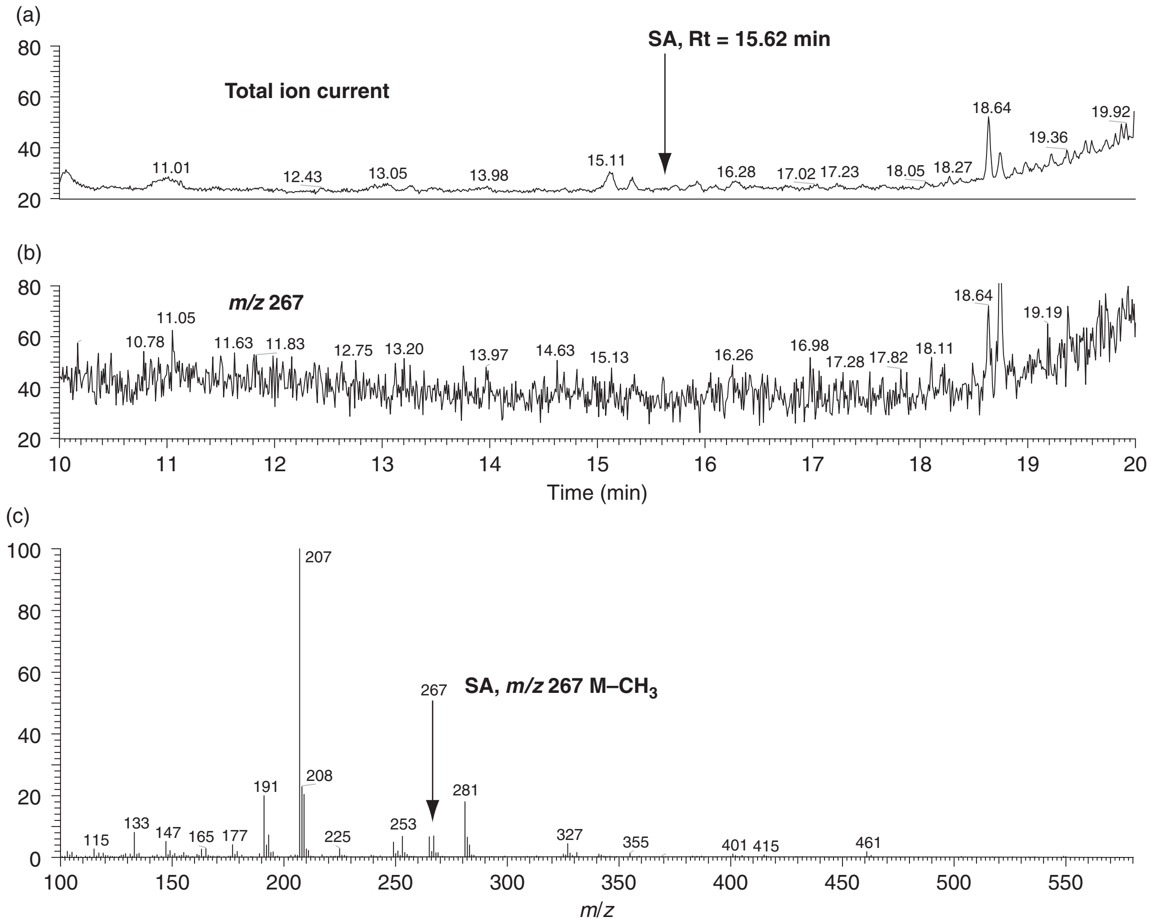


Figure 15 Quantifying salicylic acid in tomato. The full-scan mass spectra of the trimethylsilylated tomato extract contains too much background chemical noise for the salicylic acid to be satisfactorily assayed. Neither the total ion current chromatogram (a) nor the extracted ion chromatogram of *m/z* 276 (b) contains a discrete peak for salicylic acid, although *m/z* 276 is observable in the mass spectra corresponding to the retention time of the analyte (c).

sulfates) and ion-pairing reagents (e.g., trifluoroacetic acid) are also implicated in ion suppression.[160] Matrix effects may also be minimized by a simple process of sample dilution.[161]

As mentioned above, in a competitive ionization process, molecules with the lowest ionization potentials will be preferentially ionized and it is quite possible that this competition, in addition to matrix suppression, will result in the relative abundance of sample metabolites not being reflected in the MS data. Any reduction in the number of analyte ions available for analysis will have an impact on the assay with loss of sensitivity (higher limits of detection and quantitation).

In the case of an untargeted metabolomic experiment, the issue of sample cleanup is complicated by the need to retain as many of the metabolites as possible, and avoiding bias against any particular group or class of components. This is often resolved by fractionation into a number of subsamples, for example, by retention and analysis of the remaining aqueous phase after solvent extraction or fractionation by mixed mode SPE. The trade-off here is that more analytes are potentially available for assay but at the expense of time devoted to running many more analyses on the different sample fractions.[157,162]

9.10.4.5.4 Internal standards

In any analytical procedure, it is inevitable that there will be variations in instrumental parameters and in compliance with analytical protocols. It is also important to remember that in a mass spectrometer equimolar

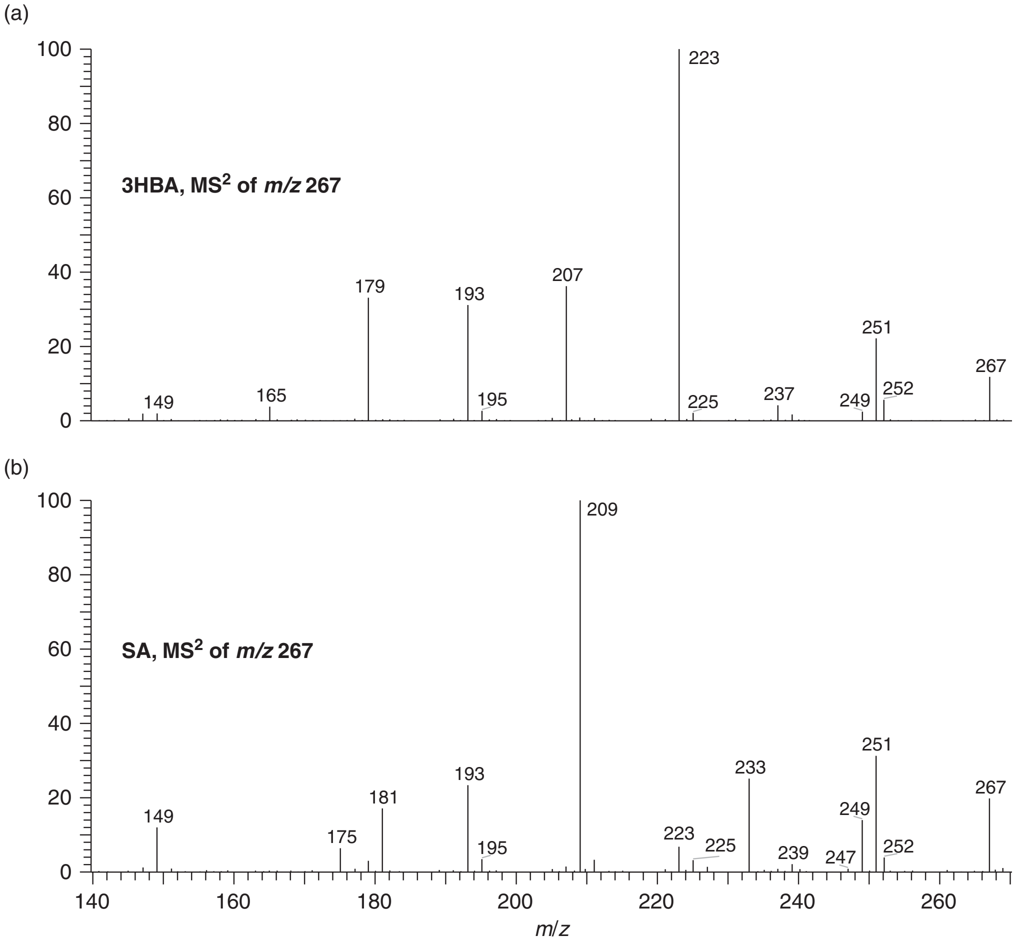


Figure 16 Quantifying salicylic acid in tomato. MS^2 of *m/z* 267 for both 3-hydroxybenzoic acid (a) and salicylic acid (b). The salicylic acid ion at *m/z* 209 was chosen for quantification against the *m/z* 223 ion from the internal standard, 3-hydroxybenzoic acid.

amounts of different compounds do not give an equal response because of variation in the ionization efficiency, which is in part dependent on the molecular structure and in part the result of competition (ion suppression) from other analytes present in the source. These procedural and instrumental variations will affect the accuracy and precision of the assay; however, they may be compensated for by the inclusion of a standard in the assay. There are four possible ways in which a standard may be incorporated into an MS-based assay.

The first is the use of a stable isotope-labeled standard (isotopomer) of the target analyte. The most common isotopes available for use include deuterium (^{2}H), ^{13}C, ^{15}N, and ^{18}O. The advantage of this approach is that the labeled standard will have identical chemical properties to the analyte and will be partitioned with the analyte throughout the analytical procedure, eliminating extraction and instrument bias and compensating for any ionization suppression by matching the ionization properties of the analyte. Thus the ratio of the amount of IS to analyte will remain constant up to the point of analysis. The mass spectrometer will then be able to independently detect the isotopically labeled standard by virtue of the heavier mass of its parent ion and fragment ions containing the labeled moiety. Quantification is then achieved by measuring the ratio of ions from the analyte and the IS, rather than an absolute value as in the use of an external standard. Then knowing the amount of standard added, the amount of the analyte present in the sample can be calculated from a comparison of the determined ion ratios (Figure 18).

Some caution needs to be exercised in using such a standard; the label should ideally be nonexchangeable and the number of incorporated isotopes must be sufficient so that there is minimal cross talk from the naturally

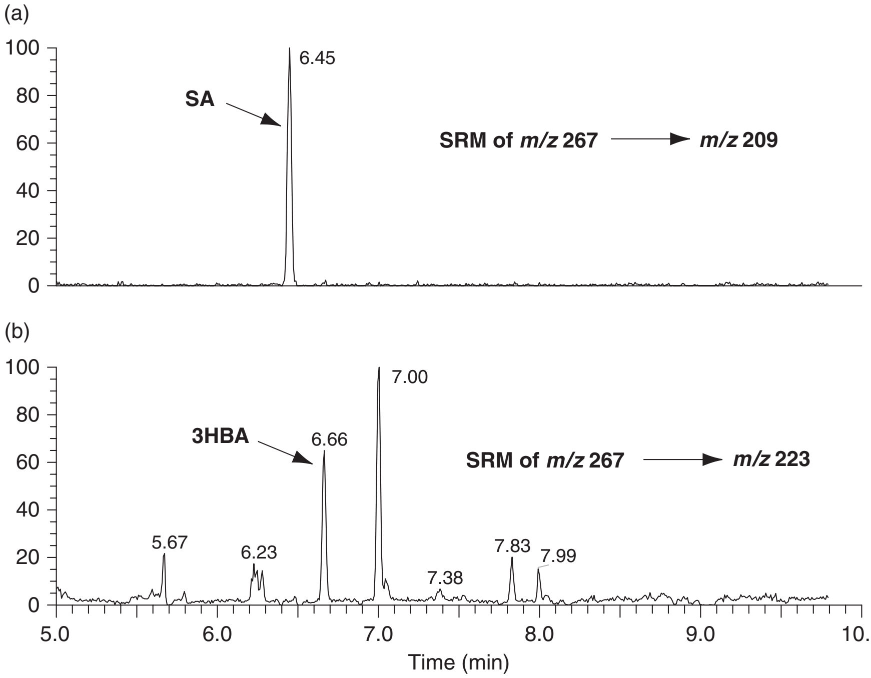


Figure 17 Quantifying salicylic acid in tomato. Selected reaction monitoring of the transition from *m/z* 267 to *m/z* 209 and from *m/z* 267 to *m/z* 223 from a tomato extract. Salicylic acid (Rt 6.45 min) (a) and 3-hydroxybenzoic acid (Rt 6.66 min) (b) can be observed at an S/N of 222 and 95, respectively.

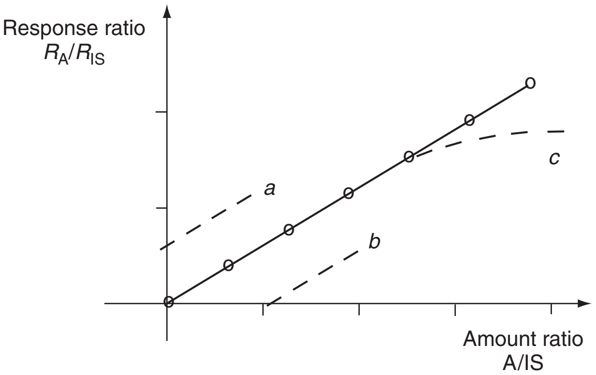


Figure 18 Calibration curves using an internal standard (IS). Analytes are quantified against an IS that has been added as early as possible in the analytical procedure. The ratios of detector responses for the analyte (R_A) and IS (R_{IS}) are plotted against the ratio of known amounts of analyte (A) and IS. When a sample is analyzed, the ratio R_A/R_{IS} is measured. Then knowing the amount of IS added into the sample, the amount of analyte present in the sample can be estimated. Curves that do not pass through the origin of the graph or which are nonlinear are diagnostic of (a) chemical interference or sample carryover, (b) sample loss during the assay due to adsorption, and (c) saturation or cross-contribution between the IS and the analyte.

occurring levels of ^{13}C and from isotopes of chlorine, bromine, and sulfur, when present. This also means that the highest degree of isotopic incorporation should be sought. Any large degree of cross talk will pose a limitation on the ultimate sensitivity of the assay. Nevertheless, isotopically labeled standards are usually regarded as approaching ideal although they are costly and of limited availability.

Where a stable isotope-labeled standard is unavailable, the analyst can use either a chemically similar homologue (e.g., incorporating an additional methylene; different *m/z* values to monitor) or a chemically similar analogue (e.g., geometric isomer; same *m/z* values to monitor) that will need to be chromatographically separated from the analyte. Obviously, it is also important that this chosen standard is not present in the sample (see, e.g., the use of 3-hydroxybenzoic acid as an IS for the quantification of salicylic acid in tomato in Figures 11–17).

In metabolomic experiments, where hundreds of analytes are to be quantified, a number of ISs representing different chemical classes of analytes are generally used (see, e.g., Jiye *et al.*[158] and Gullberg *et al.*[159]). These experiments are primarily comparative in nature as the experimenter is seeking to identify relative changes in metabolite levels and relative changes in metabolite fluxes as they occur in different experimental states.

9.10.4.5.5 Standard addition

Where there is no appropriate standard for an analyte, quantification can be made by standard addition (spiking). In this procedure, the sample is divided into several aliquots of equal volume and a series of known but increasing amounts of the analyte standard are then added to each aliquot. The samples are then diluted to the same volume yielding a series of solutions with equal concentrations of matrix but increasing concentrations of analyte. These samples are then analyzed individually for the analyte of interest and the concentration of the unknown can then be calculated from where the regression curve of the responses versus the standard additions intercepts the abscissa ($y = 0$) (Figure 19). The advantage of this method is the elimination of any chemical or physical bias between the standards and samples but this is achieved at the cost of a six- or sevenfold increase in the number of determinations required for each sample.

9.10.4.5.6 External standards

External standards, so named because they are not added to the sample, are also occasionally used but are generally only applicable to samples requiring limited preparation and for which a consistent high degree of reproducibility and good recovery can be attained. Experiments should also be completed as quickly as possible to minimize instrumental variations (e.g., ion source contamination). In brief, instrument response is plotted against the concentration (or amount) of standard analyzed and this response curve is then used to calculate analyte concentration (or amount). However, unless the matrix is well characterized, this method can be subject to matrix effects (ion suppression) and to interference from isobaric matrix components.

9.10.4.5.7 Optimization of the MS assay

In quantitative mass spectrometric assays, sources of error can be reduced to those associated with sample handling and processing, and to instrumental variation, for example, source contamination and stability of mass calibration. In general, the largest component of error is associated with sample handling and processing.[163] To a large degree, variations in protocols for purification and derivatization, poor technique, and even gross spillage of sample can be

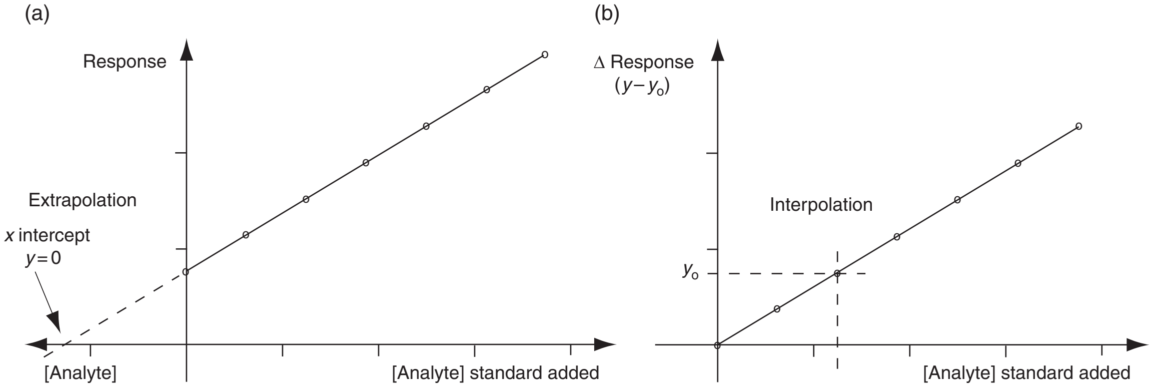


Figure 19 Standard addition calibration curves. Equal volumes of solvent containing varying amounts of standard are added (spiked) into the sample. The samples are analyzed and the analyzer response (e.g., area under the TIC or selected ion chromatogram) is plotted against the amount of standard added. The analyte concentration is estimated by extrapolating a linear least-squares regression to $y = 0$ (a). An alternative approach is to plot the difference between the spiked samples and the unspiked sample. The same calibration curve now passes through the origin and the sample analyte concentration can now be determined by interpolation with improved confidence limits[164] (b).

obviated by the use of an IS, as outlined above. Once the ratio of internal standard to analyte has been established, it will remain unchanged as long as the standard and the analyte have the same chemical properties. The integrity of the assay then depends almost entirely on the analyst's ability to accurately weigh, dissolve, dilute, and dispense the IS into the sample as required. Any errors associated with the IS will be propagated throughout the entire assay. This equates to the analyst having a basic knowledge and understanding of the analytical capabilities of balances (milligram quantities measured on analytical balance), volumetric flasks (clean and temperature equilibrated), and pipettes (calibrated, serviced, and used appropriately).[143]

It is thus critical that the IS be added to the sample at the earliest possible stage of the assay, usually the quenching or homogenization steps, and that it be allowed to equilibrate with the analyte in the sample matrix over some defined period of time (**Figure 13**). This is particularly important where there is nonspecific binding of the analyte to proteins or other cellular debris and where complete (100%) recovery cannot be achieved. The equilibration time can be established by a time-course study.[163]

Adding too much or too little IS can also limit the dynamic range of the assay, as the comparison of very large ion currents (detector saturation) with very small ion currents (poor ion statistics) will greatly increase the variance of the assay. A good guide is a threefold excess of the IS over the analyte but this may take a few trials to establish.

Other errors may be introduced into an MS assay by interference from isobaric ions. There are a number of possible remedies for this, including revising the sample preparation, changing the GC or LC column to separate away the interference, selecting an alternative structurally specific ion for the assay, and increasing the assay specificity by increasing mass resolution to monitor ions of selected elemental compositions. It is important to remember that the analyte spectrum should not be examined in isolation when choosing a set of ions for quantification but should include an appreciation of the 'background' ions that are also likely to be present, for example, ions from GC column bleed or solvent/reagent adduct ions from ESI (see Section 9.10.4.5.8). Any change in retention time and/or the shape of the chromatographic peak is likely to be indicative of interference, which is to say a lack of assay specificity.

Selected ions should be structurally specific to the analyte and should be abundant in order to maximize the assay sensitivity. In the example of the assay for salicylic acid in tomato (**Figure 14**), the ion selected for MS^2 was the structurally specific $[M-CH_3]^{+\cdot}$ ion for both the analyte and the IS. The ion at m/z 91, although intense, is a tropylium ion ($C_7H_7{}^{+\cdot}$) and would be an inappropriate selection as it would be present in most analytes containing a benzyl moiety. The ions at m/z 223 and m/z 209 in the product spectra (**Figure 16**) were chosen for the quantification because they were the most intense.

For assays based on a full-scan MS, the specificity and sensitivity can be increased by

- Careful selection of ions for quantification. In general, higher m/z values are less subject to interference. Using a different analyte derivative might assist in this. It must be borne in mind, however, that regulatory authorities will require these to be structurally specific ions.[144,145] Note that moving to SIM will not improve selectivity over that of a full scan but will improve sensitivity.
- Moving to high-resolution SIM and targeting a specific elemental composition may remove interference except when the interfering compound has the same elemental composition as the analyte. In this case, one should suspect an analyte isomer and change the sample chromatography accordingly.
- Using SRM. It is unlikely that the interfering ion will fragment in the same way as the analyte and the elimination of background chemical 'noise' by SRM will also improve sensitivity. Again, the ions selected should be structurally specific.[144,145] See, for example, the S/Ns in the TIC for the salicylate assay (**Figure 15**) and compare them with those realized in the SRM traces (**Figure 17**).
- Using high-resolution SRM and targeting specific elemental compositions in the precursor and product ions.

A comprehensive discussion of trace quantitation using MS, including error calculations, confidence limits, limits of detection (LoD), limits of quantitation (LoQ), and method validation may be found in Boyd *et al.*,[143] but also see a more general discussion of these issues as they pertain to analytical chemistry by Pritchard and Barwick.[151]

An example of a method validated according to the FDA and EU guidelines is described by Hermo *et al.*[165] These authors used LC–ToF MS to determine the levels of multiresidue antibiotic quinolines in pig livers below the maximum residue limits. They describe the optimization of their method, which is then comprehensively characterized by the determination of the linearity, the decision limit, LoD, LoQ, the precision, the accuracy, and finally the recoveries for the different residues.

9.10.4.5.8 Chemical noise and contamination

As the sensitivity of mass spectral-based assays has improved and the interest in quantifying trace analytes has increased, the problems associated with chemical noise and sample contamination have also increased. Chemical noise and contamination in an assay have the effect of reducing the S/N of the analyte signal. This places an immediate restriction on achieving the full potential of the instrumental sensitivity with the assay LoD and LoQ set higher than they might otherwise have been. While it is unlikely that chemical noise and contamination can ever be completely eliminated, they can be minimized if care is taken to avoid known sources of contamination when the assay protocols are being planned (see reviews by Ende and Spiteller[166] and Keller *et al.*,[167] on mass spectral contaminants and their origins; the supplementary data in the latter review includes a literature compilation of contaminants in an Excel spreadsheet). In general, sample contamination can be sourced to almost every part of the assay, including

- the person of the analyst (e.g., keratin proteins, fatty acids, amino acids, and cosmetic residues from hair and skin),
- solvents (e.g., degradation products, antioxidants, and stabilizers),
- reagents used in sample preparation (e.g., proteins, detergents, antioxidants, chemical bleed from 'dip sticks'),
- laboratory ware (e.g., detergent residues, plasticizers, lubricants),
- chromatography (GC or LC column degradation or bleed, and late eluting components of previous samples),
- ionization process (matrix clusters from MALDI and solvent clusters from ESI, APCI, APPI, and DESI; clusters may also include common alkali metal cations, Na^+ and K^+, in addition to other cations from the assay reagents), and
- sample carryover and cross-contamination (inadequate washing of components that are reused for each batch of samples, e.g., pipettors, recycled sample vials, GC and LC autosamplers).

Some of the precautions that can be exercised should be a normal part of good laboratory practice, and include the appropriate use of personal protection. Covering the hair and using gloves will minimize the possibility of contamination from skin and hair-derived keratin proteins, as well as amino acids, fatty acids, and cosmetic residues from the skin surface.

It is also obvious that, unless determined otherwise, the highest quality solvents should be used. This is particularly so in the case of water, which, as the initial solvent in reverse-phase chromatography, can concentrate impurities at the head of the column. It is worth remembering that with laboratory-prepared water, the outlet conductivity meter provides an estimation of residual ions in the water and does not provide a measure of any neutral organic contaminants, should they be present. Again good laboratory practice, as exemplified by the recommended periodic changes of the purification cartridges, is the best way to prevent this water becoming a source of contamination.

Other potential contaminants are, however, less obvious and these include lubricants (e.g., silicone grease), plasticizers (e.g., phthalates, phenyl phosphates, sebacates, and bisphenol A), slip agents (e.g., oleamide, erucamide, and stearamide), biocides (e.g., quarternary ammonium compounds) and polymers extracted from laboratory consumables (e.g., silicones from laboratory tubing) and membrane filters (e.g., cyclic oligomers and Nylon 66).[168,169]

Contaminants can also be sourced to reagents used in sample protocols. For example, in the extraction and purification of proteins, it is common to use detergents, which, unless they are removed from the sample that is presented for MS analysis, can represent a very persistent form of contamination that is not readily removed except by long periods of washing or by replacement of the LC column and other associated components. Detergents can also be inadvertently introduced from laboratory glassware or sample vials that have been inadequately rinsed after washing.

In addition to analyst-derived keratins, proteins/peptides can also be introduced, for example, in 'bottom-up' proteomics where the sample is digested by a proteolytic enzyme (most commonly trypsin).

This will give rise to a set of autolysis peptides from the self-digestion of the enzyme. These autolysis peptides are impossible to eliminate but can be minimized by using the highest quality autolysis-resistant enzyme. Other proteins such as bovine serum albumin (BSA) may be used in the immunopurification of specific proteins. Again, if this is an unavoidable part of the protocol, then the analyst should expect to observe peptide ions derived from these proteins.

Chromatographic materials (e.g., solid-phase extraction tubes, LC and GC columns, TLC plates) including single-use materials must be thoroughly, and appropriately, washed or conditioned to remove contaminants originating from the manufacturing process or those that may have been acquired by exposure to packing materials or to the laboratory atmosphere. In some cases, contamination is unavoidable and it is important for the analyst to be able to recognize this and to plan accordingly. For example, all GC columns and septa continuously shed volatile siloxanes, a process known as 'bleeding', as the temperature is raised. The amount of bleed is proportional to the temperature and to the amount of phase on the column. Thus columns with thicker phase coatings generally bleed more, and especially so at higher temperatures. Although modern column phase chemistry is extremely robust, columns are inevitably degraded over time with a concomitant increase in bleed. This phase degradation is accelerated by trace levels of oxygen in the carrier gas at high temperatures, so it is important to ensure that a functional oxygen trap is part of the in-line gas purification process and that column temperature limits are not exceeded.

LC columns can also bleed in the presence of the eluting solvent but for modern columns operated within their pH range this problem is generally minimal but will be exacerbated when chromatographing at high temperatures. Of greater concern is the solvent and adduct clusters generated by the atmospheric pressure ionization (ESI, APCI, APPI, and DESI).[170] The many combinations and permutations of solvent clusters, complicated by the inclusion of solvent modifiers such as acetic acid, formic acid, or triethylamine, along with the ever-present sodium and potassium cations, form a very complex chemical background against which analyses must be performed. Moreover, in the case of a solvent gradient, this chemical noise will be changing, over time, in accord with the gradient. In addition to solvent clusters, clusters also form around the eluting analytes and any contaminants picked up during the assay. A number of hardware approaches have been used to minimize the impact of this chemical noise, including orthogonal and '*z*' geometries for the spray outlet and MS inlet, the use of nebulizing and curtain gases, and ion mobility interfaces (FAIMS and TWIMS in Section 9.10.2.3.9). Tandem MS (Section 9.10.3) can be used to remove much of the remaining chemical noise and this approach can be enhanced by the use of curved collision cells to eliminate the transmission of fast neutrals to subsequent stages of tandem MS.

A considerable amount of chemical noise (mostly $<m/z$ 1000) in the form of matrix clusters is also generated by MALDI and this has generally precluded MALDI from being used to analyze small molecules. Some reduction in the occurrence and intensity of matrix clusters can be obtained by minimizing salt contamination by on-target washing and/or by sample purification (e.g., use of Zip-Tips)[167] but this must be offset against potential loss of hydrophilic analytes. Various attempts have been made to find a substitute for the MALDI matrix but to date these have lacked universal acceptance, often related to the cost, difficulty of preparation, ease of contamination, or lack of long-term stability of the alternative target surfaces (e.g., porous silica, sol gels, graphite, carbon nanotubes, fullerenes, and polymers[171]).

Finally, contamination of sample spectra can also occur by cross-contamination during sample preparation and by carryover of residual analyte from a sample analyzed earlier in the run.[172,173] Essentially, any component of the assay that is reused for each sample or batch of samples can be a source of cross-contamination or carryover. These include, for example, evaporators, pipettors, automated liquid handlers, recycled sample vials, and LC and GC autosamplers. Care needs to be taken in the selection of appropriate wash solvents that will readily solubilize the sample and analytes. This will usually be a combination of high percentage of organic solvents that may include a volatile acidic or basic modifier (e.g., formic acid or aqueous ammonia). Failure to properly wash all sample components from a chromatographic column can result in late eluting components appearing in the next, or later, analytical runs.

Unless care is exercised by the analyst, both these forms of contamination can go unnoticed and erroneous results may be reported for individual samples. Problems with cross-contamination should normally be identified during the validation phase of method development by the judicious use of blanks to test for problems with general laboratory contamination, sample preparation, and the autosampler. Carryover is assessed by injecting one or more blanks after a high concentration sample, normally at the upper limit of quantitation. If the carryover is less than 20% of the lower limit of quantitation, then this is normally deemed to be acceptable. If possible, the analyst should order the analysis from low-concentration samples to high, with high-concentration samples followed by a blank and/or additional cycles of sample syringe washing.

Trace analysis and the move to the use of smaller sample sizes represent particular challenges in that the ratio of surface area exposure to sample volume, or quantity of analyte, is increased, multiplying the possible effect and level of contamination. While mass spectral identification of contaminants will aid in identifying their source (see the literature-derived Excel database of contaminant mass spectra in the supplementary data of Keller *et al.*[172]), this is not essential. The key tool to their elimination is the appropriate use of sample blanks at each step of the analytical protocol during method development and validation.

9.10.4.6 MS Imaging

Traditional histological studies of tissue sections have been limited to either light or electron microscopy. Both these techniques have been used to obtain limited amounts of chemical information from the examined tissue. For the most part, this has been achieved through the use of a small number of specific chemical, radiographic, autoradiographic, and immunological stains. More recently, organisms have been genetically engineered to incorporate fluorescent tags into proteins (e.g., green fluorescent protein, GFP); however, these can potentially interfere with the normal functioning of the tagged protein. In general, although microscopy can yield excellent images at high resolution, there is little direct chemical information on the imaged components of the tissue surface.

Since the original idea to generate chemical images of tissue sections using MALDI-MS (Section 9.10.2.2.7),[174–177] two additional ionization techniques, SIMS (Section 9.10.2.2.8)[11,60,178] and DESI (Section 9.10.2.2.9),[179] have been added to the imaging repertoire but these have yet to gain the relative popularity of MALDI imaging. Unlike the traditional histological stains, these three MS imaging techniques require no prior assumptions about chemical identity and they are capable of sensitively visualizing a large range of small (e.g., metabolites) and large molecules (e.g., proteins) provided that they are ionizable for subsequent MS analysis, including direct molecular identification using tandem spectrometry (MS^n).

The great challenge in preparing a tissue sample for MALDI-MS imaging is that two contradictory processes must occur.[176,180] First, tissue sections are frozen in liquid nitrogen to avoid delocalization and degradation of the peptide and protein analytes. Sections are then prepared by cryosectioning and these are then mounted on a cold MALDI target. Next, matrix, either sinapinic acid for high-molecular-weight proteins or α-cyano-4-hydroxycinnamic acid for low-molecular-weight peptides and proteins ($<\sim 3$ kDa), is applied. For best image resolution and reproducibility, the matrix is usually uniformly sprayed directly onto the tissue surface. At this stage, it is important for the matrix solution to be able to solubilize, extract, and cocrystallize with the protein and peptide analytes while minimizing their delocalization. Several coatings of matrix are usually applied with a short interval between applications for the solvent to dry. This avoids the problem of large quantities of solvent potentially mobilizing the surface analytes with concomitant loss of image resolution. Images are then obtained by rastering a laser across the tissue surface, desorbing and ionizing the analytes, and generating spectra from specific locations (pixels). Virtual images based on the location of specific ions can then be generated and matched to the images of other specific ions and to light microscope images of the section. The intrinsic value of these MALDI images can be greatly enhanced if they can be precisely aligned with the image generated by traditional histopathological stains. Several approaches to this goal have been investigated and include rinsing the matrix from the tissue surface before applying the histological stain, staining the consecutive section,[181] and the use of MALDI-compatible stains.[182]

Most recently, Caprioli's group[183] have reported a novel method of dry coating tissue sections with MALDI matrix, thus minimizing the problem of the matrix solvent mobilizing the surface analytes. The dry coating procedure proved to be simple and rapid and yielded high-quality images of phospholipids.

Where it may not be convenient, or possible, to immediately flash freeze a tissue sample, ethanol-preserved paraffin-embedded specimens may also be used for MALDI imaging.[184] Thinner microtome sections can be cut from the frozen tissue following this treatment and this is an advantage where comparisons need to be made with traditional histological stains for light microscopy.

Neither DESI nor SIMS requires any special treatment of the sample surface and images are generated by rastering a microprobe spray of solvent (DESI) or a beam of energetic ions (SIMS) over the tissue section. Virtual images of desorbed secondary ions are then generated as for MALDI imaging. SIMS and MALDI imaging must both be carried out in a high vacuum. They are also complementary techniques in that SIMS is applicable to small molecules (<~500 *m*/*z*) and MALDI to large molecules (>~800 *m*/*z* to avoid matrix cluster ions) with the SIMS analysis being performed prior to the application of matrix.[59] SIMS imaging has been reported as being able to achieve lateral resolutions down to 50 nm[11] and MALDI imaging has achieved resolutions down to 10–25 μm.

Unlike MALDI (> ~ 800 Da) and SIMS (< ~ 500 Da), the new DESI technique can usefully image both small metabolite molecules and large proteins but the spatial resolution is limited to only slightly better than 400 μm[179] when sampling from tissue sections. The image resolution that DESI can achieve is determined by the cross-sectional area of the applied solvent spray as it strikes the target and this in turn is determined by solvent flow rate, solvent composition, applied voltage, and size of spray orifice. The height of the spray orifice above the surface, the angle of the incident spray (~55°), and the angle at which the desorbed and ionized analytes are sampled (0–20°) are also critical. When sampling from printed patterns on paper and thin-layer chromatography plates, image resolutions of ~40 μm have been reported.[185] Although DESI imaging will require further development before it can compete with the resolution achieved to date by SIMS and MALDI, it does offer the advantage of being applicable to surfaces and samples not readily brought within the vacuum system of the mass spectrometer.

With the release of the first generation of commercial MS imaging instruments, MS imaging is being actively applied to problems in biology and human health. If the availability of chemical images, particularly if they can be correlated with the images from the traditional histopathology stains, proves useful, this will feedback to promote further technical developments of the technique. However, future developments in a clinical or diagnostic setting will have to meet the challenge of quality assurance and information validation.[186]

9.10.4.7 Future Prospects

Over the past 100 years, MS has moved from the exclusive domain of physics to chemistry, and is now an essential tool for biologists in creating holistic approaches to studying biology. Thus, for example, developments in mass spectrometric-based proteomic and metabolomic studies directly complement high-throughput chip-based transcriptomics and the large number of genomes – bacterial, plant, and animal – that have been completely sequenced.

The increasing utility of mass spectrometers can be directly attributed to the development of new ionization methods. This is particularly exemplified in the development of ESI and MALDI, which have allowed the routine ionization of large, labile, and polar molecules. This initiated an exponential growth in the sale of mass spectrometers, which is now sustained by demands from biology and the health sciences for increasing sensitivity, resolution, and accuracy, all at a cheaper cost. As a result, the performance of mass analyzers and peripheral instrumentation has improved, size and cost have decreased, new analyzers invented (e.g., Orbitrap), old ones improved (e.g., ToF), new configurations assembled (e.g., QIT-ToF and LIT-Orbitrap), and more sophisticated computer algorithms developed to increase sample throughput and to improve the ease of operation for the nonspecialist. The improved availability of mass spectrometers has also seen their application to new areas of endeavor such as PTR-MS for environmental monitoring and MALDI-MS for microscopy, providing a new chemical dimension to the imaging of tissue sections. There are also active programs of miniaturization in progress and the prospect of handheld mass spectrometers with the new ambient ionization techniques of DESI and DART is very real.[187]

Forecasting the future is always fraught with difficulties; however, extrapolating from the immediate past, the future of MS looks to be very bright as the need to identify and quantify with increasing sensitivity and reliability is one that is not going to diminish whether it be for the purposes of research, regulation, or law enforcement.

Acknowledgments

I thank my colleagues, Sharyn Wragg for Figure 1, and Carolyn McKinlay who kept the MS lab operating during my periodic absences and who provided the myoglobin spectra.

Abbreviations

2D	two-dimensional linear ion trap
3D	three-dimensional quadrupole ion trap
APCI	atmospheric pressure chemical ionization
APPI	atmospheric pressure photoionization
BIRD	blackbody infrared dissociation
BSA	bovine serum albumin
BSTFA	*N*, *O*-bis(trimethylsilyl)trifluoroacetamide
CE	capillary electrophoresis
CI	chemical ionization
CID	collision-induced dissociation
CV	compensation voltage
DART	direct analysis in real time
DESI	desorption electrospray ionization
DMOX	dimethyloxazoline
ECD	electron capture dissociation
ECI	electron capture ionization
EDD	electron detachment dissociation
EI	electron ionization
ESI	electrospray ionization
ET	electron transfer
ETciD	supplemental low-energy CID method
ETD	electron transfer dissociation
EU	European Union
FAIMS	field asymmetric waveform ion mobility spectrometry
FAME	fatty acid methyl ester
FDA	Food and Drug Administration
FTICR MS	Fourier transform ion cyclotron resonance MS
FWHM	full-width at half-maximum height
GAG	glycosaminoglycans
GC	gas chromatography
HECD	hot ECD
HR	high-resolution
ICR	ion cyclotron resonance
IMS	ion mobility spectrometry
IR	infrared
IRMPD	infrared multiphoton dissociation
IS	internal standard
KI	Kovat’s indices
LC	liquid chromatography
LIT	linear ion trap
LMIG	liquid metal ions gun
LoD	limits of detection
LoQ	limits of quantitation
LR	low-resolution

m/z	mass to charge
MALDI	matrix-assisted laser desorption ionization
MS	mass spectrometry
MS^n	multistage mass spectrometry
MSTFA	*N*-methyl-*N*-trimethylsilyltrifluoracetamide
MuDPIT	multidimensional protein identification technology
NICI	negative ion chemical ionization
NMR	nuclear magnetic resonance
PICI	positive ion chemical ionization
PID	photon-induced dissociation
PPINICI	pulsed positive ion/negative ion CI
PTM	posttranslational modification
PTR-FTICR	proton-transfer reaction Fourier transform ion cyclotron resonance
PTR-MS	proton-transfer reaction mass spectrometry
Q	quadrupole
QIT	quadrupole ion trap
RF	radio frequency
RIC	reconstructed ion current
S/N	signal-to-noise ratio
SIFT-MS	selected ion flow tube mass spectrometry
SIM	selected ion monitoring
SIMS	secondary-ion mass spectrometry
SORI	sustained off-resonance irradiation
SPE	solid-phase extraction
SRM	selected reaction monitoring
TIC	total ion current
ToF	time-of-flight
TMCS	trimethylchlorosilane
TWIMS	traveling wave IMS
VOC	volatile organic compound

Nomenclature

B	magnetic field or magnetic sector analyzer
BE	double-focusing mass spectrometer using a magnetic sector analyzer linked to an electrostatic analyzer – reverse geometry
CE/MS	instrument in which capillary electrophoresis is directly interfaced with a mass spectrometer
Da	dalton, an atomic mass unit, commonly used in biochemistry for the mass of ions and molecules
E	electrostatic analyzer
EB	double-focusing mass spectrometer using an electrostatic analyzer linked to a magnetic sector analyzer – forward geometry
E_k	kinetic energy
ETciD	ETD combined with a supplemental low-energy CID of a peptide electron transfer product
GC/MS	instrument in which gas chromatography is directly interfaced to a mass spectrometer
LC/MS	instrument in which liquid chromatography is directly interfaced to a mass spectrometer
m	mass of an ion
metabolome	the complete set of all metabolites present in a cell at any one time. These are usually arbitrarily defined as having a molecular weight of less than 1000 Da.

mmu	millimass units (10^{-3} u)
M_r	molecular weight
MS	commonly used with respect to mass spectrometer instruments, the data they generate (the mass spectrum), and the technique mass spectrometry
MS^n	tandem mass spectrometry of *n* stages; MS^2 may be written as MS/MS
m/z	mass to charge ratio
ppb	parts per billion
ppm	parts per million
proteome	the complete set of all proteins present in a cell at any one time
q	collision cell contained within an RF-only quadrupole or multipole
R	mass resolution
u	unified atomic mass unit, one-twelfth the mass of the most abundant naturally occurring isotope of carbon, ^{12}C
z	number of charges on an ion

References

1. B. Bothner; G. Siuzdak, *ChemBioChem* **2004**, *5*, 258–260.
2. J. J. Thompson, *Rays of Positive Electricity and Their Application to Chemical Analysis*; Longmans: London, UK, 1913.
3. M. A. Grayson, Ed., *Measuring Mass: From Positive Rays to Proteins*; Chemical Heritage Press: Philadelphia, USA, 2002.
4. J. Griffiths, *Anal. Chem.* **2008**, *80*, 5678–5683.
5. J. T. Watson; O. D. Sparkman, *Introduction to Mass Spectrometry: Instrumentation, Applications and Strategies for Data Interpretation*, 4th ed.; John Wiley and Sons: Chichester, UK, 2007.
6. G. J. Trout; R. Kazlauskas, *Chem. Soc. Rev.* **2004**, *33*, 1–13.
7. M. Thevis; W. Schänzer, *Curr. Org. Chem.* **2005**, *9*, 825–848.
8. P. Hemmersbach, *J. Mass Spectrom.* **2008**, *43*, 839–853.
9. C. Fenselau; R. Caprioli, *J. Mass Spectrom.* **2003**, *38*, 1–10.
10. J. R. de Laeter, *Mass Spectrom. Rev.* **1998**, *17*, 97–125.
11. L. A. McDonnell; R. M. A. Heeren, *Mass Spectrom. Rev.* **2007**, *26*, 606–643.
12. Y. Hayasaka; G. A. Baldock; A. P. Pollnitz, *Aust. J. Grape Wine Res.* **2005**, *11*, 188–204.
13. K. Dettmer; P. A. Aronov; B. D. Hammock, *Mass Spectrom. Rev.* **2007**, *26*, 51–78.
14. E. J. Want; A. Nordström; H. Morita; G. Siuzdak, *J. Proteome Res.* **2007**, *6*, 459–468.
15. R. Aebersold; D. R. Goodlet, *Chem. Rev.* **2001**, *101*, 269–295.
16. J. C. Smith; J.-P. Lambert; F. Elisma; D. Figeys, *Anal. Chem.* **2007**, *79*, 4325–4344.
17. R. J. Simpson, *Proteins and Proteomics: A Laboratory Manual*; Cold Spring Harbour Laboratory Press: New York, USA, 2003.
18. A. G. Marshall; C. L. Hendrickson; G. S. Jackson, *Mass Spectrom. Rev.* **1998**, *17*, 1–35.
19. M. L. Vestal, *Chem. Rev.* **2001**, *101*, 361–375.
20. K. B. Tomer, *Chem. Revi.* **2001**, *101*, 297–328.
21. D. R. Knapp, *Handbook of Analytical Derivatization Reactions*; Wiley-Interscience Publication, John Wiley & Sons, Inc.: New York, USA, 1979.
22. K. Blau, J. Halket, Eds., *Handbook of Derivatives for Chromatography*, 2nd ed., John Wiley & Sons, Inc.: Chichester, UK, 1993.
23. J. M. Halket; V. G. Zaikin, *Eur. J. Mass Spectrom.* **2003**, *9*, 1–21.
24. J. M. Halket; V. G. Zaikin, *Eur. J. Mass Spectrom.* **2003**, *9*, 421–434.
25. J. M. Halket; V. G. Zaikin, *Eur. J. Mass Spectrom.* **2004**, *10*, 1–19.
26. J. M. Halket; V. G. Zaikin, *Eur. J. Mass Spectrom.* **2004**, *10*, 555–568.
27. J. M. Halket; V. G. Zaikin, *Eur. J. Mass Spectrom.* **2005**, *11*, 127–160.
28. A. J. Dempster, *Philos. Mag.* **1916**, *31*, 438–443.
29. C. J. W. Brooks, *Philos. Trans. R. Soc. London Ser. A* **1979**, *293*, 53–67.
30. S. G. Wyllie; B. A. Amos; L. Tökés, *J. Org. Chem.* **1977**, *42*, 725–732.
31. H. Budzikiewicz; C. Djerrassi; D. H. Williams, *Mass Spectrometry of Organic Compounds*; Holden-Day, Inc.: San Francisco, USA, 1967.
32. F. W. McLafferty; F. Tureček, *Interpretation of Mass Spectra*, 4th ed.; University Science Books: Sausalito, CA, USA, 1993.
33. J. H. Gross, *Mass Spectrometry. A Textbook*; Springer-Verlag: Heidelberg, Germany, 2004.
34. E. de Hoffmann; V. Stroobant, *Mass Spectrometry: Principles and Applications*, 3rd ed.; John Wiley and Sons: Chichester, UK.
35. A. A. Solovev; V. I. Kadentsev; O. S. Chizhov, *Russian. Chem. Rev.* **1979**, *48*, 631–644.
36. B. Munson; F. H. Field, *J. Am. Chem. Soc.* **1966**, *88*, 2621–2630.
37. B. Munson, *Anal. Chem.* **1977**, *49*, 772A–778A.
38. A. G. Harrison, *Chemical Ionization Mass Spectrometry*, 2nd ed.; CRC Press: Boca Raton, FL, USA, 1992.
39. K. Bieman; J. A. McCloskey, *J. Am. Chem. Soc.* **1962**, *84*, 3192–3193.

40. G. Junk; H. Svec, *J. Am. Chem. Soc.* **1962**, *85*, 839–845.
41. G. W. A. Milne; T. Axenrod; H. M. Fales, *J. Am. Chem. Soc.* **1970**, *92*, 5170–5175.
42. W. Lindinger; A. Hansel; A. Jordan, *Chem. Soc. Rev.* **1998**, *27*, 347–354.
43. C. Dehon; E. Gaüzère; J. Vaussier; M. Heninger; A. Tchapla; J. Bleton; H. Mestdagh, *Int. J. Mass Spectrom.* **2008**, *272*, 29–37.
44. D. Smith; P. Španěl, *Mass Spectrom. Rev.* **2004**, *24*, 661–700.
45. C. K. Meng; M. Mann; J. B. Fenn, *Phys. D* **1988**, *10*, 361–368.
46. V. G. Zaikin; J. M. Halket, *Eur. J. Mass Spectrom.* **2006**, *12*, 79–115.
47. E. C. Horning; M. G. Horning; D. I. Carroll; I. Dzidic; R. N. Stillwell, *Anal. Chem.* **1973**, *45*, 936–943.
48. K. A. Hanold; S. M. Fischer; P. H. Cormia; C. E. Miller; J. A. Synge, *Anal. Chem.* **2004**, *76*, 2842–2851.
49. S.-S. Cai; K. A. Hanald; J. A. Syage, *Anal. Chem.* **2007**, *79*, 2491–2498.
50. A. Bagag; A. Giuliani; O. Laprévote, *Int. J. Mass Spectrom.* **2007**, *264*, 1–9.
51. D. B. Robb; M. W. Blades, *Anal. Chim. Acta* **2008**, *627*, 34–49.
52. D. Debois; A. Giuliani; O. Laprévote, *J. Mass Spectrom.* **2006**, *41*, 1554–1560.
53. F. Hillenkamp, J. Peter-Katalinić, Eds., *MALDI MS: A Practical Guide to Instrumentation, Methods and Applications*; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2007.
54. M. Karas; D. Bachmann; F. Hillenkamp, *Int. J. Mass Spectrom. Ion Processes* **1987**, *78*, 53–68.
55. M. Karas; D. Bachmann; U. Bahr; F. Hillenkamp, Matrix-Assisted UV Laser Desorption of Non-volatile Compounds. In *Advances in Mass Spectrometry*; J. F. J. Todd, Ed.; Wiley: Chichester, 1987; Vol. 10B, p 969.
56. M. Karas; U. Bahr; U. Giessmann, *Mass Spectrom. Rev.* **1991**, *10*, 335–357.
57. F. Hillenkamp; M. Karas; R. C. Beavis; B. T. Chait, *Anal. Chem.* **1991**, *63*, 1193A–2003A.
58. S. C. C. Wong; R. Hill; P. Blenkinsop; N. P. Lockyer; D. E. Weibel; J. C. Vickerman, *Appl. Surf. Sci.* **2003**, *203*, 219–222.
59. A. Brunelle; D. Touboul; O. Laprévote, *J. Mass Spectrom.* **2005**, *40*, 985–999.
60. R. M. A. Heeren; L. A. McDonnell; E. Amstalden; S. L. Luxembourg; A. F. M. Altelaar; S. R. Piersma, *Appl. Surf. Sci.* **2006**, *252*, 6827–6835.
61. R. G. Cooks; Z. Ouyang; Z. Takáts; J. M. Wiseman, *Science* **2006**, *311*, 1566–1570.
62. Z. Takáts; J. M. Wiseman; B. Gologan; R. G. Cooks, *Science* **2004**, *306*, 471–473.
63. R. B. Cody; J. A. Larameé; H. D. Durst, *Anal. Chem.* **2005**, *77*, 2297–2302.
64. I. Cotte-Rodriguez; C. C. Mulligan; R. G. Cooks, *Anal. Chem.* **2007**, *79*, 7069–7077.
65. S. A. McLuckey; J. M. Wells, *Chem. Rev.* **2001**, *101*, 571–606.
66. N. I. Tarantin, *Phys. Part. Nucl.* **1999**, *30*, 167–194.
67. H. Wollnik, *J. Mass Spectrom.* **1999**, *34*, 991–1006.
68. A. A. Makarov, *Anal. Chem.* **2000**, *72*, 1156–1162.
69. R. E. March; J. F. J. Todd, *Quadrupole Ion Trap Mass Spectrometry*, 2nd ed.; John Wiley and Sons: Hoboken, NJ, USA, 2005.
70. J. C. Schwartz; M. W. Senko; J. E. P. Syka, *J. Am. Soc. Mass Spectrom.* **2002**, *13*, 659–669.
71. J. W. Hagar, *Rapid Commun. Mass Spectrom.* **2002**, *16*, 512–526.
72. M. Hardman; A. A. Makarov, *Anal. Chem.* **2003**, *72*, 1156–1162.
73. Q. Hu; R. J. Noll; H. Li; A. Makarov; M. Hardman; R. G. Cooks, *J. Mass Spectrom.* **2005**, *40*, 430–443.
74. M. W. Senko; J. D. Canterbury; S. Guan; A. G. Marshall, *Rapid Commun. Mass Spectrom.* **1996**, *10*, 1839–1844.
75. H.-K. Lim; J. Chen; C. Sensenhauser; K. Cook; V. Subrahmanyam, *Rapid Commun. Mass Spectrom.* **2007**, *21*, 1821–1832.
76. J. V. Olsen; L. M. F. de Godoy; G. Li; B. Macek; P. Mortensen; R. Pesch; A. Makarov; O. Lange; S. Horning; M. Mann, *Mol. Cell. Proteomics* **2005**, *4*, 2010–2021.
77. J. V. Olsen; B. Macek; O. Lange; A. Makarov; S. Horning; M. Mann, *Nat. Methods* **2007**, *4*, 709–712.
78. M. Guilhaus; D. Selby; V. Mlynski, *Mass Spectrom. Rev.* **2000**, *19*, 65–107.
79. N. Mirsaleh-Kohan; W. D. Robertson; R. N. Compton, *Mass Spectrom. Rev.* **2008**, *27*, 237–285.
80. M. P. Barrow; W. I. Burkitt; P. J. Derrick, *Analyst* **2005**, *130*, 18–28.
81. K. Breuker; M. Jin; X. Han; H. Jiang; F. W. McLafferty, *J. Am. Soc. Mass Spectrom.* **2008**, *19*, 1045–1053.
82. R. Guevremont, *J. Chromatogr. A* **2004**, *1058*, 3–19.
83. B. M. Kolakowski; Z. Mester, *Analyst* **2007**, *132*, 842–864.
84. S. D. Pringle; K. Giles; J. L. Wildgoose; J. P. Williams; S. E. Slade; K. Thalassionos; R. H. Bateman; M. T. Bowers; J. H. Scrivens, *Int. J. Mass Spectrom.* **2007**, *261*, 1–12.
85. I. V. Chernushevich; A. V. Loboda; B. A. Thompson, *J. Mass Spectrom.* **2001**, *36*, 849–865.
86. G. L. Glish; D. J. Burinski, *J. Am. Soc. Mass Spectrom.* **2008**, *19*, 161–172.
87. J. W. Hagar, *Anal. Bioanal. Chem.* **2004**, *378*, 845–850.
88. R. J. Cotter; W. Griffith; C. Jelinek, *J. Chromatogr. B* **2007**, *855*, 2–13.
89. R. Mihalca; Y. E. M. van der Burgt; L. A. McDonnell; M. Duursma; I. Cerjak; A. J. R. Heck; R. M. A. Heeren, *Rapid Commun. Mass Spectrom.* **2006**, *20*, 1838–1844.
90. L. M. Mikesh; B. Ueberheide; A. Chi; J. J. Coon; J. E. P. Syka; J. Shabanowitz; D. F. Hunt, *Biochim. Biophys. Acta* **2006**, *1764*, 1811–1822.
91. H. Han; Y. Xia; M. Yang; S. A. McLuckey, *Anal. Chem.* **2008**, *80*, 3492–3497.
92. P. Roepstorff; J. Fohlman, *Biomed. Mass Spectrom.* **1984**, *11*, 601.
93. R. A. Zubarev; K. F. Haaselmann; B. Budnik; F. Kjeldsen; F. Jensen, *Eur. J. Mass Spectrom.* **2002**, *8*, 337–349.
94. R. A. Zubarev, *Mass Spectrom. Rev.* **2003**, *22*, 57–77.
95. F. Kjeldsen; K. F. Haselmann; B. A. Budnik; F. Jensen; R. A. Zubarev, *Chem. Phys. Lett.* **2002**, *356*, 201–206.
96. M. M. Savitski; F. Kjeldsen; M. L. Nielsen; R. A. Zubarev, *J. Am. Soc. Mass Spectrom.* **2007**, *18*, 113–120.
97. B. A. Budnik; K. F. Haselmann; R. A. Zubarev, *Chem. Phys. Lett.* **2001**, *342*, 299–302.
98. F. Kjeldsen; O. A. Silivra; I. A. Ivonin; K. F. Haselmann; M. Gorshkov; R. A. Zubarev, *Chem. Eur. J.* **2005**, *11*, 1803–1812.
99. A. Kalli; K. Håkansson, *Int. J. Mass Spectrom.* **2007**, *263*, 71–81.
100. J. J. Wolff; I. J. Amster; L. Chi; R. J. Linhardt, *J. Am. Soci. Mass Spectrom.* **2006**, *18*, 234–244.
101. J. J. Wolff; T. N. Laremore; A. M. Busch; R. J. Linhardt; I. J. Amster, *J. Am. Soci. Mass Spectrom.* **2008**, *19*, 790–798.

102. J. J. Wolff; L. Chi; R. J. Linhardt; I. J. Amster, *Anal. Chem.* **2007**, *79*, 2015–2022.
103. M. Kaczorowska; H. J. Cooper, *J. Am. Soc. Mass Spectrom.* **2008**, *19*, 1312–1319.
104. J. Yang; J. Mo; J. T. Adamson; K. Håkansson, *Anal. Chem.* **2005**, *77*, 1876–1882.
105. J. Yang; K. Håkansson, *J. Am. Soc. Mass Spectrom.* **2006**, *17*, 1369–1375.
106. J. Mo; K. Håkansson, *Anal. Bioanal. Chem.* **2006**, *386*, 675–681.
107. J. E. P. Syka; J. J. Coon; M. J. Schroeder; J. Shabanowitz; D. F. Hunt, *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 9528–9533.
108. A. Chi; D. L. Bai; Y. L. Geer; J. Shabanowitz; D. F. Hunt, *Int. J. Mass Spectrom.* **2007**, *259*, 197–203.
109. R. Hartmer; D. A. Kaplan; C. R. Gebhardt; T. Ledertheil; A. Brekenfeld, *Int. J. Mass Spectrom.* **2008**, *276*, 82–90.
110. D. L. Swaney; G. C. McAlister; M. Wirtala; J. C. Schwartz; J. E. P. Syka; J. J. Coon, *Anal. Chem.* **2007**, *79*, 477–485.
111. R. A. Zubarev; A. R. Zubarev; M. M. Savitski, *J. Am. Soc. Mass Spectrom.* **2008**, *19*, 753–761.
112. M. M. Savitski; M. L. Nielsen; R. A. Zubarev, *Anal. Chem.* **2007**, *79*, 2296–2302.
113. A. Scherl; S. A. Shaffer; G. K. Taylor; P. Hernandez; R. D. Appel; P.-A. Binz; D. R. Goodlett, *J. Am. Soc. Mass Spectrom.* **2008**, *19*, 891–901.
114. M. Mann; N. L. Kelleher, *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 18132–18138.
115. H. Molina; R. Matthiesen; K. Kandasamy; A. Pandey, *Anal. Chem.* **2008**, *80*, 4825–4835.
116. F. Kjeldsen; O. B. Horning; S. S. Jensen; A. M. B. Giessing; O. N. Jensen, *J. Am. Chem. Soc.* **2008**, *19*, 1156–1162.
117. C. K. Barlow; R. A. O'Hair, *J. Mass Spectrom.* **2008**, *43*, 1301–1319.
118. J. M. Halket; V. G. Zaikin, *Eur. J. Mass Spectrom.* **2006**, *12*, 1–13.
119. V. G. Zaikin; J. M. Halket, *Eur. J. Mass Spectrom.* **2005**, *11*, 611–636.
120. C. Hopley; T. Bristow; A. Lubben; A. Simpson; E. Bull; K. Klagkou; J. Herniman; J. Langley, *Rapid Commun. Mass Spectrom.* **2008**, *22*, 1779–1786.
121. C. Eckers; J. J. Monaghan; J.-C. Wolff, *Eur. J. Mass Spectrom.* **2005**, *11*, 73–82.
122. K. E. Arthur; J.-C. Wolff; D. J. Carrier, *Rapid Commun. Mass Spectrom.* **2004**, *18*, 678–684.
123. Z. Tozuka; H. Kaneko; T. Shiraga; Y. Mitani; M. Beppu; S. Terashita; A. Kawamura; A. Kagayama, *J. Mass Spectrom.* **2003**, *38*, 793–808.
124. J.-C. Wolff; L. A. Thompson; C. Eckers, *Rapid Commun. Mass Spectrom.* **2003**, *17*, 215–221.
125. F.-F. Hsu; J. Turk, *J. Am. Soc. Mass Spectrom.* **2005**, *68*, 1510–1522.
126. F.-F. Hsu; J. Turk, *J. Am. Soc. Mass Spectrom.* **2007**, *18*, 2065–2073.
127. J. Heberle; G. Simchen, *Silylating Agents*, 2nd ed.; Fluka Chemie AG: Buchs, Switzerland, 1995.
128. O. Fiehn, *Phytochemistry* **2003**, *62*, 875–886.
129. H. Van den Dool; P. D. Kratz, *Chromatogr.* **1963**, *11*, 41 463–471.
130. J. H. Beynon, *Nature* **1954**, *174*, 735–737.
131. A. W. T. Bristow, *Mass Spectrom. Rev.* **2005**, *25*, 99–111.
132. M. L. Gross, *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 57.
133. R. J. DeLaeter; J. K. Böhlke; P. De Bièvre; H. Hidaka; H. S. Peiser; K. J. R. Rosman; P. D. P. Taylor, *Pure Appl. Chem.* **2003**, *75*, 683–800.
134. O. A. Mamer; A. Lesimple, *J. Am. Soc. Mass Spectrom.* **2004**, *15*, 626.
135. I. Ferrer; E. M. Thurman, *Rapid Commun. Mass Spectrom.* **2007**, *21*, 2538–2539.
136. J. F. Pickup; K. McPherson, *Anal. Chem.* **1976**, *48*, 1885–1890.
137. C. S. Hsu, *Anal. Chem.* **1984**, *56*, 1356–1361.
138. V. Pellegrin, *J. Chem. Educ.* **1983**, *60*, 626–632.
139. T. Kind; O. Fiehn, *BMC Bioinf.* **2006**, *7*, 234–243.
140. C. Dass, *Fundamentals of Contemporary Mass Spectrometry*; John Wiley and Sons: Hoboken, NJ, USA, 2007.
141. M. N. Eberlin, *J. Mass Spectrom.* **2006**, *41*, 141–156.
142. R. K. Boyd; J. D. Henion; M. Alexander; W. L. Budde; J. D. Gilbert; S. M. Musser; C. Palmer; E. K. Zurek, *J. Am. Soc. Mass Spectrom.* **1996**, *7*, 211–218.
143. R. K. Boyd; C. Basic; R. A. Bethem, *Trace Quantitative Analysis by Mass Spectrometry*; John Wiley and Sons: Chichester, UK, 2008.
144. FDA Guidance for Industry 118, Mass spectrometry for confirmation of the identity of animal drug residues, US Department of Health and Human Services, Food and Drug Administration, May 2003. http://www.fda.gov/ (accessed April 2009).
145. European Union COMMISSION DECISION 2002/657/EC of 12 August 2002, *Official J. Eur. Union* **2002**, *L 221*, 8–36.
146. M. W. F. Nielen; M. C. van Engelen; R. Zuiderent; R. Ramaker, *Anal. Chim. Acta* **2007**, *586*, 122–129.
147. W. Weckwerth; K. Wenzel; O. Fiehn, *Proteomics* **2004**, *4*, 78–83.
148. S. A. Trauger; E. Kalisak; J. Kalisiak; H. Morita; M. V. Weinberg; A. L. Menon; F. L. Poole, II; M. W. W. Adams; G. Siuzdak, *J. Proteome Res.* **2008**, *7*, 1027–1035.
149. G. G. Harrigan, R. Goodacre, Eds., *Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 2003.
150. FDA Guidance for Industry, Bioanalytical Method Validation, US Department of Health and Human Services, Food and Drug Administration, May 2001. http://www.fda.gov/ (accessed April 2009).
151. E. Pritchard; V. Barwick, *Quality Assurance in Analytical Chemistry*; John Wiley and Sons: Chichester, UK, 2007.
152. S. G. Villas-Bôas; S. Mas; M. Åkesson; J. Smedsgaard; J. Nielsen, *Mass Spectrom. Rev.* **2005**, *24*, 613–646.
153. A. Nordström; E. Want; T. Northen; J. Lehtiö; G. Siuzdak, *Anal. Chem.* **2008**, *80*, 421–429.
154. L. Cuadros-Rodríguez; M. G. Bagur-González; M. Sánchez-Viñas; A. González-Casado; A. M. Gómez-Sáez, *J. Chromatogr. A* **2007**, *1158*, 33–46.
155. W. Schwab, *Phytochemistry* **2003**, *62*, 837–849.
156. D. E. Raynie, *Anal. Chem.* **2004**, *76*, 4659–4664.
157. S. G. Villas-Bôas, Sampling and Sample Preparation. In *Metabolome Analysis: An Introduction*; S. G. Villas-Bôas, U. Roessner, M. A. E. Hansen, J. Smedsgaard, J. Nielsen, Eds.; John Wiley and Sons: Hoboken, NJ, USA, 2007.
158. A. Jiye; J. Trygg; J. Gullberg; A. I. Johansson; P. Jonsson; H. Antti; S. L. Marklund; T. Moritz, *Anal. Chem.* **2005**, *77*, 8086–8094.

159. J. Gullberg; P. Jonsson; A. Nordström; M. Sjöström; T. Moritz, *Anal. Biochem.* **2004**, *331*, 283–295.
160. R. King; R. Bonfiglio; C. Fernandez-Metzler; C. Miller-Stein; T. Olah, *J. Am. Chem. Soc.* **2000**, *11*, 942–950.
161. J. Schuhmacher; D. Zimmer; F. Tesche; V. Pickard, *Rapid Commun. Mass Spectrom.* **2003**, *17*, 1950–1957.
162. W. Weckwerth, Ed., *Metabolomics: Methods and Protocols*; Humana Press: Totowa, NJ, USA, 2007.
163. B. J. Millard, *Quantitative Mass Spectrometry*; Heydon and Sons Ltd.: London, UK, 1978.
164. P. C. Meier; R. E. Zünd, *Statistical Methods in Analytical Chemistry*; John Wiley and Sons: New York, NY, USA, 1993.
165. M. P. Hermo; D. Barrón; J. Barbosa, *J. Chromatogr. A* **2008**, *1201*, 1–14.
166. M. Ende; G. Spiteller, *Mass Spectrom. Rev.* **1982**, *1*, 29–62.
167. B. O. Keller; J. Sui; A. B. Young; R. M. Whittal, *Anal. Chim. Acta* **2008**, *627*, 71–81.
168. Q&A News Article, *Nature* **2008**, *453*, 964.
169. G. R. McDonald; A. L. Hudson; S. M. J. Dunn; H. You; G. B. Baker; R. M. Whittal; J. W. Martin; A. Jha; D. E. Edmondson; A. Holt, *Science* **2008**, *322*, 917.
170. X. Guo; A. P. Bruins; T. R. Covey, *Rapid Commun. Mass Spectrom.* **2006**, *20*, 3145–3150.
171. M. Najam-ul-Haq; M. Rainer; C. W. Huck; P. Hausberger; H. Kraushaar; G. K. Bonn, *Anal. Chem.* **2008**, *80*, 7467–7472.
172. N. C. Hughes; E. Y. K. Wong; J. Fan; N. Bajaj, *AAPS J.* **2007**, *9*, E353–E360.
173. J. W. Dolan, *LCGC North Am.* **2006**, *24*, 754–760.
174. M. Stoeckli; P. Chaurand; D. E. Hallahan; R. M. Caprioli, *Nat. Med.* **2001**, *7*, 493–496.
175. R. M. Caprioli; T. B. Farmer; J. Gile, *Anal. Chem.* **1997**, *69*, 4751–4760.
176. B. Spengler, Microprobing and Imaging MALDI for Biomarker Detection. In *MALDI MS: A Practical Guide to Instrumentation, Methods and Applications*; F. Hillenkamp, J. Peter-Katalinic, Eds.; Wiley-VCH Verlag GmbH and Co.KGaA: Weinheim, Germany, 2007.
177. W. M. Hardesty; R. M. Caprioli, *Anal. Bioanal. Chem.* **2008**, *391*, 899–903.
178. J. S. Fletcher; N. P. Lockyer; S. Vaidyanathan; J. C. Vickerman, *Anal. Chem.* **2007**, *79*, 2199–2206.
179. D. R. Ifa; J. M. Wiseman; Q. Song; R. G. Cooks, *Int. J. Mass Spectrom.* **2007**, *259*, 8–15.
180. R. L. Caldwell; R. M. Caprioli, *Mol. Cell. Proteomics* **2005**, *4*, 394–401.
181. D. S. Cornett; J. A. Mobley; E. C. Dias; M. Anderson; C. L. Arteaga; M. E. Sanders; R. M. Caprioli, *Mol. Cell. Proteomics* **2006**, *5*, 1975–1983.
182. P. Chaurand; S. A. Schwartz; D. Billheimer; B. J. Xu; A. Crecelius; R. M. Caprioli, *Anal. Chem.* **2004**, *76*, 1145–1155.
183. S. M. Puolitaival; K. E. Burnum; D. S. Cornett; R. M. Caprioli, *J. Am. Soc. Mass Spectrom.* **2008**, *19*, 882–886.
184. P. Chaurand; J. C. Latham; K. B. Lane; J. A. Mobley; V. V. Polosukhin; P. S. Wirth; L. B. Nanney; R. M. Caprioli, *J. Proteome Res.* **2008**, *7*, 3543–3555.
185. V. Kertesz; G. J. Van Berkel, *Rapid Communi. Mass Spectrom.* **2008**, *22*, 2639–2644.
186. G. Maddalo; F. Petrucci; M. Iezzi; T. Pannellini; P. Del Boccio; D. Ciavardelli; A. Biroccio; F. Forli; C. Di Ilio; E. Ballone; A. Urbani; G. Federici, *Clin. Chim. Acta* **2005**, *357*, 210–218.
187. Z. Ouyang; L. Gao; M. Fico; W. J. Chappell; R. J. Noll; R. G. Cooks, *Eur. J. Mass Spectrom.* **2007**, *13*, 13–18.

Biographical Sketch

Charles H. Hocart, descendent of Guernsey and Yorkshire stock, completed a B.Sc. (Hon) degree (Chemistry, 1977) and an M.Sc. (Clinical Biochemistry, 1979) at the University of Western Australia, where his long-term interest in the use of mass spectrometry to solve biological problems was aroused. His Ph.D. studies with Professor Berthold Halpern (Wollongong) using GC/MS to look for markers of metabolic diseases were unfortunately curtailed by Halpern's premature death. He then moved to Canberra to study the cytokinin family of plant hormones under the supervision of Professor Stuart Letham (Australian National University, 1981–85). Two postdoctoral fellowships then followed with Professor Jim McCloskey (Utah, 1986–87) and Professor Urs Schlunegger (Bern, 1988–99) for identifying modified nucleotides from Archaebacteria, and for studying the negative ion mass spectrometry of cytokinins, respectively, before he again started to work with Letham, courtesy of a National Research Fellowship (ANU, 1989–93). A 3-year period as an analyst and evaluator at the Therapeutic Drug Administration then led to his current appointment managing the RSBS Mass Spectrometry Facility (ANU, 1996–present). Research collaborations have included the archaeology of betel nut and kava in the Pacific islands (Dr. Barry Fankhauser); the identification and quantification of cell wall polysaccharides in cellulose-deficient *Arabidopsis thaliana* mutants (Professor Richard Williamson); the chemical signals regulating the interaction of rice with *Rhizobium* isolates from the Nile Delta (Professor Barry Rolfe); and most recently, the generation of biodiesel from the native leguminous tree, *Pongamia pinnata*, and from microalgae (Professor Peter Gresshoff and Dr. Michael Djordjevic).

9.11 Applications of Modern Mass Spectrometry Techniques in Natural Products Chemistry

Roland D. Kersten, Michael J. Meehan, and Pieter C. Dorrestein, University of California, San Diego, La Jolla, CA, USA

9.11.1 Introduction

The current emergence of drug resistance and higher incidence of diagnosed illnesses, such as cancer, Alzheimer's disease, and diabetes, coupled with an increasing world population have resulted in an increased interest in the study of natural products and their biosynthetic machineries. More than 50% of all therapeutics have origins in natural products with many more currently in clinical trials.[1] The recognition that many organisms have the metabolic capacity to produce a large number of natural products is leading to increased availability of sequenced genomes, which is further resulting in the betterment of the instruments used to study them. There are many tools to characterize natural products and their biosynthetic machinery including nuclear magnetic resonance (NMR), high-performance liquid chromatography (HPLC), ultraviolet–visible spectroscopy (UV–vis), infrared (IR), circular dichroism (CD), and various x-ray diffraction techniques (X-ray). Standing out among these common methods is mass spectrometry (MS). MS has become so essential in the study of these systems that there is not a scientific journal in the world that would accept the characterization of a new natural product in the absence of MS data. Nevertheless, unlike NMR where new methods for the characterization of molecules are frequently developed, the use of creative MS has been rather limited within the general biosynthetic natural product community. MS is usually done as an afterthought; once one has already obtained activities by other means. This, however, is changing and MS is moving to the forefront of many investigations. The reason for this paradigm shift is the ever-changing landscape of modern MS tools. This chapter will emphasize how modern MS has been utilized to uncover the hidden features of natural product biosynthesis.

The past few years have seen a substantial increase in the capacity of commercial MS instruments. These changes are partially driven by the clinical 'omics' community, but are also found to be very useful in other areas of science. Unlike in the past where promising developments in MS would take a decade or longer to reach the general public, new developments in MS are being quickly commercialized on user-friendly instruments. For example, the tandem mass spectrometry (MS/MS or MS^2) method of electron transfer dissociation (ETD) was originally published in 2004 and the first commercial version came out in 2005.[2] ETD and its, by reactivity only, related cousin the electron capture dissociation[3] represent just two of the many recent advances in MS. The past 5 years alone have born witness to an explosion of advances in resolution, sources acquisition speeds, data processing, and ionization sources.[4–19] The current rate of development of MS tools indicates that this chapter too will have aged by the time this review is published. Therefore, this review will not only serve as a snapshot of widely used MS approaches in the biosynthetic investigations of natural products, but also aims to provide a glimpse into the short- and long-term future capacities of MS in the field.

The emphasis of this review is placed on two structural classes of natural products: polyketides and nonribosomal peptides (NRPs). The MS of these biosynthetic pathways is most advanced and will be covered in detail. In the following sections we will describe the current methods and applications used to study the biosynthetic pathways of natural products and provide a glimpse into upcoming techniques. In addition, a brief introduction to experimental design using high-end MS to study the biosynthesis of other natural metabolites, such as ribosomally encoded pathways and cofactors, is described.

9.11.1.1 Introduction to NRPS and PKS Biosynthesis

Many important therapeutics, in use in clinics today, are biosynthesized by the nonribosomal peptide synthetase (NRPS) and polyketide synthase (PKS) paradigm. For example, many of the antibiotics (penicillin, cephalosporin, vancomycin, erythromycin, etc.), immunosuppressors (cyclosporine, rapamycin), antiviral agents (luzopeptin A), antitumor agents (bleomycin), and toxins (thaxtomin) are NRPS and PKS derived.[20–22] **Figure 1** displays a small selection of natural products that are NRPS and PKS derived and illustrates the diversity of molecular structures generated by these biosynthetic paradigms.

NRPS biosynthesis differs substantially from ribosomally encoded peptides. In the NRPS biosynthetic strategy, the substrates and intermediates are covalently linked to an active site serine on the thiolation (T) domain of an NRPS via the phosphopantetheinyl arm.[23] This posttranslational event is accomplished by a phosphopantetheinyl transferase (PPTase), which primes the active site serine with a portion of coenzyme A (CoA) to generate the holo form of the T domain (**Figure 2(b)**).[24] In traditional NRPS biosynthesis, amino

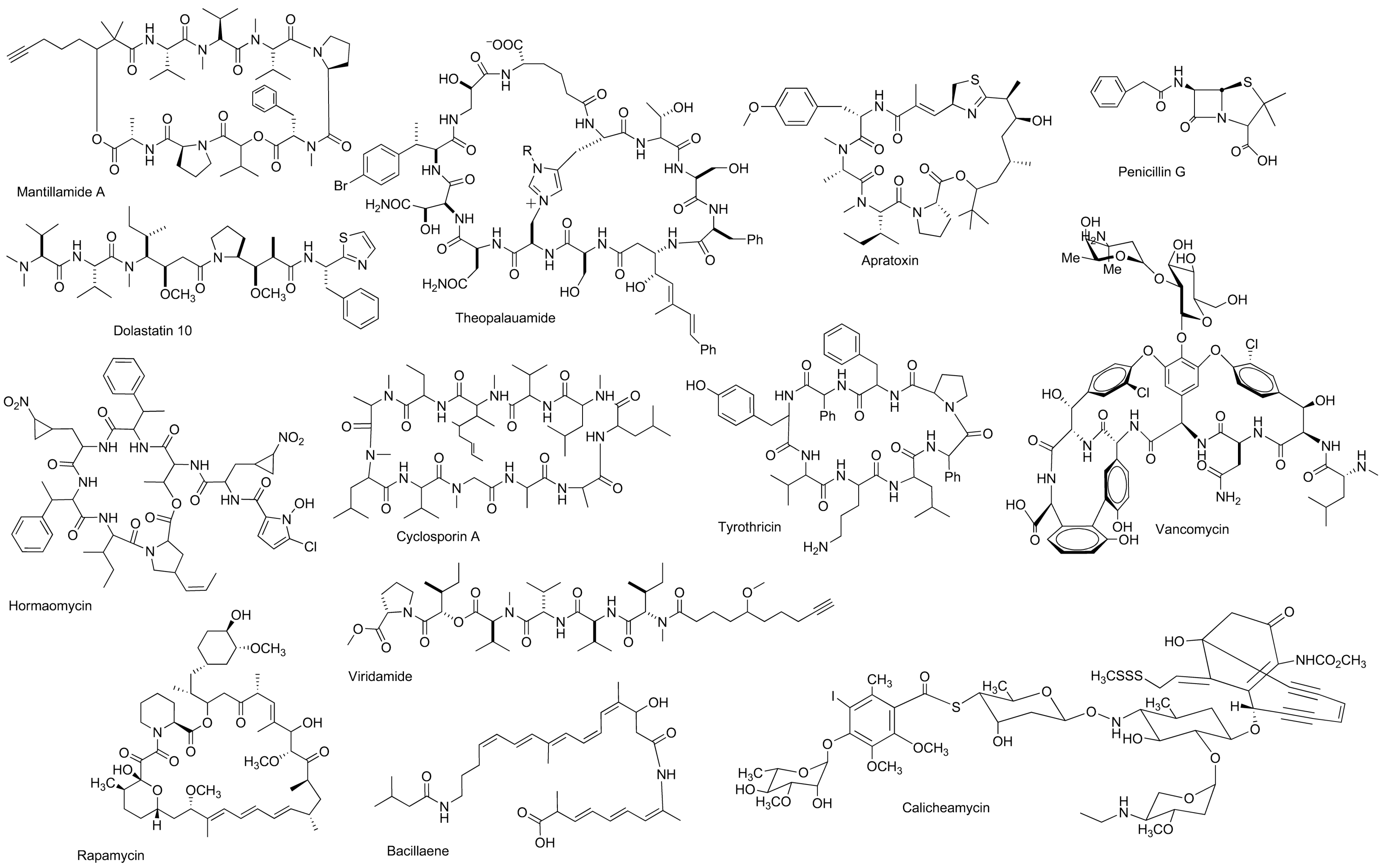


Figure 1 Some examples showing the diversity of NRPS- and PKS-derived natural products.

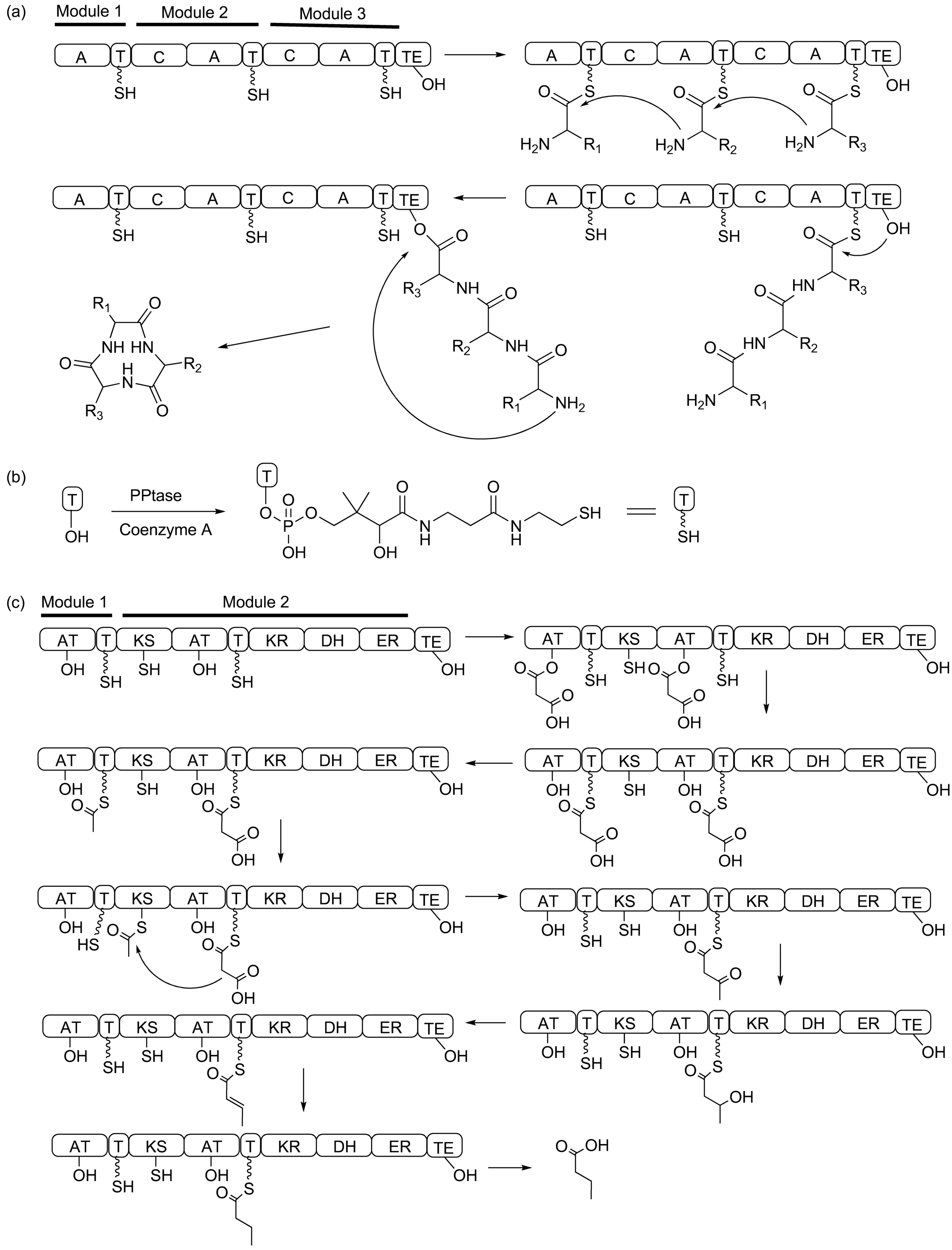


Figure 2 Overview of NRPS and PKS biosynthesis. (a) NRPS biosynthesis. (b) Priming of apo T domain with coenzyme A by a phosphopantetheinyl transferase. (c) PKS biosynthesis.

acids are activated by adenosine triphosphate (ATP). Activation is achieved by the formation of an adenylated carboxylate of an amino acid that is encoded by the first adenylation (A) domain (**Figure 2(a)**).[25] The activated amino acid then undergoes nucleophilic displacement by the thiol terminus of the phosphopantetheine (PPant) arm tethered to the first T domain. During the elongation step, the resulting aminoacyl-enzyme is then condensed with a second acyl–enzyme species on a downstream carrier domain to form a linear peptide, in a reaction catalyzed by a condensation (C) domain. Further modifications often take place while the amino acid or the growing peptide is still attached to a carrier domain (e.g., oxidation, chlorination, or other types of modifications).[26] Finally, the mature peptide is released from the last carrier domain via cyclization or removed by a termination domain.[27] The most common off-loading domain is a thioesterase (TE),[28] although other termination domains exist. While there are currently 22 known ribosomally encoded amino acids, there are hundreds of nonribosomally encoded amino acids expanding the array of structures (and therefore pharmacologically active compounds) that can be generated via the NRPS paradigm.[29]

Just like NRPS biosynthesis, polyketide biosynthesis often takes place on multidomain megasynthases.[23] PKSs also carry their substrates and intermediates on T domains, also referred to as acyl carrier proteins (ACPs). The substrates in PKS biosynthesis are not amino acids or free carboxylic acids as it is observed with NRPS systems, but rather they are CoA-activated carboxylic acids. The most common substrate for polyketide biosynthesis is malonyl-CoA. In order to divert a small subfraction of the cellular pool of malonyl-CoA, PKSs activate the malonyl by a nucleophilic attack of the active site serine from the acyltransferase (AT) domain forming a covalent oxo-ester linkage (**Figure 2(c)**). Subsequently, a transacylation takes place onto the T domain linking the malonyl to the thiol of the phosphopantetheinyl functionality via a thioester bond. While the malonyl is still attached to the T domain, it is decarboxylated by the ketosynthase (KS) domain before a second transacylation takes place to the active site cysteine of the KS domain generating an acetyl-*S*-KS intermediate, something that to date is very hard to capture directly by MS. At this point, a second malonyl group is loaded onto the upstream T domain, a reaction again catalyzed by an AT domain. As the KS domain decarboxylates the second malonyl, a Claisen condensation takes place with the acetyl group on the KS domain forming acetoacetate, the first unit in an elongating polyketide. Like NRPS biosynthetic pathways, PKS biosynthesis can incorporate a variety of tailoring reactions.[23] Tailoring reactions such as chlorination, oxidation, and cyclization may be observed.[30–32] By far the most common tailoring steps in PKS biosynthesis are ketoreduction, dehydration, and enoyl-reductions. Ketoreduction is catalyzed by the ketoreductase (KR) domain and uses the cofactor nicotinamide adenine dinucleotide (phosphate) hydride (NAD(P)H) to convert the β-carbonyl to an alcohol. This same alcohol can be dehydrated by a dehydratase (DH) domain. The DH domain catalyzes the deprotonation of the α-proton and protonates the leaving hydroxyl group in an α,β-unsaturated thioester. This thioester can then be further reduced to a completely saturated bond by the enoylreductase (ER). This reaction utilizes the cofactor NAD(P)H as required in the ketoreduction.

While there are many natural products that are either NRPS or PKS derived, there are multidomain megasynthases that contain both NRPS and PKS biosynthetic features. Examples of such biosynthetic motifs are epothilone, jamaicamide, and the endiyne C-1027.[33–35] At NRPS–PKS interfaces in these systems, the condensation of malonyl takes place with an amino acid that was activated by an A domain. Alternatively, an amino acid loaded on a T domain condenses with an elongating polyketide. Since the substrates and intermediates in NRPS, PKS, or hybrid NRPS/PKS biosynthesis introduce mass changes on T domains, they are ideal candidates for investigation by MS.

9.11.1.2 New Tools in the Characterization of Multidomain and Phosphopantetheinylated Proteins

For many years, it was very difficult to study multidomain, posttranslationally, or transiently modified proteins. These studies required several years of efforts by MS experts. But as instrumentation improved it has enabled additional investigators to carry out this type of research. Many proteins, in particular NRPS and PKS, are of substantial size. Some of the largest single open reading frames are responsible for the generation of mega dalton polypeptides. An example of this are the 2.5 MDa NRPSs of the syringomycin pathway and the megasynthase of the biosynthetic pathway of an 18-mer peptaibol.[36,37] In the majority of cases, it is not possible to look at intact proteins of this size, so they must be digested before the active sites can be analyzed in terms of tethered biosynthetic substrates and intermediates. However, when these proteins are digested, the complexity increases and the active sites have to be found within a haystack of data. Many modern proteomic programs such as InSpecT, Sequest, OMSSA, Spectumill, Mascott, and MassLynx fail to find these active site peptides or domains of posttranslationally modified megasynthases from complete or limited digests.[38,39] There are three main reasons for this: (1) Despite the surge in the development of proteomic platforms for the analysis of digested proteins, all proteomic programs have a difficult time identifying ions based on the fragmentation data of parent ion that have charges that are greater than 4+. Most of these programs, by default, ignore 4+ ions or larger as this would increase the number of false positives. We estimate that approximately 30% of all the data in a liquid chromatography–mass spectrometry (LC–MS) of a complete tryptic digest would belong to these higher charges. In addition, because of the poorer nature of 1+ ion fragmentation, these too are often not identified. (2) While many of these programs can identify nonlabile modifications, they have to be 'trained' for labile modifications should it be possible to find them at all. For example, most of these programs have a specific scoring function for the analysis of phosphopeptides and none of these programs have a scoring function for labile modifications such as the phosphopantetheinylation found on fatty acid synthase (FAS), NRPS, and PKS multidomain proteins. (3) For a given LC–MS/MS or MS^2 run for a digest, we typically obtain 10 000–20 000 MS/MS or MS^2 spectra, yet on a good run only 2000–3000 MS/MS or MS^2 spectra are annotated. This begs the question what the remaining spectra are and clearly indicate that we are not yet identifying everything we can from the data. Thus there is an enormous opportunity for the development of programs that can capture the remaining uncharacterized spectra, including 4+ or greater charge ions or ions with different labile modifications.

Even though automated programs to facilitate the analysis of these large multidomain proteins involved in the biosynthesis of natural products do not exist, it has become possible to investigate such systems by MS. After a short protease digestion, most of the active sites are often 5–20 kDa. It is not yet reasonable to map the active sites by MS/MS or MS^2 data as it is done in proteomic investigations.[40] However, with the emergence of high-resolution instruments that have routine mass accuracies within 10 ppm, it became possible to find the active sites by the intact mass of the peptides alone giving rise to a manageable number of false positive matches. Any false match is eliminated by MS/MS or MS^2. When one performs MS/MSor MS^2, many fragment ions are detected that should match up to expected fragments for the active sites. The false positive active sites matches are eliminated on the basis that their fragments will not match up to the sequence of the active site, while all of the true positive candidate fragments match quite well. In 1999, Kelleher and coworkers were the first to recognize the usefulness of Fourier transform–ion cyclotron resonance–mass spectrometry (FT–ICR–MS) resolution and mass accuracy for the investigations of multidomain NRPS proteins.[41] At that time, FT–ICR–MS instruments were the only instruments with this degree of mass accuracy and the best of these instruments were custom built and not readily available to the general public. As instrumentation and the development of efficient strategies to analyze multidomain NRPS/PKS proteins are advancing, more and more laboratories can carry out these types of investigations, including laboratories that only have access to low-resolution instruments.[42–46] In the following sections, we describe the current and new MS approaches to analyze some of these biosynthetic pathways since a 2006 review on this topic.[47]

There are several key advances that have been in use since 2006 for the investigations of NRPS and PKS proteins: (1) The use of larger-sized magnets (e.g., 12 Tesla) enables active site mapping on an LC time-scale.[48,49] (2) It was discovered that the labile posttranslational modification of T domains can be ejected during thermal activation methods and that this ejection can be used to 'observe' substrates and intermediates on the active site thiol of a phosphopantetheinyl functionality. This assay, called the phosphopantetheinyl ejection assay (PEA), has found many uses and has been adopted by several other laboratories already.[48–58] (3) The PEA can also be applied to low-resolution instruments making this assay accessible to many other researchers who work with phosphopantetheinylated proteins.[43,44] (4) Finally, there are now instances describing other methods for analyzing the phosphopantetheinylated proteins, such as the use of a phosphatase to remove the phopshopantetheinyl functionality so that it may be characterized by MS. All of these new MS capabilities are highlighted in the next sections.

9.11.1.3 A Brief Introduction to FT–ICR–MS

The application of FT–ICR–MS analysis to the characterization of NRPS and PKS proteins is critical as this type of analysis has accelerated our understanding of these complex systems over the past decade. It is important to introduce some of the fundamentals of FT–ICR–MS in order to understand the advantage of this technique in NRPS/PKS research. It should be noted that this is not meant as an authorative review on the subject but as a simplified introduction to the complexities of FT–ICR–MS for a reader who has never been exposed to this type of instrumentation. For a deeper insight, there are several very good FT–ICR–MS reviews on the subject that should be consulted.[59–61] FT–ICR–MS is an image current-based detection strategy.[59–61] This means that ions are detected by a perturbation of a current that is captured as a frequency. To accomplish this, charged ions are introduced into a detection cell about the size of a soda can. Inside of this soda can-sized cell, there are excitation plates on which an alternating current is applied to generate a cyclotron motion to the charged ions. The basis of FT–ICR–MS is this ion cyclotron motion, which arises from the interaction of an ion with a unidirectional magnetic field. The ion experiences a force, the Lorentz force, which causes this ion to travel in a circular orbit perpendicular to this magnetic field. The radius of the orbital motion is defined by the magnetic field strength. While the charged ion is in an orbital motion it passes the detector plates where the ion perturbs the current. This current can be very accurately measured as a time domain. The resulting time domain (free induction decay (FID)) can be subjected to Fourier transform to obtain the frequency of the ions undergoing the orbital motion. This frequency can then be converted to a mass measurement with Equation (1).

$$f_{\mathrm{cyc}} = \frac{zB}{2\pi m} \tag{1}$$

where f_{cyc} is the cyclotron frequency, z the charge of the ion, B the magnetic field strength, and m the ion mass. The magnetic field of the spectrometer is held constant, provided by an ultrahighly stable superconducting magnet, and the mass-to-charge ratio of the ion (m/z) is determined by measuring its f_{cyc}. Since frequency measurement is inherently accurate and can be measured more accurately than other physical properties,[59] the FT–ICR mass spectrometer offers superb mass resolution and mass accuracy. Equation (1) also shows that by increasing the magnet field strength, the resolving power and scan speed increase in a linear fashion.[59] The ability to collect a good FID is important for high resolution. Truncating the FID results not only in an increased scan rate but also in a loss of fine information. This is one of the reasons why FT–ICR–MS instruments needs very high vacuum, which is $10\mathrm{E}^{-10}$ Torr inside the analyzer cell. Without high vacuum, the ions collide with other ions from air or gas and the FID beads out, essentially truncating the FID. The effect of truncation on the resulting resolution is shown in **Figure 3**. Many commercial instruments may not use the wording 'truncation of FID' and they may simply refer to resolutions of 100 000, 50 000, or others. To accomplish the different resolutions, these commercial manufacturers truncate the FID or expand the time domain collection. How is this relevant experimentally? The larger the FID, the greater the resolution, but the longer the scan time is for a single scan event. Thus there is a trade-off when collecting data at high resolution. For example, if one collects data at the highest resolution a single scan can take place from seconds to minutes per scan. This would not be optimal to use on an LC timescale. Therefore, one must weigh the importance of scan rate versus the required resolution carefully when performing LC–MS with an FT–ICR–MS instrument.

9.11.1.4 Interpretation of FT–ICR–MS

Due to the fact that most people investigating natural product biosynthesis do not routinely use FT–ICR–MS in their research, the interpretation of a high-resolution mass spectrum of a protein domain is highlighted in this section. The broadband spectrum of a freestanding thiolation protein Pks4 from the bikaverin biosynthetic pathway is shown in **Figure 4(a)** as a mixture of its apo and holo form.[62] This protein has a mass of 14 394 Da. The mass range for FT–ICR–MS analysis of proteins involved in the biosynthesis of natural products must fall between 200 and 2000 m/z. In order to enable the visualization of these ions within this standard m/z window, ions of Pks4, or other protein domains that are larger than 2000 Da, need to be multiply charged (z), which is experimentally accomplished by application of electrospray ionization (ESI). In the case of Pks4, we see that the same protein has multiple charge states ranging from 15 charges on the left side of the spectrum to charge 11 on

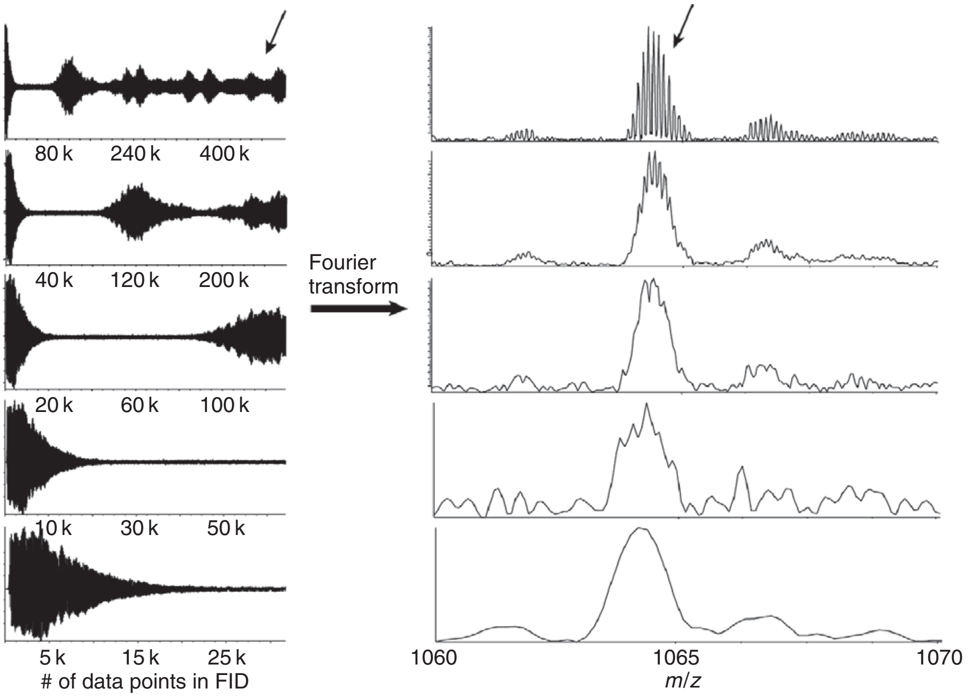


Figure 3 Correlation between the time domain duration and mass resolution. The more data points the FID (left) consists of, the higher the resolution of the mass signals (right) obtained by Fourier transform. The arrow on the left points to the beads in the time domain signal that accounts for the fine information (i.e., isotopes) shown by the arrow on the right.

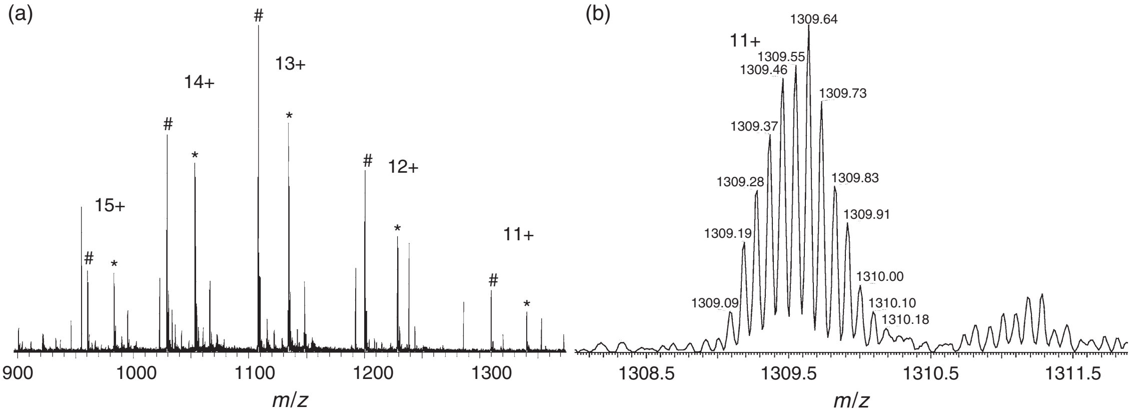


Figure 4 ESI–FT–ICR mass spectrum of freestanding T domain Pks4 from bikaverin biosynthetic pathway. (a) Multiple charge states of apo (#) and holo (*) Pks4. (b) Single charge state (11+) of Pks4 shows isotopic resolution obtained by FT–ICR–MS analysis.

the right side of the spectrum. The mass is calculated by taking the observed *m*/*z* and multiplying this value with the observed charge. Details on how to calculate the mass of such a protein or protein domain by manual means are reviewed by Dorrestein and Kelleher.[47] Pks4 is a relatively small protein and proteins of a size <20 kDa are usually well suited for FT–ICR–MS analysis. Larger proteins pose more challenges to MS characterization. For instance, if a protein of 100 kDa is analyzed, it is not uncommon to find between 100 and 150 charges on the protein and that the charge profile has >100 different charge states as opposed to the 5 observed for Pks4. Figure 4(b) shows an enlarged picture of a single charge state. When one zooms in on a charge state observed in Figure 4(a), an isotopic profile of the ion becomes visible. This type of characteristic isotopic profile exists because larger proteins, or protein domains, have many carbons, nitrogens, and oxygens. The natural abundance of isotopes other than the monoisotopic masses (e.g., ^{12}C, ^{14}N, and ^{16}O) is the reason why an isotopic profile is observed for a protein. The main forms of isotopes that contribute to the isotopic profile of a protein are the natural ^{13}C and ^{15}N that are present at ~1 and 0.36%, respectively. FT–ICR–MS has the resolving power that enables the visualization of the isotopes of proteins up to 110 kDa in mass.[63] It should be noted however that, in practice, the larger the protein, the more difficult it is to get isotopic resolution and that the theoretical maximum isotopic resolution of large multidomain proteins with higher field magnets (>12T) has not yet been achieved experimentally.[59]

9.11.1.5 LC–FT–ICR–MS Analysis of NRPS and PKS Proteins

Before 2006, all NRPS and PKS proteins were investigated via direct infusion of a protein or protein domain that was purified offline by HPLC, or by other C_{18}, C_8, or C_4 forms of peptide/protein purification such as Ziptips or traps. The best results were obtained using nanospray infusion, such as an Advion nanospray robot or similar nanospray devices. The advantage of a nanospray over direct microspray infusion via a syringe, a commonly used infusion method, is twofold. First, nanospray creates finer droplets than traditional forms of electrospray, making the desolvation of the droplets emitted from the spray needle or nozzle easier, and many more ions enter into the gas phase improving the detection. Second, with just 5–10 μl of a purified sample, one can analyze this sample for over an hour, sometimes up to 3 h. With a syringe infusion approach, the infusion rate is usually 2 μl min^{-1}. This is a significant limitation in situations where one has limited sample or a low concentration sample. A single FT–ICR–MS scan may take 8–120 s depending on the selected settings, and typically, 20–200 scans are required to attain a good spectrum. Therefore, nanospray becomes a critical component in the analytical platform for the biosynthetic investigations of NRPS and PKS systems. The reason why so many scans are required is illustrated in **Figure 5**. **Figure 5** shows that even when the sample ionizes well, a single scan is not sufficient for accurate data. As shown for Pks4, it takes ~10 scans to generate an accurate isotopic profile for the protein. Many more scans may be required if the sample's concentration is low, does not ionize well, or contains competing ions. Depending on the conditions, acquisition of data may take minutes to hours, and is unlike NMR where some experiments run overnight. While dependent on many factors, in practice, for every two-fold signal-to-noise improvement needed in order to collect the data, the data acquisition time needs to be increased four-fold. However, as the magnet size of an FT–ICR instrument increases, the sensitivity, and therefore the scan rate, increases compared to when the data are obtained at the same resolution on an instrument with a smaller magnet.

NRPS/PKS active site mapping on the LC timescale is possible by online LC–FT–ICR–MS. While this approach is still difficult to perform with a 7T magnet, the most common magnet size at this time, it becomes routine with a 12T magnet. The challenges with online active site mapping are the inherent signal limitation of the electrospray and the long accumulation times of FT–ICR–MS instruments. The first online LC–FT–ICR–MS analysis of a thiolation active site was described by the Marahiel group in 2006.[42] In their LC trace, they were able to observe a 1097.995 Da ion that corresponded to the posttranslationally phosphopantetheinylated form of the T domain of tyrocidine synthase B.[42] In addition, they were able to observe the amino acid Phe and dipeptide Phe–Phe loaded onto the active site thiol of the phosphopantetheinyl group. If the digestion had resulted in a 15 kDa protein domain containing the active site, these data would have been most likely much more difficult to obtain on an LC timescale. Since then, the Kelleher laboratory has built a 12T FT–ICR–MS instrument enabling routine online LC–FT–ICR–MS analysis and detection of NRPS and PKS active sites.[58,64] Hopefully, instruments with such high sensitivities and with this type of resolving power will become available to others who are tackling these types of systems for biosynthetic interrogation. Yet, as our understanding of the gas-phase fragmentation behavior of active site-tethered substrates and intermediates advances, it may not even be necessary to have high-resolution instruments for these types of experiments. One such advance is the PEA for the characterization of phopshopantetheinylated proteins.[50]

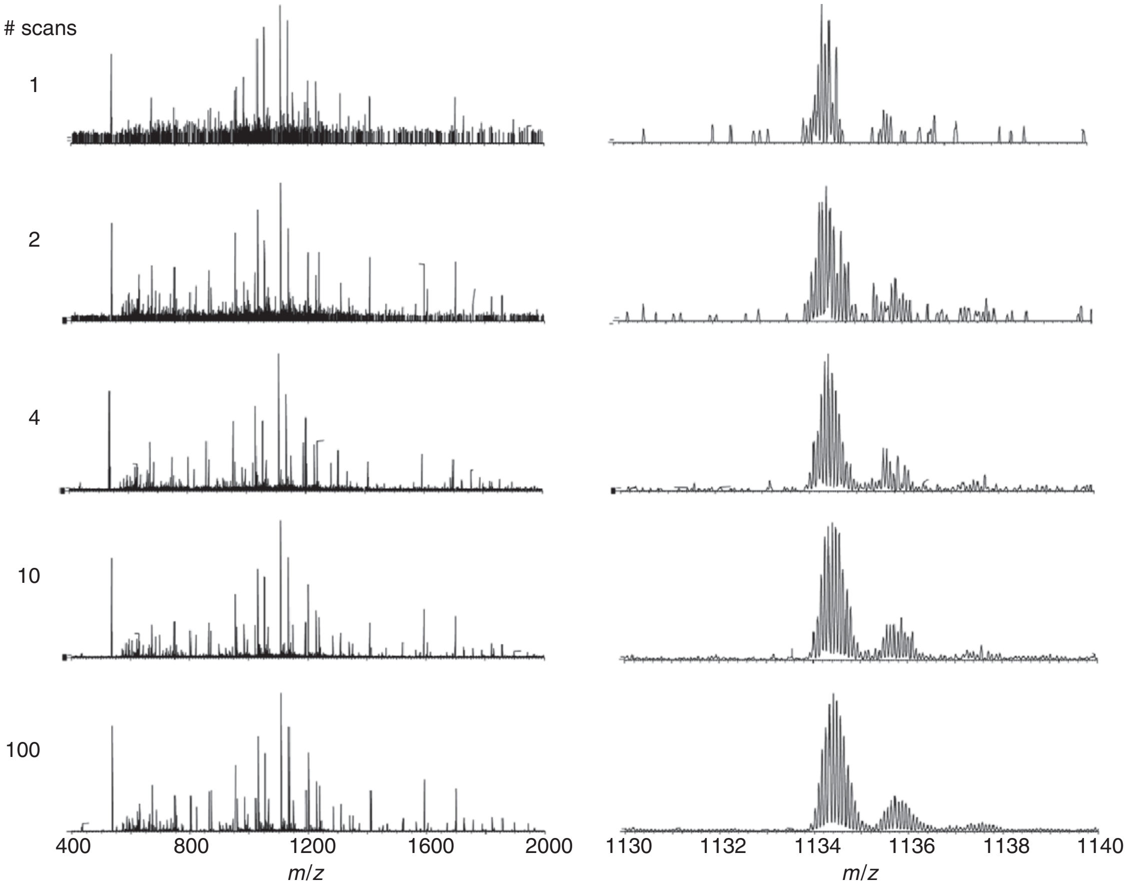


Figure 5 Correlation of FT–ICR–MS scan number and mass signal quality. The more MS scans are acquired (left), the higher the signal-to-noise ratio in the corresponding mass spectrum (right).

9.11.1.6 The Phosphopantetheinyl Ejection Assay

T domains of many biosynthetic pathways are phosphopantetheinylated. The phosphopantetheinyl posttranslational modification of serine is in many ways a novel phosphopeptide. It has been well recognized that the C–O connection of the phosphodiester bond in phosphopeptides is preferentially broken when they are subjected to thermal fragmentation methods (described in Section 9.11.1.7).[65–69] This C–O bond is energetically the most labile connection in such a peptide. Two main mechanisms are proposed for the ejection of a neutral loss phosphate (98 Da) from a phosphopeptide and are shown in **Figure 6**.[2,68] The first mechanism results in the end products that would be expected for a McLafferty-type rearrangement, the dehydroalanine and the uncharged phosphate. In this mechanism the phosphate deprotonates the α-proton on the serine, ejecting the phosphate as a neutral ion. It is not yet known if the rearrangement is a McLafferty-type homolytic rearrangement or a heterolytic cleavage as drawn in **Figure 6**. The second mechanism for neutral ion loss observed in phosphopeptides was proposed by Hunt and coworkers[2] in 2004. In this mechanism depicted in **Figure 6(b)**, a five-membered oxazoline is formed, a reaction that is promoted by the formation of protonated phosphate in the gas phase leading to the cleavage of the C–O connection and release of phosphoric acid as a neutral ion. The support for the understanding that both neutral loss mechanisms are operational has been provided by isotope labeling studies, although these studies did show that the oxazoline mechanism was favored.[68] A phosphopantetheinyl serine modification has a related C–O phosphoester connection but it also has a second C–O connection bearing the active site thiol of the T domain. These C–O linkages on the phosphodiester are the atom connections that most readily fragment in a fashion similar to phosphopeptide

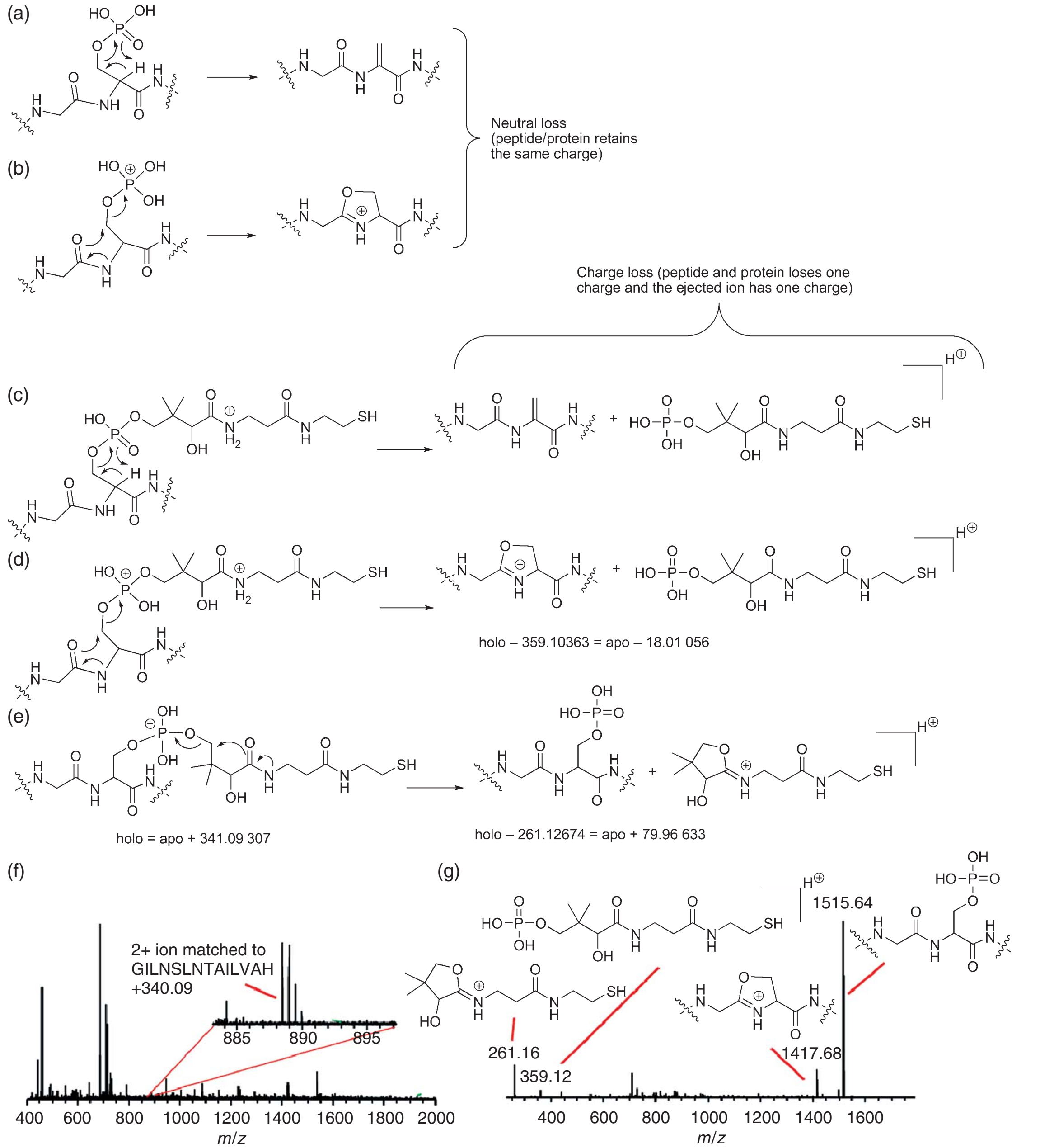

(a)
(b)
Neutral loss (peptide/protein retains the same charge)
Charge loss (peptide and protein loses one charge and the ejected ion has one charge)
(c)
(d)
holo – 359.10363 = apo – 18.01 056
(e)
holo = apo + 341.09 307
holo – 261.12674 = apo + 79.96 633
(f)
2+ ion matched to GILNSLNTAILVAH +340.09
885
890
895
400 600 800 1000 1200 1400 1600 1800 2000
m/z
(g)
1515.64
261.16
359.12
1417.68
400 600 800 1000 1200 1400 1600
m/z

Figure 6 The phosphopantetheinyl ejection assay (PEA). (a) McLafferty-type ejection mechanism. (b) Oxazoline ejection mechanism. McLafferty-type ejection (c) and oxazoline ejection (d) on phosphopantetheinylated protein yield charge-reduced protein minus water and phospho-PPant ejection ion. (e) Iminolactone PPant ejection on phosphopantetheinylated protein yields charge-reduced phospho-apo protein and PPant ion. PEA on CouN5 fragment in broadband FT mass spectrum (f), MS^2 spectrum (g) shows PPant ion (261.16 *m/z*) and phospho-PPant ion (359.12 *m/z*).

neutral losses observed with phosphopeptides. In addition, these C–O linkages will be preferentially fragmented over the fragmentation of amide linkages normally expected of peptides. While the phosphoester ejection from a phosphopeptide results in a neutral loss, the ejection of the phosphodiester from a phosphopantetheinylated peptide or protein results in a loss of a charge on the peptide, resulting in the formation of the charge-reduced apo protein minus water in the case of the McLafferty-type ejection (**Figure 6(c)**), the oxazoline ejection (**Figure 6(d)**), and the thiazoline mechanism, or the charge-reduced apo protein plus phosphate when an iminolactone is ejected (**Figure 6(e)**).[50] At this time the iminolactone phosphopantetheinyl ejection appears to be the ion that is the most abundant and therefore the most useful in the investigation of substrates and intermediates tethered to the active site thiol of the PPant. **Figures 6(f)** and **6(g)** show examples of phosphopantetheinyl ejection of a peptide obtained via digestion of the T protein CouN5 from the coumermycin biosynthetic pathway. The intact ion corresponding to GILNSLNTAILVAH was subjected to collisionally-induced dissociation (CID) and clearly showed all the different ejected PPant ions (observed 261.16 *m*/*z* and 359.12 *m*/*z*). This figure shows that these ions are the most prevalent ones generated in this spectrum. While they are the most abundant ions in ~70% of the phosphopantetheinylated proteins or domains investigated to date, these ejected ions are not always this abundant. Nonetheless, the PEA is a welcome addition to the arsenal of tools used to interrogate NRPS and PKS proteins.

PEA can also be carried out to monitor time courses as shown in **Figure 7**. In this reaction, the condensation of malonyl-*S*-PigH with pyrrolyl-*S*-PigG catalyzed by PigJ was directly monitored by the PEA of both of the two T domains of PigH in a time-dependent manner and shows that changes can be monitored using PEA. This implies that kinetic information may be obtained from monitoring the ejected ion only. However, caution should be taken to prevent the overinterpretation of the kinetics from the observed ejected ions. While changes in a time-dependent manner can be observed using PEA, there are three concerns that should be considered if one wants to obtain true kinetics using this method. First, the ejected ions may have different ionization efficiencies. Altering the ratios of two forms of a phosphopantetheinylated protein and then analyzing this by PEA may not respond in a linear behavior. Some of the discrepancy of what the ejected ions report is evidenced in **Figure 7**, as the intact spectrum on the left of **Figure 7** does not correlate with the PEA relative ratios. In

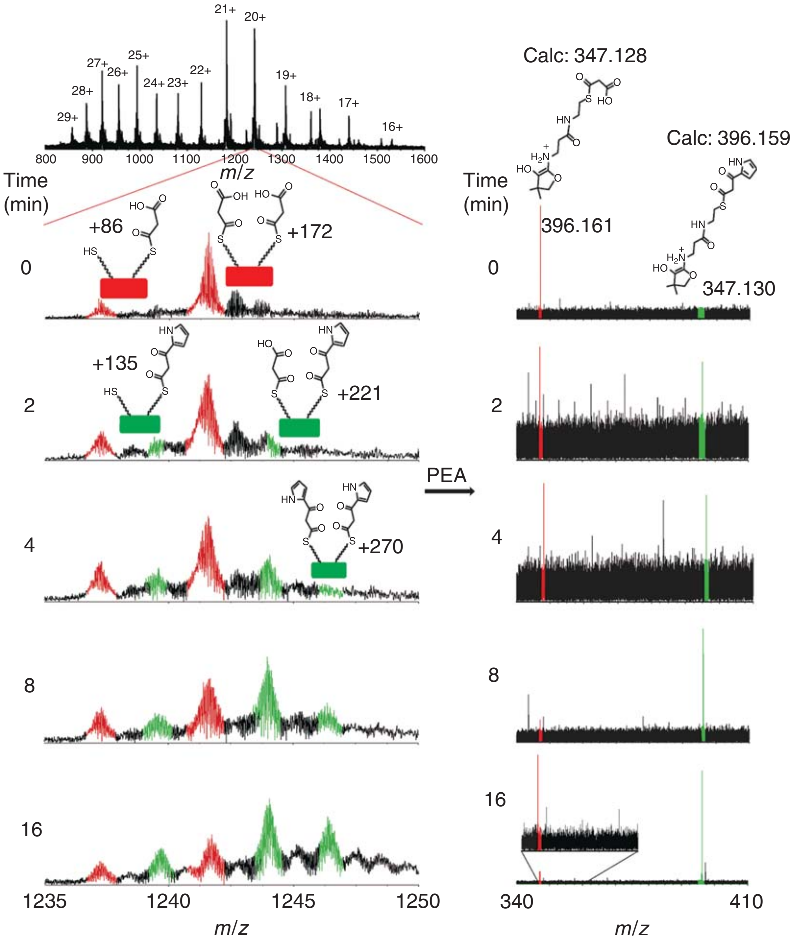


Figure 7 Monitoring time courses by PEA. The condensation of malonyl-*S*-PigH with pyrrolyl-*S*-PigG catalyzed by PigJ is monitored in a time course by broadband FTMS (left) and PEA (right) on the two active sites of PigH.

addition to the differences in ionization efficiencies, the substrate is often partially eliminated from the ejected ion as well. Therefore, the strength of the thioester of the substrate or biosynthetic intermediate will affect the ratios of the ejected ions observed. Because PEA is usually performed on a single charge state of the protein or peptide, the ratio of the ejected ion will vary as well, depending on the ionization efficiency of the parent ion. That change in the ratios between ions of different charge states is directly evident from the Pks4 data in Figure 4(a). Looking at the apo versus holo ratio of each charge state of Pks4, the ratio of apo (annotated with # in Figure 4(a)) versus the phosphopantetheinylated form (annotated with * in Figure 4(a)) of the protein changes from 0.95 to 0.8. This same phenomenon is observed with different substrate-loaded or biosynthetic intermediate-loaded forms of T domains. While the ratios of different charge states will affect the ejected ions, the manner in which the ions are accelerated by CID also affect the PEA ratios, especially when a small parent ion isolation window is applied. Therefore, it is recommended that the isolation window width for CID, when multiple species are analyzed, is greater than the isotopic profile width of any of the ions that one is interested in fragmenting. Otherwise, different fragmentation energies are applied to the different parent ions and differential ejection can be observed. Despite these inherent caveats, PEA has been used to provide some kinetic information[55,69] and follows the generally accepted accuracy of 10–20% that can be obtained with the investigations of NRPS and PKS protein domains.[70]

9.11.1.7 How Is the PEA Accomplished?

PEA is accomplished using thermal activation methods. There are a large number of thermal activation methods that could be used for this. Source fragmentation, blackbody infrared radiation dissociation (BIRD), sustained off-resonance irradiation–collisionally activated dissociation (SORI–CAD), infrared radiation multi-photon dissociation (IRMPD), and CID are examples of such thermal MS/MS or MS^2 methods.[71–76] Currently, source fragmentation, CID, and IRMPD are the only methods that have been utilized for the PEA, and because of this, these are the only three that are covered in this section of this chapter. Before we begin to look at how and when one should use these methods, it is important to describe the common instrumentation configurations used in NRPS and PKS studies. Since most of PEAs have been performed on FT–ICR–MS instruments, the two main commercially FT–ICR–MS configurations are described in this section. In the first configuration, after the sample is introduced, the ions pass through a heated capillary inlet to a quadrupole where ions can be isolated and then passed on to a linear ion trap (linear IT). The ions can then be detected in the ion trap but at low resolution. Alternatively, the ions can be passed to the ICR cell and the ions can be detected with high resolution (Figure 8). Thermo Finnigan hybrid instruments are typically configured in this fashion. In this instrument configuration, thermal fragmentation of ions can take place by exciting the ions at the source by increasing the voltage and colliding the ions with air. On the other hand, CID can be accomplished via excitation of the ions in the ion trap resulting in helium gas collisions. Finally, the instrument can also be equipped with an optional IRMPD and the ions will be fragmented directly in the cell of the instrument. Other commercial FT–ICR instruments from Varian (formerly IonSpec) and Bruker often have a different configuration. Both of these instruments have a sample inlet, usually with a heated capillary, an optional ion funnel (not depicted) to capture as many ions as possible, followed by an isolation quadrupole and an accumulation octupole before passing the ions on to the cell of the ICR instrument. Thermal activation can be accomplished (1) at the source via collisions with air, (2) in the accumulation octupole via collisions with helium or other inert

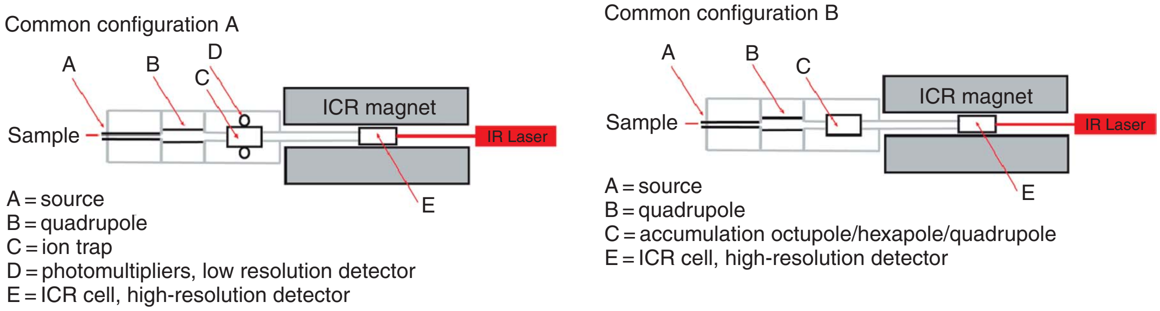

Common configuration A
A
B
D
C
ICR magnet
Sample
IR Laser
E
A = source
B = quadrupole
C = ion trap
D = photomultipliers, low resolution detector
E = ICR cell, high-resolution detector
Common configuration B
A
B
C
ICR magnet
Sample
IR Laser
E
A = source
B = quadrupole
C = accumulation octupole/hexapole/quadrupole
E = ICR cell, high-resolution detector

Figure 8 The two main FT–ICR–MS configurations that have been utilized in the investigations of NRPS and PKS systems: LTQ–FT–ICR–MS configuration (left) and accumulation multipole FT–ICR–MS configuration (right).

gas, and (3) inside the cell using a pulsed laser (IRMPD). Below we describe the types of experiments that can be performed by each setup.

9.11.1.7.1 The LTQ–FT–ICR–MS configuration for PEA (common configuration A)

A hybrid instrument such as the hybrid linear trap quadrupole–Fourier transform–ion cyclotron resonance–mass spectrometry (LTQ–FT–ICR–MS) configuration (**Figure 8**) allows one to perform some very interesting experiments with respect to NRPS and PKS multidomain proteins. First, we can fragment at the source of the instrument, resulting in phosphopantetheinyl ejection for any ion that enters the instrument. The advantage of this approach is that we are not just isolating a single charge state for fragmentation but all the charge states from a protein, thereby increasing the signal intensity of the ejected ion. The disadvantage of this approach is that one is likely to get many more signals in the low *m*/*z* region, requiring additional confirmation of the signals observed. This confirmation can be accomplished using substrates with stable isotopes or an additional round of fragmentation in a data-dependent manner as done for proteomics experiments. This is of particular importance when the ejection of the PPant is performed on low-resolution instruments as described in Section 9.11.1.8. An additional disadvantage of using source fragmentation for PEA is that the ejected ion cannot be correlated with the parent ion. This is important when one wants to map active sites. While source fragmentation is one of the approaches to obtain PPant ejection, the most common approach is CID. In CID, ions of interest are accelerated and collide with gas ions (helium being the most commonly used). During this collision, most of the kinetic energy is subsequently converted to thermal energy and if the phosphopantetheinylated ion undergoing the collisions has enough vibrational energy amassed, which likely requires multiple collisions it will dissociate into two ions. CID can be accomplished in an inert gas-filled quadrupole or in an ion trap. In the case of a Thermo Finnigan instrument (configuration A), CID is accomplished in the linear ion-trap portion of the instrument. Once dissociated, the fragment ions can be observed using a photomultiplier or the ions can be passed on to the ICR cell for high-resolution analysis. Unfortunately, ion traps suffer from a major but well-documented limitation that is defined as the 1/3rd rule (**Figure 9**).[77,78] When the activation *q*, that is, the energy that is responsible for accelerating all of the ions to be fragmented, is raised, low *m*/*z* product ions start to lose stable trajectories causing them to be ejected from the trap.[77,78] For example, when an activation *q* is set to 0.25 and the parent ion is isolated at 1200 *m*/*z*, one cannot observe fragment ions below 400 *m*/*z*. Some improvements in terms of the detectability of the low *m*/*z* scanning range can be gained by lowering the activation *q*, but this change necessitates an increase in the time for activation from 30 to 100 ms in order to obtain significant fragmentation. On the author's LTQ–FT–ICR–MS instrument, when *q* is set to 0.2, the 1/3rd rule becomes a 1/4th rule, thereby increasing the size of the ions that can be fragmented, and ultimately detected, for the PEA assay. A partial solution for this limitation of ion traps is a related software-controlled mechanism called pulsed-Q dissociation (PQD). Although this is not yet available for commercial LTQ–FT–ICR–MS instruments, it is available for two-dimensional linear trap quadrupoles (LTQs), three-dimensional linear quadrupole (LCQs), and hybrid linear trap quadrupole-orbitraps (LTQ–ORBIs).[79,80] Most ion traps in existence today are not equipped with a PQD software upgrade so that its utility is limited at this time. While it is possible to equip an LTQ–FT–ICR–MS instrument with IRMPD, it has not yet been applied toward PEA on such instruments.

9.11.1.7.2 The accumulation multipole setup for PEA (common configuration B)

The other common setup for FT–ICR–MS instrumentations is configuration B (**Figure 8**), which differs from the ion trap configuration by having an accumulation multipole instead of a linear ion trap. Such an instrument is also capable of PEA by conducting fragmentation at the source of the instrument. The combination of the quadrupole-accumulation multipole setup is very useful in PEA. The advantage is that one can select out ions of interest in the quadrupole and then accumulate ions in the accumulation octupole as shown in **Figure 10(b)**. The ions can then be excited for CID inside the accumulation multipole or they can be passed on to the cell of the instrument. Inside the cell of the instrument, the isolated ions can be subjected to PEA using infrared radiation (IRMPD). While both CID and IRMPD are thermal activation methods and the resulting fragmentation are quite similar it is not yet investigated if one results in improved PPant ejection over the other. The advantage of CID or IRMPD on an accumulated ion signal is that the intensity of the ejected ion is much more intense. The disadvantage is that it can take many seconds, sometimes up to 60 s, to collect a single scan on an 8.4T instrument. Therefore this configuration is typically not amenable to the LC timescale.

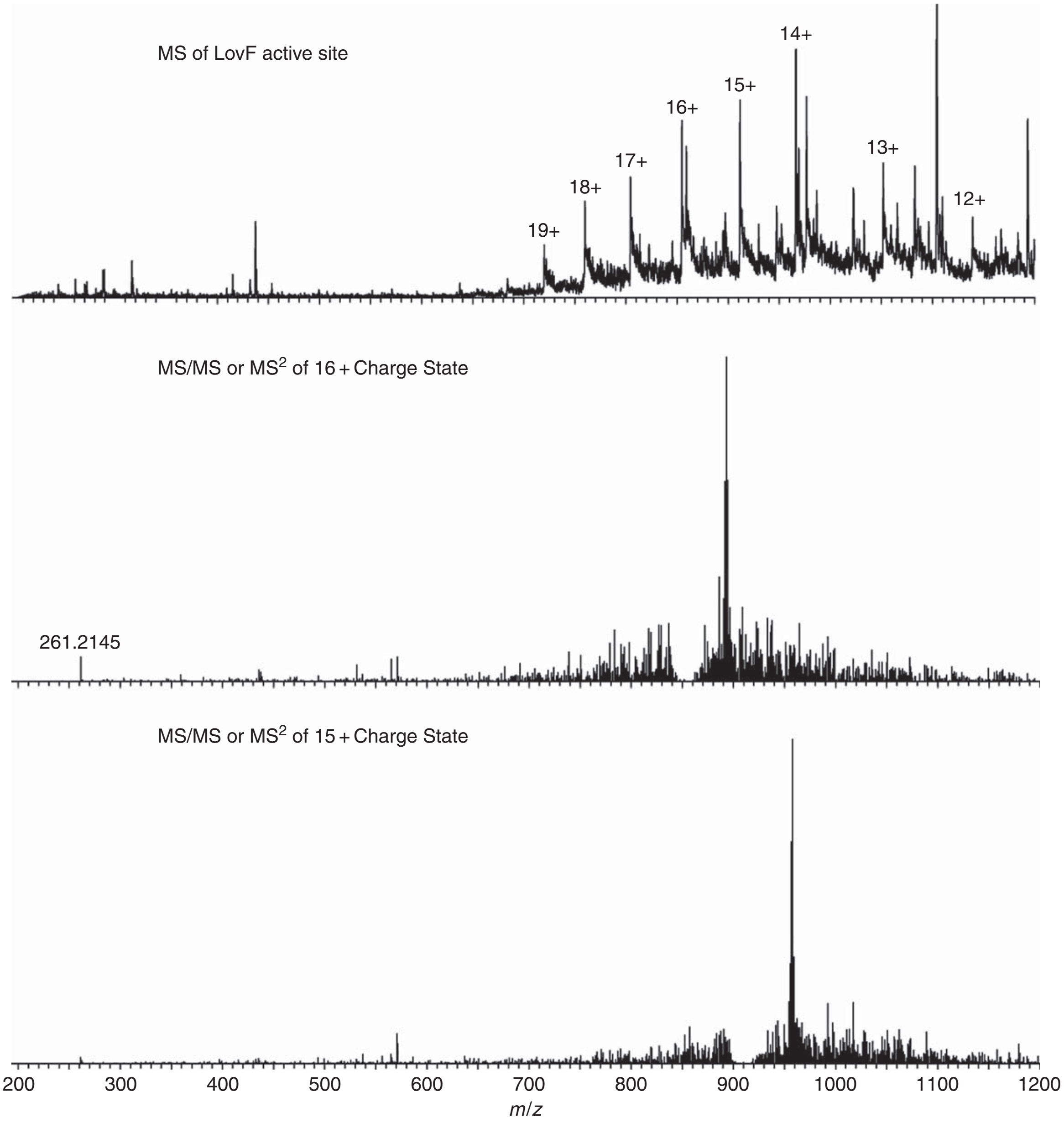


Figure 9 Illustration of the 1/3rd rule. The top panel shows multiple charge states of the phosphopantetheinylated active site containing peptide from PKS LovB. The 16+ ion at 854.636 *m/z* and the 15+ ion at 911.636 *m/z* of the peptide were subjected to CID. The middle panel shows that PPant ejection (261 *m/z*) is clearly detected in the MS/MS or MS^2 spectrum resulting from CID of the 16+ ion. The bottom panel shows that PPant ejection resulting from fragmentation of the 15+ ion cannot be detected due to limitations defined by the 1/3rd rule.

9.11.1.8 PEA on Non-ICR Instruments: Low-Resolution Phosphopantetheinyl Ejection

PEA is an important discovery in the characterization of phosphopantetheinylated proteins. It allows one to analyze nonradioactively labeled substrates thereby making many more substrates accessible to biosynthetic studies of these types of proteins. In addition, this approach often eliminates the need for preparing synthetic standards for comparison of hydrolyzed thioester intermediates, which may be challenging and time consuming to accomplish. Moreover, PEA immediately shows that the substrate or intermediate is connected to the PPant posttranslational modification and not elsewhere on the protein. This is not possible to detect with radioactivity or any other

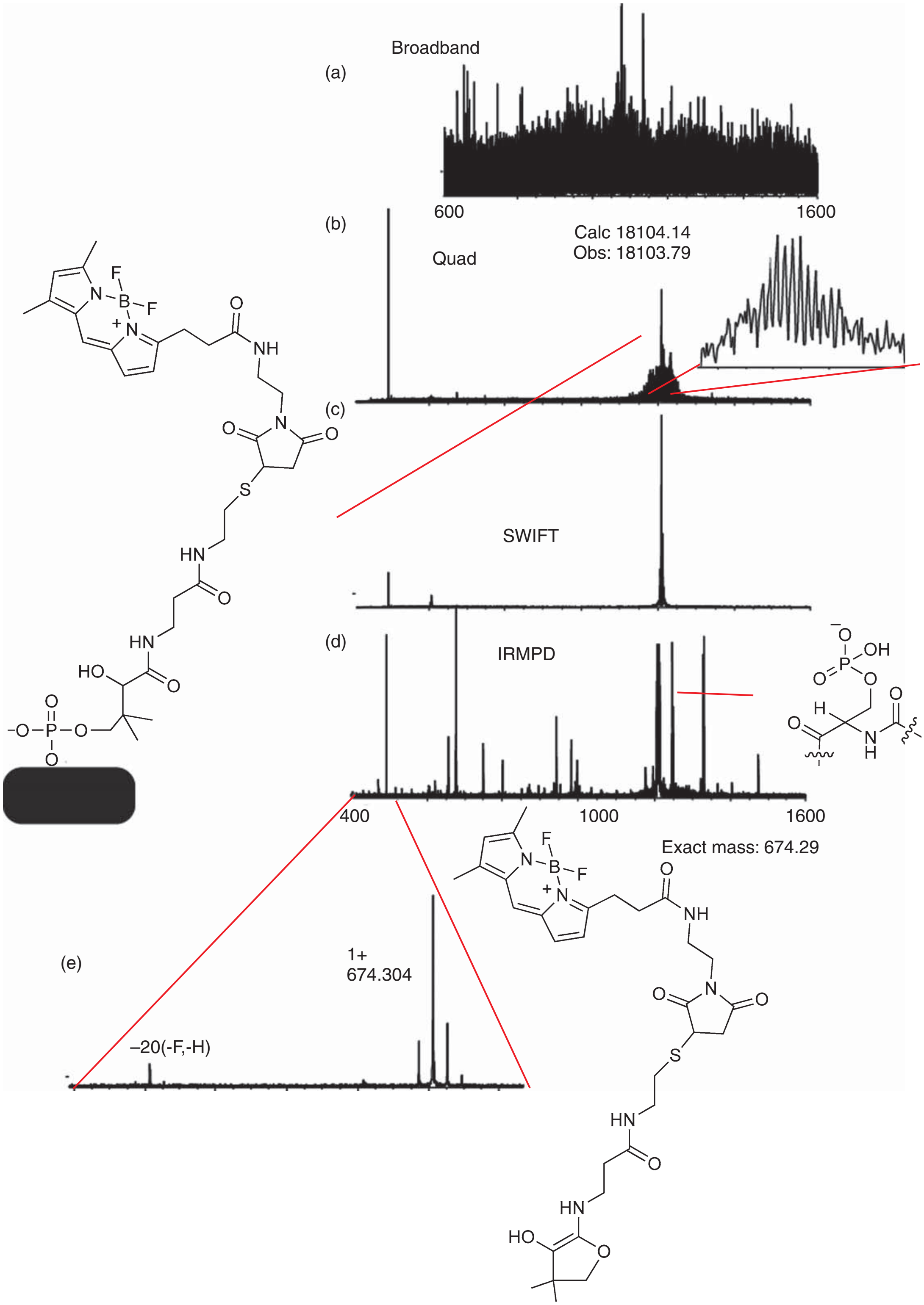


Figure 10 The accumulation multipole setup for PEA. Ions of interest are selected by the quadrupole (a), accumulated in the accumulation octupole (b), isolated by SWIFT (c), and subjected to PEA by IRMPD (d,e) in the analyzer cell.

conventional assay used to study these types of systems. An additional advantage that PEA has over radioactive assays is that it provides a mass signature and therefore unexpected intermediates can be observed. Even though PEA is such a useful tool in the investigation of substrates and biosynthetic intermediates, it has not been widely used by other investigators because it requires costly FT–ICR–MS instrumentation and highly skilled researchers to use them. Because most MS instruments are able to perform source fragmentation, CID, or both, this method should be readily be applicable to low-resolution instruments. Since properly calibrated low-resolution instruments have mass accuracies well within 0.5 Da, they can be used to differentiate molecular species within 1 Da for the 1+ ejected PPant ions, making such an instrument a very useful tool for the characterization of phosphopantetheinylated proteins. As PEA can be applied to low-resolution instruments, it is an assay that is accessible to most researchers studying phosphopantetheinylated proteins. The challenge of this assay is that even though the ejected ion in the low *m*/*z* range is often the most abundant one, there are other ions in this region of the spectrum. Due to these other ions and to noisier low-resolution detectors, which are not based on image current measurement, it can be more difficult to confirm the candidate ejected ion than anticipated. This confirmation can be carried out using labeling studies or directly by PPant fragmentation, a new MS^3 (additional fragmentation of ions generated by MS/MS or MS^2) method, if an instrument has the ability to perform MS^3. When this MS^3 method was originally reported, detection of 12 diagnostic ions were reported for the second round of MS/MS or MS^2 on the ejected ion, but there are many more fragments with lower abundance, including an important and abundant ejected ion fragment that reports on the substrates loaded onto the thiol of the phosphopantetheinyl functionality (**Figure 11**, Dorrestein, unpublished observations).[81] Thus far, the confirmation of the ejected ions has only been accomplished on an ion trap instrument in two different modes. In the first mode, the phosphopantetheinylated protein is observed by MS, the phosphopantetheinylated ions are isolated in the ion trap and subsequently subjected to CID (MS/MS or MS^2). Then the ejected ion is isolated and fragmented again using CID (MS^3). This will result in the diagnostic fragment pattern for the phosphopantetheinylated peptide ejection ions. The second mode relies on the initial use of source fragmentation followed by CID on the ejected ions in the low *m*/*z* region (MS^2). In this case, the method can be performed in a data-dependent manner, as it is done for proteomic experiments, to find the characteristic phosphopantetheinyl 'MS^3' signature from any of the ions that enter the instrument. The main disadvantage of the source fragmentation

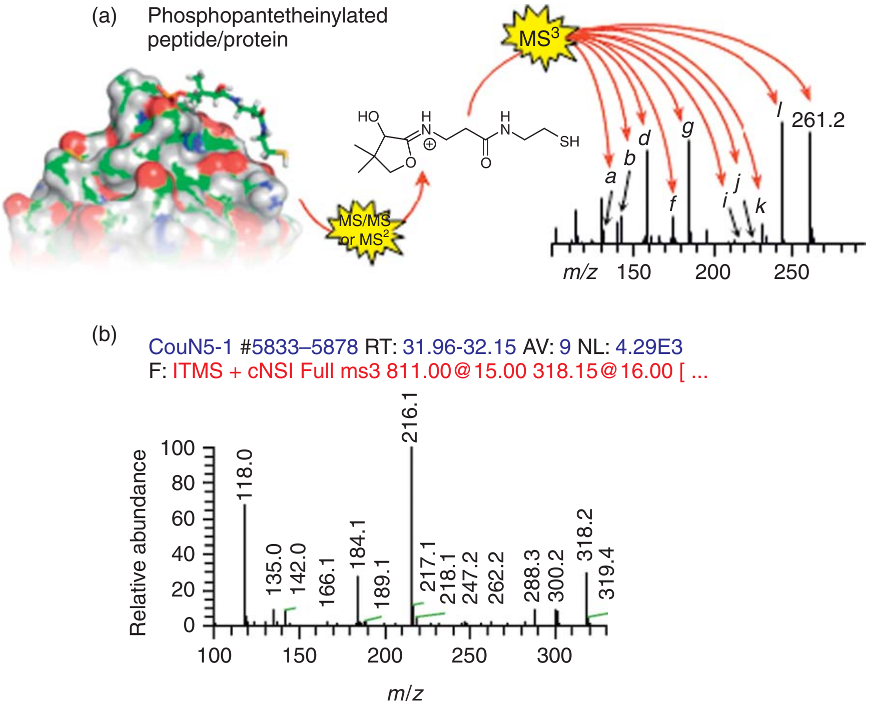


Figure 11 The PPant fragmentation method. (a) A phosphopantetheinylated protein is first subjected to the PEA and, subsequently, the PPant ejected ion is fragmented again to yield diagnostic MS/MS or MS^2 peaks (b).

method, as mentioned earlier, is that all the information of the precursor ion is lost, but it should still enable the analysis of phosphopantetheinylated proteins on nonion trap, low-resolution instruments that have only one stage of MS capabilities by additional confirmation of the PEA ions.

9.11.1.9 Low-Resolution Capillary LC–MS on Ion Traps

To date, FT–ICR–MS instruments are the main instruments used for PEA but they are not the best instruments to analyze NRPS/PKS on an LC timescale because, currently, their scan rates are very long. The scan rates of most low-resolution instruments are shorter and, therefore, much better suited for interfacing with LC. In proteomics, it has become standard to use small-bore columns (30–100 μmol l^{-1} inner diameter) with nanoflow (200–500 nl min^{-1}) gradients, but this has not yet found much use in the investigations of biosynthetic pathways. More recently, with the ability to perform PEA, these LC–MS approaches have emerged as useful tools in the investigation of phosphopantetheinylated peptides and proteins. The advantage of LC with a 100 μmol l^{-1} column as opposed to traditional 1, 2.1, or 4.6 mm HPLC columns is that it uses much less material. In the case of a 4.6 mm column, injection of >100 μg of sample is not uncommon to attain good signals, while in the case of a 100 μmol l^{-1} diameter column, one typically uses 0.1–1 μg of material. This approach has recently been used to observe the phosphopantetheinylation of a carrier domain from the hemolytic toxin pathway from *Streptococcus agalactiae* and its corresponding phosphopantetheinyl ejection.[81]

9.11.1.10 Mass Spectrometry of Intact NRPS and PKS Multidomain Proteins

There have only been a few reports on the MS of intact multidomain NRPS and PKS proteins.[82] The NRPS proteins GrsA and NikP1 have been analyzed by FT–ICR–MS but not with isotopic resolution. In addition, the 6-deoxyerythronolide B synthase (DEBS) PKS didomain has been investigated by MS.[82] In most cases, a mass shift can be observed upon phosphopantetheinylation or substrate loading. The main challenge with intact NRPS and PKS protein analysis is to obtain sufficient quantities of very pure proteins. A small amount of contamination by a small peptide or small domain will compete for the signal very efficiently, making it nearly impossible to observe the larger ion. While the examples of GrsA, NikP1, and DEBS illustrate that it is possible to interrogate these larger proteins, it should even be possible to analyze entire NRPS and PKS multiprotein complexes such as the ones observed for bacillaene.[83] These NRPS/PKS multiprotein complexes are similar in size to intact viral particles or intact ribosomes that have been investigated by MS.[84,85] The MS of intact protein and multiprotein complexes is an area that remains wide open for exploration.

9.11.1.11 Mass Spectrometry of Phosphopantetheinylated but Non-NRPS and Non-PKS Proteins

In many respects, the phosphopantetheinyl functionality is similar to a phosphoserine and treatments that work with phosphorylated proteinaceous materials will often work with phosphopantetheinylated materials as well. For instance, phosphatases are responsible for the removal of phosphates but were recently shown to remove the phosphopantetheinyl functionality as well.[86] While this assay has not yet been utilized on the biosynthesis of natural products, this assay has been applied to the investigation of the phosphopantetheinylated protein 10-formyltetrahydrofolate dehydrogenase, which is involved in the formation of formyltetrahydrofolate.[86] Donato *et al.* used the phosphatase assay to confirm the presence of the phosphopantetheinylation. The characterized phosphopantetheinyl modification suggests that there may be many other phosphopantetheinylated proteins that have not been identified. This assumption has recently been confirmed using a phage display approach.[87] In the study by Donato *et al.* a common matrix-assisted laser desorption/ionization–time-of-flight (MALDI–TOF) instrument was used for analysis. One caution about their interpretation of the spectra should be noted and it is a direct result that the experiments were performed by a novice and not an MS expert. In their spectra, these researchers observed a mass of 360.08 Da and in their text they report that the mass of the ion should be 358.33 Da, which is not correct. We recalculated the mass of the hydrolyzed ion and anticipate that the mass of this hydrolyzed species should be 359.104 Da in the positive mode while in the negative mode it is 357.089 Da. In addition to the monomer unit, the authors saw a dimer with a mass of 550.60 Da where both phosphates had been removed. According to our calculations, this dimer should have a mass of 555.252 Da for a

1+ ion or 553.237 Da for the 1− ion, while they reported a calculated mass of 552 Da for this ion. Their measured mass error is likely a result of poor calibration or relying on old calibration files, as most people who use MALDI–TOF instruments in core facilities do. However, it is unclear why their calculated masses were not spot-on. In the end, while it is important for people to realize the mass errors in this particular report if they wish to repeat the experiments, this point of caution does not invalidate the overall conclusion provided in this paper. The MS still supports that 10-formyltetrahydrofolate dehydrogenase is a phosphopantetheinylated protein and therefore is an addition to the ever-expanding population of phosphopantetheinylated proteins such as the citrate synthase glycine cleavage system T protein (GcvT) in the glycine cleavage system of select organisms. Furthermore, this paper described a new approach to the characterization of phosphopantetheinylated proteins by MS. It remains to be determined if this approach is widely applicable to other phosphopantetheinylated proteins and if it can be applied to substrate- and intermediate-loaded PPant arms, for example, of NRPS and PKS systems. The major concern in this case is the lability of the thioester and the long phosphatase incubation times. A typical half-life of a thioester is 200–400 min^{-1} even under stabilizing acidic conditions.[88]

9.11.1.12 The Development of Recognition Software for PEA on an LC Timescale

While there are many proteomics programs designed to find peptides or neutral ion losses such as phosphoric acid from phosphopeptides, there are no programs or algorithms developed that can identify phosphopantetheinylated peptides. The main challenges in the annotation of phosphopantetheinylated peptides by such programs is the observation that the ejected ions are often very abundant and, therefore, limiting to the intensity and number of normal fragment ions typically encountered with peptides. In the case of phosphopeptides, there were enough examples available that software could be trained to recognize such peptides.[89] In the case of phosphopantetheinylated proteins, there are a limited number of such data training sets with which to train new software. In addition, phosphopantetheinylated peptides have two phosphoester linkages while phosphopeptides only have one, making it more likely that this ion is ejected instead of amide cleavage. Finally, the thiol of the phosphopantetheinyl functionality is modified with substrates and intermediates making it even harder to identify and find phosphopantetheinylated peptides. Such modifications to phosphopantetheinylated peptides will need to be taken into account when conducting searches for active sites. Current efforts are underway to overcome some of these limitations. Once solutions to this problem are obtained, it will become possible to study the biosynthesis of the therapeutics that are biosynthesized on phosphopantetheinylated proteins at their native levels using proteomic approaches.

9.11.2 Applications of Mass Spectrometry on NRPS Systems

NRPs are important bioactive and medicinally applied natural products, including compounds such as cephalosporins, penicillins, and vancomycin. These natural products are biosynthesized by NRPSs, one of the two thiotemplate biosynthetic machineries found in fungal and microbial secondary metabolism. As mentioned before, two factors made NRPS characterization difficult until the late 1990s: (1) their large size (up to 700 kDa) and (2) their complex multimodular structure comprising a diversity of catalytic and carrier domains. Consequently, the characterization of NRPS intermediates and substrates from enzymatic assays was a laborious task exemplified by the dissection of the gramicidin biosynthetic system by Stein *et al.*[90] Herein, multiple steps of radioactive labeling of carrier protein active sites, digests, HPLC separations, and low-resolution mass spectrometric analysis were required to identify all biosynthetic substrates from milligram quantities of the corresponding NRPS proteins GS1 and GS2. In 1999, Shaw-Reid *et al.*[41] applied electrospray ionization Fourier transform mass spectrometry (ESI–FTMS) to characterize enterobactin NRPS enzymology, which simplified and accelerated the identification process of NRPS intermediates significantly. In addition, they reduced the required protein amount to microgram quantities. Since then, MS has held its initial promise as a key method to investigate NRP biosynthesis by the improvement of instrumentation of large molecule MS in combination with advanced molecular biology techniques and the development of innovative MS-based assays to study NRPS – such as the PEA[50] and the substrate screening assay.[91] Many methods and investigations of NRPS biosynthetic pathways and their tailoring reactions are summarized in the comprehensive 2006 review by Dorrestein and Kelleher.[47] This section will highlight the recent applications of the modern MS approaches applied to NRPS proteins since this 2006 review and the progress that has been made since then.

In the 2006 review, the basic experimental procedures to detect intermediates on NRPS carrier proteins are explained and the mechanistic insights for the application of those thiotemplate systems gained by MS are described. This section connects to this review by summarizing recent MS-based studies of nonribosomal biosynthetic machineries from 2006 until January 2009 and emphasizes the application of the new MS methods described in the previous section of this review. As described above and in the review by Dorrestein and Kelleher, the major advantages of MS as an investigative tool of NRPS systems are: (1) *In vitro* substrate identification of NRPS domains by mapping of T domain active sites. The direct detection of T domain-tethered substrates and intermediates allows the investigation of substrate specificity of NRPS catalytic domains and tailoring enzymes. (2) Parallel detection of NRPS T domains and rapid quench methods allow the investigation of relative T domain occupancy by intermediates, the investigation of intermediate flux and pseudokinetic interrogation of NRPS downstream processing. (3) Substrate screening methods allow verification of predicted substrates and characterization of orphan NRPS gene clusters by fast substrate identification from a complex substrate pool.

Since most current papers that are published on NRPS systems utilize some form of MS, we have divided the application of MS methods to understand NRPS pathways into five different categories: (1) investigation of substrate specificity in NRPS systems, (2) characterization of new NRPS enzymology and of deviations from colinearity, (3) characterization of tailoring reactions, (4) characterization of multistage assembly line action, and (5) time courses.

In the following sections, it is explained briefly what makes NRPS biochemistry accessible to mass spectrometric investigation and a new active site mapping approach for NRPS and PKS systems, the online LC–MS–PEA assay. Subsequently, each of the above categories of MS applications in NRPS biosynthesis research is presented based on recent research.

9.11.2.1 MS Accessibility of NRPS Systems

NRPS enzymology has been outlined in several excellent reviews.[23,92] There are three main features of NRPS that make these biosynthetic systems amenable for MS characterization. The first feature is that all substrates and intermediates are covalently tethered to carrier domains. The covalent tethering of intermediates allows their detection by isolation of the carrier protein active sites. Therefore, NRP chain elongation and tailoring reactions along the biosynthetic pathway can be characterized by detection of corresponding mass changes of active site-bound intermediates. Second, NRPs are usually biosynthesized along a thiotemplate with multiple carrier domains. The timing of biosynthetic events on an NRPS assembly line can be dissected by parallel mass spectrometric detection of those multiple carrier proteins and their covalently bound intermediates. Third, substrates and intermediates are covalently bound on the carrier proteins as thioesters. The weakness of the thioester C–O bond enables the cleavage of intermediates from carrier proteins as PPant species and, therefore, enables the PEA (see Section 9.11.1.6).

9.11.2.2 Active Site Screening and PPant Ejection Assay in NRPS Investigation

One of the most important steps in NRPS characterization by MS is active site mapping, that is, isolation and confirmation of a carrier protein active site in the mass spectrometer. Currently, there are two experimental types of techniques to map active sites by MS – offline-LC active site mapping and online-LC active site mapping. The general offline-LC active site mapping approach and recent techniques to accelerate active site detection are reviewed by Dorrestein and Kelleher.[47] Since 2006, online LC–MS and the PEA have been applied to speed up active site mapping even more. The LC–MS–PEA active site mapping approach has been used to map NRPS active sites by the Kelleher group[51,93] and the Marahiel group.[42,52] It is generally pursued as follows and as shown in **Figure 12**.

The first step is the priming of the NRPS active site and a subsequent limited tryptic digest of the protein. The digested sample is loaded on a reverse-phase liquid chromatography (RPLC) C18 column, which is directly connected to the inlet of an FT mass spectrometer. During online LC separation, the eluent is analyzed by MS and MS^2 on an LC timescale. In the mass spectrometer the eluent is first analyzed by broadband Fourier transform mass spectrometry (FTMS). Then, peaks in the resulting broadband FT mass spectrum are

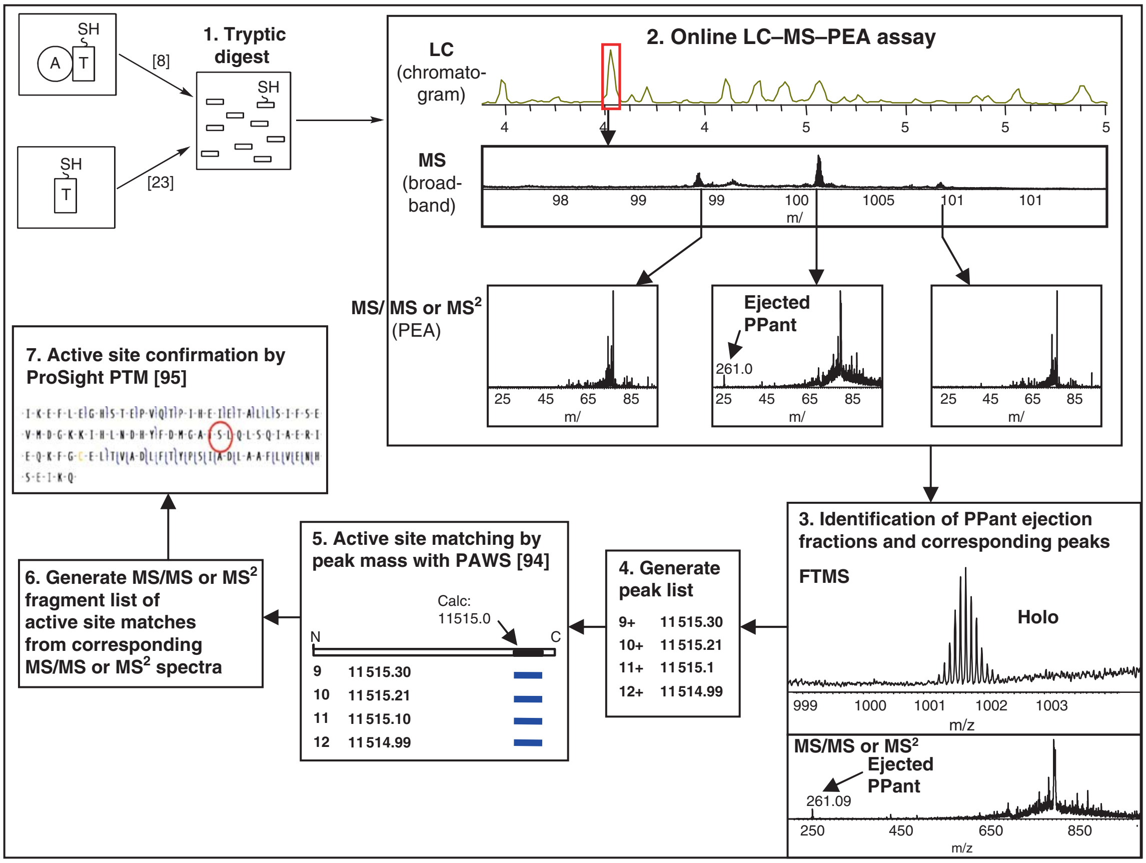

SH
A
T
[8]
SH
T
[23]
1. Tryptic digest
SH
2. Online LC–MS–PEA assay
LC (chromatogram)
4
4
4
5
5
5
5
MS (broadband)
98
99
99
100
1005
101
101
m/
MS/MS or MS^2 (PEA)
25 45 65 85
m/
Ejected PPant
261.0
25 45 65 85
m/
25 45 65 85
m/
7. Active site confirmation by ProSight PTM [95]
I-K-E-F-L-E-G-H-S-T-E-P-V-Q-T-P-I-H-E-I-E-T-A-L-L-S-I-F-S-E
V-M-D-G-K-K-I-H-L-N-D-H-Y-F-D-M-G-A-S-L-Q-L-S-Q-I-A-E-R-I
E-Q-K-F-G-C-E-L-T-V-A-D-L-F-T-Y-P-S-I-A-D-L-A-A-F-L-V-E-N-H
S-E-I-K-Q
6. Generate MS/MS or MS^2 fragment list of active site matches from corresponding MS/MS or MS^2 spectra
5. Active site matching by peak mass with PAWS [94]
Calc: 11515.0
N
C
9 11515.30
10 11515.21
11 11515.10
12 11514.99
4. Generate peak list
9+ 11515.30
10+ 11515.21
11+ 11515.1
12+ 11514.99
3. Identification of PPant ejection fractions and corresponding peaks
FTMS
Holo
999 1000 1001 1002 1003
m/z
MS/MS or MS^2
Ejected PPant
261.09
250 450 650 850
m/z

Figure 12 Active site mapping by online LC–FT–ICR–MS–PEA assay.

fragmented selectively and successively and MS^2 data for each mass peak are collected for a programmed number of scans. This MS^2 analysis of eluting mass peaks is called data-dependent analysis because the eluent MS^2 analysis is only carried out on occurring mass signals. The data-dependent fragmentation of eluting peaks enables the PEA simultaneously. The streamline collection of MS and MS^2 data during the LC–MS–PEA assay shortens the time of MS measurement for active site mapping to ~1 h per experiment. The data analysis for mapping the active site within the LC chromatogram first utilizes PPant ejection data to characterize elution fractions that comprise active site fragments. The occurrence of the ejected PPant species 261.127 m/z and 359.104 m/z indicates that the fragmented peak within an elution fraction was phosphopantetheinylated and, therefore, might be an eluting active site fragment. The corresponding mass of the putative active site fragment can be determined as described by Dorrestein and Kelleher[47] from the FTMS broadband spectrum and mapped in the NRPS amino acid sequence by PAWS.[94] If an active site fragment could be identified, the MS^2 peaks of the identified active site fragment are analyzed with ProSight PTM[95] in order to confirm the mapped sequence by detected b- and y-ions.

Ultimately, the LC–MS–PEA active site mapping approach has advantages and disadvantages in comparison to recent offline-LC active site mapping approaches. Major advantages are the acceleration and simplification by automatization of active site mapping and the substantial decrease of required protein quantities. A disadvantage is the limited accumulation time of ions to obtain high-quality MS data and, thus, the limited number of peaks that can be selected and fragmented during data-dependent MS^2 analysis. In addition, the required instrumentation is also more expensive as the collection of the FT–ICR–MS data on an LC timescale from a complex protein sample demands stronger magnets for faster accumulation time. Therefore, the LC–MS–PEA assay for active site mapping is currently ideally accomplished on a 12T FT–ICR–MS instrument[93] but it was already conducted on a 7T FT–ICR–MS instrument.[51]

Offline-LC active site mapping might still be ideal for active site mapping of complex NRPS systems because active site fragments can appear as low intensity peaks relative to the other coeluting ions in the FT broadband mass spectrum and, consequently, then might not trigger data-dependent fragmentation and analysis on the LC timescale in the online-LC approach. In offline-LC active site mapping experiments, collected fractions can be directly infused into the FT–ICR–MS and data can be collected for an extended period of time in order to increase the quality of weak peptide signals. Offline-LC approaches have gained more reliability from the PEA that is an additional step to confirm a primed active site fragment by ejected PPant species. Current examples of applied offline-LC active site mapping are investigations of NRPS systems of the prodigiosin biosynthesis,[96] the microcystin biosynthesis,[97] and the vibrobactin biosynthesis.[98]

9.11.2.3 Investigation of Substrate Specificity in NRPS Systems

One of the main applications of MS in NRPS research is the investigation of NRPS substrate specificity. Herein, MS is used for two purposes: substrate identification and determination of substrate tolerance of different NRPS domains. Both approaches are referred to as substrate specificity assays in this section. MS-based substrate identification is the *in vitro* and *in vivo* characterization of the native substrates utilized by a catalytic NRPS domain or tailoring enzyme during biosynthesis of an NRP natural product. MS-based determination of substrate tolerance is the *in vitro* characterization of alternative substrates that a catalytic NRPS domain or tailoring enzyme uses besides its native substrates.

Prior to MS-based substrate specificity assays, certain NRPS substrate specificities can be predicted by bioinformatics. Adenylation domain substrates can be predicted based on their '10 letter code'[99,100] by substrate prediction tools such as the NRPS predictor.[101] Methyltransferases can be predicted in their substrates and methylation sites by bioinformatic analysis too.[102] In addition, substrates of catalytic NRPS domains and tailoring enzymes can be predicted by the structure of the known NRP natural product. Either way, predicted substrates of NRPS domains need to be experimentally verified. A traditional technique to determine substrate specificity of an A domain is the adeonsine triphosphate–pyrophosphate (ATP–PP$_i$) exchange assay. The ATP–PP$_i$ exchange assay characterizes substrates indirectly by observing the radioactive pyrophosphate incorporation into ATP from a reverse reaction with pyrophosphate and the acyl–adenylate of the substrate.[103] Because the PP$_i$ exchange measures the back exchange of pyrophosphate into ATP, the determined substrate can deviate from the true substrate as it may be only the kinetically most competent substrate of the reverse adenylation reaction. In contrast to this assay, MS has become a more reliable tool to identify NRPS substrates because it determines the true substrate specificity by detection of the complete adenylation reaction product, that is, the substrate tethered on a T domain.

Before selected publications are presented, a general guideline of MS-based substrate specificity assays for specific NRPS domains is given. This guideline emphasizes three experimental aspects of an MS-based NRPS substrate specificity assay. First, which MS instrumentation should be applied? Second, which substrates have to be considered for a catalytic NRPS domain or tailoring enzyme? Third, which substrate pool can be utilized to identify a native or alternative substrate?

The applied MS instrumentation for a substrate specificity assay depends on the size of the carrier domain construct whose active site is monitored. If multidomain constructs or T domain constructs >20 kDa are investigated, active site mapping is required and, therefore, ESI–FTMS instruments have to be applied because of their established active site mapping capabilities described above. If the monitored T domain constructs are freestanding and <20 kDa, no active site mapping is required and low-resolution MS instruments can be utilized for characterization of loaded substrates too.

Each type of catalytic NRPS domain and tailoring enzyme has characteristic substrates. Herein, a differentiation can be made between biosynthetic substrates and carrier protein substrates. A biosynthetic substrate is a building block or intermediate of the NRP natural product. For instance, an amino acid can be the biosynthetic substrate of an A domain or an NRP intermediate can be the biosynthetic substrate of a tailoring enzyme. A carrier protein substrate is the T domain recognized by a catalytic NRPS domain or tailoring enzyme for substrate loading or turnover of tethered biosynthetic substrates. In the following, biosynthetic substrates and carrier protein substrates of A domains, C domains, tailoring enzymes, and aminoacyl transferases (**Figure 13**) are summarized.

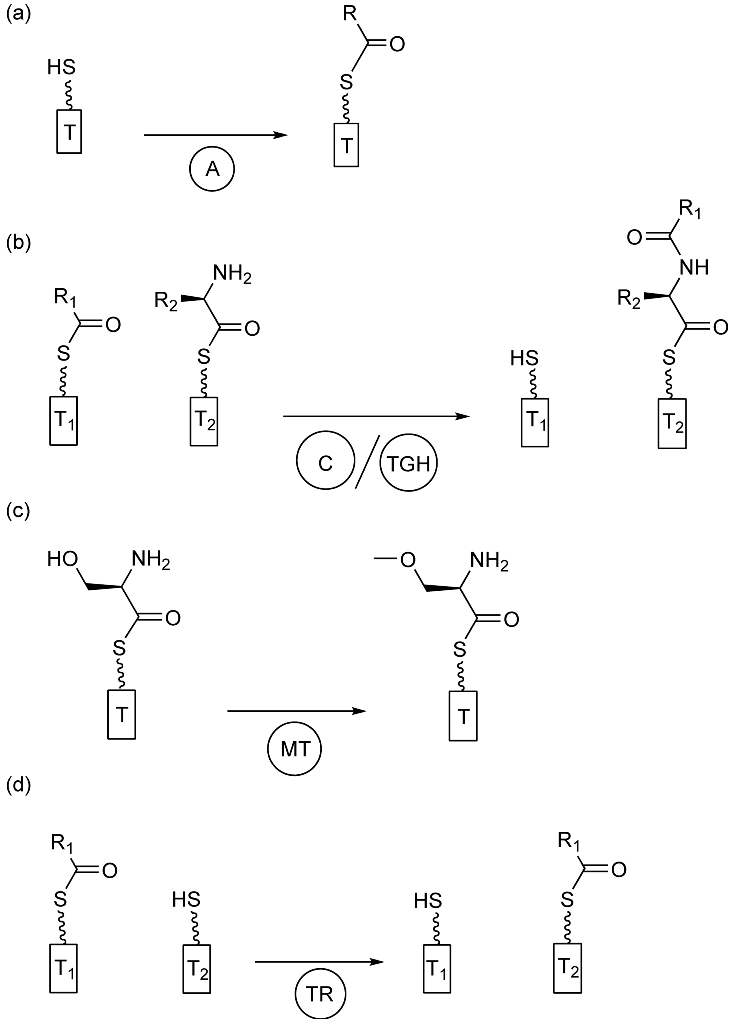


Figure 13 General substrate specificity assays. (a) Adenylation domain. (b) Condensation/TGH domain. (c) Tailoring enzyme, for example, *O*-methyltransferase. (d) Aminoacyl transferase.

Adenylation domains utilize free amino acids,[97] aryl acids,[104] or fatty acids,[51] biosynthetic substrates, and one T domain for tethering their substrates on the thiotemplate. C domains have an upstream nucleophile and a downstream electrophile as biosynthetic substrates, and two T domain substrates: one upstream and one downstream T domain.[93,105] Transferases, such as the aminoacyl transferase CmaE in the crotonine biosynthetic pathway,[106] are similar to C domains in that they use two T domain substrates but they have only one biosynthetic substrate tethered to the downstream T domain. Tailoring enzymes can have two forms of biosynthetic substrates: T domain-bound substrates[52,55] or non-T domain-bound substrates. Non-T domain-bound substrates can be free carbon acids,[46,107] CoA-activated species, or the analogue of the natural product lacking the assayed chemical modification.

MS-based substrate specificity assays that are applied to NRPS domains are mainly aimed to characterize native and alternative biosynthetic substrates. Table 1 presents biosynthetic substrate pools for substrate screening of catalytic NRPS domains and tailoring enzymes. Biosynthetic substrate pools are specified for the two purposes of assaying NRPS substrate specificity as mentioned above: (a) substrate identification and (b) determination of substrate tolerance. The biosynthetic substrate pools for each specific domain and for each experimental purpose are explained as follows.

Table 1 Biosynthetic substrate specificity assays for different NRPS domains and tailoring enzymes: Substrate pools for two experimental purposes

Domain	*Purpose of substrate specificity assay*	*Biosynthetic substrate pool*	*References*
Adenylation domain	Substrate identification	(1) Multiple biosynthetic substrates, e.g., algal lysate	91 – *in vitro* screening 97 – *in vivo* screening
		(2) Predicted native biosynthetic substrate	104, 107, 108 – *in vitro* screening 97 – *in vivo* screening
	Substrate tolerance	(1) Multiple biosynthetic substrates excluding native substrate	
		(2) Single biosynthetic substrate except native substrate	107 – *in vitro* screening 97 – *in vivo* screening
Condensation domain/ TGH domain	Substrate identification	Predicted native nucleophile and electrophile tethered to upstream and downstream T domain, respectively	105
	Substrate tolerance	Nonnative nucleophile and electrophile tethered to upstream and downstream T domain, respectively	93
Tailoring enzyme	Substrate identification	(1) Predicted native substrate tethered to T domain	52, 55, 105
		(2) Predicted native substrate	52, 105, 109
		(3) NRP natural product analogue lacking predicted chemical modification	52, 105
	Substrate tolerance	(1) Nonnative substrate tethered to T domain	52, 55
		(2) Nonnative substrate	109
		(3) NRP natural product derivatives	
Transferase	Substrate identification	Predicted native substrate tethered to downstream T domain	106
	Substrate tolerance	Nonnative substrate tethered to downstream T domain	106

Adenylation domains are the only domain type that can be studied by substrate screening with multiple biosynthetic substrates in one assay. All other NRPS domains are screened for biosynthetic substrates in a 'one-assay-one-substrate' approach. If the investigated biosynthetic substrate is T domain bound, it has to be tethered to a T domain substrate by a promiscuous PPTase, for example, Sfp, before the enzymatic reaction. If it is non-T domain bound, it is incubated with the enzyme without prior loading on a carrier protein. Biosynthetic substrate identification of A domains can be done either with a multiple substrate pool, for example, algal lysate, or with the predicted native biosynthetic substrate only. As noted above, substrate specificity assays are usually conducted *in vitro*, but recently, *in vivo* substrate specificity assays for an A domain were done with McyG adenylation–thiolation (A–T) didomain by coexpression with PPTase Svp in *Escherichia coli* and the application of varying growth media.[97] Biosynthetic substrate tolerance of A domains can be investigated *in vitro* by incubation with a multiple substrate pool lacking the native substrate or single substrates except the native substrate.

Biosynthetic substrate identification of C domains is pursued by tethering the predicted native nucleophile and electrophile to the corresponding native upstream and downstream T domains, respectively, and detection of the condensation product on the upstream active site after the enzymatic reaction. Biosynthetic substrate tolerance of C domains is characterized by the same approach as for substrate identification, except that nucleophiles and electrophiles different than from the biosynthetic pathway are screened. Electrophile biosynthetic substrate tolerance is screened with alternative electrophiles on the downstream T domain in the presence of the native nucleophile on the upstream T domain. Nucleophile biosynthetic substrate tolerance is screened with alternative nucleophiles on the upstream T domain in presence of the native electrophile on the downstream T domain.

Aminoacyl transferases can be characterized in their native biosynthetic substrate by tethering the predicted biosynthetic substrate on the downstream T domain and MS detection of substrate transfer to the upstream T domain upon incubation with the transferase. Biosynthetic substrate tolerance of transferases is tested by loading biosynthetic substrates differing from the native substrate on the downstream T domain and by the same approach as for substrate identification.

MS-based substrate identification assays of tailoring enzymes depend on whether the biosynthetic substrate is T domain bound or not. If the biosynthetic substrate is T domain bound, the predicted native substrate is loaded on the T domain and the mass change upon the tailoring reaction is detected by ESI–FTMS. If the biosynthetic substrate is non-T domain bound, conversion of the predicted native substrate by tailoring reaction can be detected by low-resolution MS. Biosynthetic substrate tolerance assays for tailoring enzymes are conducted in a 'one-assay-one-substrate' approach like substrate identification assays but with alternative biosynthetic substrates.

MS-based assays that are aimed to characterize the specificity of catalytic NRPS domains and tailoring enzymes for carrier protein substrates can be done on high-resolution mass spectrometers or, for small substrate T domains (<20 kDa),[105] on low-resolution mass spectrometers. For investigation of T domain substrate tolerance, the native T domain substrates of a catalytic NRPS domain or tailoring enzyme are exchanged by different T domains, for example, from different NRPS systems. In addition, the tolerance of T domain order can be tested for C domains and aminoacyl transferases by reverse-ordered tethering of native biosynthetic substrates to the native T domains and MS detection of the reaction product on the assayed upstream active site.[93]

Most of the latest publications on NRPS substrate specificity are focused on A domain specificity because their substrate screening is straightforward in terms of biosynthetic substrate form (free amino acids/fatty acids/aryl acids) and T domain substrates (one T domain). Four studies focus on substrate specificity of NRPS loading modules of microcystin biosynthesis,[97] mycosubtilin biosynthesis,[51] daptomycin biosynthesis,[108] and leinamycin biosynthesis.[108] The A domains of microcystin, mycosubtilin, and daptomycin biosynthesis initiation showed fatty acid specificity. The initial domain from leinamycin biosynthesis has D-amino acid specificity. Another paper presents the elucidation of aryl acid-specific AsbC adenylation enzyme from petrobactin biosynthesis.[104]

The first example in which MS was utilized to determine the specificity of an A domain from an NRPS gene cluster was published in a joint effort by the Moore and Kelleher laboratories. Hicks *et al.*[97] investigated the substrate specificity of the loading module McyG of microcystin synthetase. Microcystin is a cyclic NRPS–PKS hybrid toxin derived from various cyanobacteria genera. The initiation module McyG comprises an A–T didomain (**Figure 14(b)**), which was predicted based on the '10 letter code' to activate and load

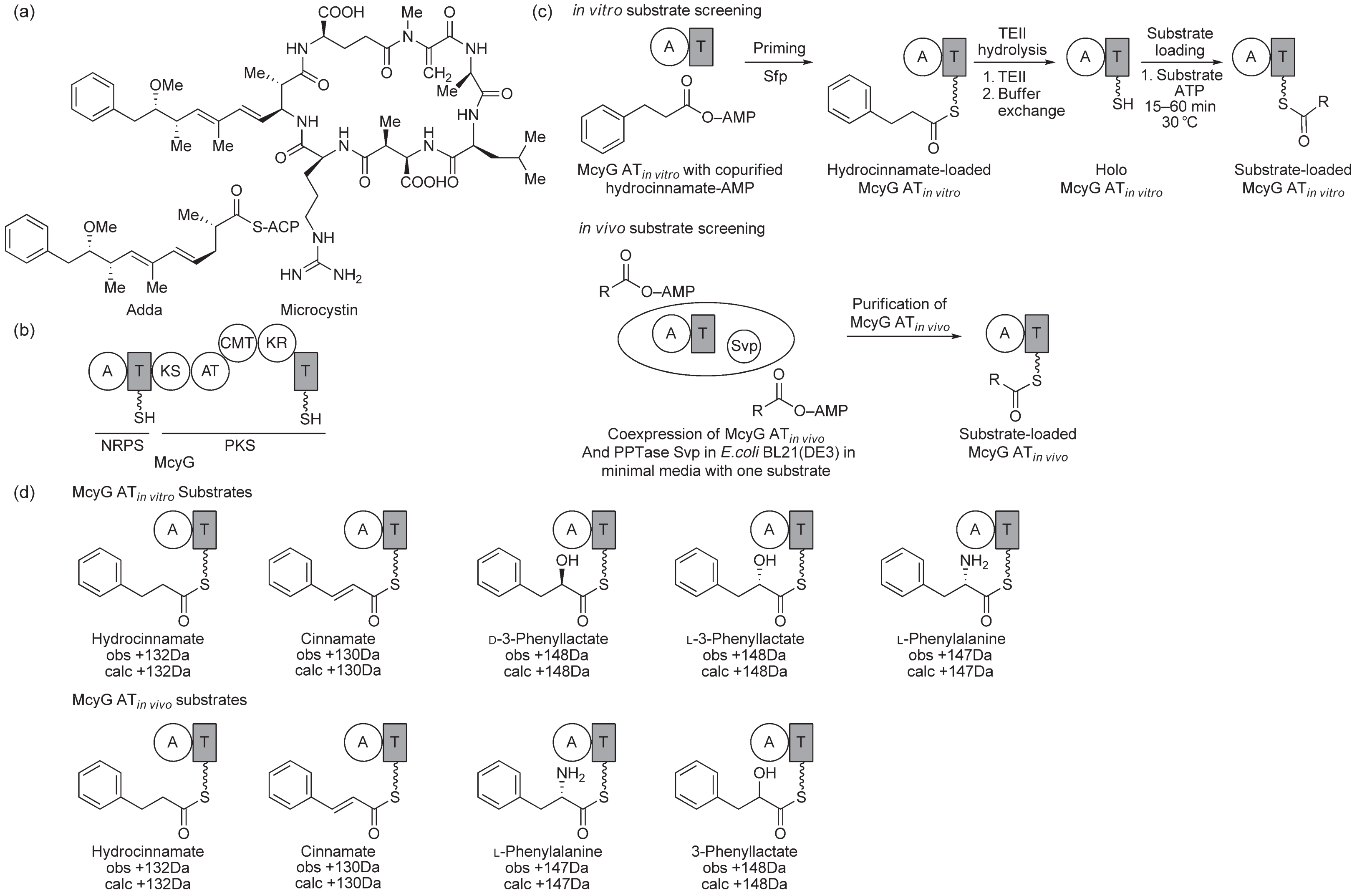


Figure 14 *In vivo* and *in vitro* substrate screening of loading module of microcystin biosynthesis. (a) Microcystin and Adda. (b) Loading protein McyG of microcystin synthetase. (c) *In vitro* and *in vivo* substrate screening assays of McyG AT. (d) Characterized substrates of McyG AT$_{in\ vivo}$ and McyG AT$_{in\ vitro}$ by ESI–FTMS (observed and calculated mass shifts from holo McyG AT active site).

phenylacetate as a starter unit for subsequent polyketide extension and formation of the aromatic β-amino acid (2S,3S,8S,9S)-3-amino-9-methoxy-2,6,8-trimethyl-10-phenyl-4,6-decadienoic acid (Adda) within the microcystin structure (**Figure 14(a)**). FT–ICR–MS was utilized to determine the substrate specificity of this domain. Two McyG AT constructs were used for substrate identification – holo AT$_{in\ vivo}$ and holo AT$_{in\ vitro}$ The holo AT$_{in\ vivo}$ construct was generated by heterologous coexpression of McyG AT with Svp PPTase and allowed *in vivo* substrate screening and, subsequently, the first detection of *in vivo* McyG intermediates (**Figure 14(c)**). *In vivo* and *in vitro* phosphopantetheinylation of McyG A–T was confirmed by MALDI–TOF MS, which allowed detection of intact McyG AT (>78 kDa). Nonetheless, no holo McyG AT$_{in\ vivo}$ and low holo AT$_{in\ vitro}$ could be detected. Furthermore a posttranslational modification with a higher mass was detected on both McyG species and was proposed as a copurified substrate bound to the McyG AT active site. This putative substrate was characterized by electrospray ionization–Fourier transform–ion cyclotron resonance–mass spectrometry (ESI–FT–ICR–MS) on the mapped T active site to have a mass that corresponded to hydrocinnamate. Subsequent hydrolysis and small-molecule MS by gas chromatography/electron impact–mass spectrometry (GC/EI–MS) confirmed this hypothesis. This is a surprising finding because it is not clear how hydrocinnamate could be used in the microcystin pathway. Because *in vivo* feeding studies could not confirm the expected phenylacetate substrate, specificity of McyG AT, Hicks *et al.* also tested alternative substrates of McyG by ATP–PP$_i$ exchange assay. Surprisingly, the A–T didomain accepted a wide range of phenylpropanoids including hydrocinnamate and was in agreement with these original findings by MS. To follow up on these observations, the McyG AT substrate specificity was further investigated by MS methods. *In vitro* substrate screening was done by hydrolysis of hydrocinnamate to free the thiol of the phosphopantetheinyl arm of the purified AT by a type II thioesterase. Once the free thiol was obtained, the sample was buffer exchanged so that the released dihydrocinnamate is removed from the reaction mixture to allow subsequent mass spectrometric activity screens with other phenylpropanoids (**Figure 14(c)**). This confirmed the substrate tolerance of McyG loading module by loading of five phenylpropanoids with different efficiencies (**Figure 14(d)**). In addition, *in vivo* substrate screening of holo McyG AT$_{in\ vivo}$ was carried out. This was done by *E. coli* growth in complex medium leading to the observed preference of hydrocinnamate loading. However, supplementation of complex growth medium with excess cinnamate led to preference of cinnamate loading on McyG AT and growth in defined media led to alternative substrate loading of 3-phenyllactate and 3-phenylalanine (**Figure 14(d)**). In summary, this study proved substrate tolerance of McyG for phenylpropanoids but not the predicted phenylacetate. This study, as well as an additional recent study that identified that the substrate phenylalanine of GrsA copurifed with a heterologously expressed A domain, indicate that this may not be an uncommon phenomenon in the investigations of NRPS proteins *in vitro*. In such a case, MS will be critical to understand what is loaded onto T domains. The characterization of *in vivo* intermediates by MS as introduced by Hicks *et al.* is the first example that demonstrates the *in vivo* reconstitution of NRPS systems in complement to MS-based *in vitro* reconstitution approaches. Therefore, it is the first example that connects *in vivo* substrate loading, albeit the protein is overproduced.

In a paper by Hansen *et al.*[51] both biochemical assays and FT–ICR–MS were utilized for characterization of the substrate specificity for another NRPS initiation module: the loading module of mycosubtilin biosynthesis. Mycosubtilin, a potent antifungal natural product of the iturin class of cyclic lipopeptides, is a β-amino fatty acid-containing octapeptide (**Figure 15(a)**). The initial protein of mycosubtilin biosynthesis is MycA, which is predicted to comprise an acyl ligase (AL) and a T domain as a loading module (**Figure 15(b)**). This didomain is called MycA10 and is a loading module with fatty acid specificity. Although being an AL, MycA10 is considered in this NRPS section because of its similar chemistry to NRPS A domains and its upstream position to an NRPS assembly line. Hansen *et al.* showed by *in vitro* assays with radiolabeled decanoic acid that the AL activates its substrate by adenylation. Substrate tolerance of the AL for fatty acids with a 10–16 carbon-comprising chain was detected by a radiolabeled chase experiment. The loading of a fatty acid on MycA was further characterized by the online LC–MS–PEA assay. AL-T$_1$ didomain was incubated with ATP and decanoic acid for 1.25 h and subsequently limited digested by trypsin. The digestion mixture was separated by reverse-phase high-performance liquid chromatography (RP–HPLC) and T domain-tethered decanoic acid was detected by LC–MS and data-dependent MS/MS or MS^2 of intact peptides for T domain active site mapping and by PEA (**Figure 15(c)**). This was accomplished using a 4.6 mm diameter C18 column on a 7 tesla FT–ICR–MS

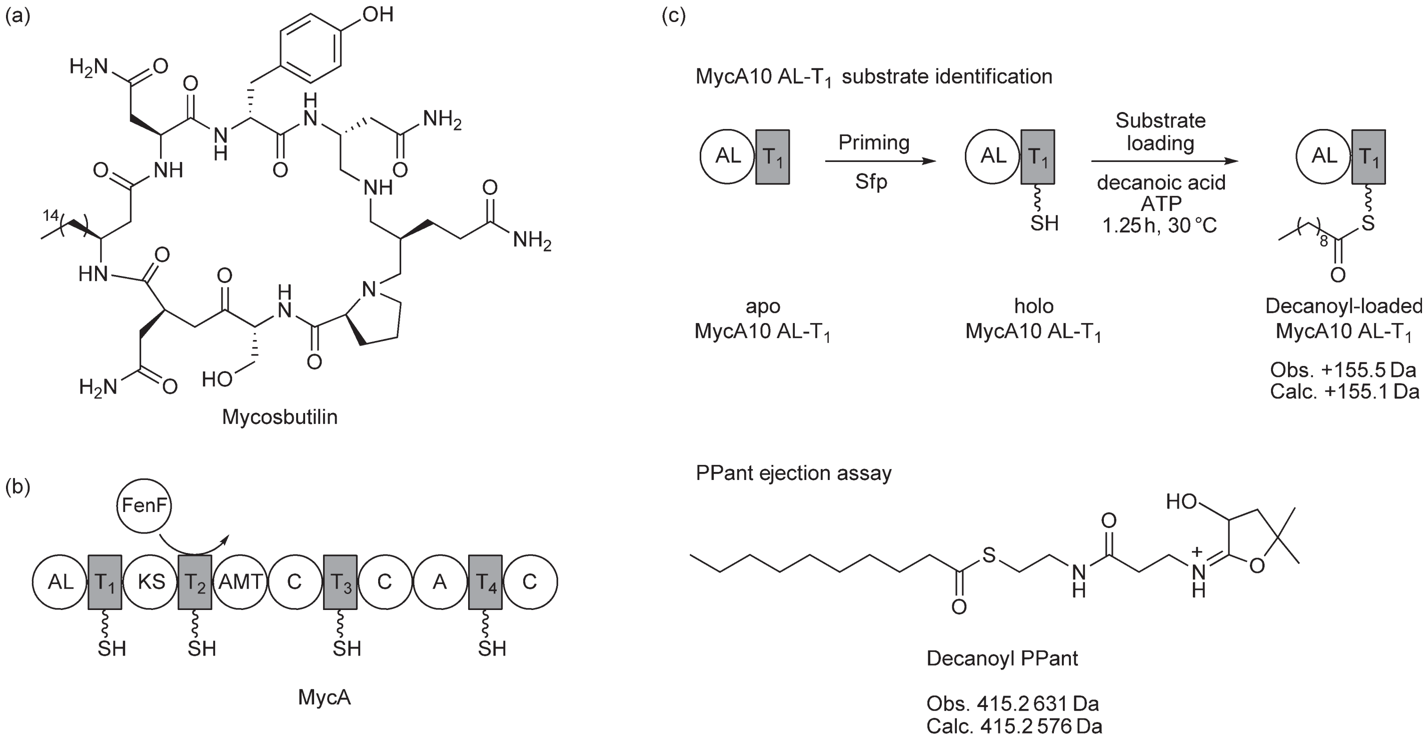


Figure 15 Substrate identification of McyA10 AL-T_1 loading module of mycosubtilin biosynthesis. (a) Mycosubtilin. (b) Loading protein MycA of mycosubtilin biosynthesis. (c) MycA10 AL-T_1 substrate identification by ESI–FTMS (observed and calculated mass shift from T_1 active site) and PEA.

instrument. This study characterized the loading module of mycosubtilin biosynthesis, MycA10, as an A domain by biochemical assays and confirmed a predicted fatty acid substrate specificity by ESI–FT–ICR–MS.

Another example of a substrate identification of an A domain is a study by Wittmann *et al.*[109] about the lipidation of daptomycin. Daptomycin is a clinically important semisynthetic derivative of the A21978 branched cyclic lipopeptide isolated from *Streptomyces roseosporus*. It comprises a 13 amino acid peptide core coupled to a decanoic acid moiety (**Figure 16(a)**). Wittmann *et al.* characterized the putative adenylating enzyme DptE as a fatty acid adenosine monophosphate (AMP) ligase that activates the natural substrate decanoic acid by adenylation and, subsequently, tethers it to the freestanding T domain DptF (**Figure 16(a)**). Thiolation activity of DptF was assayed by Bodipy labeling[110] and ESI–FTMS detection of holo DptF. Lipidation activity of DptE was characterized by detection of decanoic acid adenylation and loading on DptF (**Figure 16(b)**). Substrate tolerance of the adenylation enzyme DptE for alternative biosynthetic substrates and T domain substrates was investigated by ATP–PP_i exchange assay. Herein, Wittmann *et al.* showed that DptE can load various fatty acid substrates on DptF but not on other T domains.

Tang *et al.*[108] characterized the substrate specificity of the NRPS initiation module of leinamycin biosynthesis by ATP–PP_i exchange assay and offline-HPLC electrospray ionization mass spectrometry (ESI–MS). Leinamycin is a hybrid NRP–polyketide natural product isolated from *Streptomyces atroolivaceus* S-140. It comprises a D-alanine and shows potent antitumor activity. The NRPS initiation domain of leinamycin biosynthesis consists of a freestanding A domain, LnmQ, and a freestanding T domain, LnmP (**Figure 17(a)**), which enabled detection of apo-, holo-, and substrate-loaded T domain by low-resolution ESI–MS. The MS-based substrate screening assays were conducted by testing one substrate per assay. D-Alanine and glycine loading on LnmP was detected by ESI–MS and ATP–PP_i exchange (**Figure 17(b)**). Hereby, D-alanine was identified as the native A domain substrate because it resulted in a higher ATP–PP_i exchange activity by LnmQ, whereas glycine was characterized as an alternative substrate of the investigated A domain. LnmQ is the first known A domain with D-amino acid stereospecificity.

Finally, the characterization of adenylation enzyme AsbC within the petrobactin biosynthetic pathway by Pfleger *et al.*[104] is an example of an adenylation enzyme with aryl acid specificity. Although petrobactin is an NRPS-independent siderophore from *Bacillus anthracis* (**Figure 18(a)**), it is covered in this NRPS section too, because of the NRPS homologue enzymology of adenylation enzyme AbsC and freestanding T domain AsbD (**Figure 18(b)**). The native substrate and substrate tolerance were determined by single substrate screens and LC–ion trap MS. Therefore, the detection of AsbC specificity for aryl acid 3,4-dihydroxybenzoic acid (3,4-DHBA) is another example for substrate identification by low-resolution MS in combination with the ATP–PP_i exchange assay. AsbC substrate tolerance for multiple substituted benzoic acid derivatives could be detected by the same approach (**Figure 18(c)**).

Substrate specificity of a transglutaminase homologue (TGH) domain, which is a new type of NRPS C domain was characterized by MS application within two studies from the Walsh group.[93,105] First, Fortin *et al.* identified the substrates of TGH domain AdmF from the andrimid biosynthetic pathway by electrospray ionization–quadrupole–time-of-flight–mass spectrometry (ESI–Q–TOF MS). Then, Magarvey *et al.* investigated AdmF substrate tolerance by ESI–FTMS. Andrimid is a hybrid NRP–polyketide antibiotic isolated from various bacteria that shows nanomolar inhibition of the bacterial acetyl-CoA carboxylase. The six-module hybrid NRPS–PKS assembly line is of interest because of six interfaces between NRPS and PKS enzymology and a new type of amide bond forming C domain. At the first NRPS–PKS interface, an octatrienoyl group tethered on the T domain AdmA and (*S*)-β-Phe tethered on the T domain AdmI are condensed by TGH AdmF to give the intermediate octatrienoyl-β-Phe-*S*-AdmI (**Figure 19(a)**).

Fortin *et al.*[105] determined the native AdmF substrates by condensation product formation on the upstream T domain of AdmF, which is AdmI. The predicted electrophile was confirmed as an octatrienoyl moiety tethered to the downstream T domain AdmA. The predicted nucleophile was confirmed as a β-phenylalanine bound to the upstream T domain AdmI. The AdmF reaction products octatrienoyl-β-Phe-*S*-AdmI and holo AdmA were detected by ESI–Q–TOF MS (**Figure 19(a)**).

In the study by Magarvey *et al.*[93] substrate tolerance of multiple components of the NRPS–PKS system for andrimid biosynthesis was studied by the application of biochemical radiolabel assays and MS assays. Magarvey *et al.* characterized the stereoselective formation of (*S*)-β-Phe from L-Phe by aminomutase AdmH by incubation with radiolabeled substrates and HPLC analysis. The (*S*)-β-Phe specificity of the downstream, freestanding A

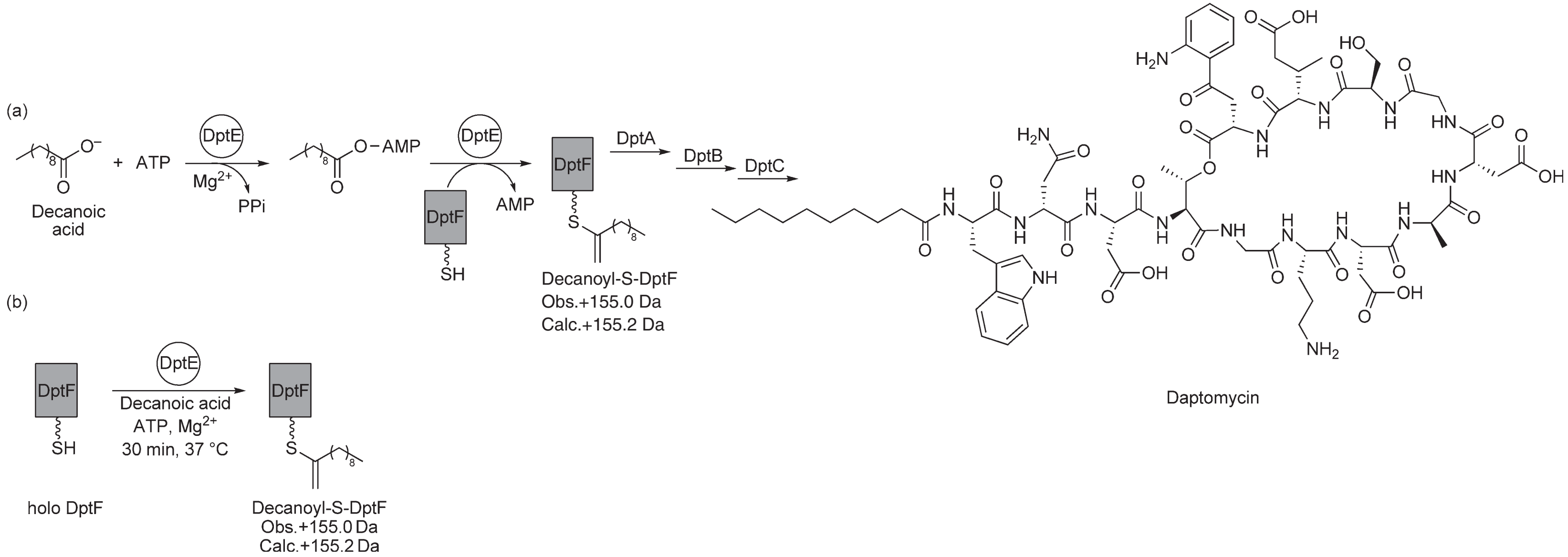


Figure 16 Lipidation of daptomycin. (a) Role of DptE and DptF in daptomycin lipidation. DptE adenylates decanoic acid and tethers it on T domain DptF for insertion into daptomycin. (b) DptE substrate identification assay (observed and calculated mass shifts from holo DptF characterized by ESI–FTMS).

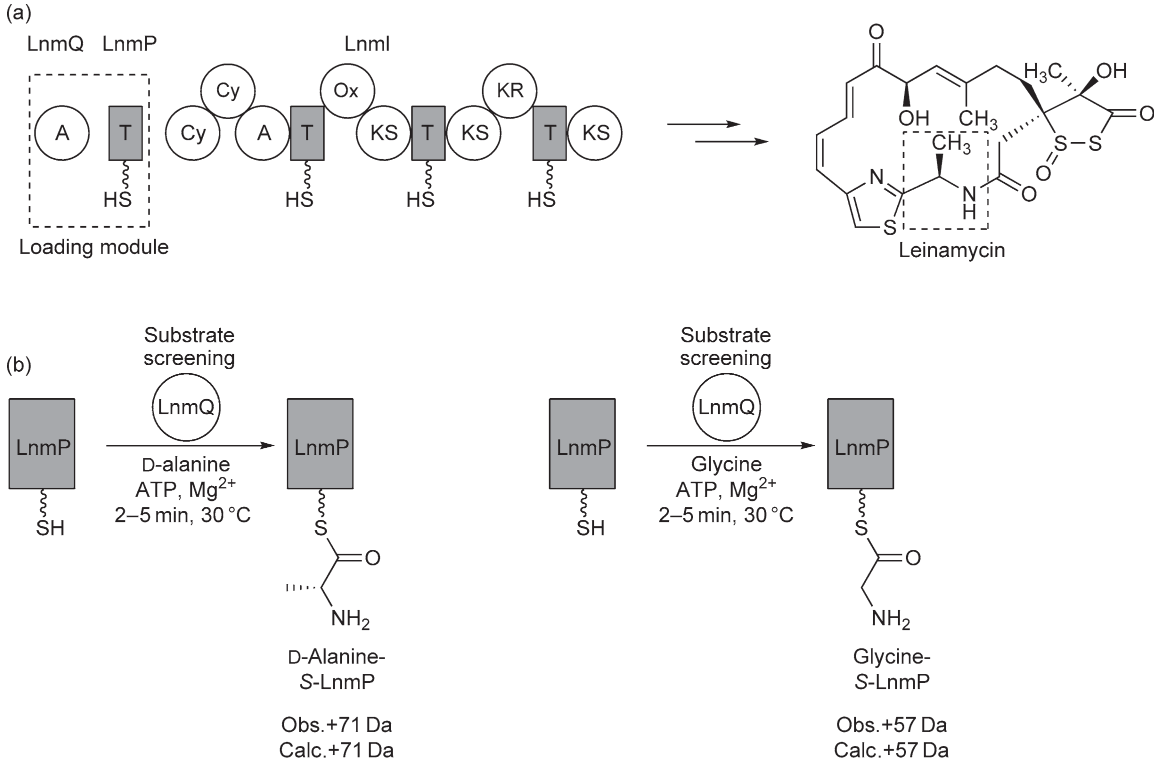


Figure 17 The loading module of leinamycin biosynthesis. (a) Loading module components (LnmQ and LnmP) of leinamycin synthetase and leinamycin structure. (b) Substrate screening assays of LnmQ. D-Alanine and glycine loading was detected by ESI–MS (observed and calculated mass shifts of holo LnmP).

domain AdmJ, which adenylates and loads (*S*)-*β*-Phe only on the T domain AdmI, was identified by ATP–PP$_i$ exchange assay. Thus, AdmH and AdmJ are considered as specificity gatekeepers within the early stages of andrimid biosynthesis. In addition, the substrate tolerance of TGH AdmF was characterized in terms of nucleophile specificity, electrophile specificity, and T domain specificity. It was shown that AdmF was specific for the T domains AdmA and AdmI in any order (**Figure 19(b)**). AdmF was characterized by MS to accept a wide array of acyl chain donors (**Figure 19(d)**) and nucleophile amines (**Figure 19(c)**) for amide bond formation, which shows its relaxed substrate specificity. All AdmF reaction products were identified by broadband FTMS of T domain-tethered intermediates and PEA through the online LC–MS–PEA assay. Therefore, AdmF was characterized as a promiscuous enzyme.

The substrate identification and determination of substrate tolerance of tailoring enzymes by MS is exemplified by two recent studies from the Marahiel group about calcium-dependent antibiotic (CDA) biosynthesis.[46,52] CDA is another nonribosomal lipopeptide, like daptomycin and mycosubtilin, with bioactivity against multidrug-resistant pathogens. CDA comprises an 11 amino acid chain cyclized to a 10-membered ring and it contains two characteristic functionalities – a unique 2,3-epoxyhexanoyl moiety and a *β*-hydroxyasparagine residue (**Figure 20(a)**).

In the first study, Kopp *et al.*[52] investigated epoxidation enzymes within CDA biosynthesis to form the 2,3-epoxyhexanoyl moiety. Two putative oxygenases, HxcO and HcmO, were cloned, expressed, and characterized in terms of substrate specificity by ESI–FTMS. HxcO was predicted as an acyl-CoA dehydrogenase that catalyzes C2–C3 bond dehydrogenation of an alkanoic acid substrate and subsequent epoxidation of the C2–C3 double bond. HcmO was predicted as a flavin-dependent monooxygenase that epoxidizes an alk-2-enoic acid substrate. First, the substrate forms of the putative epoxidation enzymes were determined. Three forms of the HxcO-predicted native substrate hexanoic acid were tested for epoxidation by HxcO: hexanoyl-CoA, hexanoyl-CDA analogue, and T domain-bound hexanoic acid. The free substrate and natural product analogue

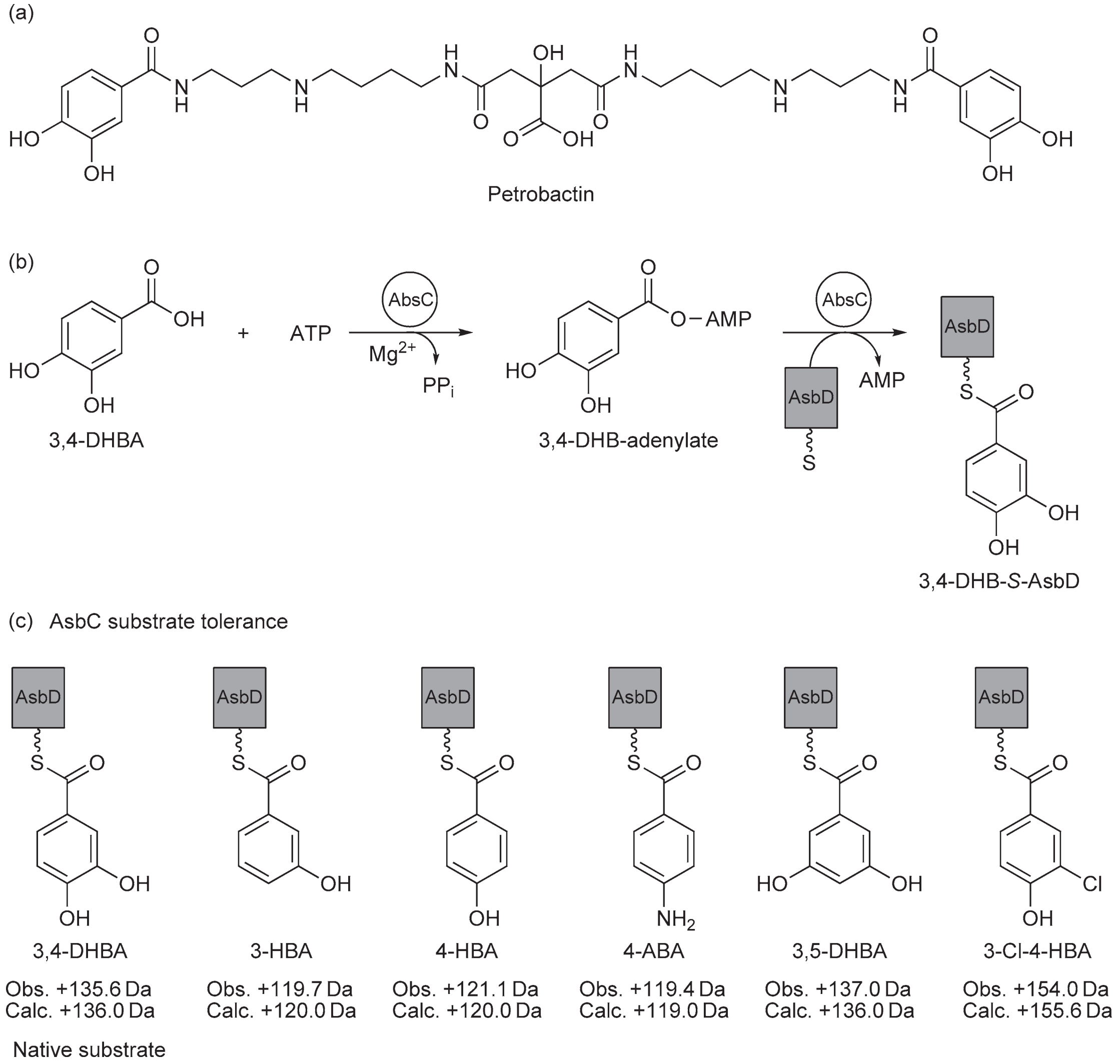


Figure 18 Adenylation enzyme AsbC from petrobactin biosynthesis has aryl acid specificity. (a) Petrobactin. (b) AsbC enzymology. AsbC adenylates native substrate 3,4-DHBA and tethers it to thiolation domain AsbD. (c) AsbC substrate tolerance characterized by LC–IT–MS (observed and calculated mass shifts of AsbD).

were not epoxidized but the T domain tethered substrate was. For HcmO, hexenoyl-CoA and T domain-bound hex-2-enoic acid were assayed as HcmO substrates and also only T domain-bound substrate epoxidation was detected. Epoxidation of the hexanoyl-*S*-T by HxcO and epoxidation of hex-2-enoyl-*S*-T by HcmO was characterized by online HPLC–ESI–FTMS and PEA (**Figure 20(b)**). In addition, substrate tolerance of HxcO and HcmO was investigated by testing epoxidation of various T domain-bound alkanoic and alkenoic acids, respectively. Herein, one alternative substrate was screened per assay and epoxidation of the substrates was characterized as before. Kopp *et al.* showed that HxcO could epoxidize various T domain-bound fatty acid substrates with different chain lengths whereas HcmO showed only epoxidation of one alternative substrate (crotonyl-*S*-T), which is similar to the HcmO natural substrate (**Figure 20(c)**). Overall, this study is a very good reference of MS-based substrate specificity assays for tailoring enzymes.

The second study on CDA tailoring enzymes is the substrate identification of the nonheme Fe^{2+}/α-ketoglutarate-dependent oxygenase AsnO, which was predicted to catalyze Cβ-hydroxylation of Asn9 side chain to yield the CDA functionality β-hydroxyasparagine. Strieker *et al.*[46] showed in this study by MS that AsnO is not hydroxylating a CDA analogue lacking the corresponding hydroxyl group or a T

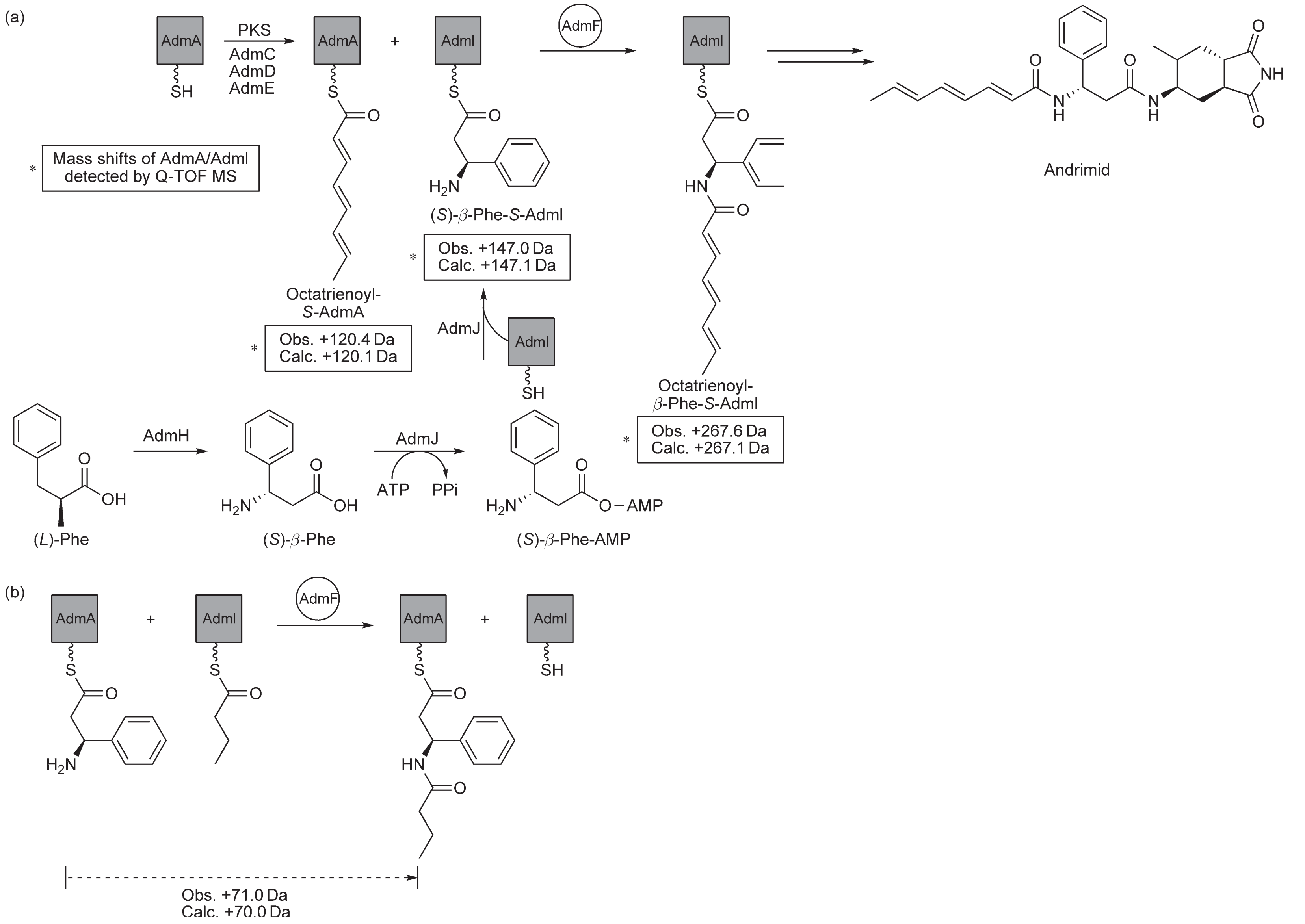
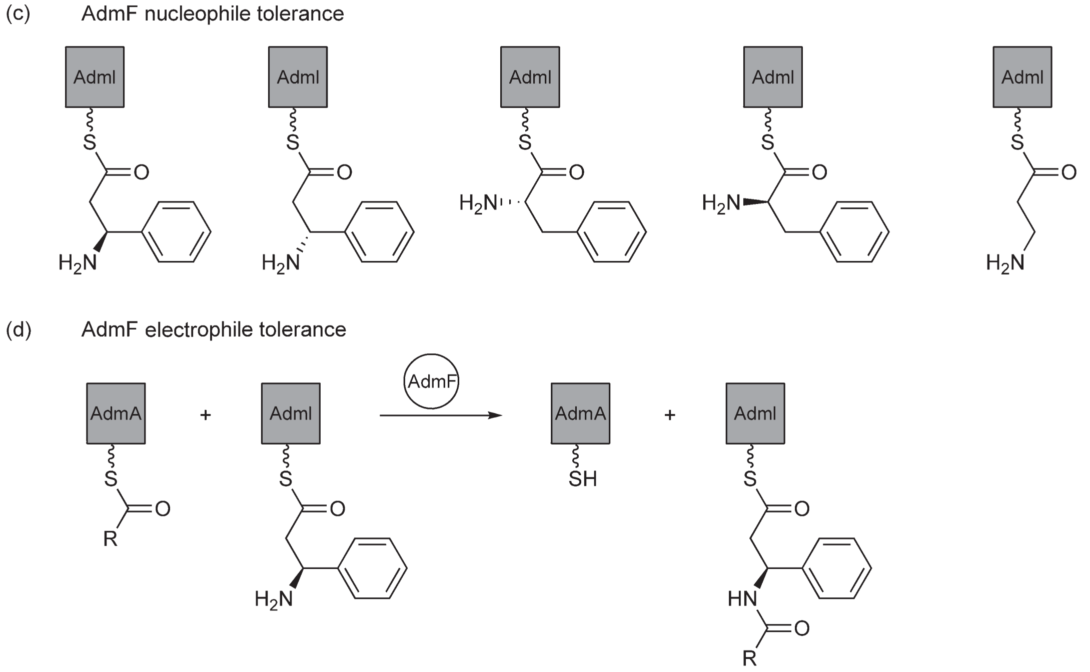


Figure 19 (Continued)



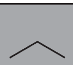
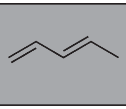
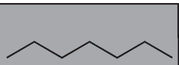
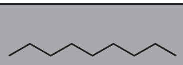
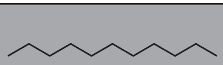
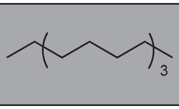
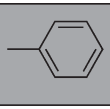
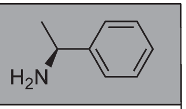
	R	CH$_3$								
Mass shift of AdmI (Da)	Obs.	189.1	218.1	241.1	276.1	302.2	328.2	384.3	265.1	294.1
	Calc.	189.1	217.1	241.1	273.2	301.2	329.2	385.3	265.1	294.1

Figure 19 Investigation of substrate specificity in early stages of andrimid biosynthesis. (a) Early stages in andrimid biosynthesis and characterization of AdmF reaction. AdmF condensation product octatrienoyl-β-Phe-*S*-AdmI was detected by ESI–Q–TOF MS (observed and calculated mass shifts of AdmI/AdmA). (b) AdmF tolerance for reverse T domain order. AdmF condensation product was detected on AdmA by ESI–FTMS. (c) AdmF nucleophile tolerance. Depicted nucleophiles were characterized by ESI–FTMS to undergo AdmF condensation with butyryl-*S*-AdmA. (d) AdmF electrophile tolerance. Electrophiles depicted in the table were characterized by ESI–FTMS to undergo AdmF-catalyzed condensation with (*S*)-β-Phe.

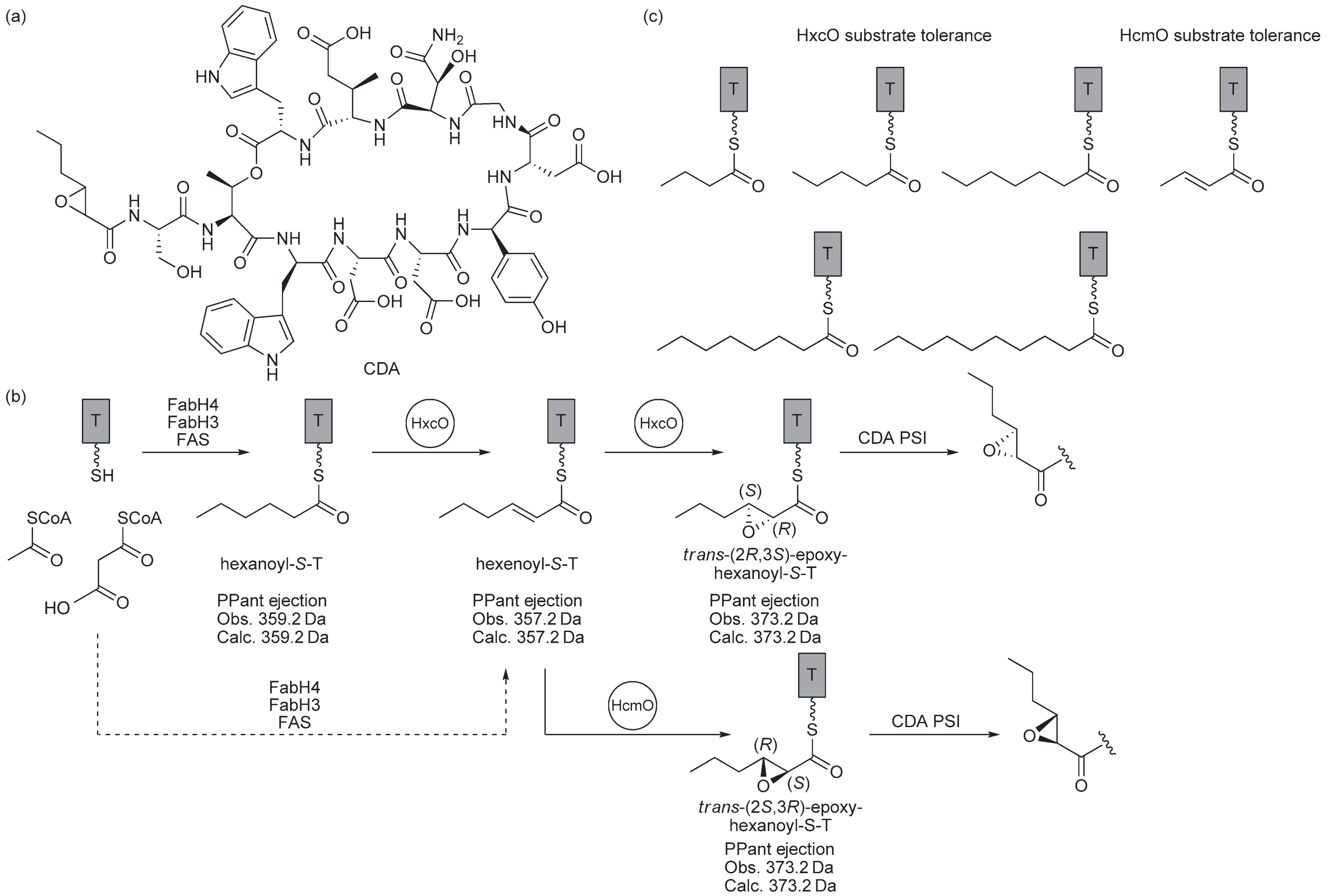


Figure 20 CDA epoxidation by tailoring enzymes HxcO/HcmO. (a) CDA. (b) Proposed mechanism of HxcO/HcmO epoxidation to form *trans*-2,3-epoxyhexanoyl moiety, reaction intermediates, and products were characterized by online LC–FTMS and PEA. The stereochemistry of HxcO/HcmO reaction products was characterized by the amide ligation assay.[52] (c) HxcO and HcmO substrate tolerance characterized By LC-IT-MS.

domain-bound L-/D-asparagine because no +16 Da mass shifts upon hydroxylation were detected after AsnO reactions with these putative biosynthetic substrates. Therefore, the most likely substrate was a free amino acid, that is, L- or D-asparagine. X-ray could identify the free amino acid L-asparagine as the native AsnO substrate.

Further recent MS applications to investigate substrate specificity were the characterization of aminoacyl-transferase CmaE in coronamic acid biosynthesis pathway[106] (**Figure 21(a)**). Purified CmaE transferred various chemically differing aminoacyl groups between various T domains, which were detected by MALDI–TOF MS, and therefore CmaE promiscuity was identified (**Figure 21(b)**).

The studies by Fortin *et al.*,[105] Strieter *et al.*,[106] Pfleger,[104] and Tang[108] showed that it is not always necessary to apply high-resolution MS for substrate identification on NRPS systems and for determination of their substrate tolerance.

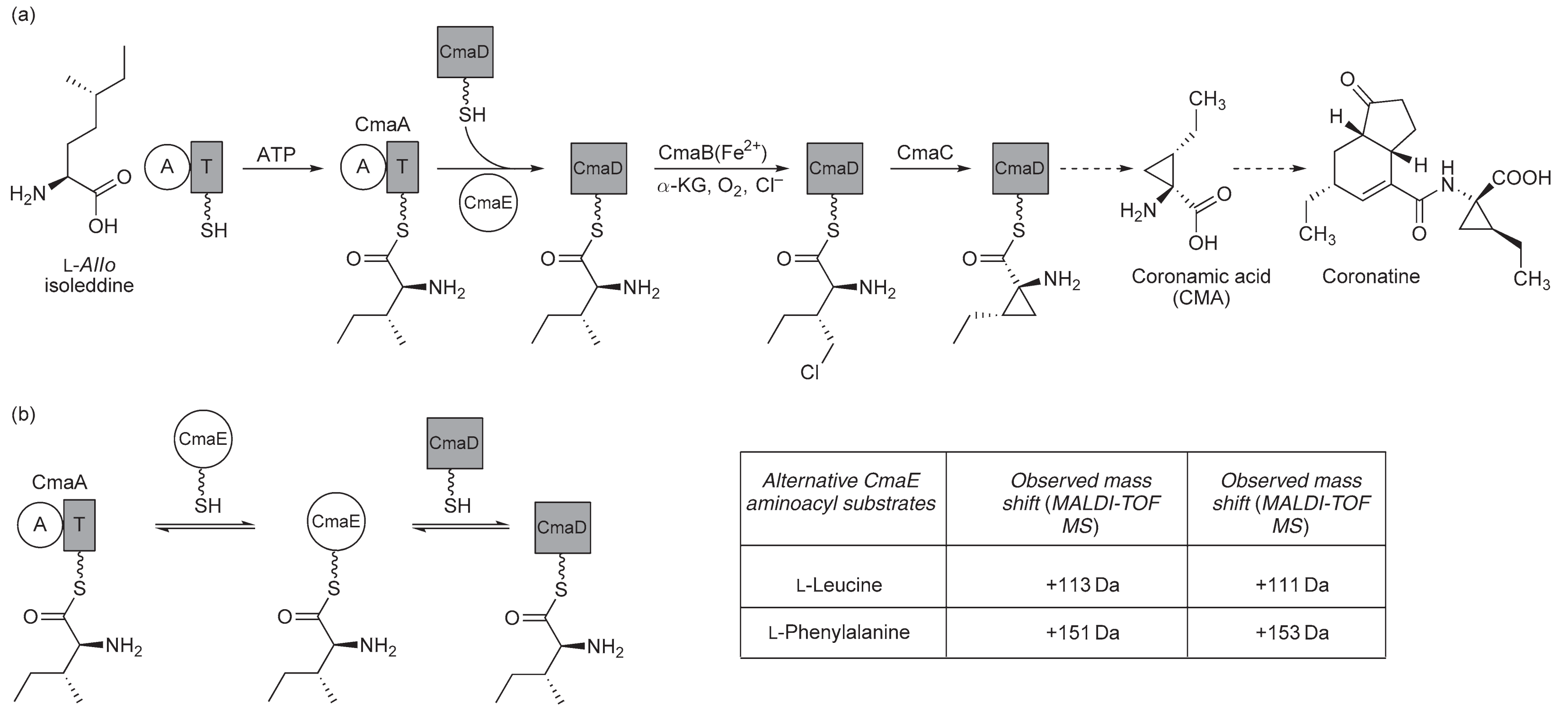
9.11.2.4 Dissection of New NRPS Enzymology and of Deviations from NRPS Colinearity

Several recent studies applied MS for mechanistic insights into new NRPS enzymology and tailoring reactions. As mentioned earlier, a new type of thiotemplate C domain, the TGH AdmF, was predicted within the andrimid biosynthesis. Before the described study by Magarvey *et al.*[93] characterizing the AdmF promiscuity, Fortin *et al.*[105] dissected amide bond formation catalyzed by AdmF applying HPLC and ESI–Q–TOF MS for detection of condensation product octatrienoyl-β-Phe tethered on AdmI T domain. In contrast to regular NRPS C domains, the TGH C domain performs covalent catalysis by an acyl–enzyme intermediate on an active site cysteine (C90) within a catalytic triad Cys–His–Asp (**Figure 22**). This mechanism of a new type of thiotemplate C domain was dissected by radiolabeling of the acyl–AdmF intermediate and its inactivation by site-directed mutagenesis (C90A).

In another mechanistic study of NRPS catalytic domains, Stein *et al.*[42] differentiated epimerization (E) domains of NRPS assembly lines into aminoacyl E domains and peptidyl E domains by monitoring intermodular transfer. The characterized differences in intermodular transfer activity by the two types of E domains from the tyrocidine biosynthesis (one from an initiation module and one from an elongation module) were done by formation of the two dimodule constructs TycB$_{2\text{-}3}$-AT-CATE/COM$_{\text{tycA}}$ with the initiation E domain of tyrocidine biosynthesis at its C-terminus and TycB$_{2\text{-}3}$-AT-CAT/E$_{\text{tycA}}$ with an elongation E domain of tyrocidine biosynthesis at its C-terminus. For both the constructs, Phe–Phe dipeptide formation rate was characterized by ESI–FTMS quantification of TyrB$_{2\text{-}3}$ T$_3$ domain-loaded intermediates. TyrB$_{2\text{-}3}$ T$_3$ domain active site was mapped by preliminary fluorophore labeling, HPLC separation, and online LC–ESI–FTMS analysis of labeled fractions. For investigation of intermodular transfer of aminoacyl or dipeptidyl groups by E domains, a reporter construct TyrB$_1$-CAT/TE$_{\text{srf}}$ was formed. TyrB$_1$-CAT could be recognized by both E domains and it monitored aminoacyl transfer by dipeptide formation (**Figure 23**(**a**,1)) and peptidyl transfer by tripeptide formation (**Figure 23**(**a**,2)). Both products were detected and quantified as offloaded molecules by HPLC–ESI–TOF MS after incubation of each of the E domain constructs with the reporter construct. The E domain of the initial module showed more activity of aminoacyl transfer to the downstream T domain and is called aminoacyl E domain. This type of E domain causes misinitiation within an NRPS assembly line by unselective formation of a D-amino acid and its subsequent intermodular transfer. The E domain of the TycB elongation module showed only selective D-Phe formation in Phe–Phe dipeptidyl intermediate and subsequent transfer of L-Phe–D-Phe to the reporter construct. This type of E domain is called peptidyl E domain, which is a gatekeeper for intermediate downstream processing in nonribosomal biosynthetic machineries (**Figure 23(b)**). This approach by Stein *et al.* demonstrates how mechanistic insights can be gained through formation of well-chosen NRPS constructs and mass spectrometric analysis.

The recent studies of Hansen *et al.*[51] and Wittmann *et al.*[109] revealed a new mechanism for lipidation of lipopeptide biosynthesis such as mycosubtilin or daptomycin biosynthesis by application of ESI–FTMS. Both papers describe that fatty acid incorporation is catalyzed by an A domain with fatty acid specificity in the loading module of the mycosubtilin NRPS (**Figure 15**) or by a preassembly line A and T domain in daptomycin biosynthesis (**Figure 16**).

In addition, Tang *et al.*[108] characterized a new mechanism for D-amino acid incorporation into NRP chains by NRPS, which was previously thought to happen only by epimerization of T domain-tethered L-amino acid



Alternative CmaE aminoacyl substrates	*Observed mass shift (MALDI-TOF MS)*	*Observed mass shift (MALDI-TOF MS)*
L-Leucine	+113 Da	+111 Da
L-Phenylalanine	+151 Da	+153 Da

Figure 21 Characterization of aminoacyl transferase CmaE in coronamic acid biosynthesis pathway. (a) Within the biosynthetic pathway, CmaE carries out substrate shuttling from the CmaA T domain to the CmaD T domain. (b) CmaE substrate tolerance was characterized by MALDI–TOF MS (observed and calculated mass shift of CmaD in the table). In addition, evidence of reversible aminoacyl transfer by CmaE was detected.

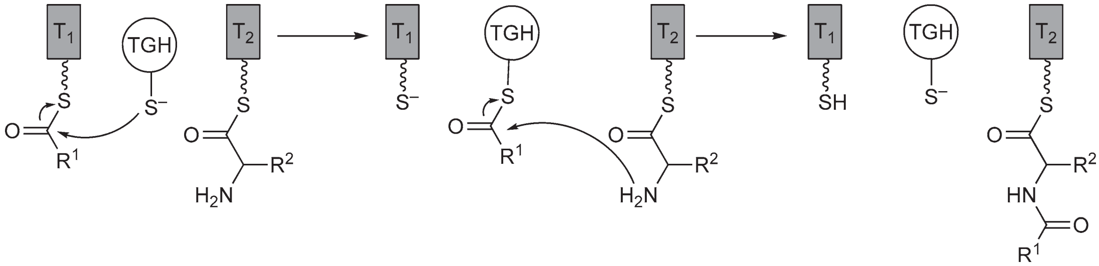


Figure 22 Condensation mechanism of transglutaminase homologue domain by covalent atalysis.[105]

substrates via ${}^{L}C_{D}$ domains or E domains. The initiation module of leinamycin biosynthesis comprises an A domain that has D-alanine specificity (**Figure 17**) and, therefore, it constitutes a new direct mechanism of D-amino acid incorporation.

Wenzel *et al.*[111] applied MS to investigate a deviation from NRPS colinearity in the myxochromide S biosynthetic pathway – the first reported NRPS module skipping process. Myxochromides are lipopeptides isolated from several myxobacteria. Myxochromides A are structurally similar lipohexapeptides produced by *Myxococcus xanthus* and myxochromides S are structurally similar lipopentapeptides produced by *Stigmatella aurantiaca*. Both the myxochromide types are biosynthesized by hybrid PKS/NRPS that are identical in module and domain arrangement. Because myxochromides S contain only a five amino acid chain despite a six module NRPS (**Figure 24(a)**), Wenzel *et al.* investigated the absence of an L-proline in myxochromides S that is integrated by module 4 of the correlating MchC_{A} NRPS into myxochromides A. Adenylation activity was detected for both A4 domains from MchC_{A} NRPS and MchC_{S} NRPS by ATP–PP_{i} exchange assay with a slightly lower activity of A4 from MchC_{s}. The T domain activity was investigated by coexpression of each T4 with PPTase MtaA in *E. coli* and ESI–FTMS and MALDI–TOF MS analysis of the GST-tagged T domain constructs. Only phosphopantetheinylation of T4 from MchC_{A} was detected by the corresponding +340 Da shift of the apo T domain peak. Although the GST-T domain constructs were ~35 kDa in size, no active site mapping was necessary, which shows that very pure NRPS constructs larger than 20 kDa can be characterized by ESI–FTMS or MALDI–TOF MS. The lacking MchC_{S} T4 domain activity is due to a point mutation in the carrier domain active site that results in the complete deactivation of the module 4 in myxochromides S biosynthetic assembly line. Therefore, Wenzel *et al.* proposed a module skipping mechanism based on the characterized 'loss-of-function' point mutation in the T4 domain, which is the first skipping process described in a multimodular NRPS (**Figure 24(b)**).

9.11.2.5 Characterization of Tailoring Reactions

Recently, two new tailoring reactions were analyzed using MS.[55,107] Other NRP tailoring mechanisms that have been characterized by utilization of MS and protein crystallography have been reviewed elsewhere.[47,112]

Gatto *et al.*[107] characterized the mechanism of L-pipecolic acid formation by cyclodeaminase RapL from L-lysine within rapamycin biosynthesis, which is a hybrid NRP–polyketide antibiotic (**Figure 25(a)**). RapL was characterized by biochemical assays to require cofactor nicotinamide adenine dinucleotide (NAD^{+}) and an oxidative cyclodeamination reaction mechanism corresponding to ornithine cyclodeamination was proposed based on ESI–FTMS analysis of RapL reaction products (**Figure 25(b)**).

Another cyclization mechanism was investigated by Kelly *et al.*[55] Cyclopropane ring formation by CmaC catalysis from a γ-chloro-L-*allo*-Ile intermediate within coronamic acid assembly line was characterized. Coronamic acid is a fragment of coronatine, a hybrid NRP–polyketide phytotoxin (**Figure 21(a)**). In this study, mechanistic insights were gained by ESI–Q–FT–ICR–MS detection of T domain CmaD tethered intermediates of CmaC reaction. Isotopically labeled substrates at specific positions combined with the high mass accuracy of ESI–FT–ICR–MS allowed the dissection of CmaC-catalyzed propane ring formation by detection of γ-chloride loss (**Figure 26(a)**) and α-hydrogen exchange (**Figure 26(b)**). Therefore, a new mechanism of cyclopropane formation was proposed in which CmaC-Zn^{2+}-mediated carbanion formation is

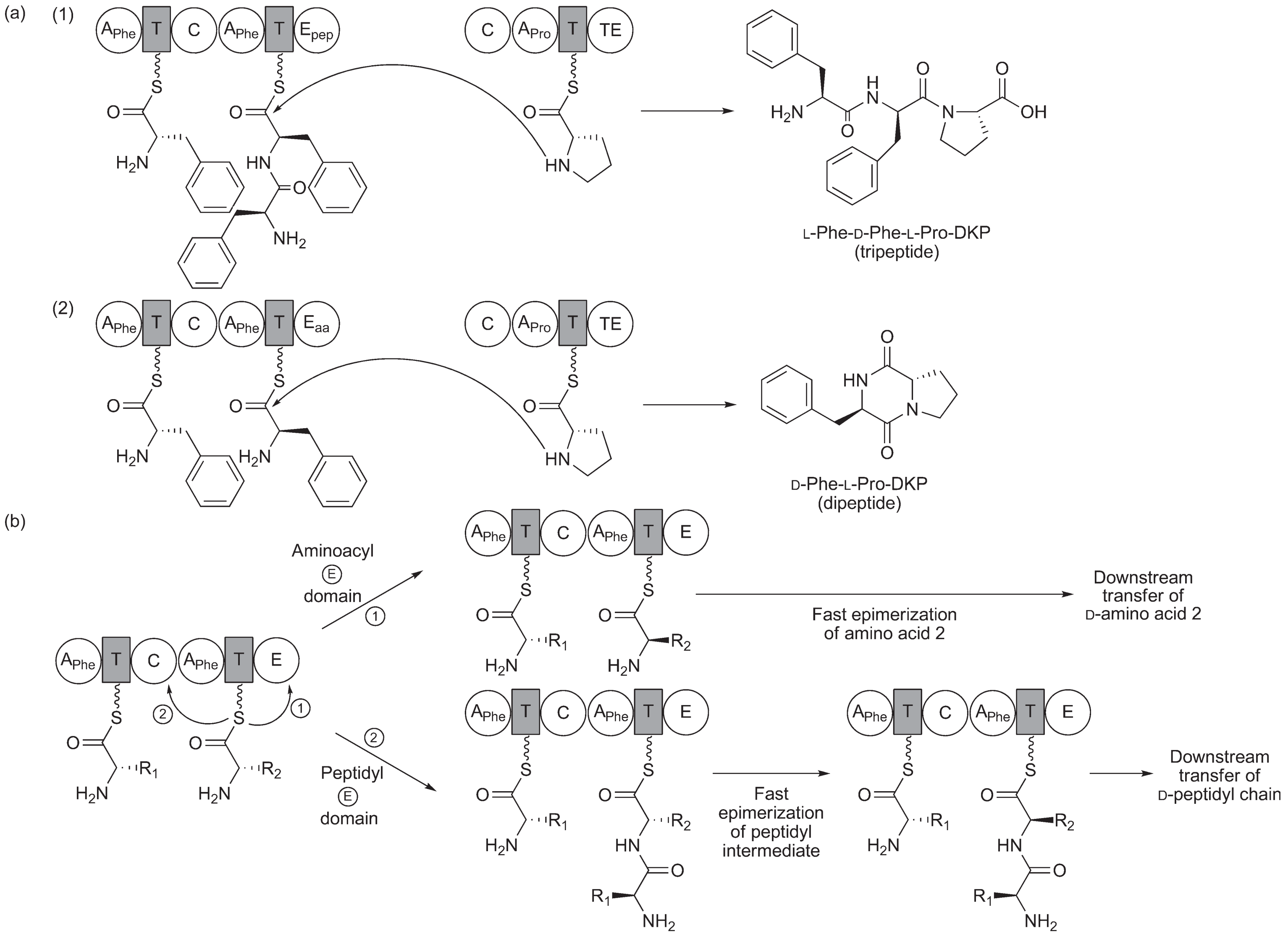


Figure 23 Differentiation of initiation and elongation epimerization domains by intermodular transfer activity. (a) Intermodular transfer activity assay. Peptidyl or aminoacyl specificity of epimerization domains is monitored by tripeptide formation or dipeptide formation, respectively. (b) Proposed mechanism of aminoacyl epimerization domain and peptidyl epimerization domain in NRPS.[42]

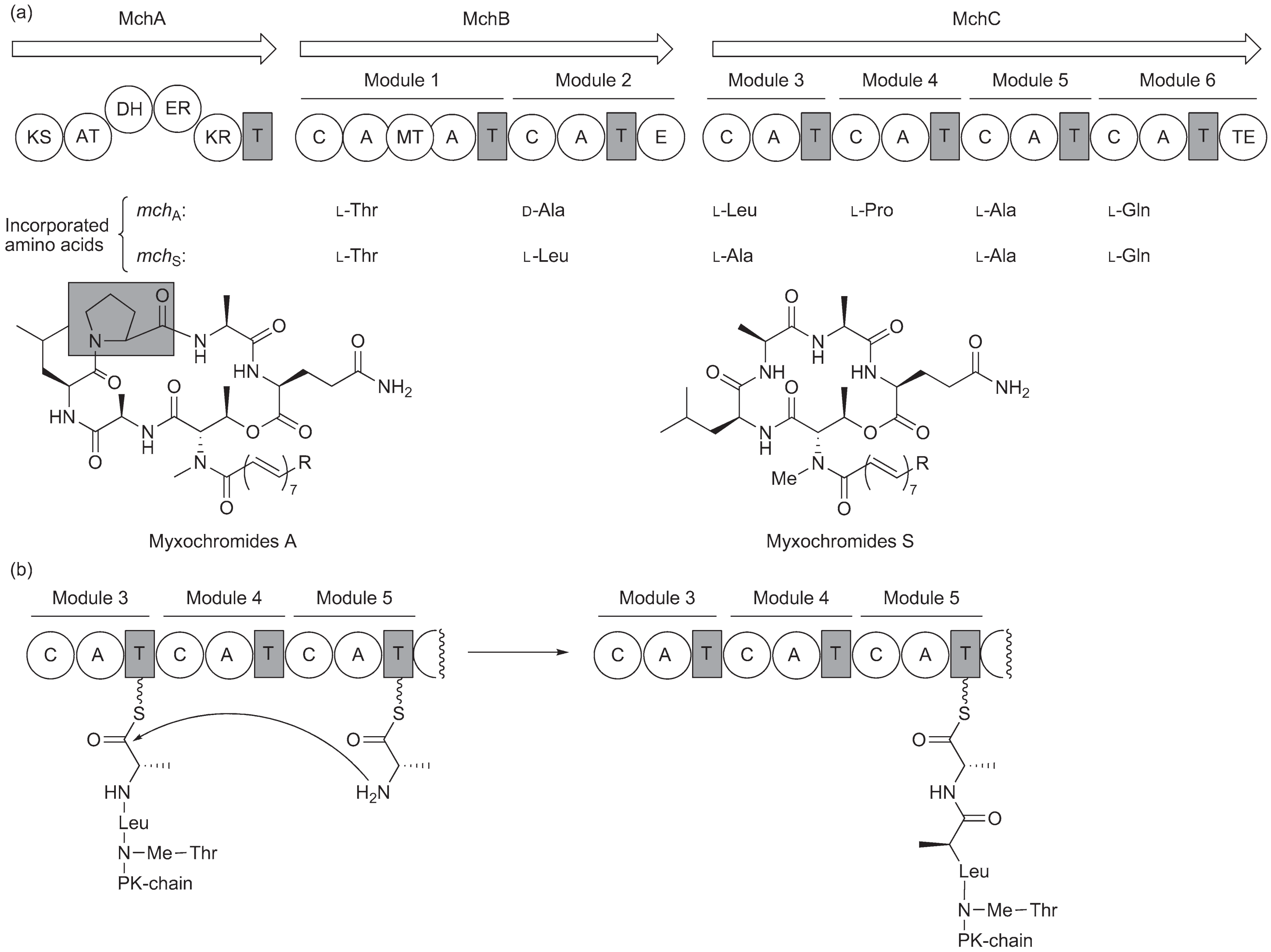


Figure 24 NRPS module skipping mechanism revealed by myxochromides A and S biosynthesis. (a) Myxochromides A and S structures and module and domain arrangement in their biosynthetic machineries. (b) Proposed NRPS-module skipping mechanism by 'loss-of-function' mutation in T domain of the skipped module.

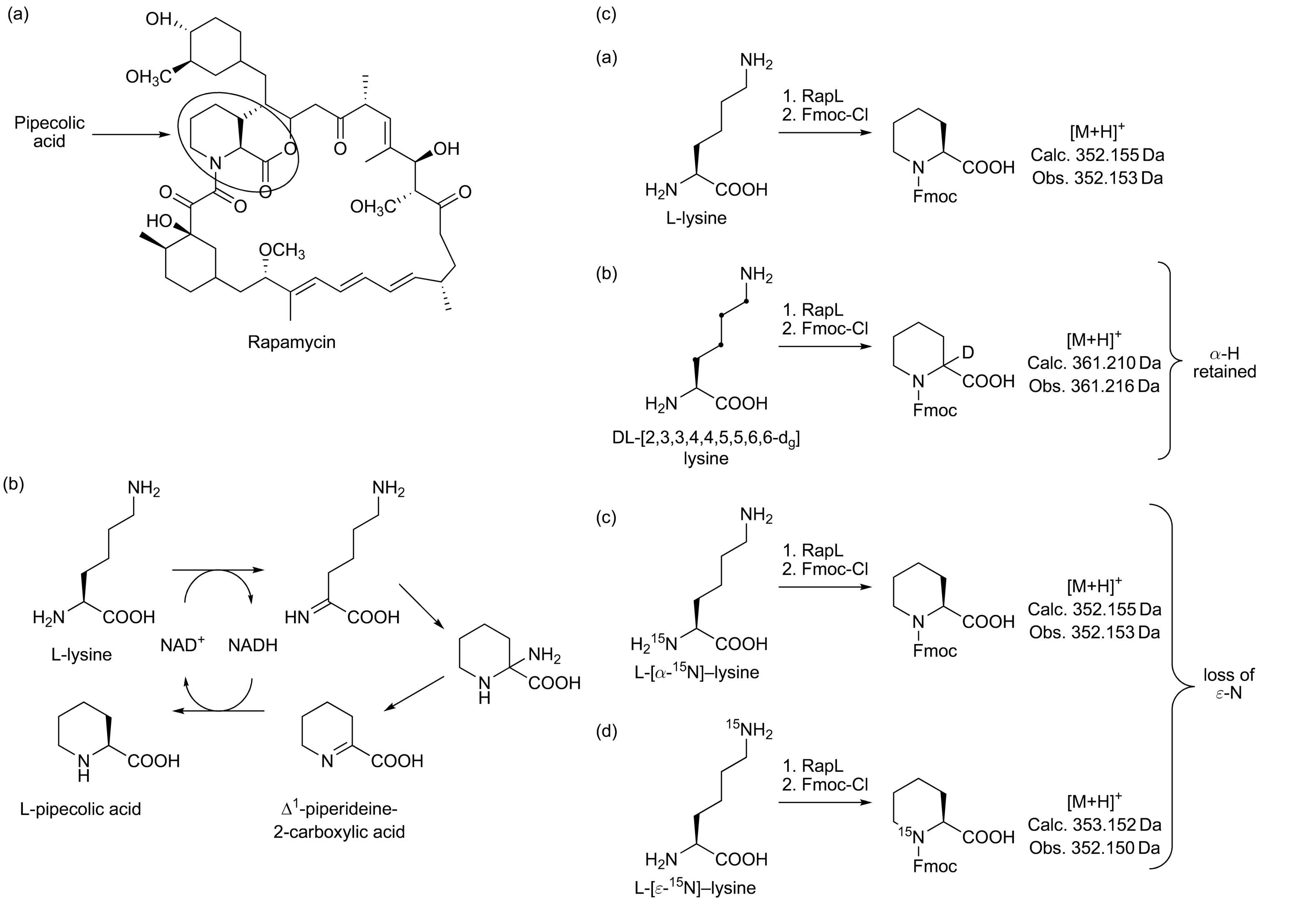


Figure 25 L-Pipecolic acid formation by cyclodeaminase RapL in rapamycin biosynthesis. (a) Rapamycin and incorporated pipecolic acid moiety. (b) Proposed oxidative cyclodeamination mechanism of pipecolic acid formation from L-lysine. (c) RapL activity assays and exact ESI–FTMS analysis of derivatized reaction products revealing mechanistic insights such as α-H retainment and loss of ε-N.

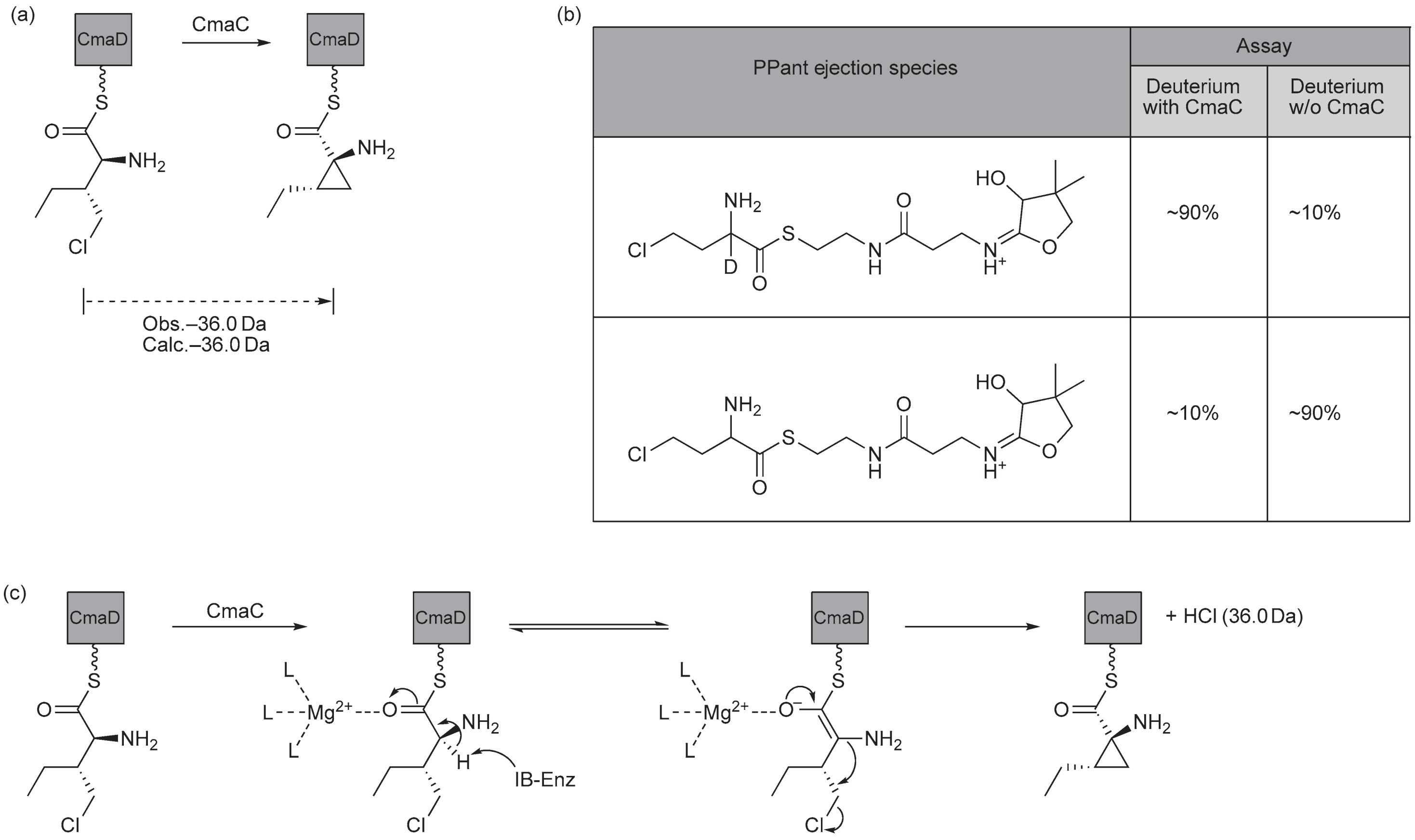


Figure 26 Cyclopropane ring formation by CmaC in coronamic acid biosynthesis. (a) Characterization of γ-Cl loss by ESI–FTMS. (b) Characterization of α-H exchange by deuterium solvent exchange and PEA. CmaC promotes deuterium incorporation at α-C position. (c) Proposed cyclopropane ring formation mechanism.[55]

followed by nucleophilic substitution of a γ-Cl-leaving group (**Figure 26(c)**). Alternative mechanisms such as azetidine formation were experimentally denied. Before CmaC-catalyzed cyclization, γ-chloro-L-*allo*-Ile is transferred from CmaA T domain to CmaD T domain by predicted aminoacyl transferase CmaE. In a subsequent study, Strieter *et al.*[106] investigated this shuttling step within the coronamic acid biosynthesis by application of MALDI–TOF MS. As described above, CmaE shuttling promiscuity was characterized by MS. In addition, the reversibility of CmaE transfer within the coronamic acid biosynthesis was also detected by MALDI–TOF MS. L-Valine, tethered on upstream CmaD T domain, was detected after several minutes incubation with CmaE on downstream holo CmaA T domain. This suggested reversible aminoacyl transfer by CmaE.

A novel epoxidation tailoring reaction is described by the work of Kopp *et al.*[52] about the formation of the 2,3-epoxyhexanoyl moiety in CDA. Online LC–ESI–FT–ICR–MS and the PEA characterized that both the putative oxygenases HxcO and HcmO of the CDA gene cluster are involved in the fatty acid epoxidation. It is shown that HxcO catalyzes the C2–C3 dehydrogenation of hexanoyl-*S*-T and subsequent hexenoyl-*S*-T epoxidation to yield 2,3-epoxyhexanoyl-*S*-T as the main product. Hexenoyl-*S*-T was detected as an HxcO reaction side product. HcmO utilizes only hexenoyl-*S*-T to give the 2,3-epoxyhexanoyl-*S*-T. Therefore, Kopp *et al.* concluded that HxcO is the main fatty acid epoxidation catalyst within CDA biosynthesis and HcmO epoxidizes the hexenoyl side product of the HxcO reaction to the epoxyhexanoyl moiety (**Figure 20(b)**). Interestingly, it was shown by a novel amide ligation assay with an amine enantiomer and subsequent chiral HPLC analysis that the HxcO and HcmO epoxides have opposite stereochemistry. This is an example for a limitation of MS for characterization of biosynthetic pathways.

9.11.2.6 Characterization of Multistage Assembly Line Action

The characterization of multistage assembly line action by high-resolution MS was developed in particular on yersiniabactin assembly line components. Five of its active sites were detected in parallel to monitor intermediate downstream processing on a biosynthetic thiotemplate.[47,113]

In a recent study by Garneau-Tsodikova *et al.*,[96] the early stages of prodigiosin biosynthesis were characterized by ESI–FTMS detection of intermediates on two active sites. Although the biosynthetic steps could only be characterized in a single turnover fashion, it is an example for dissection of multistage assembly line action by MS. Prodigiosin of the prodiginine class of natural products comprises three pyrrole rings and is derived by *Serratia marcescens*. Each of its three pyrrole rings is proposed to be formed by a different mechanism. In the first stages of prodigiosine biosynthesis, dipyrrole formation is proposed as depicted in **Figure 27(a)**. Adenylation domain PigI adenylates and loads L-proline onto freestanding PigG carrier protein. L-Prolyl-*S*-PigG is double dehydrogenated by flavoprotein desaturase PigA yielding pyrrolyl-*S*-PigG intermediate. The pyrrolyl group is transferred to the PigJ active site, a ketosynthase with a chain length factor (CLF) partner domain. The ketosynthase catalyzes decarboxylation of a malonyl moiety to a carbanion tethered on the T_1 active site (we changed the original notation from this paper as ACP_1 to T_1 to reflect the notation used in this review) of the tridomain PigH and subsequent formation of pyrrolyl-β-ketoacyl-*S*-PigH. PigH comprises a putative PLP-containing seryltransferase (SerT), which catalyzes the formation of the second pyrrole ring by serine insertion. Garneau-Tsodikova *et al.* could confirm most of these biosynthetic steps by *in vitro* reconstitution assays and ESI–FTMS detection of corresponding intermediates on PigG T domain and PigH T_1. On FTMS-mapped PigG active site, tethered L-proline and L-pyrrolyl intermediates could be detected after PigI adenylation or PigA double dehydrogenation, respectively. In addition, simultaneous substrate loading of both PigH active sites (T_1 and T_2) could be confirmed (**Figure 27(b)**). On FTMS-mapped PigH T_1 active site, malonyl and pyrrolyl-β-ketoacyl intermediates were detected upon PigL malonyl loading or PigJ Claisen condensation, respectively. Serine insertion by SerT and bipyrrole formation have not been detected by ESI–FTMS yet. The study shows the dissection of predicted NRPS–PKS assembly line action only by ESI–FTMS and manifests its potential as a stand-alone investigative tool of biosynthetic thiotemplates. Once again, an NRPS–PKS interface is investigated.

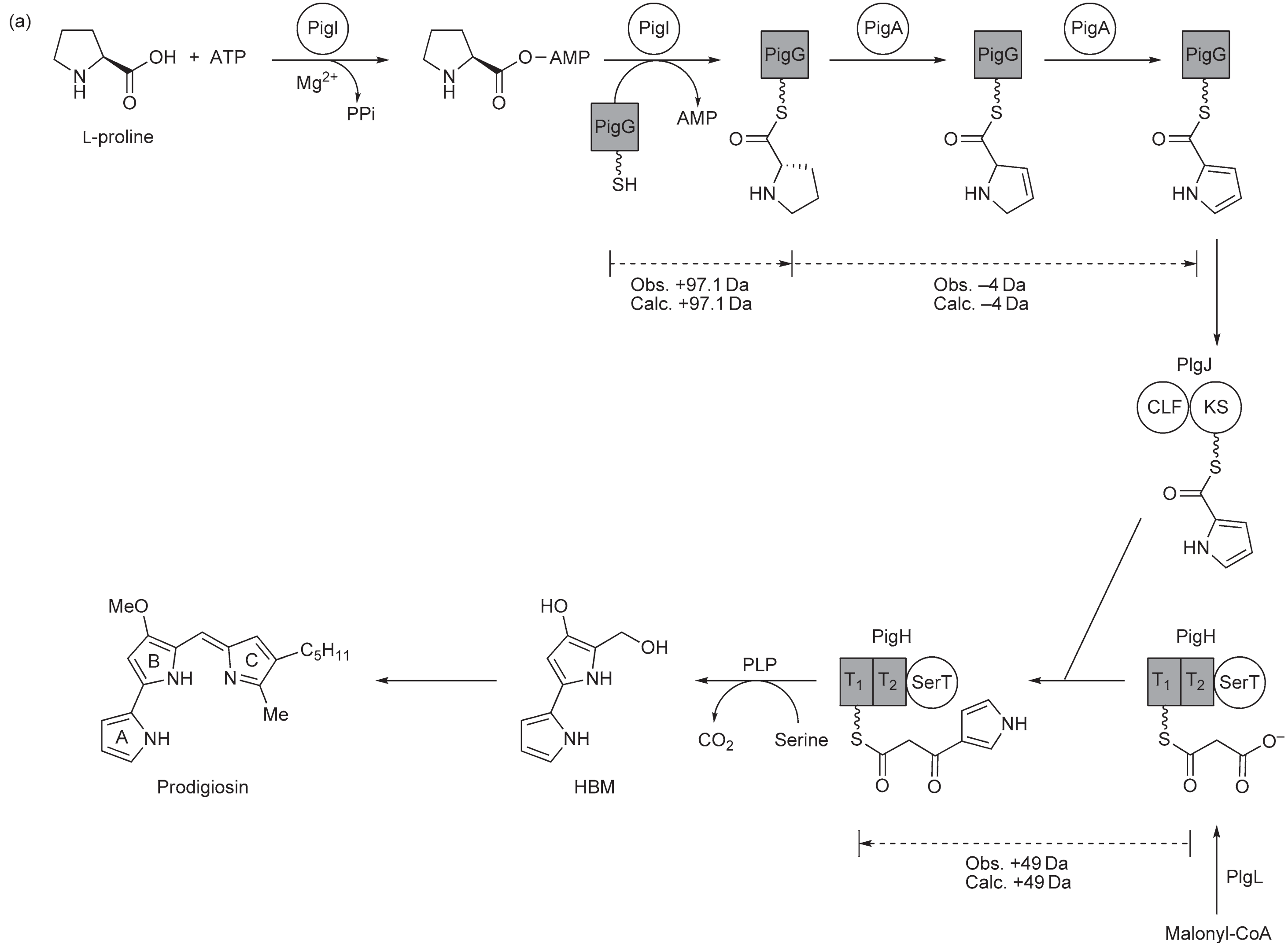


Figure 27 (Continued)

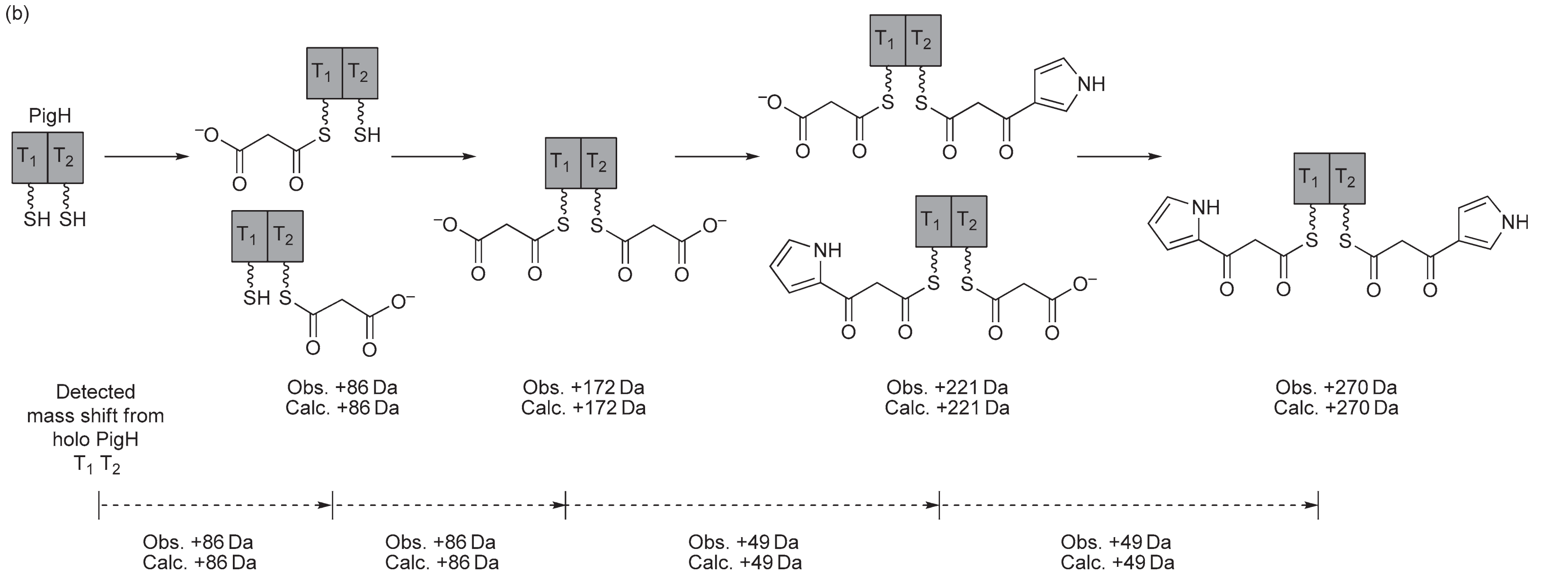
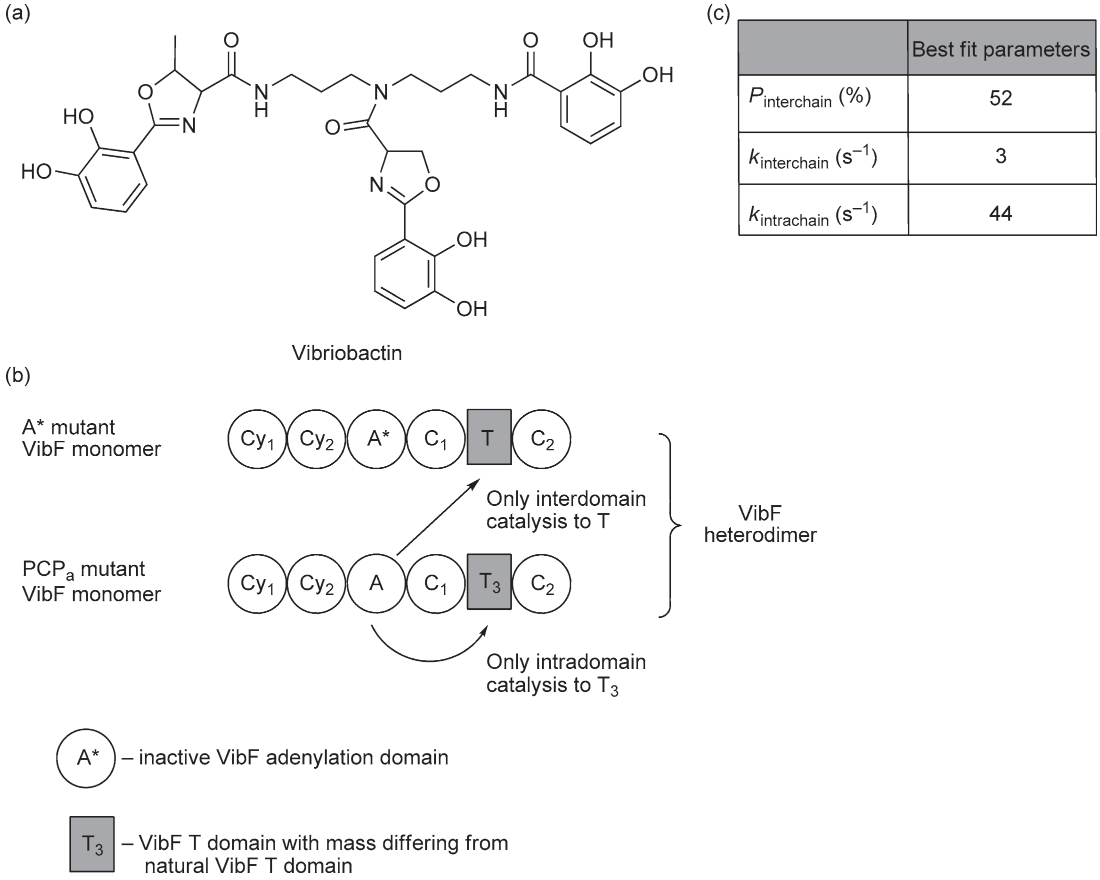


Figure 27 Monitoring multiple active sites in prodigiosin biosynthesis. (a) Early stages of prodigiosin biosynthesis dissected by *in vitro* reconstitution and ESI–FT–ICR–MS (observed and calculated mass shifts of PigG T domain and PigH T_1 domain active sites). (b) Characterization of substrate loading on both PigH active sites T_1 and T_2 by ESI–FT–ICR–MS (observed and calculated mass shifts of PigH T_1T_2).

9.11.2.7 Time Courses

The investigation of time courses of NRPS assembly line processing by MS was realized by semiquantitative approaches until 2006, as highlighted by Dorrestein and Kelleher.[47] The problem of obtaining true-kinetic time courses for intermediate flux on thiotemplates is the limited complexity of peptide mixtures,which can be analyzed by ESI–FTMS and different ionization behavior of the same active site species loaded with different intermediates causing deviations from true intermediate quantities. So far kinetic time course experiments on NRPS systems have been considered as pseudokinetic. An example is a study by Hicks *et al.*[98] in which interchain and intrachain acylation in dimeric VibF of vibrobactin NRP biosynthesis was detected and semiquantitative rates for both processes in presteady state were calculated. By detection of time-dependent holo T domain decrease during L-threonine loading by VibF A domain, the relative occupancy of T domains with L-threonyl intermediate was indirectly measured because of unsuccessful L-threonyl-*S*-T domain detection by ESI–FTMS. Interchain and intrachain acylation in the homodimeric VibF was investigated by formation a of a heterodimeric construct with an inactive A domain on one chain and two T domains with differing masses (**Figure 28(b)**). Acylation rates were determined by numeric modeling of time-dependent occupancy curves of the interchain or intrachain T domains. The intrachain acylation rate was significantly faster than the interchain rate and an equal flux of intermediates was detected for both alternative pathways (**Figure 28(c)**). This was the first approach to gain pseudokinetic data of competitive pathways in a dimeric NRPS system.



	Best fit parameters
$P_{\text{interchain}}$ (%)	52
$k_{\text{interchain}}$ (s^{-1})	3
$k_{\text{intrachain}}$ (s^{-1})	44

Figure 28 Investigation of interchain and intrachain acylation in dimeric VibF. (a) Vibrobactin. (b) VibF heterodimer for interchain and intrachain kinetic assay and ESI–FTMS analysis. (c) Obtained best-fit kinetic parameters for VibF intrachain and interchain acylation, $P_{\text{interchain}}$ – probability of interchain flux, $k_{\text{interchain}}$ – interchain acylation rate, $k_{\text{intrachain}}$ – intrachain acylation rate.

This section highlighted recent applications of modern MS to dissect biosynthesis on NRPS systems and NRPS–PKS interfaces by *in vitro* reconstitution. MS is increasingly applied for *in vitro* characterization of gatekeeping and promiscuity of NRPS components and a first *in vivo* NRP intermediate was detected by MS. In addition, MS is a valuable method to dissect single reaction mechanisms, for example, of tailoring reactions or to dissect multistage NRPS assembly line action. The investigation of NRPS time courses has been accomplished to date only in a pseudokinetic fashion. As mentioned above, several laboratories gained excess to FTMS instrumentation since its first application to an NRPS system in 1999 by Kelleher and coworkers.[41] This trend is shown here by the presented NRPS studies. Future diversification and a wider application of FTMS methods in an NRPS context can be expected. Additionally, several of the presented studies applied low-resolution MS, in particular MALDI–TOF MS. These studies do not include NRPS active site mapping, which is only possible with high-resolution MS.

9.11.2.8 Orphan Gene Cluster Characterization by Mass Spectrometry

As outlined recently,[47] the ability of ESI–FTMS to characterize NRPS substrates, intermediates, and tailoring reactions could be applied for elucidation of natural product chemistry from orphan NRPS gene clusters. Orphan gene clusters are gene clusters with unknown natural products. Various strategies have been developed to identify their secondary metabolites. The majority of these so-called genome mining approaches, which are reviewed elsewhere,[114] are aimed to isolate the unknown natural product. In contrast to these discovery strategies, the *in vitro* reconstitution approach is aimed to dissect the structure of the unknown natural product by characterization of the recombinant orphan biosynthetic enzymes. Herein, MS would complement recent *in vitro* reconstitution tools and could provide the main information in the structure elucidation process by reliable identification of substrates and chemical modifications. MS-based *in vitro* reconstitution of an orphan NRPS gene cluster could be pursued as follows and depicted as in **Figure 29**.

First, the functions of all orphan gene products are predicted by BLAST or other protein analysis tools. NRPS and tailoring enzymes are differentiated. In addition, the substrates of the A domains can be predicted from their '10 letter codes'[99,100] by bioinformatic tools such as NRPS Predictor[101] although this is not a prerequisite for the *in vitro* reconstitution approach but a routine in NRPS characterization.

Second, all A domains are expressed heterologously and screened for their native substrates by MS-based substrate screening.[91] The substrate screening assay depends on the A domain construct: If an AT didomain or larger construct is screened, active site mapping or PEA-based 'top down' analyses are required to identify the substrates. If the A domain loads onto a freestanding T domain, substrates can be characterized directly by the T domain mass shift without prior active site mapping. The T domain substrate should be the pathway-specific T domain corresponding to the A domain, biosynthetic substrates can be screened in a defined way, for example, a specific amino acid mixture, or in a undefined way, for example, algal lysate. Substrate identification from an A domain of an orphan gene cluster has been achieved on two A–T didomains, PksN and PksJ, from the *pksX* pathway from *Bacillus subtilis* by Dorrestein *et al.*[91] Problems in this step can be a difficult and laborious protein expression and A domain inactivity upon heterologous expression.

Third, based on the identified A domain substrates, NRP intermediates and a complete NRP scaffold are synthesized by solid-phase peptide synthesis.[115] These small peptides will serve as substrates to elucidate tailoring enzymes.

Fourth, the tailoring enzymes are characterized in their substrates, their chemical modifications on the NRP-scaffold, and their reaction mechanisms. The appropriate substrate of each tailoring enzyme has to be identified. The substrate can be a T domain-tethered substrate or intermediate, CoA-activated acid, or an analogue of the complete natural product. Subsequently, the chemical modification is dissected by *in vitro* tailoring reaction of the corresponding substrate or intermediate and FTMS detection of its mass shift. Additionally, chemical modification can be localized within the NRP by FT–MS^2-based structure elucidation. Herein, the tailoring reaction substrate and product are fragmented in the mass spectrometer separately and the NRP fragment with the chemical modification is identified by the corresponding mass shift.

Based on this MS dissection of the native substrates and chemical modifications, a related or even identical structure of the unknown natural product can be synthesized and tested in terms of bioactivity and physiological target. The result from the *in vitro* reconstitution should be confirmed by another genome mining

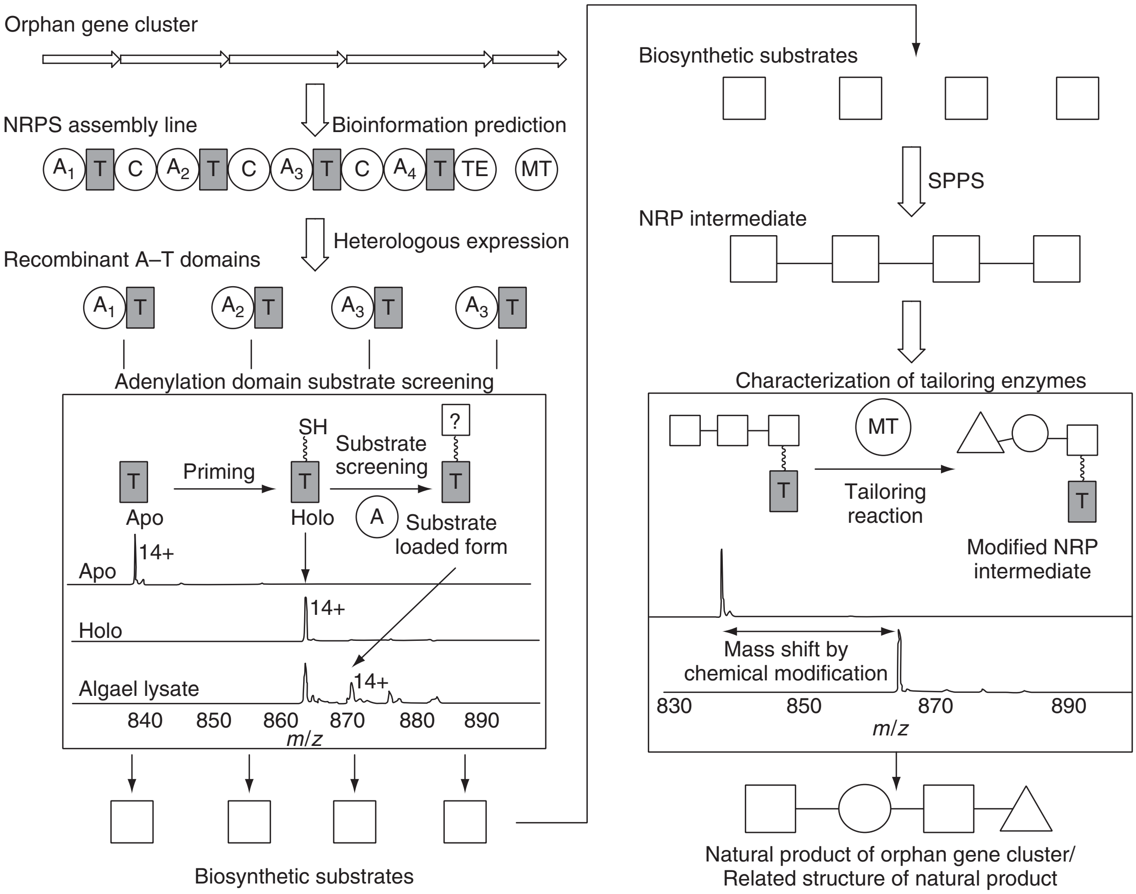


Figure 29 MS-based *in vitro* reconstitution of orphan NRPS gene clusters. Substrates and chemical modifications of unknown NRP natural products can be dissected by FT–ICR–MS methods such as substrate screening and PEA. Based on these biosynthetic informations a structure related to the actual natural product can be drawn.

approach aimed to discover the natural product, for example, genomeisotopic approach based on the identified substrates. The disadvantage of the MS-based *in vitro* reconstitution of orphan gene clusters is on the one hand the laborious task of heterologous protein expression and on the other hand the requirement of high-resolution MS, which is recommended because of its high mass accuracy and MS^n features. The advantage of the approach is that any NRPS gene cluster – silent or expressed – can be identified in its natural product chemistry. Advances in bioinformatic prediction tools, gene synthesis, and heterologous protein expression should lay a more rigid foundation for the role of MS as a major genome mining tool.

9.11.3 Applications of Mass Spectrometry on PKS Systems

Polyketides represent a source of numerous pharmacologically and commercially useful compounds.[116] The appeal of polyketides is that they are a structurally diverse class of compounds, yet these complex molecules can be synthesized from much smaller, simpler acyl-CoAs. Polyketides are synthesized by large multidomain megasynthase PKSs.[23,117] These megaenzymes efficiently carry out the addition of the acyl-CoAs to form elongated intermediates that undergo a variety of different enzymatic tailoring steps. PKSs are classified as type I, type II, or type III. Type I PKSs are a single protein consisting of a linear arrangement of the various catalytic and carrier domains. Type II PKSs consist of the various domains that exist as individual proteins that interact with each other. Type III PKSs function without the use of a T domain. In addition, polyketides can be classified as either being modular or iterative. Modular PKSs have multiple domains that function in an assembly-line fashion in which the substrate is bound to a carrier protein of the first domain, undergoes modifications, and then is transferred to a carrier protein within the next domain. The growing intermediate is passed from one catalytic domain to the next and undergoes elongation and additional modifications at each domain until a full-length intermediate is released. Iterative PKSs possess only a single module consisting of a few catalytic and carrier domains that are reused over and over during the elongation of the intermediate. The intermediate undergoes cycles of addition and modification. Type I PKSs and type II PKSs have been the subject of recent investigations and the fact that they possess the phosphopantetheinyl functionality on their T domains make them suitable targets for studies involving MS. Using high-resolution FTMS, the intermediates of PKS biosynthetic pathways can be detected while still covalently bound to active site of the T domain via PPant. Confirmation of the exact mass of PPant-bound intermediates can be confirmed by subsequent PEA.

9.11.3.1 Bacillaene Biosynthesis: *Bacillus subtilis* HMG-CoA/Trans-Enoylreductase and α-/β-Ketoreductase

9.11.3.1.1 Orphan gene cluster pksX

The *pksX* gene cluster of the *Bacillus subtilus* encodes a hybrid PKS–NRPS that produces a previously unknown secondary metabolite. This orphan gene cluster was found to share many similarities with gene clusters involved in the biosynthesis of curacin, jamacamide, pederin, as well as others. Of particular interest were several biosynthetic tailoring enzymes expressed such as a trans-acting AT (PksC). Also of interest were several freestanding proteins, such as a T domain (AcpK), a ketosynthase (PksF), a 3-hydroxy-3-methylglutaryl (HMG)-CoA synthase (PksG), two enoyl-CoA hydratases (PksH and PksI), and a larger protein containing multiple T domains (PksL, **Figure 30(a)**). In addition to various biochemical techniques, high-resolution MS was applied to designate functional roles to these various proteins encoded by the *pksX* gene cluster. The genes for PksC, AcpK, PksF, PksG, PksH, PksI, and the region of the gene encoding the pair of tandem T domains PksL were heterologously expressed in *E. coli* and then purified. The AcpK protein and the tandem T domains (PksL-T_2) were phosphopantetheinylated *in vivo* by coexpressed Sfp.[57]

Substrate loading onto AcpK and subsequent alterations to the PPant-tethered intermediate that were hypothesized to be carried out by the array of proteins encoded by *pksX* were determined by FTMS. PksC, when incubated with holo AcpK and malonyl-CoA, resulted in a +86 Da shift of the AcpK protein, corresponding to an 8.6 m/z shift of the 10+ of AcpK ion. This mass difference was consistent with the formation of malonyl-*S*-AcpK. This observation, coupled with detection of a +86 Da mass shift of PksC itself when incubated with malonyl-CoA, confirmed the malonyl-AT function of PksC. The function of PksF was determined by incubating the protein with malonyl-*S*-AcpK. The resulting 44 Da loss in mass, representing decarboxylation of malonyl-*S*-AcpK to acetoacetyl-*S*-AcpK, was detected with FTMS and helped verify the function of PksF as that of a ketosynthase. To probe the possible function of PksG as that of an HMG-CoA synthase (HCS), acetoacetyl-*S*-PksL-T2 was incubated with acetyl-*S*-AcpK and PksG. The addition of PksG facilitated the formation HMG-*S*-PksL-T_2, at one or both tandem T domain active sites, observed as a +60 Da mass shift by FTMS. IRMPD resulted in a PPant ejection ion with a mass of 503.152 Da, 60.022 Da larger than the mass of acetoacetyl (Acac)-loaded PPant (**Figure 30(b)**). The functional characterization of PksH and PksI activity was again facilitated by high-resolution MS and confirmed by PPant ejection analysis. HMG-*S*-PksL-T_2 was incubated first with PksH alone, then PksI alone. On its own, PksH did not yield any change in the PPant-tethered intermediate. Incubation of HMG-*S*-PksL-T_2 with PksI, on the other hand, resulted in a loss of 18 Da representing dehydration of the intermediate. Incubation of HMG-*S*-PksL-T_2 with both PksH and PksI resulted in a mass shift of 62 Da. This corresponds with the dehydration (−18 Da) and subsequent decarboxylation (−44 Da) of the HMG intermediate. These findings were further validated by observation of mass shifts in the PPant ejection ions generated by IRMPD of the various PksL-T_2-bound intermediates (**Figure 30(c)**). These studies highlighted the utility of FTMS and the PEA as means to elucidate the functions of the various products of an orphan PKS gene cluster such as *pksX* by *in vitro* reconstitution.

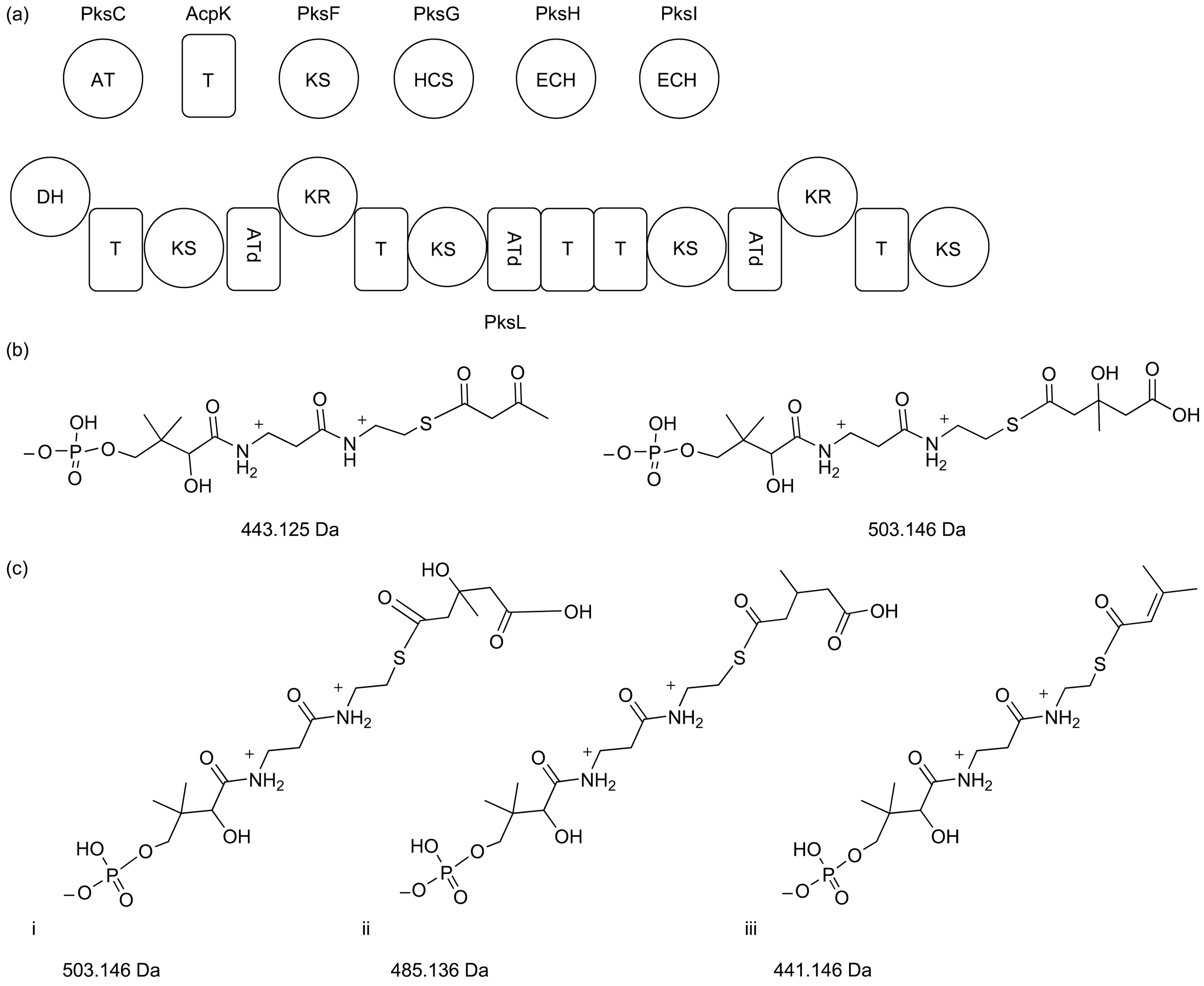


Figure 30 Bacillaene biosynthesis. (a) Annotation of the function of multiple proteins encoded by the *pksX* gene. (b) Determination of PksG (HCS) function. Acetoacetyl (Acac)-*S*-AcpK phosphopanthetheine ejection ion was observed to have a mass of 443.125 Da. Incubation of Acac-*S*-AcpK with PksG results in an increase of 60.02 Da, consistent with formation of HMG-*S*-AcpK and confirmation of HCS activity by PksG. (c) Functional determination of PksH and PksI. HMG-*S*-PksL-T_2 was incubated with PksI, and PksI and PksH. Phosphopantetheinyl ejection ions were generated using IRMPD. (1) The ejection ion from HMG-*S*-PksL-T_2 has a mass of 503.146 Da. (2) Incubation with PksI results in a loss of 18 Da consistent with dehydration. (3) Incubation with PksI and PksH results in a loss of 62 Da, corresponding to dehydration followed by decarboxylation. Incubation of HMG-*S*-PksL-T_2 did not result in any mass changes.

9.11.3.1.2 Trans-enoylreductase and α- and β-ketone reduction

More recently, the ultimate biosynthetic product of the *pksX* gene cluster was identified and its structure was elucidated.[118] This product was identified as dihydrobacillaene, which is later converted to bacillaene (**Figure 31(a)**) by PksS.[119] Dihydrobacillaene is produced by numerous enzymes and tailoring domains discussed previously, as well as an additional four megasynthase complexes. The first of these megasynthases, PksJ, contains two NRPS modules followed by two PKS modules. Recent investigation into the dihydrobacillaene biosynthetic pathway involved the determination of the origin of the α-hydroxyacyl N-cap.[58] While the first two modules of PksJ seemed capable of accepting α-hydroxyisocaproate (α-HIC) directly, before being transferred to the third PksJ module, researchers found that PksJ preferentially loads α-ketoisocaproate (α-KIC) first, then transfers the intermediate to the third module. Module 3 of PksJ possesses a pair of tandem T domains, which receive the α-KIC-containing intermediate. The presence of a KR domain and the tandem T domains within the third PksJ module led to the hypothesis that the α-KIC-containing intermediate is reduced to α-HIC within the third module by the single KR domain acting on both the α-keto and β-keto groups of the α-KIC intermediate (**Figure 31(b)**).

To test the functions of the KR domain, investigators heterologously overexpressed the region of PksJ containing the KR domain and the tandem T domains, PksJ(KR-T_3-T_4). To assess the typical function of the PksJ-KR, as a β-ketoreductase, apo PksJ(KR-T_3-T_4) was incubated with Acac-CoA and Sfp in order to generate Acac-*S*-PksJ(KR-T_3-T_4). Reduction was carried out using either nicotinamide adenine dinucleotide hydride (NADH) or nicotinamide adenine dinucleotide phosphate hydride (NADPH), followed by ArgC digestion. The T domain active site fragments incubated with NAD(P)H were compared to controls. A small shift in mass of the active site was detected in samples treated with NADH/NADPH, and the mass shift of +2.0 166 Da was verified by carrying out source-induced dissociation for PPant ejection (**Figure 31(c)**). This mass shift is consistent with β-ketoreduction of Acac-*S*-PksJ(KR-T_3-T_4) to β-hydroxybutyrl-*S*-PksJ(T_3T_4), thus confirming the function of the PksJ module 3-KR as a β-ketoreductase. While a β-ketoreduction represents the rule for KR function, α-ketoreduction represents the exception. In order to test the ability of PksJ module 3-KR to reduce the distant α-ketone of α-KIC, researchers utilized α-ketoisocaproyl-γ-aminobutyrate (α-KIC-GABA) as a model substrate. Reduction of α-KIC-GABA-*S*-PksJ(KR-T_3-T_4) occurred in an NAD(P)H-dependent manner and was detected by protease digestion followed by FTMS and PPant ejection analysis (**Figure 31(c)**.

As noted previously, there were several trans-acting elements of the dihydrobacillaene biosynthetic pathway. The exact function of one such element from the *pksX* gene cluster, PksE, was analyzed by FTMS.[49] PksE was proposed to act as enoyl reductase (ER), reducing the C14′–C15′ bond during dihydrobacillaene biosynthesis by PksJ. To test this activity, 2-butenoyl-*S*-Pks(T_3T_4) was generated, incubated with NAD(P)H, digested, and analyzed by LC–FTMS. An NAD(P)H-dependent reduction of the alkene bond was detected by a +2 Da shift of the active-site-containing peptide of 2-butenoyl-*S*-Pks(T_3T_4) and could be confirmed by PPant ejection. The *pksX* gene cluster possesses many atypical features. Initially, FTMS proved to be useful in determining the function of several major components of the dihydrobacillaene biosynthetic pathway. More recent research has highlighted the usefulness of FTMS and the PEA as a means of exploring noncanonical features of complex NRPS/PKS systems, in particular the function of ketoreductases and enoylreductases that impart small (2 Da) changes in the intermediate. The PEA is a reliable method for verifying these slight changes.

9.11.3.2 Curacin A Biosynthesis: ECH1 and ECH2

The application of high-resolution MS has allowed characterization of enzymes involved in the biosynthesis of curacin A, a mixed polyketide–NRP produced by *Lyngbya majuscule*, which possesses cytotoxic properties.[120] Researchers were intrigued by the unusual structure of this molecule (**Figure 32(a)**) and during their investigations of the enzymes involved in curacin A biosynthesis, an HCS-like gene cassette was identified, similar to that of the *pksX* gene cluster in *B. subtilis*, as well as others. The HCS-like cassette, involved in curacin A biosynthesis, encodes five separate enzymes that include a T domain, a ketosynthase, an HCS, and two separate enoyl-CoA synthases.[121] This set of enzymes was hypothesized to be responsible for the formation of

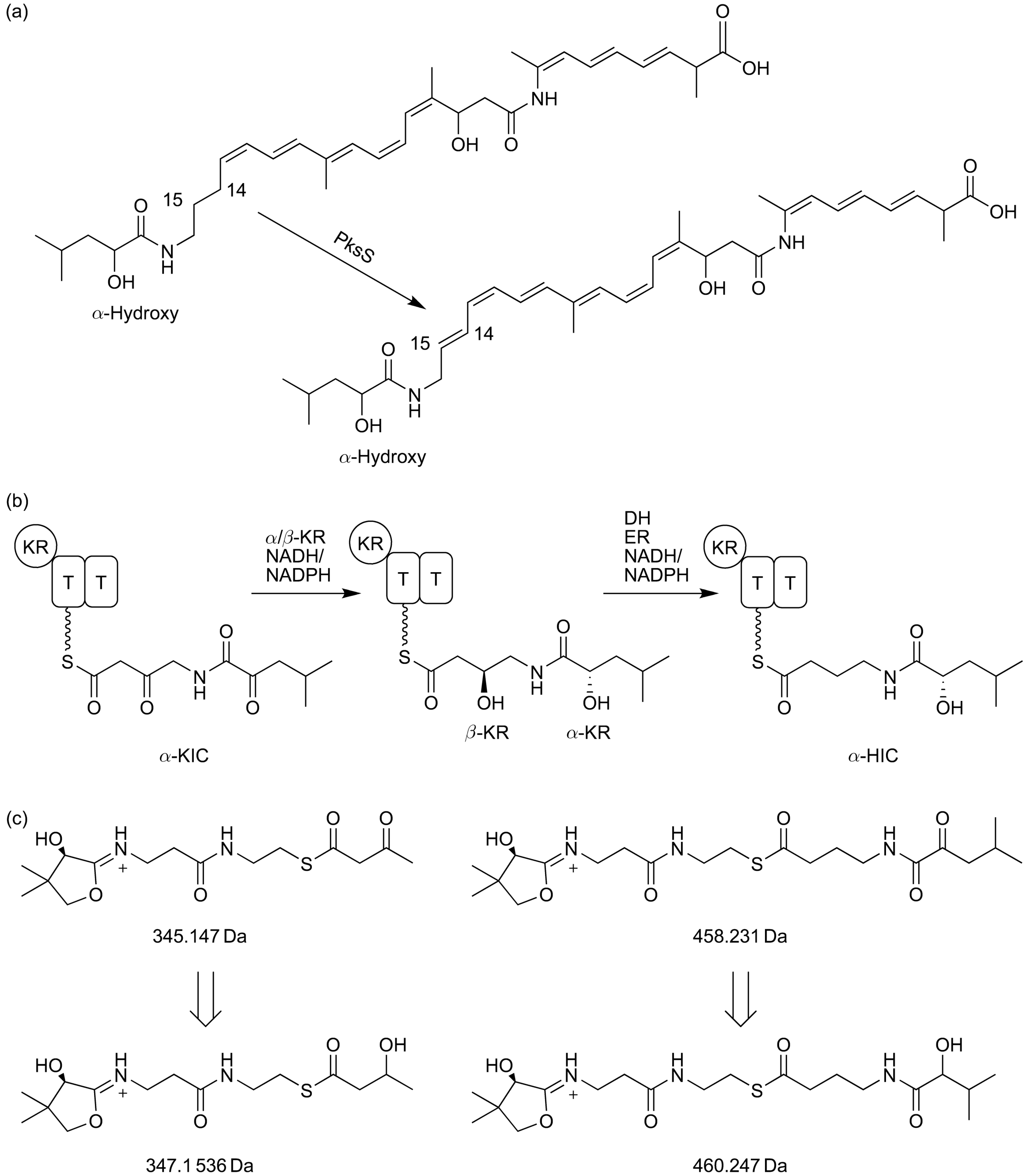


Figure 31 Bacillaene biosynthesis. (a) Structure of dihydrobacillaene and bacillaene. PksJ oxidizes the C14′–C15′ bond after dihydrobacillaene has been synthesized. Also, note the α-hydroxyacyl N-cap. This particular N-capping has been reported very rarely. (b) α- and β-Ketoreduction of α-KIC to α-HIC. The KR domain of the first PKS module in PksJ is capable of reducing both the α-KIC amide and the β-ketone in an NAD(P)H-dependent fashion. The order in which these two reductions occur is unknown. Ultimately, keto-reduction is followed by dehydration and enoyl reduction. (c) Theoretical structure of PPant ejection ions used to analyze PksJ ketoreduction. Right: PPant ejection ion resulting from IRMPD of Acac-*S*-PksJ(T_3-T_4) incubated with PksJ. Mass shift of +2.017 Da corresponds with reduction of the β-ketone. Left: PPant ejection ion resulting from IRMPD of α-KIC-GABA-*S*-PksJ(T_3-T_4) incubated with PksJ. Shift of +2.015 Da is observed in PPant ejection ions.

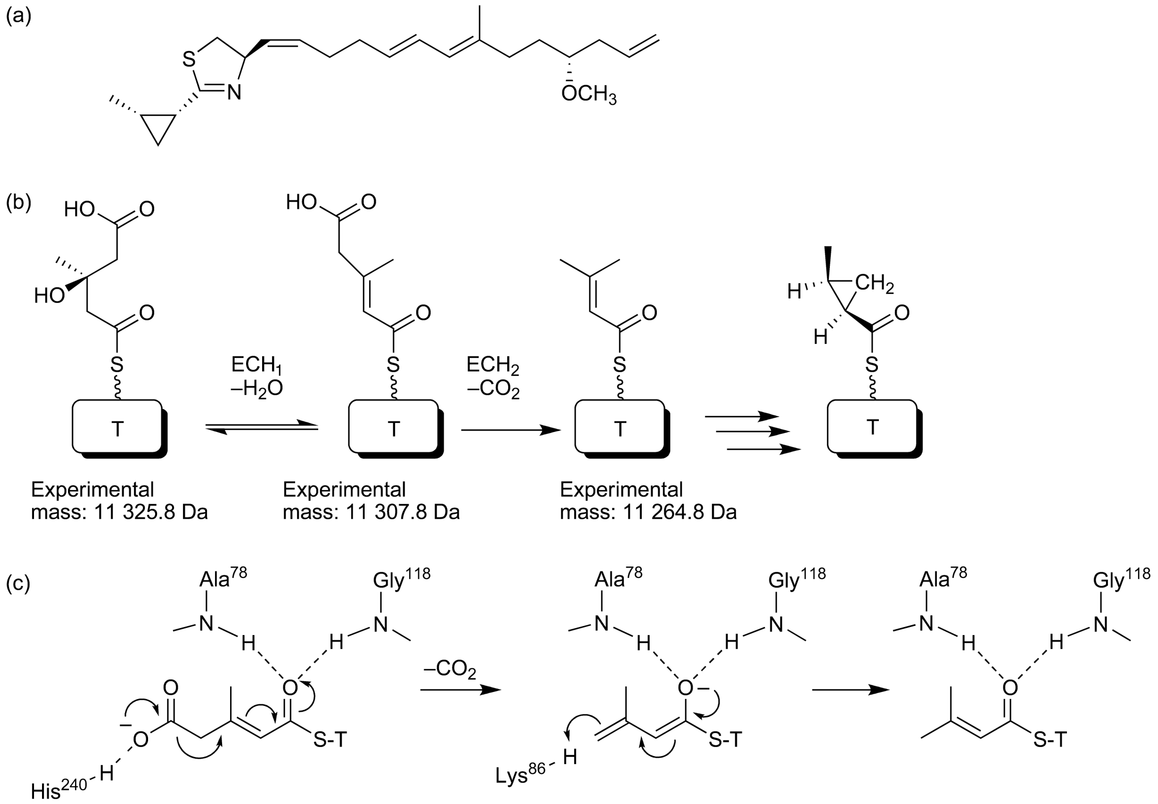

(a)
S
N
OCH3
(b)
HO
O
HO
O
S
T
ECH1
–H2O
ECH2
–CO2
CH2
H
Experimental mass: 11 325.8 Da
Experimental mass: 11 307.8 Da
Experimental mass: 11 264.8 Da
(c)
Ala78
Gly118
N
H
O
–CO2
S-T
His240
Lys86

Figure 32 Curacin biosynthesis. (a) Curacin A structure. (b) T domain bound intermediates involved in cyclopropyl ring formation. 3-Hydroxy-3-methylglutaryl (HMG)-*S*-T undergoes dehydration catalyzed by ECH1 to produce 3-methylglutaconyl-*S*-T. ECH2 catalyzes the subsequent decarboxylation to yield 3-methylcrotonyl-*S*-T. (c) Proposed mechanism of decarboxylation of 3-methylglutaconyl-*S*-T by ECH2. The His240 residue of ECH2 acts to position the substrate and prime its decarboxylation. Lys86 donates a proton to the enolate anion. Point mutations of these two residues substantially diminished the production of 3-methylcrotonyl-*S*-T, as detected by FT–ICR–MS.

the cyclopropanyl ring of curacin A. In particular, the functions of the two enoyl-CoA synthases ECH1 (CurE) and ECH2 (CurF) were determined using high-resolution MS.

As members of the functionally diverse crotonase superfamily, ECH1 and ECH2 were expected to have different roles in the formation of the cyclopropanyl ring precursor of curacin A.[122] To probe the function of ECH1 and ECH2, the T domain (CurB) was first incubated with (*S*)-HMG-CoA, in order to covalently attach HMG to the PPant arm of the T domain active site. Next, the individual activities of each enzyme were determined by incubating ECH1, ECH2, or both enzymes with HMG-*S*-T. ESI–FTMS was used to analyze the products of these reactions by detecting the mass differences observed in the various reactions compared to HMG-*S*-T alone. The 12+ ion was used to detect these mass differences. The most abundant mass of HMG-*S*-T was determined to be 11325.8 Da. Incubation with ECH2 alone did not result in any new products. However, incubation of HMG-*S*-T with ECH1 resulted in the detection of a new product with a mass of 11307.8 Da, and incubation of HMG-*S*-T with ECH1 and ECH2 yielded two products with masses of 11 307.8 and 11 264.8 Da. These differences observed by FTMS corresponded to losses of 18 and 62 Da. This provided evidence that ECH1 functions to dehydrate HMG-*S*-T (–18 Da) to form 3-methylglutaconyl-*S*-T. The 3-methylglutaconyl-*S*-T undergoes subsequent decarboxylation, catalyzed by ECH2, to form 3-methylcrotonyl-*S*-T (**Figure 32(b)**). After identifying the exact functions of ECH1 and ECH2, researchers determined the crystal structure of the N-terminal domain ECH2 (CurF).[123] Structural alignments of CurF ECH2 with other members of the crotonase superfamily revealed several key features of the enzyme active site. Crystallization of ECH2 complexed with product analogues was not successful, so computational modeling was used to identify three polar side chains within the active site chamber that possessed potential catalytic function: Tyr82, Lys86, and His240. The previously established ECH1/ECH2 enzymatic assay was carried out using ECH2 mutants containing Y82F, K86A, K86Q, H240A, and H240Q. The wild-type and mutant ECH2 enzymes were incubated with ECH1 and (*S*)-HMG-*S*-T (CurA-S-T(II)), then the different incubated reaction mixtures were run on a C4 column and eluted with acetonitrile. The samples were analyzed by ESI–FTMS after being redissolved in an electrospray solution (55% acetonitrile:45% water, with 0.05% formic acid and 0.05% TFA). FTMS was used to detect the presence of 12+ charged (*S*)-HMG-*S*-T, 3-methylglutaconyl-*S*-T, and 3-methylcrotonyl-*S*-T as a means of confirming the effect of the site-directed mutagenesis of the various residues. Using this ECH1/ECH2 assay and FTMS, it was determined that substitution of Tyr82 resulted in only minimal reduction in the production of 3-methylcrotonyl-*S*-ACP from HMG-*S*-T, while substitutions of Lys86 and His240 resulted in drastic decreases in product formation. The identification of essential active site residues allowed researchers to propose a mechanism of action for ECH2 (CurF, **Figure 32(c)**). The work carried out with the curacin A biosynthetic pathway highlights the importance of FT–ICR–MS in both the determination of intermediates in NRPS/PKS pathways as well as evaluation of PKS domain functions.

9.11.3.3 Enediyne Biosynthesis: SgcE

MS has proven to be essential in investigations of the biosynthetic pathways of C-1027, an enediyne antitumor antibiotic. C-1027 is an extremely cytotoxic compound, isolated from *Streptomyces globisporus*, consisting of a binding protein (CagA) and a reactive nine-membered enediyne core containing a pair of conjugated acetylenic groups (**Figure 33(a)**). The enediyne core is the key to the cytotoxicity of C-1027 because when it is released from the protein complex it can form a transient biradical species that can induce single-strand and double-strand breaks in DNA molecules.[124] Characterization of the biosynthetic gene cluster revealed the presence of a type I iterative PKS (SgcE) that catalyzes the formation of the nine-membered enediyene core from acyl-CoAs. A unique feature of this particular PKS is the presence of an integrated C-terminal PPTase that covalently attaches 4′-phosphopantetheine to the active site serine of the T domain. Typically, PPTases are freestanding components in PKS biosynthetic pathways. *In vivo* experiments involving the inactivation of proposed SgcE active sites were carried out, resulting in loss of C-1027 production by SgcE. In order to confirm the function of both the T domain and the PPTase domain, high-resolution MS was employed.[54] SgcE was digested briefly with trypsin, HPLC purified, and the fragments were analyzed by FTMS. A 3+ charge peptide fragment containing the T domain active site was determined to contain a mass shift of 340.1 Da, corresponding to the addition of PPant. By using CID to fragment this ion, the PPant modification was localized to the serine residue at position 974. The same 3+ T domain active site peptide was subjected to IRMPD and the characteristic PPant ejection ion was observed, along with the dehydroalanine form and the phosphoserine form of the peptide. The PEA was used to verify the role of the various active sites in the phosphopantetheinylation of the T domain of SgcE. Mutational inactivation of the active site of the T domain and the inactivation either of the two key residues of the PPTase domain resulted in loss of phosphopantetheinylation of the T domain, which was confirmed by the absence of the 3+ T domain active site peptide fragment with the 340.1 Da mass shift. This provided clear evidence that the C-terminal domain does in fact serve the role of a PPTase.

The information gathered during the analysis of the domain functions of SgcE can be used in a hypothesis-driven approach for mapping higher charge states of the active site peptides and to identify substrate- and intermediate-loaded forms of the T domain active site. During the investigations into the function of the PPTase domain of SgcE, the tryptic peptide containing the phosphpantetheinylated active site of the T domain was identified and found to have a mass of 4134.0 Da. Recent work was carried out in the Dorrestein laboratory to detect substrate loading onto the SgcE T domain active site. Armed with the knowledge that the active site containing peptide has a mass of 4134 Da and a 3+ charge, calculations were carried out to predetermine the expected *m*/*z* values at which malonyl, acetoacyl, and acetyl loaded forms of the active site peptide would be found. Preliminary substrate loading assays were performed using SgcE. After a 1 h incubation with malonyl-CoA, SgcE was digested for 10 min using trypsin, and then analyzed by capillary–LC–FTMS. MS data were collected using FTMS at a resolution of 50k, and the top 10 most abundant peptides in each scan were fragmented by CID in a data-dependent fashion. This method was ideal because the mass spectra could quickly be searched for the hypothetical substrate-loaded peptides and MS/MS or MS^2 data were already collected to

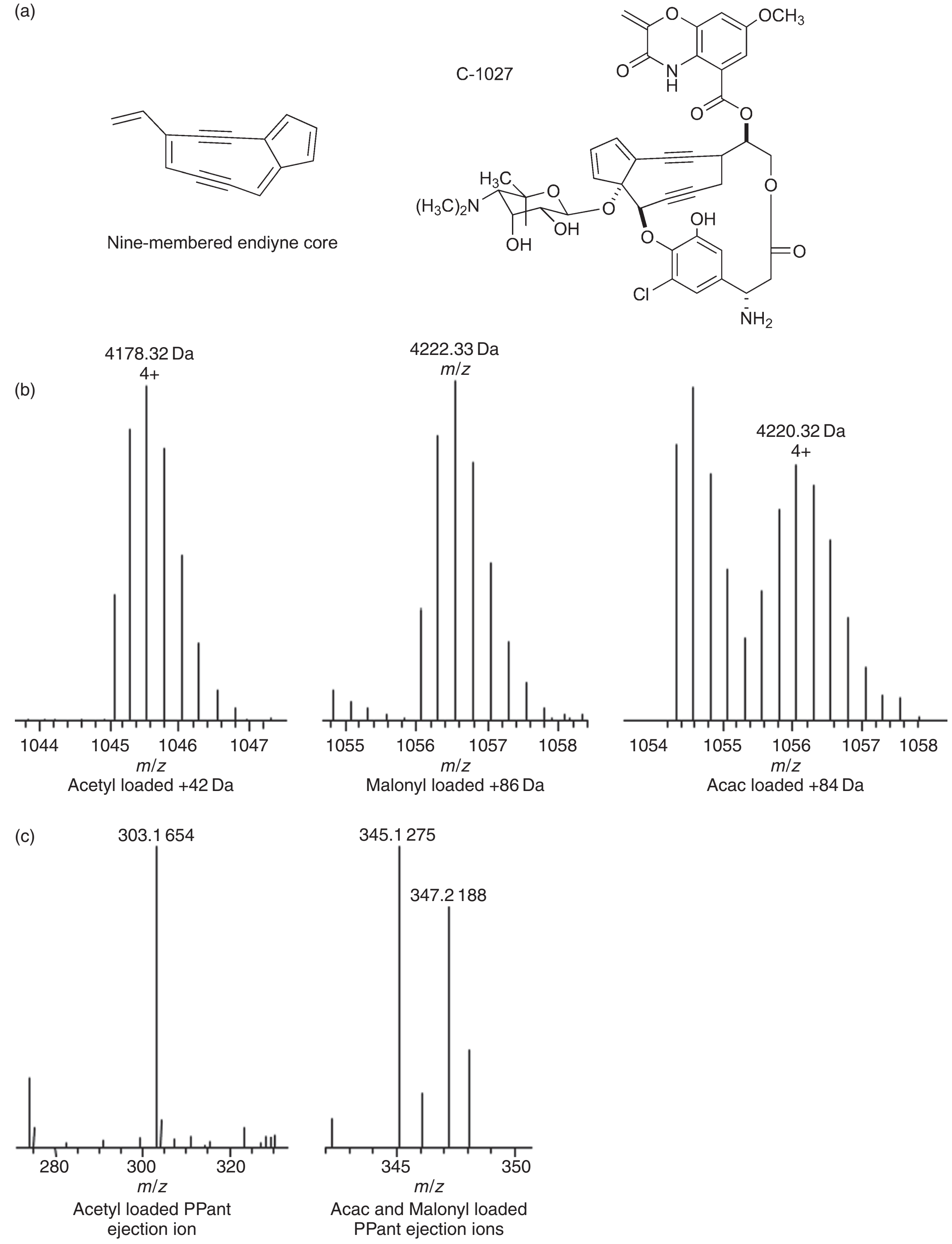


Figure 33 SgcE analysis. (a) The structure of the nine-membered endiyne core (left) produced by SgcE and the complete structure of the C-1027 molecule (right). (b) Active site containing tryptic peptide signals with detected mass shifts corresponding to the loaded substrate/intermediate: acetyl (left), malonyl (middle), and acetoacetyl (left). (c) Phosphopantetheinyl ejection ions. Ejection ion confirms the mass of the bound intermediate. The malonyl intermediate undergoes decarboxylation resulting in the formation of acetyl-(*S*)-T. Malonyl and acectoacetyl-ACP active site fragments coelute during LC–MS. The result is that the *m/z* of the two species are close enough that they are both fragmented and their resulting ejection ions can be detected in the same MS/MS or MS^2 spectrum.

confirm the mass of the loaded substrate by identifying the PPant ejection ion. Using this method, the 4+ phosphopantetheinylated T domain active site peptides were detected having mass shifts of +42 Da (acetyl), +86 Da (malonyl), and +84 Da (acetoacetyl, **Figure 33(b)**). The active site-containing peptide loaded with malonyl has only a 2 Da difference in mass from the active site peptide loaded with acetoacetate, and they partially coelute. While FTMS is capable of resolving the signals of the malonyl and acetoacyl loaded peptides, the PPant ejection ions were clearly detected by MS/MS or MS^2 as shown in **Figure 33(c)**.

Given that LC–FTMS followed by data-dependent fragmentation is capable of detecting the loaded PPant ejection ions, this method can be quite useful for detecting PPant ejection signatures in other PKS systems as well. However, in systems in which phophopantetheinylated active site containing peptides are unknown, or during proteome-level screening, subsequent data analysis of LC–MS data in search of PPant ejection ions can be extremely challenging due to the overwhelming amount of MS/MS or MS^2 data and the difficulty that the preexisting MS/MS or MS^2 analysis programs have with multiple charged precursor peptides. Currently, collaborative work is being carried between the Dorrestein laboratory and the Bafna research group in the computer science department of UCSD to develop a program capable of rapidly scanning LC–FT–MS/MS or MS^2 data for the presence of PPant ejection ions (loaded or not), the corresponding dehydroalanine (−18 Da, $z-1$) form and the phosphoserine (+80, $z-1$) form of the active site peptide in the MS/MS or MS^2 data. This software has already successfully identified PPant ejection signatures in data collected from a number of PKS and NRPS systems. As development of this software progresses, searching LC–FTMS data for PPant ejection will be dramatically increased, and this possesses the potential for *de novo* detection of PPant ejection in highly complex samples that are being analyzed by LC–MS in proteomics screening experiments.

9.11.3.4 Deconstructive Analysis of PksA

Recent efforts have been made to better understand eukaryotic iterative polyketide synthases (IPKSs) and how these biosynthetic pathways ensure the formation of specific products. As opposed to much larger modular PKSs that carry out the biosynthesis and tailoring of their products in an assembly-line fashion, iterative PKSs contain many fewer catalytic domains and the same domains are used multiple times prior to yielding the final product. It has been unclear what exactly determines the number of catalytic cycles the growing intermediate of an IPKS undergoes before it is released. To clarify the 'global division of labor' between the domains of an IPKS, researchers studied PksA by dissecting out the individual domains and reassembling them *in vitro.*[48] By expressing the various domains of the PksA as smaller units and recombining them in different combinations, the hope was to paint a better picture of the function of each domain in the biosynthesis of aflatoxin B_1. Various mono-, di-, and tridomains were expressed. The various domains of PksA were then mixed back with the PksA starter-CoA:acyl carrier protein acyltransferase (SAT)-KS-malonyl-CoA:acyl carrier protein acyltransferase (MAT) domain, and the different products of the reactions between the substrate, SAT-KS-MAT, and the other domains were analyzed. In theory, the full complement of the 6 PksA domains (**Figure 34(a)**) would be required in order to efficiently produce norsolorinic acid, an isolable precursor of aflatoxin B_1, from a starter hexanoyl-CoA and seven malonyl-CoAs.

The products formed by the various combinations of the deconstructed PksA domains, incubated with substrate, were monitored by HPLC. It was found that the combination of SAT-KS-MAT with the putative product template (PT) domain and the T domain yielded significant quantities of a product, which was not norsolorinic acid. Addition of the thioesterase/Claisen-like cyclase (TE/CLC) to the mixture resulted in substantial formation of norsolorinic acid. These findings revealed that product formation was not occurring to an appreciable extent in the absence of the PT domain. Therefore, it was hypothesized that the PT domain acts as a cyclase/aromatase and catalyzes the formation of the first two rings of a putative intermediate. This putative intermediate can then undergo spontaneously C–O ring closure in the absence of the TE/CLC domain to form the shunt product napthopyrone, or it can undergo C–C ring closure in the presence of the TE/CLC domain to from the norsolorinic acid anthrone precursor (**Figure 34(b)**).

In order to verify the proposed structures of the phosphopantetheinyl-bound intermediates, an LC–FTMS assay was employed to detect PPant ejection ions containing bound intermediate. Reactions with SAT-KS-MAT and either the T domain or a PT-T didomain were carried out with both hexanoyl-CoA and malonyl-CoA. The reactions were carried out for varying times, followed by limited trypsin digestion (15 min). The

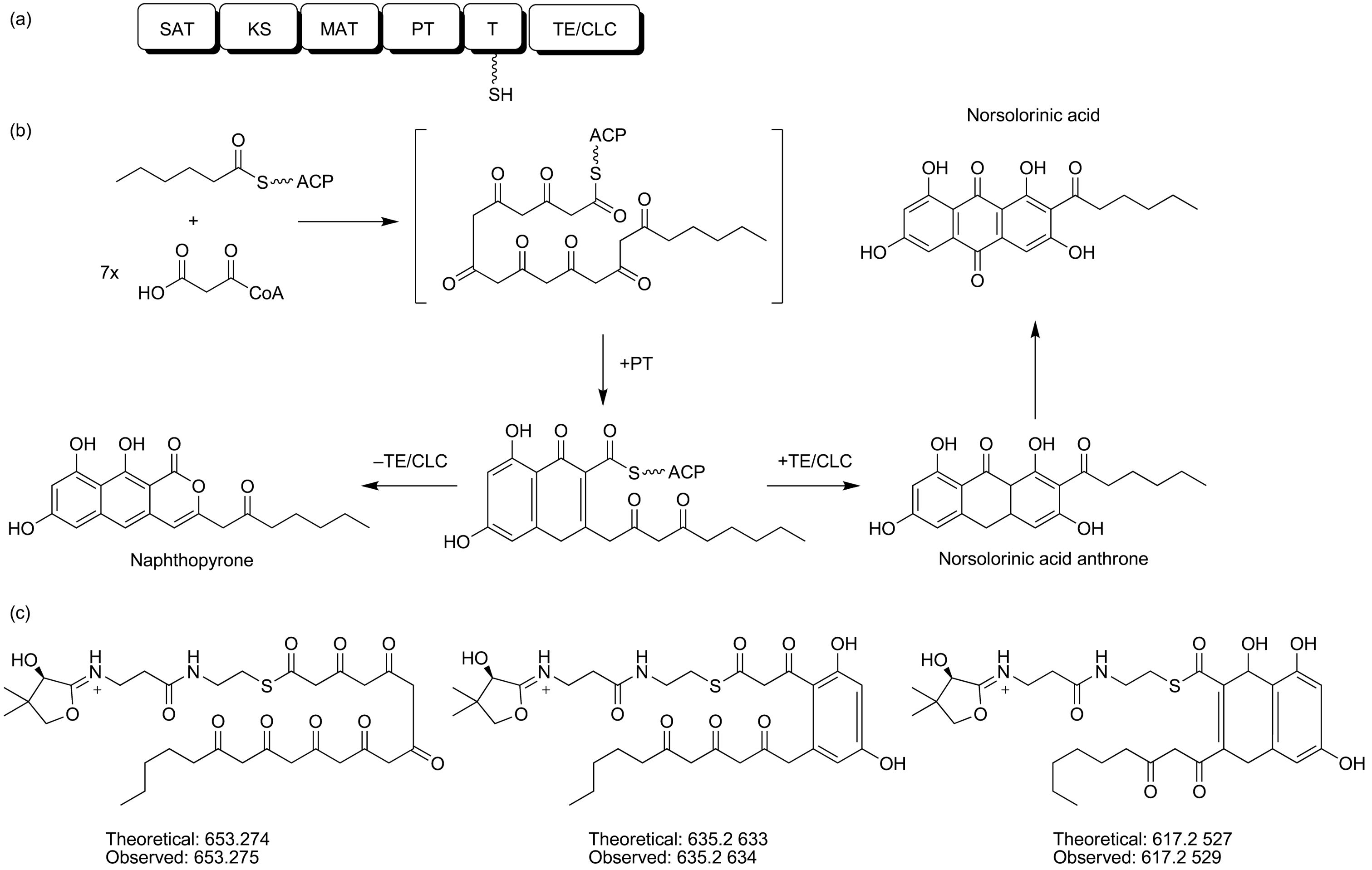


Figure 34 PksA deconstruction. (a) Enzymatic domain architecture of PksA. (b) PksA utilizes a starter hexanoyl-CoA and seven malonyl-CoAs to produce the covalently linked intermediate (brackets). The PT domain acts as an aromatase/cyclase facilitating the closure of the first two rings on the intermediate. In the absence of the TE/CLC domain the intermediate undergoes C–O cyclization to spontaneously form the naphthopyrone. In the presence of the TE/CLC domain, the intermediate undergoes C–C cyclization to from the norsolorinic acid anthrone, which autooxidizes to form norsolorinic acid. (c) Observed PPant ejection ions confirming the structures of the proposed intermediates bound to the active site of the PksA T domain. The first intermediate (left) was detected on the T domain active site after incubation of SAT-KS-MAT with T domain alone. Incubation of SAT-KS-MAT with PT-T results in the formation of the intermediates containing first a single-cyclization product (middle) followed by a double-cyclization product (right).

digested reactions were run separately on a C4 column over a 1 h water/acetonitrile gradient and injected directly into an FTMS. The phosphopantetheinylated active site peptides possessing various intermediates were observed and the structures were confirmed by MS/MS or MS^2 and PPant ejection. The fully extended intermediate generated by a single starter hexanoyl unit and seven malonyl units was detected bound to the T domain active site. In addition to this PPant-bound intermediate, the intermediate was detected as a single dehydrated compound with a signal aromatic ring or it was detected as a double dehydrated compound with two closed rings. The MS/MS or MS^2 PPant ejection ions reveal mass shifts with less than 1.5 ppm mass accuracy. These ejection ions were subjected to an additional round of fragmentation and the resulting ions were mapped to the structure of the intermediate (**Figure 34(c)**). The T domain was observed to accept both hexanoyl-CoA and malonyl-CoA, but PksA preferentially loads hexanoyl-CoA as the starter unit. The MS/MS or MS^2 PPant ejection ions reveal mass shifts with less than 1.5 ppm mass accuracy. The high mass accuracy of the FTMS allowed for the accurate determination of the mass of PPant-bound intermediates, even for relatively large intermediates such as the full extended octa-ketide intermediate of norsolorinic acid.

9.11.4 Prospective Applications of Current Natural Product MS Methods on NRPS and PKS Systems

Although more high-resolution mass spectrometrists are entering the field, there are currently just a handful of investigators that are developing novel approaches to investigate biosynthetic pathways of secondary metabolites. McLafferty has certainly been one of the early driving forces in the 1990s in this area but since that time others, including Kelleher, Hakansson, and Leary, began to use high-resolution MS creatively to investigate biosynthetic pathways. Some of these creative approaches are highlighted in this section.

9.11.4.1 Isotopically Depleted Proteins and Peptides

There are several factors that add to the complexity of a mass spectrum and that thus make identification of a specific protein difficult. For instance, it is not uncommon that a protein forms phosphate, sodium, or potassium adducts and that it exists in multiple oxidized forms. In addition, other protein or peptides may be ionized at the same time. These various forms of a specific protein and contaminants entering the MS instrument can make a single mass spectrum very complex. This complexity may make it tremendously difficult to identify the ions of interest with a high degree of certainty. For example, the biosynthesis of lacticin 481 by the protein LctM from the ribosomally encoded peptide LtcA requires multiple phosphorylations and dehydrations followed by thioether bridging (**Figure 35(a)**). It was not possible to discern the details of the mechanism observed for lacticin 481 biosynthesis when LtcA contained normal abundance isotopic levels because it was prone to oxidation during this process. Therefore LtcA was grown in isotopically ^{13}C- and ^{15}N-depleted material and the maturation of LtcA by LtcM was analyzed by FT–ICR–MS. **Figure 35(b)** shows the time course of the conversion of LtcA to lacticidin 481. Very little buildup of biosynthetic intermediates is observed, which characterizes LtcM as a processive enzyme. The abovementioned example shows how useful ^{13}C- and ^{15}N-depleted materials can be if the spectrum is too complex to reliably interpret the data. The depleted protein strategy is also useful when the protein is of reasonable size and yet one is interested in observing small mass changes such as dehydrogenation, deamidation, or transfer of an isotopically labeled substrate to the recipient protein. One such scenario has been found in the thiazole phosphate biosynthetic pathway in prokaryotes.

One of the pathways to thiazole phosphate, the active portion of thiamin utilizes a ubiquitin-like protein called ThiS, an E1-type protein called ThiF, a flavin-dependent oxidase ThiO, the thiazole synthase ThiG, a sulfur source, ATP, 1-deoxy-D-xylulose-5-phosphate, and glycine.[125] This pathway is initiated by the adenylation of ThiS by ThiF (**Figure 36**). Once the acyl–adenylate is formed, a sulfur, presumably from a cysteine desulferase-mediated reaction, displaces the adenylate to form a C-terminal ThiS-thiocarboxylate. Subsequently, it reacts with deoxy-D-xylulose-phosphate (DXP) that is transiently but covalently tethered to the thiazole synthase ThiG **4**. The sulfur from ThiS is transferred to the C3 carbon of DXP **5**. During this process, one of the hydroxyl groups from DXP is transferred to the C-terminal end of ThiS forming a stable

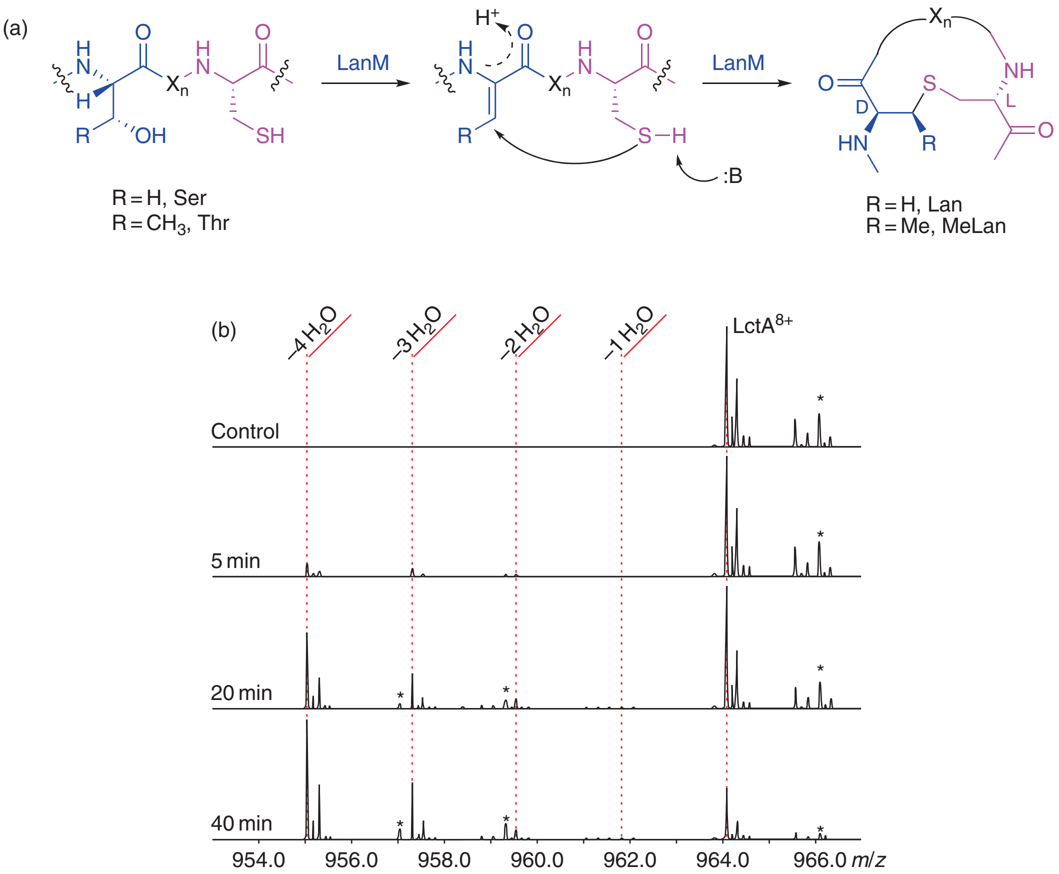


Figure 35 Conversion of LctA to the mature form of lacticidin 481 catalyzed by LtcM. (a) Thioether bridge formation in lantibiotic biosynthesis leading to dehydration (−18 Da). (b) MS time course of isotopically depleted LctA conversion to lacticidin 481 by LtcM shows that little dehydration intermediates are observed.

thioenolate intermediate on ThiG **6**. Once ThiO oxidizes glycine **15** and transfers the corresponding oxidized glycine **16**, it cyclizes and forms a stable carboxy-thiazolephosphate species **11** that is decarboxylated when it is coupled to the pyrimidine portion of thiamin.[126] This pathway is arguably one of the most complex biosynthetic pathways, involving several protein substrate interactions described in the literature to date.

^{13}C- and ^{15}N-depleted ThiS was necessary to visualize oxygen transfer shown as steps 5–8 in Figure 36. Oxygen transfer from DXP to the C-terminal end of ThiS was monitored by the incorporation of ^{18}O from ^{18}O-labeled DXP. Since ^{18}O-DXP was generated enzymatically, it was only partially ^{18}O-labeled. Therefore the incorporation was limited. When nondepleted material was used and the conversion of ThiS-thiocarboxylate to ThiS was monitored, a small shift was observed in the spectral profile but it could not be conclusively demonstrated that ^{18}O had indeed been incorporated. There were three reasons for this: (1) there was only partial incorporation, (2) there was still some ThiS-carboxylate sample left in the ThiS-thiocarboxylate sample at the onset of the reaction, and (3) there was no good way to deconvolute the contribution of an ^{18}O to the overall spectral profile. However when ^{13}C- and ^{15}N-depleted ThiS was generated, the shift of 2 Da was readily visualized. (Figure 37). A protein that is depleted shows a rather different isotopic envelope compared to the one described for Pks4 in Figure 4. In the depleted protein the main species observed is the monoisotopic ion. The second most abundant ion is due to the contribution of ^{18}O from naturally abundant oxygen isotopes. The +2 isotope dramatically increases in intensity when ThiS-thiocarboxylate is incubated with ^{18}O-DXP. Subsequent MS/MS or MS^2 demonstrated that this incorporation was at the C-terminal end of ThiS. This provides a good example of the utility of ^{13}C- and ^{15}N-depleted protein strategy.

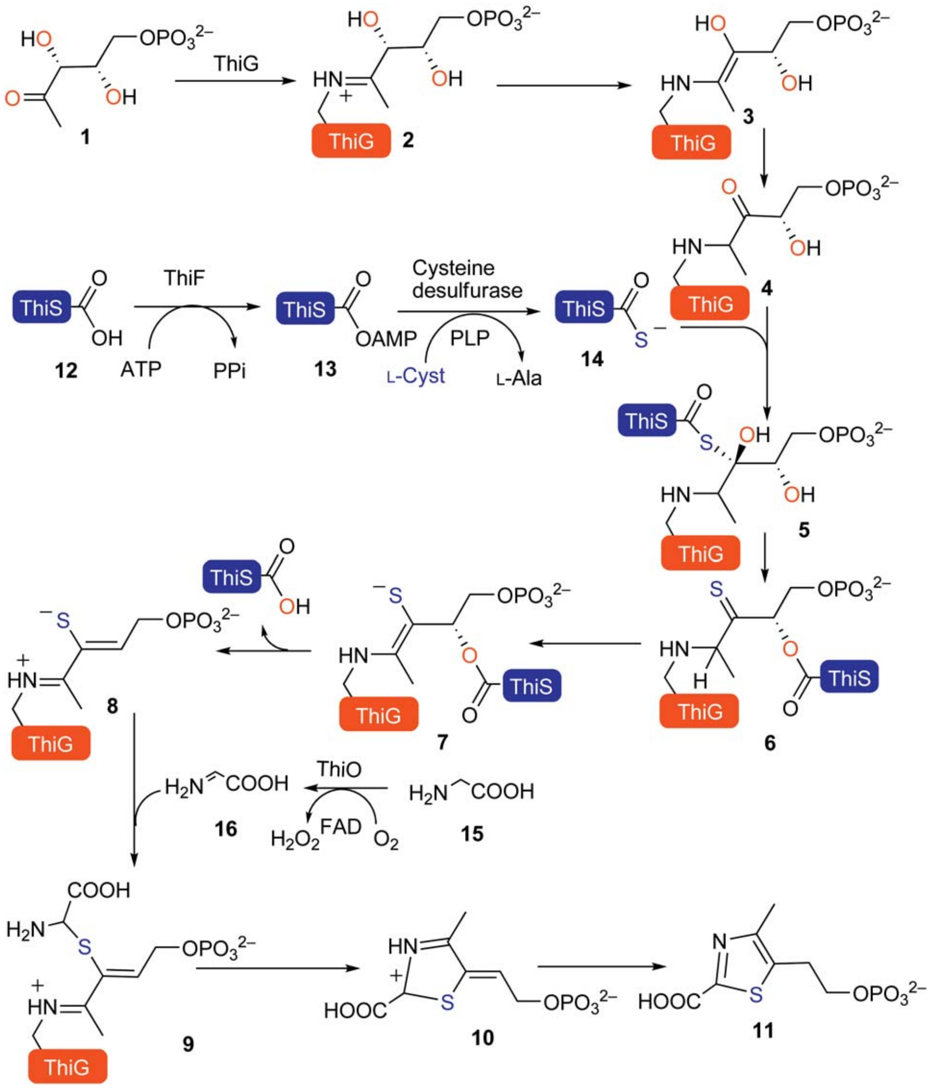


Figure 36 A thiazole phosphate biosynthetic pathway that has been studied by application of isotopically depleted proteins and mass spectrometry. The blue boxes represent ThiS.

9.11.4.2 Trapping Reactive Intermediates

In addition to the unusual transfer of a hydroxyl group from a substrate (DXP) to the C-terminal end on a protein, there are several reactive intermediates on the thiazole synthase pathway.[127] In order to visualize these intermediates they had to be trapped before they could be visualized by MS. Imine **2** or its Amadori rearrangement product **4** in Figure 35 were trapped with $NaBH_4$. $NaBH_4$ reduction resulted in an irreversible linkage to the thiazole synthase ThiG. Similar type of sugar substrate–imine linkage has also been observed for other biosynthetic pathways such as the ones found in pyridoxal phosphate biosynthesis.[128–130] Trapping intermediates like this will no doubt be important in the investigations of NRPS and PKS biosynthetic pathways as well.

9.11.4.3 High-Resolution Mass Spectrometry of Noncovalent Interactions

Another promising MS approach is the investigation of noncovalent complexes. A beautiful example of this was provided by Leary in a study of 5′-adenylylsulfate (APS) reductase, a 4F–4S iron–sulfur cluster-containing protein.[131] APS reductase catalyzes the reduction of APS to sulfite. Leary and coworkers were able to observe

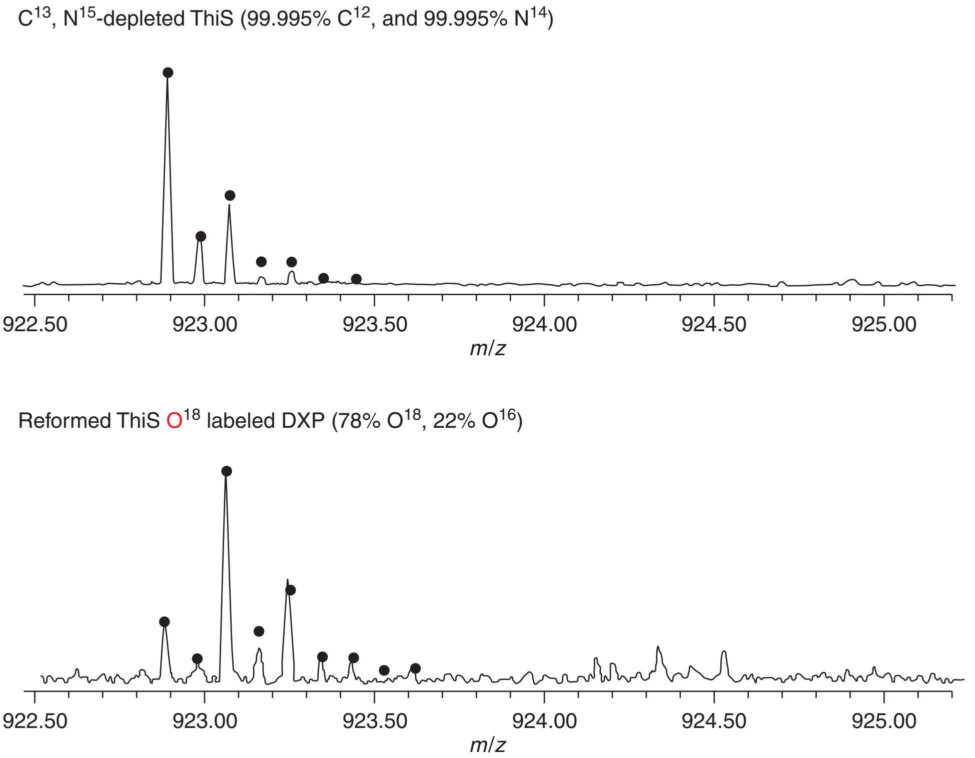


Figure 37 FT–ICR–MS signal of isotopically depleted ThiS (top) and the effect of ^{18}O incorporation (bottom) in ThiS.

the intact iron–sulfur cluster and even to show the interaction of the protein with APS, AMP, and thioredoxin, a cofactor in this biosynthetic process. This result of visualizing the intact APS reductase with its noncovalent substrates is remarkable because it was always believed that iron–sulfur clusters are unstable. Undoubtingly, the application of this approach to NRPS and PKS systems would reveal new insights into their biosynthetic pathways.

9.11.5 Up-and-Coming Advances in Mass Spectrometry Tools for the Investigation of Natural Products and Their Biosynthetic Pathways

In the last few years, mass spectrometers have undergone a revolution. Just 5 years ago, FT–ICR–MS was the only type of MS that was able to obtain sufficient mass accuracy and resolution that enabled the investigations of NRPS and PKS systems in any detail. Since then new methods have evolved, instruments have improved, and additional mass spectrometers have become available that have a mass accuracy within 10 ppm. An example is the ORBI-trap, which is the only other FT analyzer and which measures an image current too. This instrument should be capable of doing any of the experiments described in this chapter. An ORBI-trap can achieve resolutions >100 000. In addition to the ORBI-trap, the most recent Q–TOFs have a resolution of 60 000 and a mass accuracy of sub-parts per million. This mass accuracy rivals the mass accuracy that can be achieved by modern FT–ICR–MS instrumentation. As high-resolution instruments are becoming commonplace, we anticipate that many other investigators will begin using the approaches described in this chapter as well.

In addition to higher mass accuracy instruments, there are also new MS tools that are going to be very exciting in the investigations of NRPS and PKS systems or biosynthetic pathways in general. One such tool is ion mobility. Ion mobility will enable us to look at large complexes and the changes that take place on such complexes. It is also likely that the PEA can be taken to another level as it may be possible to separate the ejected ion from the rest of the peptides in the gas phase. All these exciting developments in MS will ensure that the investigations of these important natural products will be possible with higher accuracy, with smaller samples, and at a faster throughput. Maybe, someday in the next few decades, we may be able to perform these biosynthetic studies at single cell levels.

Acknowledgments

Pieter C. Dorrestein appreciates all the funding he has received from the ACS and Moores Cancer Center, Beckman Foundation, PhRMA Foundation, V-Foundation, and NIH (1R01GM086283 and 1R01GM085128), and the Dorrestein laboratory members and all his colleagues and collaborators on the numerous projects described in this chapter. In addition, the authors thank Michael Thomas for providing the adenylation T didomain construct from **Figure 10**, Yi Tang for providing Pks4, and Ben Shen for providing SgcE. RDK appreciates his USA scholarship from the Studienstiftung des deutschen Volkes.

Abbreviations

3,4-DHBA	3,4-dihydroxybenzoic acid
A (domain)	adenylation
Acac	acetoacetyl
ACP	acyl carrier protein
AL	acyl ligase
AMP	adenosine monophosphate
APS	5′-adenylylsulfate
A–T (didomain)	adenylation–thiolation (didomain)
AT	acyltransferase (domain)
ATP	adenosine triphosphate
ATP–PPi	adeonsine triphosphate–pyrophosphate exchange
BIRD	blackbody infrared radiative dissociation
C (domain)	condensation (domain)
CD	circular dichroism
CDA	calcium-dependent antibiotic
CID	collisionally-induced dissociation
CLC	Claisen-like cyclase
CLF	chain length factor
CoA	coenzyme A
DH (domain)	dehydratase (domain)
DXP	deoxy-D-xylulose-phosphate
E (domain)	epimerization (domain)
ER (domain)	enoyl reductase (domain)
ESI	electrospray ionization
ESI–FT–ICR–MS	electrospray ionization–Fourier transform–ion cyclotron resonance–mass spectrometry
ESI–FTMS	electrospray ionization–Fourier transform mass spectrometry
ESI–MS	electrospray ionization–mass spectrometry
ESI–Q–TOF MS	electrospray ionization–quadrupole–time-of-flight–mass spectrometry
ETD	electron transfer dissociation
FAS	fatty acid synthase
FID	free induction decay
FT–ICR–MS	Fourier transform–ion cyclotron resonance–mass spectrometry

FTMS/FT–MS	Fourier transform–mass spectrometry
GC/EI–MS	gas chromatography/electron impact–mass spectrometry
HCS	HMG-CoA synthase
HIC	hydroxyl isocaproate
HMG	3-hydroxy-3-methylglutaryl
HPLC	high-performance liquid chromatography
IPKS	iterative polyketide synthase
IR	infrared
IRMPD	infrared multiphoton dissociation
KIC	ketoisocaproate
KR (domain)	ketoreductase (domain)
KS (domain)	ketosynthase (domain)
LC	liquid chromatography
LC–MS	liquid chromatography–mass spectrometry
LC–MS–PEA	liquid chromatography–mass spectrometry–phosphopantetheinyl ejection assay
LCQ	three-dimensional linear quadrupole
Linear IT	linear ion trap
LTQ	two-dimensional linear trap quadrupole
LTQ–FT–ICR–MS	hybrid linear trap quadrupole–Fourier transform–ion cyclotron resonance–mass spectrometry
LTQ–ORBI	hybrid linear trap quadrupole–orbitrap
MALDI	matrix-assisted laser desorption/ionization
MALDI–TOF	matrix-assisted laser desorption/ionization–time-of-flight
MAT (domain)	malonyl-CoA:acyl carrier protein acyltransferase
MDa	megadalton
MS	mass spectrometry
MS/MS (or MS^2)	tandem mass spectrometry
MS^3	additional fragmentation of ions generated by MS/MS or MS^2
NAD(P)H	nicotinamide adenine dinucleotide (phosphate) hydride
NAD$^+$	nicotinamide adenine dinucleotide
NADH	nicotinamide adenine dinucleotide hydride
NADPH	nicotinamide adenine dinucleotide phosphate hydride
NMR	nuclear magnetic resonance
NRP	nonribosomal peptide
NRPS	nonribosomal peptide synthetase
NRPS/PKS	hybrid nonribosomal peptide synthetase/polyketide synthase
PEA	phosphopantetheinyl ejection assay
PKS	polyketide synthase
PPant	phosphopantetheine
PPTase	phosphopantetheinyl transferase
PQD	pulsed-Q dissociation
Q–TOF	quadrupole–time-of-flight
RP–HPLC	reverse-phase high-performance liquid chromatography
RPLC	reverse-phase liquid chromatography
SAT (domain)	starter-CoA:acyl carrier protein acyltransferase
SerT	seryltransferase
SORI–CAD	sustained off-resonance irradiation–collisionally activated dissociation
T (domain)	thiolation (domain)
TE (domain)	thioesterase
TE/CLC (domain)	thioesterase/Claisen-like cyclase
TGH	transglutaminase homologue

UV–vis	ultraviolet–visible spectroscopy
X-ray	X-ray diffraction techniques

Nomenclature

Asn	asparagine
Cys	cysteine
Da	dalton
His	histidine
kDa	kilo dalton
Lys	lysine
m/z	mass/charge
nl min^{-1}	nanoliters per minute
Phe	phenylalanine
ppm	parts per million
–S–	bonding sulfur of thioester
T (unit)	tesla
Tyr	tyrosine
***z* (unit)**	charge
μg	microgram
μl min^{-1}	microliter per minute
μmol l^{-1}	micromolar

References

1. D. J. Newman; G. M. Cragg, *J. Nat. Prod.* **2007**, *70*, 461–477.
2. J. E. Syka; J. J. Coon; M. J. Schroeder; J. Shabanowitz; D. F. Hunt, *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 9528–9533.
3. R. A. Zubarev; D. M. Horn; E. K. Fridriksson; N. L. Kelleher; N. A. Kruger; M. A. Lewis; B. K. Carpenter; F. W. McLafferty, *Anal. Chem.* **2000**, *72*, 563–573.
4. H. J. Cooper; K. Håkansson; A. G. Marshall, *Mass Spectrom. Rev.* **2005**, *24*, 201–222.
5. J. Pól; P. Novák; M. Volný; G. H. Kruppa; R. Kostiainen; K. Lemr; V. Havlíček, *Eur. J. Mass Spectrom.* **2008**, *14*, 391–399.
6. A. A. Shvartsburg; R. D. Smith, *Anal. Chem.* **2008**, *80*, 9689–9699.
7. N. E. Manicke; T. Kistler; D. R. Ifa; R. G. Cooks; Z. Ouyang, *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 321–325.
8. Z. Takáts; J. M. Wiseman; B. Gologan; R. G. Cooks, *Science* **2004**, *306*, 471–473.
9. N. D. Udeshi; P. D. Compton; J. Shabanowitz; D. F. Hunt; K. L. Rose, *Nat. Protoc.* **2008**, *3*, 1709–1717.
10. G. C. McAlister; W. T. Berggren; J. Griep-Raming; S. Horning; A. Makarov; D. Phanstiel; G. Stafford; D. L. Swaney; J. E. Syka; V. Zabrouskov; J. J. Coon, *J. Proteome Res.* **2008**, *7*, 3127–3136.
11. T. R. Northen; J. C. Lee; L. Hoang; J. Raymond; D. R. Hwang; S. M. Yannone; C. H. Wong; G. Siuzdak, *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 3678–3683.
12. T. R. Northen; O. Yanes; M. T. Northen; D. Marrinucci; W. Uritboonthai; J. Apon; S. L. Golledge; A. Nordström; G. Siuzdak, *Nature* **2007**, *449*, 1033–1036.
13. H. Han; Y. Xia; M. Yang; S. A. McLuckey, *Anal. Chem.* **2008**, *80*, 3492–3497.
14. Y. Nakata; Y. Honda; S. Ninomiya; T. Seki; T. Aoki; J. Matsuo, *J. Mass Spectrom.* **2009**, *44*, 128–136.
15. A. Brunelle; O. Laprévote, *Anal. Bioanal. Chem.* **2009**, *393*, 31–35.
16. H. Hazama; J. Aoki; H. Nagao; R. Suzuki; T. Tashima; K. Fujii; K. Masuda; K. Awazu; M. Toyoda; Y. Naito, *Appl. Surf. Sci.* **2008**, *255*, 1257–1263.
17. P. Chaurand; S. A. Schwartz; R. M. Caprioli, *Curr. Opin. Chem. Biol.* **2002**, *6*, 676–681.
18. Q. Hu; R. J. Noll; H. Li; A. Makarov; M. Hardman; R. Graham Cooks, *J. Mass Spectrom.* **2005**, *40*, 430–443.
19. J. Van der Greef; R. Van der Heijden; E. R. Verheij, *Adv. Mass Spectrom.* **2004**, *16*, 145–165.
20. A. L. Demain, *Appl. Microbiol. Biotechnol.* **1999**, *52*, 455–463.
21. J. Dubois; D. Guenard; F. Gueritte, *Expert Opin. Ther. Pat.* **2003**, *13*, 1809–1823.
22. L. Du; B. Shen, *Curr. Opin. Drug Discov. Devel.* **2001**, *4*, 215–228.

23. M. A. Fischbach; C. T. Walsh, *Chem. Rev.* **2006**, *106*, 3468–3496.
24. Z. Zhou; P. Cironi; A. J. Lin; Y. Xu; S. Hrvatin; D. E. Golan; P. A. Silver; C. T. Walsh; J. Yin, *Chem. Biol.* **2007**, *2*, 337–346.
25. U. Linne; A. Schäfer; M. T. Stubbs; M. A. Marahiel, *FEBS Lett.* **2007**, *581*, 905–910.
26. C. T. Walsh; H. Chen; T. A. Keating; B. K. Hubbard; H. C. Losey; L. Luo; C. G. Marshall; D. A. Miller; H. M. Patel, *Curr. Opin. Chem. Biol.* **2001**, *5*, 525–534.
27. F. Kopp; M. A. Marahiel, *Curr. Opin. Biotechnol.* **2007**, *18*, 513–520.
28. K. M. Hoyer; C. Mahlert; M. A. Marahiel, *Chem. Biol.* **2007**, *14*, 13–22.
29. Y. Hu; V. Phelan; I. Ntai; C. M. Farnet; E. Zazopoulos; B. O. Bachmann, *Chem. Biol.* **2007**, *14*, 691–701.
30. L. C. Blasiak; F. H. Vaillancourt; C. T. Walsh; C. L. Drennan, *Nature* **2006**, *440*, 368–371.
31. N. Peric-Concha; B. Borovicka; P. F. Long; D. Hranueli; P. G. Waterman; I. S. Hunter, *J. Biol. Chem.* **2005**, *280*, 37455–37460.
32. A. Li; J. Piel, *Chem. Biol.* **2002**, *9*, 1017–1026.
33. L. Tang; S. Shah; L. Chung; J. Carney; L. Katz; C. Khosla; B. Julien, *Science* **2000**, *287*, 640–642.
34. D. J. Edwards; B. L. Marquez; L. M. Nogle; K. McPhail; D. E. Goeger; M. A. Roberts; W. H. Gerwick, *Chem. Biol.* **2004**, *11*, 817–833.
35. W. Liu; S. D. Christenson; S. Standage; B. Shen, *Science* **2002**, *297*, 1170–1173.
36. E. Guenzi; G. Galli; I. Grgurina; D. C. Gross; G. Grandi, *J. Biol. Chem.* **1998**, *273*, 32857–32863.
37. X. Wei; F. Yang; D. C. Straney, *Can. J. Microbiol.* **2005**, *51*, 423–429.
38. S. Tanner; H. Shu; A. Frank; L. C. Wang; E. Zandi; M. Mumby; P. A. Pevzner; V. Bafna, *Anal. Chem.* **2005**, *77*, 4626–4639.
39. L. McHugh; J. W. Arthur, *PLoS Comput. Biol.* **2008**, *4*, e12.
40. F. X. Wu; P. Gagné; A. Droit; G. G. Poirier, *Bioinformatics* **2008**, *9*, S13.
41. C. A. Shaw-Reid; N. L. Kelleher; H. C. Losey; A. M. Gehring; C. Berg; C. T. Walsh, *Chem. Biol.* **1999**, *6*, 385–400.
42. D. B. Stein; U. Linne; M. Hahn; M. A. Marahiel, *ChemBioChem* **2006**, *7*, 1807–1814.
43. R. Gerber; L. Lou; L. Du, *J. Am. Chem. Soc.* **2009**, *131*, 3148–3149.
44. C. Qiao; D. J. Wilson; E. M. Bennett; C. C. Aldrich, *J. Am. Chem. Soc.* **2007**, *129*, 6350–6351.
45. L. Gu; T. W. Geders; B. Wang; W. H. Gerwick; K. Håkansson; J. L. Smith; D. H. Sherman, *Science* **2007**, *318*, 970–974.
46. M. Strieker; F. Kopp; C. Mahlert; L. O. Essen; M. A. Marahiel, *Chem. Biol.* **2007**, *2*, 187–196.
47. P. C. Dorrestein; N. L. Kelleher, *Nat. Prod. Rep.* **2006**, *23*, 893–918.
48. J. M. Crawford; P. M. Thomas; J. R. Scheerer; A. L. Vagstad; N. L. Kelleher; C. A. Townsend, *Science* **2008**, *320*, 243–246.
49. S. B. Bumpus; N. A. Magarvey; N. L. Kelleher; C. T. Walsh; C. T. Calderone, *J. Am. Chem. Soc.* **2008**, *130*, 11614–11616.
50. P. C. Dorrestein; S. B. Bumpus; C. T. Calderone; S. Garneau-Tsodikova; Z. D. Aron; P. D. Straight; R. Kolter; C. T. Walsh; N. L. Kelleher, *Biochemistry* **2006**, *45*, 12756–12766.
51. D. B. Hansen; S. B. Bumpus; Z. D. Aron; N. L. Kelleher; C. T. Walsh, *J. Am. Chem. Soc.* **2007**, *129*, 6366–6367.
52. F. Kopp; U. Linne; M. Oberthür; M. A. Marahiel, *J. Am. Chem. Soc.* **2008**, *130*, 2656–2666.
53. Y. A. Chan; M. T. Boyne, II; A. M. Podevels; A. K. Klimowicz; J. Handelsman; N. L. Kelleher; M. G. Thomas, *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 14349–14354.
54. J. Zhang; S. G. Van Lanen; J. Ju; W. Liu; P. C. Dorrestein; W. Li; N. L. Kelleher; B. A. Shen, *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 1460–1465.
55. W. L. Kelly; M. T. Boyne, II; E. Yeh; D. A. Vosburg; D. P. Galonić; N. L. Kelleher; C. T. Walsh, *Biochemistry* **2007**, *46*, 359–368.
56. C. T. Calderone; D. F. Iwig; P. C. Dorrestein; N. L. Kelleher; C. T. Walsh, *Chem. Biol.* **2007**, *14*, 835–846.
57. C. T. Calderone; W. E. Kowtoniuk; N. L. Kelleher; C. T. Walsh; P. C. Dorrestein, *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 8977–8982.
58. C. T. Calderone; S. B. Bumpus; N. L. Kelleher; C. T. Walsh; N. A. Magarvey, *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 12809–12814.
59. A. G. Marshall; C. L. Hendrickson, *Annu. Rev. Anal. Chem.* **2008**, *1*, 579–599.
60. A. G. Marshall; C. L. Hendrickson; G. S. Jackson, *Mass Spectrom. Rev.* **1998**, *17*, 1–35.
61. T. Liu; M. E. Belov; N. Jaitly; W. J. Qian; R. D. Smith, *Chem. Rev.* **2007**, *107*, 3621–3653.
62. W. Zhang; Y. Li; Y. Tang, *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 20683–20688.
63. N. L. Kelleher; M. W. Senko; M. M. Siegel; F. W. McLafferty, *J. Am. Soc. Mass Spectrom.* **1997**, *8*, 380–383.
64. N. L. Kelleher, University of Illinois at Urbana-Champaign, Urbana, IL. Unpublished work, 2006.
65. P. J. Ulintz; A. K. Yocum; B. Bodenmiller; R. Aebersold; P. C. Andrews; A. I. Nesvizhskii, *J. Proteome Res.* **2009**, *8*, 887–899.
66. A. Tholey; J. Reed; W. D. Lehmann, *J. Mass Spectrom.* **1999**, *34*, 117–123.
67. M. Edelson-Averbukh; R. Pipkorn; W. D. Lehmann, *Anal. Chem.* **2006**, *78*, 1249–1256.
68. A. M. Palumbo; J. J. Tepe; G. E. Reid, *J. Proteome Res.* **2008**, *7*, 771–779.
69. S. G. Van Lanen; S. Lin; P. C. Dorrestein; N. L. Kelleher; B. Shen, *J. Biol. Chem.* **2006**, *281*, 29633–29640.
70. M. P. Barrow; W. I. Burkitt; P. J. Derrick, *Analyst* **2005**, *130*, 18–28.
71. L. Sleno; D. A. Volmer, *J. Mass Spectrom.* **2004**, *39*, 1091–1112.
72. W. D. Price; P. D. Schnier; E. R. Williams, *Anal. Chem.* **1996**, *68*, 859.
73. J. W. Gauthier; T. R. Trautman; D. B. Jacobson, *Anal. Chim. Acta.* **1991**, *246*, 211.
74. D. P. Little; J. P. Speir; M. W. Senko; P. B. O'Connor; F. W. McLafferty, *Anal. Chem.* **1994**, *66*, 2809–2815.
75. W. Li; C. L. Hendrickson; M. R. Emmett; A. G. Marshall, *Anal. Chem.* **1999**, *71*, 4397–4402.
76. M. W. Senko; J. P. Speir; F. W. McLafferty, *Anal. Chem.* **1994**, *66*, 2801–2808.
77. A. H. Payne; G. L. Glish, *Anal. Chem.* **2001**, *73*, 3542–3548.
78. A. H. Racine; A. H. Payne; P. M. Remes; G. L. Glish, *Anal. Chem.* **2006**, *78*, 4609–4614.
79. J. C. Schwartz; J. E. P Syka; S. T. Quarmby, In Proceedings of the 53rd ASMS Conference on Mass Spectrometry and Allied Topics, San Antonio, TX, 5–9 June 2005.
80. T. Schlabach; T. Zhang; K. Miller; R. Kiyonami, The 2006 ABRF Conference, Long Beach, CA, 11–14 February 2006.
81. D. Meluzzi; W. H. Zheng; M. Hensler; V. Nizet; P. C. Dorrestein, *Bioorg. Med. Chem. Lett.* **2008**, *18*, 3107–3111.
82. L. M. Hicks; M. T. Mazur; L. M. Miller; P. C. Dorrestein; N. A. Schnarr; C. Khosla; N. L. Kelleher, *ChemBioChem* **2006**, *7*, 904–907.
83. P. D. Straight; M. A. Fischbach; C. T. Walsh; D. Z. Rudner; R. Kolter, *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 305–310.

84. B. Bothner; G. Siuzdak, *ChemBioChem* **2004**, *5*, 258–260.
85. A. R. McKay; B. T. Ruotolo; L. L. Ilag; C. V. Robinson, *J. Am. Chem. Soc.* **2006**, *128*, 11433–11442.
86. H. Donato; N. I. Krupenko; Y. Tsybovsky; S. A. Krupenko, *J. Biol. Chem.* **2007**, *282*, 34159–34166.
87. J. Yin; P. D. Straight; S. Hrvatin; P. C. Dorrestein; S. B. Bumpus; C. Jao; N. L. Kelleher; R. Kolter; C. T. Walsh, *Chem. Biol.* **2007**, *14*, 303–312.
88. R. S. Brown; A. Aman, *J. Org. Chem.* **1997**, *62*, 4816–4820.
89. S. H. Payne; M. Yau; M. B. Smolka; S. Tanner; H. Zhou; V. Bafna, *J. Proteome Res.* **2008**, *7*, 3373–3381.
90. T. Stein; J. Vater; V. Kruft; A. Otto; B. Wittmann-Liebold; P. Franke; M. Panico; R. McDowell; H. R. Morris, *J. Biol. Chem.* **1996**, *271*, 15428–15435.
91. P. C. Dorrestein; J. Blackhall; P. D. Straight; M. A. Fischbach; S. Garneau-Tsodikova; D. J. Edwards; S. McLaughlin; M. Lin; W. H. Gerwick; R. Kolter; C. T. Walsh; N. L. Kelleher, *Biochemistry* **2006**, *45*, 1537–1546.
92. K. J. Weissman; R. Müller, *ChemBioChem* **2008**, *9*, 826–848.
93. N. A. Magarvey; P. D. Fortin; P. M. Thomas; N. L. Kelleher; C. T. Walsh, *Chem. Biol.* **2008**, *3*, 542–554.
94. PAWS is freeware originally developed by ProteoMetrics, LLC and now made available by Genomic Solutions (http://bioinformatics.genomicsolutions.com/).
95. http://prosightptm.scs.uiuc.edu (accessed 1 September 2006).
96. S. Garneau-Tsodikova; P. C. Dorrestein; N. L. Kelleher; C. T. Walsh, *J. Am. Chem. Soc.* **2006**, *128*, 12600–12601.
97. L. M. Hicks; M. C. Moffitt; L. L. Beer; B. S. Moore; N. L. Kelleher, *Chem. Biol.* **2006**, *1*, 93–102.
98. L. M. Hicks; C. J. Balibar; C. T. Walsh; N. L. Kelleher; N. J. Hillson, *Biophys. J.* **2006**, *91*, 2609–2619.
99. T. Stachelhaus; H. D. Mootz; M. A. Marahiel, *Chem. Biol.* **1999**, *6*, 493–505.
100. G. L. Challis; J. Ravel; C. A. Townsend, *Chem. Biol.* **2000**, *7*, 211–224.
101. C. Rausch; T. Weber; O. Kohlbacher; W. Wohlleben; D. H. Huson, *Nucleic Acids Res.* **2005**, *33*, 5799–5808.
102. M. Z. Ansari; J. Sharma; R. S. Gokhale; D. Mohanty, *Bioinformatics* **2008**, *9*, 454.
103. P. Berg; F. H. Bergmann; E. J. Ofengand; M. Dieckmann, *J. Biol. Chem.* **1961**, *236*, 1726–1734.
104. B. F. Pfleger; J. Y. Lee; R. V. Somu; C. C. Aldrich; P. C. Hanna; D. H. Sherman, *Biochemistry* **2007**, *46*, 4147–4157.
105. P. D. Fortin; C. T. Walsh; N. A. Magarvey, *Nature* **2007**, *448*, 824–827.
106. E. R. Strieter; F. H. Vaillancourt; C. T. Walsh, *Biochemistry* **2007**, *46*, 7549–7557.
107. G. J. Gatto, Jr.; M. T. Boyne, II; N. L. Kelleher; C. T. Walsh, *J. Am. Chem. Soc.* **2006**, *128*, 3838–3847.
108. G. L. Tang; Y. Q. Cheng; B. Shen, *J. Biol. Chem.* **2007**, *282*, 20273–20282.
109. M. Wittmann; U. Linne; V. Pohlmann; M. A. Marahiel, *FEBS J.* **2008**, *275*, 5343–5354.
110. J. J. La Clair; T. L. Foley; T. R. Schegg; C. M. Regan; M. D. Burkart, *Chem. Biol.* **2004**, *11*, 195–201.
111. S. C. Wenzel; P. Meiser; T. M. Binz; T. Mahmud; R. Müller, *Angew. Chem. Int. Ed. Engl.* **2006**, *45*, 2296–2301.
112. S. A. Samel; M. A. Marahiel; L. O. Essen, *Mol. Biosyst.* **2008**, *4*, 387–393.
113. S. M. McLoughlin; N. L. Kelleher, *J. Am. Chem. Soc.* **2005**, *127*, 14984–14985.
114. M. Zerikly; G. L. Challis, *ChemBioChem* **2009**, *10*, 625–633.
115. R. B. Merrifield, *J. Am. Chem. Soc.* **1963**, *85*, 2149–2154.
116. D. Hoffmeister; N. P. Keller, *Nat. Prod. Rep.* **2007**, *24*, 393–416.
117. S. G. Van Lanen; B. Shen, *Curr. Opin. Drug Discov. Devel.* **2008**, *11*, 186–195.
118. R. A. Butcher; F. C. Schroeder; M. A. Fischbach; P. D. Straight; R. Kolter; C. T. Walsh; J. Clardy, *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 1506–1509.
119. J. J. Reddick; S. A. Antolak; G. M. Raner, *Biochem. Biophys. Res. Commun.* **2007**, *358*, 363–367.
120. P. Verdier-Pinard; J. Y. Lai; H. D. Yoo; J. Yu; B. Marquez; D. G. Nagle; M. Nambu; J. D. White; J. R. Falck; W. H. Gerwick; B. W. Day; E. Hamel, *Mol. Pharmacol.* **1998**, *53*, 62–76.
121. H. M. Holden; M. M. Benning; T. Haller; J. A. Gerlt, *Acc. Chem. Res.* **2001**, *34*, 145–157.
122. L. Gu; J. Jia; H. Liu; K. Håkansson; W. H. Gerwick; D. H. Sherman, *J. Am. Chem. Soc.* **2006**, *128*, 9014–9015.
123. T. W. Geders; L. Gu; J. C. Mowers; H. Liu; W. H. Gerwick; K. Håkansson; D. H. Sherman; J. L. Smith, *J. Biol. Chem.* **2007**, *282*, 35954–35963.
124. Y. Sugimoto; T. Otani; S. Oie; K. Wierzba; Y. Yamada, *J. Antibiot.* **1990**, *43*, 417–421.
125. P. C. Dorrestein; H. Zhai; F. W. McLafferty; T. P. Begley, *Chem. Biol.* **2004**, *11*, 1373–1381.
126. A. Hazra; A. Chatterjee; T. P. Begley, *J. Am. Chem. Soc.* **2009**, *131*, 3225–3229.
127. P. C. Dorrestein; H. Huili Zhai; S. V. Taylor; F. W. McLafferty; T. P. Begley, *J. Am. Chem. Soc.* **2004**, *126*, 3091–3096.
128. K. E. Burns; Y. Xiang; C. L. Kinsland; F. W. McLafferty; T. P. Begley, *J. Am. Chem. Soc.* **2005**, *127*, 3682–3683.
129. J. W. Hanes; K. E. Burns; D. G. Hilmey; A. Chatterjee; P. C. Dorrestein; T. P. Begley, *J. Am. Chem. Soc.* **2008**, *130*, 3043–3052.
130. T. Raschle; D. Arigoni; R. Brunisholz; H. Rechsteiner; N. Amrhein; T. B. Fitzpatrick, *J. Biol. Chem.* **2007**, *282*, 6098–6105.
131. H. Gao; J. Leary; K. S. Carroll; C. R. Bertozzi; H. Chen, *J. Am. Soc. Mass Spectrom.* **2007**, *18*, 167–178.

Biographical Sketches

Roland D. Kersten was born in Berlin, Germany, in 1985. He obtained his Diplom in 2008 in biochemistry from the Free University, Berlin. The final 9 months of Diplom thesis, on the characterization of orphan gene clusters from the marine organism *Salinipora tropica*, was supervised by Dr. Dorrestein at the University of California, San Diego. Following the completion of his Diplom, Roland continued to answer natural product biosynthetic questions by modern mass spectrometry. Roland is currently working in the Dorrestein laboratory as a research associate. In the summer of 2009, he will be joining the graduate program at the Scripps Institution of Oceanography, San Diego, where he will continue to implement mass spectrometry as a genome mining tool for natural product discovery under the dual mentorship of Dr. Bradley Moore and Dr. Pieter C. Dorrestein.

Michael J. Meehan was born in Downer's Grove, Illinois. He received his bachelor's degree in biology from the University of San Francisco (USF) in 2001. During his studies at USF, Michael was primarily interested in immunology, and he also worked for Dr. Eugene V. Benton on a joint project between the USF Department of Physics and the Loma Linda University Cancer Institute, studying the interaction of high-energy proton beams with tissue-equivalent materials by atomic force microscopy. Michael entered the master's program in chemistry at UCSD in the fall of 2007 and joined the laboratory of Dr. Pieter Dorrestein in early 2008. During his time in the Dorrestein laboratory, Michael's research focused on mapping of active sites of polyketide synthases and detection of intermediates on these enzymes using high-resolution mass spectrometry. Michael will complete his master's thesis by the summer of 2009.

Dr. Pieter C. Dorrestein was born in Utrecht, the Netherlands, in 1974. He received his undergraduate chemistry training at the Northern Arizona University in 1998 under the tutelage of Dr. John MacDonald. Pieter remained in John's laboratory as a masters student for an additional year. In 1999, Pieter moved to the Department of Chemistry and Chemical Biology at Cornell University, where he received his Ph.D. in 2004. His Ph.D., on the biosynthetic investigations of thiamin, was supervised by Dr. Tadgh Begley. Following his Ph.D., Pieter, as an NIH NRSA Kirchstein fellow, combined his interests in mass spectrometry and natural products. During this time Pieter used high-resolution mass spectrometry methods to investigate the biosynthesis of therapeutic agents under the supervision and cosponsorship of Dr. Neil Kelleher (University of Illinois) and Dr. Christopher T. Walsh (Harvard Medical School). In September 2006, Pieter joined the UCSD Skaggs School of Pharmacy and Pharmaceutical Sciences and the Departments of Pharmacology, Chemistry, and Biochemistry. His current research aims to develop new mass spectrometry approaches to detect and characterize natural products as well as their biosynthesis. In addition, his laboratory aims to characterize the function of posttranslational modifications involved in disease processes. Since his arrival in UCSD, Pieter has received several honors, including the Beckman Young Investigator Award, the V-Foundation Scholar Award, the PhRMA Foundation Award, the Eli Lilly Young Investigator Award in Analytical Chemistry, and enjoys funding from the National Institute of Health.

9.12 Mass Spectrometry: Structure Determination of Proteins and Peptides

Chhabil Dass, The University of Memphis, Memphis, TN, USA

9.12.1 Introduction

Proteins are essential biomolecules formed by joining 20 naturally occurring L-α-amino acids in a specific sequence. Proteins play crucial roles in virtually all biological processes. Just about every cell and organ of our body use proteins to create the chemistry of life. Although genes carry the information of what proteins need to be produced, it is proteins that make life. Some proteins act as enzymes to catalyze specific reactions, whereas others are antibodies, receptors, hormones, or major components of muscles, bones, skin, and hairs. Proteins control the expression of genes and participate in the transport and storage of small molecules and in the generation and transmission of nerve impulses. The protein chain is held together by peptide bonds formed between the α-carboxylic group of one amino acid and the α-amino group of another amino acid. All amino acids contain a central carbon atom to which are attached a hydrogen atom, an amino group, a carboxylic group, and a side chain. It is the side chain that imparts a distinct character to an amino acid and to a protein. In the context of the protein structure, amino acids are called residues. By convention, the chain begins with the free amino group-containing residue, called the N-terminus, and end with a free α-carboxylic acid group-containing residue, called the C-terminus. Although there is no rigid rule, chains with less than 40 amino acids are considered to be peptides.

Proteins are the end products of gene expression. Production of a protein in a cell starts with transcription of the genetic information from the gene to the messenger ribonucleic acid (mRNA), followed by translation of this information into the sequence of a protein. During transport, proteins undergo a variety of posttranslational modifications (PTMs). The ultimate structure of a protein, known as conformation, is a well-defined three-dimensional (3D) arrangement of the constituent amino acids. Depending upon the environment, a protein may organize itself into one of the following four levels of structures:[1] the primary structure describes the linear sequence of the constituent amino acids and location of disulfide bonds, plus any other covalent modifications of amino acids. The secondary structure represents the folding of short segments of proteins into defined structural elements, such as α-helices, β-sheets, and turns that are stabilized by hydrogen bonds between –C=O and H–N– groups in the backbone. The tertiary structure describes how the secondary structural elements of a single protein chain interact with each other to fold into a unique 3D structure required to manifest the function of that protein. Tertiary structures are stabilized by disulfide bridges, hydrogen bonds, salt bridges, and hydrophobic interactions. Finally, the quaternary structure is found in proteins that are made from more than one polypeptide chain; these subunits combine together through polar and hydrophobic interactions to form a multimeric complex. In addition to these four levels, discrete portions of a protein may fold into domains, which are compact globular units or structural motifs usually endowed with their own biological functions.

Knowledge of the amino acid sequence of a protein is paramount in understanding biological events, the molecular basis of its biological activity, and predicting its 3D structure. Identifying the exact molecular form of a protein is also important from the viewpoint of human health and in unraveling evolutionary events. Proteins from a common ancestor have enhanced sequence homology.

9.12.1.1 Scope of the Present Work

The analysis of proteins and peptides is a highly visible application of mass spectrometry. This chapter will shed light on and describe in detail the widely accepted proteomics-based technology for large-scale identification of proteins. Proteomics is defined as the study of the entire protein complement of the genome. *De novo* sequencing of peptides is another area where mass spectrometry has made tremendous strides; a substantial portion of this chapter is devoted to this topic. Other pertinent mass spectrometry applications discussed in this chapter with which a biologist must be familiar are detecting phosphorylation, glycosylation, higher-order structures of proteins, and protein–ligand noncovalent interactions. A great deal of the success of mass spectrometry in these fields is the result of the development of 'soft' ionization techniques, electrospray ionization (ESI),[2] and matrix-assisted laser desorption/ionization (MALDI),[3,4] as well as the emergence of advanced instruments with high-mass, high-sensitivity, and high-throughput capabilities (see two volumes on mass spectrometry by the author,[5,6] and Chapter 9.10 for a discussion of mass spectrometry instrumentation).

9.12.2 Proteomics: A Field of Protein Characterization

Present-day technology for large-scale characterization of proteins in biological samples by mass spectrometry is in proteomics.[7] This is a relatively new field of biomedical research that deals with systematic study of gene products. The concept of a proteome has been defined as the entire protein complement expressed at a given time by a genome or by a cell or tissue type under a given condition (i.e., proteome is protein + genome).[8] A proteome, however, is more complicated than a genome because enormous numbers of proteins can be obtained from a single gene. In the past, conducting a global-scale study of genes and proteins was a formidable undertaking; experimental techniques were limited to the analysis of a single or a small subset of genes or proteins. To achieve success in large-scale protein identification in biological media, one must meet the challenges posed by the enormous complexity of biological samples and diverse physicochemical properties, their dynamic characters, and the wide range of expression levels of proteins present. Now, with the increasing availability of genomic sequences, the development of bioinformatics, and several pertinent experimental techniques, including mass spectrometry approaches and sample fractionation techniques, simultaneous analysis of a large number of proteins in a complex biological sample has become practical through proteomics. This new experimental paradigm is able to decipher the information contained in genomic sequences in terms of the structure, function, and control of biological processes.

Proteomics has gained wide acceptance as a means of studying biological systems because it provides a more detailed picture of the events that take place in the cell. The proteome of a living system is highly dynamic, as the types of expressed proteins, their expression levels and subcellular location, and the types of PTMs are all in constant flux and greatly dependent on the environment and physiological state of the cell, meaning whether or not a cell is healthy or whether or not the cell as been subject to insult. In contrast, the genome is a static system and is essentially identical in every cell of an organism; thus, the genome is unable to describe biological processes completely. For example, mRNA levels do not provide a clear picture of the isoforms of the translated proteins, their levels, and their regulatory status. Furthermore, this impressive technology has opened the door to identify proteins as potential therapeutic agents and as biomarkers of disease and stress due to external stimuli at much reduced time and cost.

Mass spectrometry has become a major player in proteome analysis because of its integration with high-resolution separation techniques and protein databases and its inherent high sensitivity, high structure specificity, high-mass capability, and opportunity for automation. Short analysis times and straightforward sample preparation steps are the other advantages of mass spectrometry-based proteomics.

9.12.2.1 Subareas of Proteomics

Proteomics can be classified into the following subareas: (1) characterization proteomics, also known as mining or profiling – a survey analysis that provides identity of all proteins presents in a cell, tissue, or biofluid, (2) differential proteomics – a study of differentially expressed proteins under different physiological states (e.g., healthy versus diseased subjects, or drug- and toxin-treated subjects), (3) functional proteomics – a field of identification of a network of proteins that interact with each other to carry out a specific cellular function, (4) structural proteomics – a study of 3D structures and dynamics of proteins and of protein complexes, and (5) posttranslational modification proteomics – a study of posttranslationally modified proteins, notably those that have undergone phosphorylation (phosphoproteomics) or glycosylation (glycoproteomics).

9.12.2.2 Basic Concept and Essential Steps of Proteome Analysis

Mining proteins in a tissue, fluid, or cell culture is a major area of application of proteome analysis. The underlying principle is that the accurate mass or the amino acid sequence of a peptide with six or more amino acid residues is a unique signature of a protein. Thus, by matching this unique mass or the sequence-specific ions with entries in a database, we can identify the protein in question. The concept of proteome analysis is simple, but its execution is challenging, primarily owing to the complexity of any given proteome; a typical cell may contain more than 20 000 expressed proteins with a broad dynamic range of expression. In addition, a protein may have several isoforms and may have undergone PTM. Keeping these issues in mind, any successful protocol for proteome analysis must follow these essential steps: (1) sample preparation, (2) separation of a complex mixture of proteins, (3) cleavage of proteins, (4) mass spectrometry analysis, and (5) database searching.[7]

9.12.2.3 Preparation of Protein Sample

A biological sample must be processed so that proteins can be extracted with high efficiency; to this end, a tissue sample is minced in the presence of a suitable lysis buffer that might contain detergents (e.g., SDS, CHAPS, Tween to solubilize membrane proteins), denaturing agents (urea or acids; to disrupt protein–protein interactions and higher-order structures), reducing agents (2-mercaptoethanol or dithiothreitol to reduce disulfide bonds), enzymes (e.g., DNase, RNase; to digest nucleic acids, carbohydrates, and lipids), and a cocktail of enzyme inhibitors (to prevent proteolysis of proteins). The minced tissue sample is homogenized in a blender.

9.12.2.4 Fractionation of a Complex Protein Mixture

Because of the diversity and the complexity of proteins in biological samples, a high degree of fractionation into individual proteins or a simple mixture has become a necessary requirement. Several diverse properties of proteins, such as shape, charge, hydrophobicity, and affinities for other molecules, are exploited to achieve optimal separation and purification. A generic separation scheme is shown in **Figure 1**.[6] The principal separation techniques used are one-dimensional (1D)-gel electrophoresis (1-DE), 2-DE, and two-dimensional (2D)-liquid chromatography (LC). To enhance the probability of mass spectrometry identification of low copy number proteins in biological samples, the samples must be preferentially enriched by ammonium sulfate precipitation or affinity chromatography. To facilitate the detection of low-abundance proteins in biological fluids (blood, plasma, cerebrospinal fluid (CSF)), major blood proteins, such as albumin, must be removed. A variety of immunodepletion columns can be purchased for this purpose.[9] When the sample is passed through the column, highly abundant proteins are collectively retained as a bound fraction, and the low-abundance proteins are eluted in the flow-through. Other options for enhancing the detection of low-abundance proteins are (1) fractionation of the whole organ lysate into distinct organelles (e.g., nuclear membrane, mitochondria, endoplasmic reticulum, microsomes, and cytoplasm) using differential centrifugation and density gradient centrifugation[10] and (2) prefractionation of the sample with free-flow fractionation (FFE).[11]

The concept of a chemical tag or a purification handle has also gained wide acceptance in the selective enrichment of certain proteins. A small peptide is attached to the protein to be purified, and enhanced affinity of

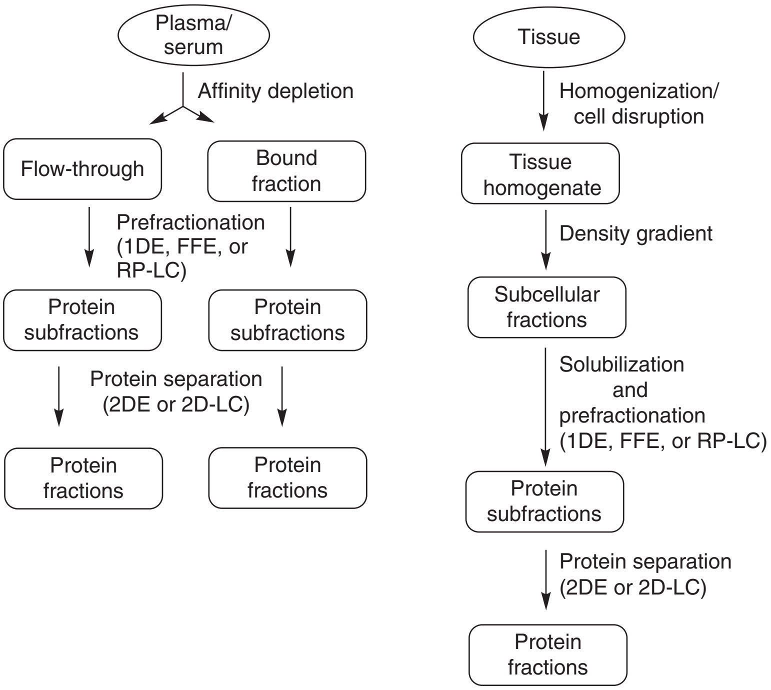


Figure 1 A general scheme for separation of proteins from biofluids and tissues. Reproduced from C. Dass, *Fundamentals of Contemporary Mass Spectrometry*; Wiley-Interscience: Hoboken, NJ, 2007, with permission from Wiley-Interscience, Copyright 2007.

the attached tag for ion-exchange, immunoaffinity, or metal ion affinity columns is used to selectively isolate the tagged proteins.[12] A well-known example is the attachment of hexahistidine to proteins and subsequent isolation of the fusion proteins by exploiting the affinity of hexahistidine with immobilized metal ions affinity chromatography (IMAC).

9.12.2.4.1 Two-dimensional gel electrophoresis

2-DE is the most powerful and robust technique that has been used for the last three decades to separate a complex mixture of proteins from biological specimens.[13] It is the 2D approach that combines two complementary techniques, isoelectric focusing (IEF) and sodium dodecyl sulfate (SDS)–polyacrylamide gel electrophoresis (PAGE). In IEF, separation of proteins occurs on the basis of their inherent charge or isoelectric point (p*I* values; i.e., the point at which the net charge of the protein is zero). The protein sample is mixed with an appropriate cocktail of carrier ampholytes and applied to a standard immobilized pH gradient gel (IPG) strip. After overnight rehydration of the strip, appropriate electrical potential is applied to the strip to effect separation of the proteins.

The second dimension separates proteins on the basis of their molecular size. The focused IEF strips are equilibrated with the SDS–Tris buffer containing a reducing agent (e.g., 2-mercaptoethanol or dithiothreitol) and an alkylating agent (e.g., iodoacetamide) and placed at the top of the SDS–polyacrylamide gel. After completion of the separation, the gel is stained with colloidal Coomassie Blue (CCB), silver staining reagent, or more sensitive fluorescent dyes such as SYPRO Ruby, Deep Purple, or cyanine dyes to visualize the separated proteins. The stained images are captured with a fluorescence imager. The software (e.g., PDQuest) is available to simplify the image analysis process. The protein spots are picked up with an automatic spot picker for digestion and subsequent mass spectrometry analysis. With this orthogonal separation scheme, a complex mixture of thousands of proteins can be separated in a single experiment. Another noteworthy advantage of this technique is that 2D plates can serve as a medium to store the separated proteins. On the downside, it is time consuming, labor intensive and expensive, and lacks reproducibility and the ability for automation.

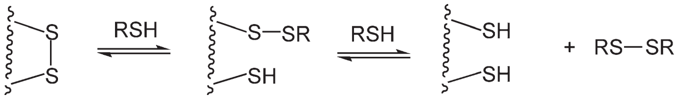
9.12.2.4.2 Multidimensional liquid chromatography

Although liquid chromatography techniques have become quite popular in the separation of peptides in complex protein digests, they are yet to make an impact for the separation of protein samples for proteome-wide applications. It is envisioned that in the future their application for protein separation will increase. Various combinations of reversed-phase (RP)-HPLC with ion-exchange, size-exclusion, chromatofocusing (CF), IEF, and capillary electrophoresis (CE) have emerged for 2D separation of complex mixtures of proteins and peptides. A recent addition in this field is the use of CF as the first dimension and RP-HPLC as the second-dimension separation device.[14] CF is a column-based liquid-phase separation technique, in which proteins are fractionated on the basis of differences in their p*I* values in a weak ion-exchange column.

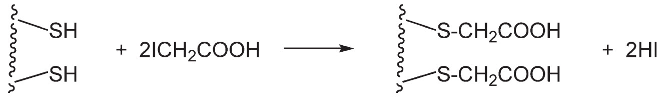
9.12.2.5 Site-Specific Cleavage of Proteins

The next step in proteome analysis is the cleavage of proteins into smaller peptides. This step is necessary, because current mass spectrometry instruments are better equipped for the analysis of small peptides with respect to accuracy, sensitivity, and tandem MS capability. With current instruments, the useable sequence ion data can be obtained from peptides with a chain length of 6–20 amino acids.

Auto-oxidation of cysteine residues during cleavage of the disulfide bridge-containing proteins is a potential concern. This concern can be addressed by first reducing those proteins at alkaline pH (~8.0) with either 2-mercaptoethanol or dithiothreitol (Equation (1)) and then alkylating with iodoacetic acid to *S*-carboxymethyl derivatives (Equation (2)) The reduction–alkylation process also disrupts the 3D structure of proteins to allow more sites accessible for cleavage.



(1)



(2)

A diverse collection of chemical reagents and endoproteases with broad specificity has become available for cleaving proteins (see **Table 1**). An ideal cleaving agent is the one that provides the highest yields of peptides of

Table 1 Typical protein-cleaving agents

Cleaving agent	*Specificity*	*Digestion conditions (buffer; pH; temperature (°C))*
Chemical agents		
Cyanogen bromide	Met-X	70% TFA
N-chlorosuccinimide	Trp-X	50% acetic acid
N-bromosuccinimide	Trp-X	50% acetic acid
Highly specific proteases		
Trypsin	Arg-X, Lys-X	50 mmol l^{-1} NH_4HCO_3; 8.5; 37
Endoproteinase Glu-C	Glu-X	50 mmol l^{-1} NH_4HCO_3; 7.6; 37
Endoproteinase Arg-C	Arg-X	50 mmol l^{-1} NH_4HCO_3; 8.0; 37
Endoproteinase Lys-C	Lys-X	50 mmol l^{-1} NH_4HCO_3; 8.5; 37
Endoproteinase Asp-N	X-Asp	50 mmol l^{-1} NH_4HCO_3; 7.6; 37
Nonspecific proteases		
Chymotrypsin	Phe-X, Tyr-X, Trp-X, Leu-X	50 mmol l^{-1} NH_4HCO_3; 8.5; 37
Thermolysin	X-Phe, X-Leu, X-Ile, X-Met	50 mmol l^{-1} NH_4HCO_3; 8.5; 37 X-Val, X-Ala
Pepsin	Phe-X, Tyr-X, Trp-X, Leu-X,	0.01 mol l^{-1} HCl; 2.0; 37 Met-X
Elastase	Broad specificity	5 mmol l^{-1} NH_4HCO_3; 37

Reproduced from C. Dass, *Fundamentals of Contemporary Mass Spectrometry*; Wiley-Interscience: Hoboken, NJ, 2007, with permission from Wiley-Interscience, Copyright 2007.

optimal length. Cyanogen bromide (CNBr) is the most frequently used chemical agent that has the specificity to cleave an amide bond on the C-terminal side of methionine. A downside of this reaction is that owing to infrequent occurrence of methionine in proteins, the CNBr-cleaved segments are often large, which may not be amenable to mass spectrometry analysis. CNBr converts methionine to homoserine, which at acidic pH exists as a lactone, the mass of which is 48.1 Da less than the methionine residue. *N*-chloro- and *N*-bromosuccinimide are the other two chemical agents that have been used in place of CNBr; their cleavage specificity is the bond C-terminal to tryptophan.

Among the endoproteases listed in **Table 1**, trypsin is more or less a universal choice for mass spectrometry-based proteome analysis; it cleaves a bond on the C-terminal side of lysine and arginine, thus producing peptides that contain a basic residue at the C-terminus. Ionization of tryptic peptides by ESI mainly produces doubly charged ions, which upon collision-induced dissociation (CID) yields highly useable sequence ion information. Digestion with endoproteinase Lys-C, endoproteinase Arg-C, endoproteinase Glu-C (also known as *Staphylococcus aureus* V8 protease), or endoproteinase Asp-N can yield complementary larger size peptides. The first three enzymes have specificities to the bond C-terminal to lysine, arginine, and glutamic acid residues, respectively, whereas endoproteinase Asp-N cleaves the amino side of aspartic acid. Another enzyme that cleaves the N-terminal side of the amide bond is thermolysin, which has broad specificity for leucine, isoleucine, methionine, phenylalanine, and tryptophan.

Digestion is usually performed in a solution at specified conditions of pH, temperature, and buffer (see **Table 1**) and in a denaturing environment to ensure complete endpoint digestion. Volatile buffers such as ammonium carbonate and ammonium bicarbonate are preferred because they can be easily removed by lyophilization. A practical method for the removal of a nonvolatile buffer and salts is to use solid-phase extraction (SPE) cartridges prior to mass spectrometry analysis. One can also use immobilized trypsin packed into a small-diameter PEEK (polyetheretherketone) column or covalently attached to an activated MALDI probe for on-probe digestion.[15]

Because 2-DE is a premier technology for separation of proteins, extra attention has been paid to optimize cleavage of gel-separated proteins. 'In-gel' digestion is the leading approach. The spot is excised, destained, and reacted with trypsin; the resulting peptides are eluted by washing the spot. SDS and other contaminants need to be removed prior to mass spectrometry analysis. The other two options are proteolysis of intact proteins after their elution from the gel or after electroblotting them on a polymer membrane such as poly(vinylidene difluoride), carboxymethylcellulose, or Immobilon-CD. The immobilized proteins are visualized and destained. The spot is excised from the membrane, digested with a protease, and the peptides produced are extracted for mass spectrometric analysis. The use of membranes provides a high recovery and cleaner chemical background during mass spectrometry analysis.

9.12.3 Mass Spectrometry Analysis of Proteomes

9.12.3.1 Bottom-Up and Top-Down Proteomics

Mass spectrometry technologies for proteome analysis can be classified in two broad categories (**Figure 2**): (1) 'bottom-up' proteomics and (2) 'top-down' proteomics.[16] The former, discussed in detail in the next section, is the mainstream proteomics approach where in which mass spectrometry is performed on the peptide level to obtain the molecular mass or amino acid sequence of the cleaved peptides. In top-down proteomics proteins are not cleaved into smaller segments; instead, the intact protein is analyzed using Fourier-transform (FT)–ion cyclotron resonance (ICR)-based high-resolution tandem mass spectrometers to obtain the molecular mass and amino acid sequence. The protein is characterized through database search using these compound-specific parameters. The top-down approach is better suited for revealing the identity of modified proteins.

9.12.3.2 Molecular Mass Measurement of Proteins

The molecular mass of an intact protein provides a frame within which the final structure must fit. The molecular mass information is an important analytical parameter required in many situations. Top-down proteomics relies on the molecular mass of intact proteins. This information is also needed to verify the correctness of the translated sequence of a protein and to identify point mutations, to find the extent of PTMs, and to determine the number of cysteine residues and disulfide bonds. ESI and MALDI are the methods of

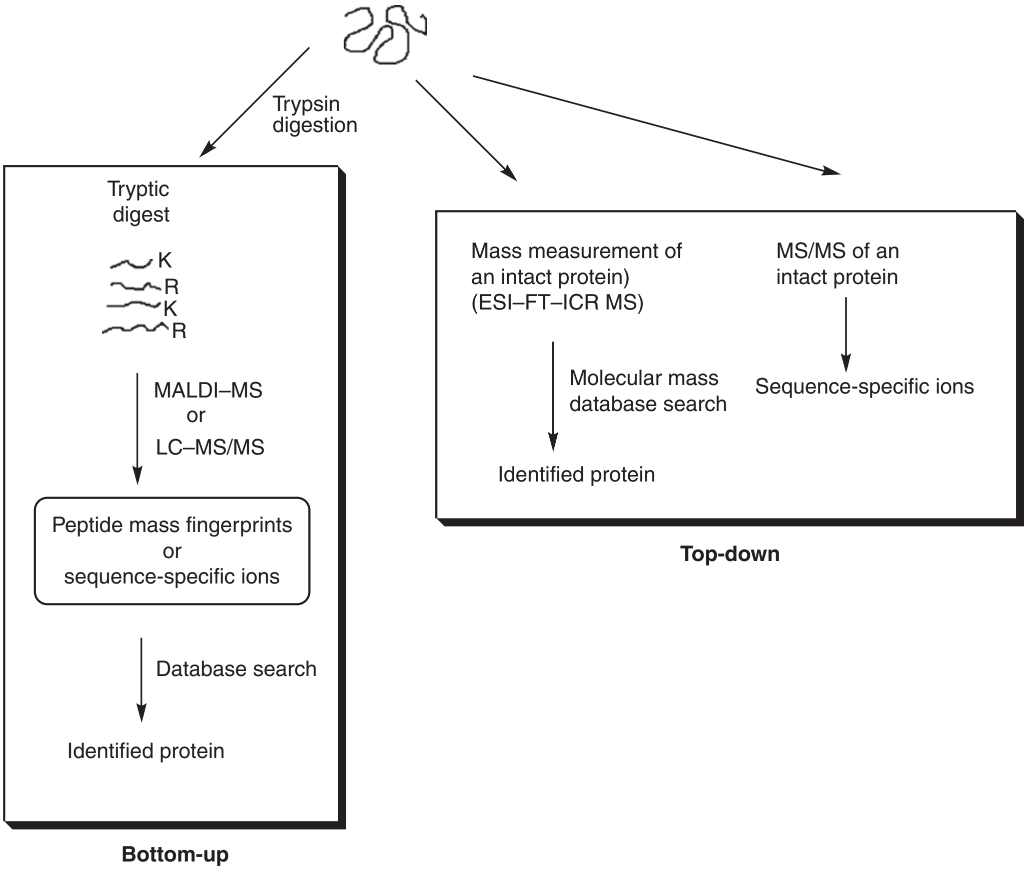

Trypsin digestion
Tryptic digest
K
R
K
R
MALDI–MS or LC–MS/MS
Peptide mass fingerprints or sequence-specific ions
Database search
Identified protein
Bottom-up
Mass measurement of an intact protein) (ESI–FT–ICR MS)
Molecular mass database search
Identified protein
MS/MS of an intact protein
Sequence-specific ions
Top-down

Figure 2 Basic principles of bottom-up and top-down proteomics. Reproduced from C. Dass, *Fundamentals of Contemporary Mass Spectrometry*; Wiley-Interscience: Hoboken, NJ, 2007, with permission from Wiley-Interscience, Copyright 2007.

choice for determining the molecular mass of intact proteins. The ESI of high-mass biopolymers produces an envelope of multiply charged ions of proteins, the deconvolution of which gives the accurate molecular mass of the target protein. MALDI produces a simpler spectrum that contains mainly the singly protonated ion. Direct mass measurement of proteins from 2D-gel plates is also feasible.

9.12.3.3 Bottom-Up Proteomics

A general outline of the ‘bottom-up’ mass spectrometry approach for proteome analysis is presented in **Figure 3**. In general, the mass spectrometry is performed at the peptide level after digesting the protein to obtain the molecular mass and amino acid sequence-specific ions, which are correlated with similar information in the protein or nucleotide database.[7,16] Based upon these measurements, the following approaches have evolved.

9.12.3.3.1 Peptide-mass fingerprinting

The concept of protein identification by peptide-mass fingerprinting (also known as peptide-mass mapping) is simple (see **Figure 3**(**a**)).[17] When a protein is digested with a protease, a set of peptides, unique to this protein, is produced.[17–24] The molecular mass of each peptide is accurately determined. All entries in a protein database are virtually treated with the same protease to provide a long list of theoretical peptide masses. If a nucleotide database is to be searched, then all entries in it need to be first translated into expressed proteins. The correlation of the measured mass of a peptide with the theoretical peptide masses in the database will identify the protein that was the precursor of the target peptide. Quite often, a database search provides several matches (‘hits’) that are close to the real peptide/protein. The number of hits can be reduced by increasing the accuracy of mass measurement and increasing the number of peptides in the search. Usually the peptide masses should be measured within ±0.05 Da of

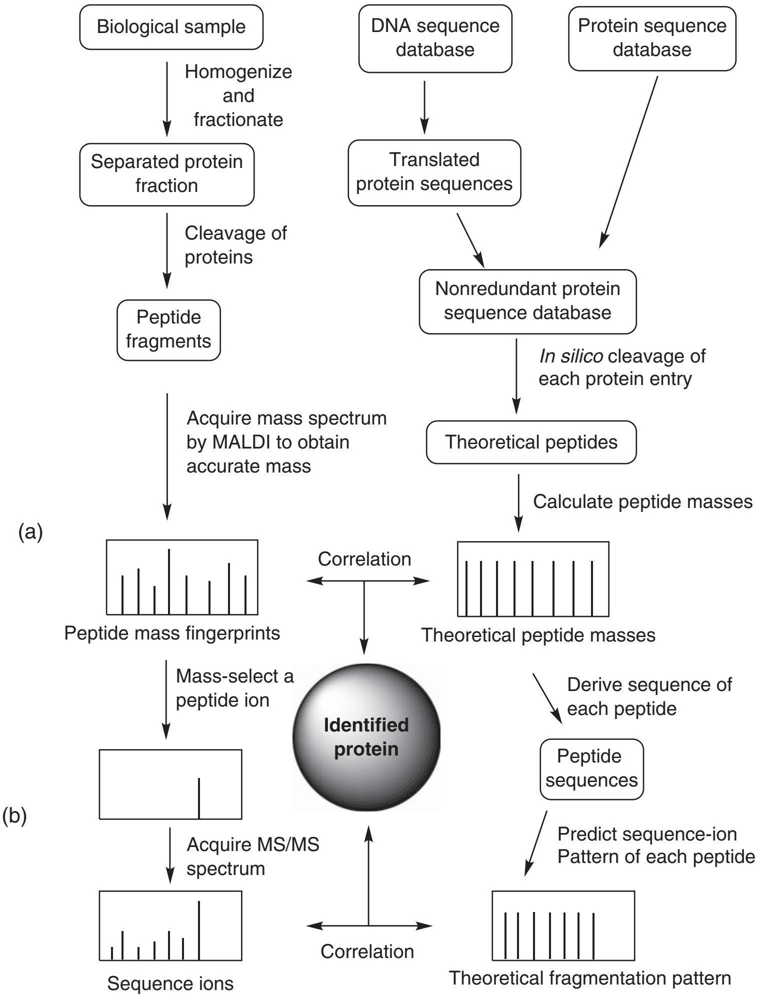

Biological sample
Homogenize and fractionate
Separated protein fraction
Cleavage of proteins
Peptide fragments
Acquire mass spectrum by MALDI to obtain accurate mass
DNA sequence database
Translated protein sequences
Protein sequence database
Nonredundant protein sequence database
In silico cleavage of each protein entry
Theoretical peptides
Calculate peptide masses
(a)
Correlation
Peptide mass fingerprints
Theoretical peptide masses
Identified protein
Mass-select a peptide ion
Derive sequence of each peptide
Peptide sequences
(b)
Acquire MS/MS spectrum
Predict sequence-ion Pattern of each peptide
Correlation
Sequence ions
Theoretical fragmentation pattern

Figure 3 A general outline of proteomics approach for profiling proteins: (a) peptide fingerprinting-based proteomics, (b) sequence ion-based proteomics. Reproduced from C. Dass, *Principles and Practice of Biological Mass Spectrometry*; Wiley-Interscience: Hoboken, NJ, 2000, with permission from Wiley-Interscience, Copyright 2001.

the actual mass; at least four to five peptide masses need to be included in the search. At the other extreme, if no match is found, then it is a new protein. MALDI–MS is the preferred method for measuring the masses of the proteolytic peptides because of the advantage that the digests can be analyzed directly without prior fractionation. Because of this, peptide-mass fingerprinting and 2-DE–MALDI–TOF (time-of-flight)–MS combination has been the most visible platform for proteome analysis for a long time the most visible platform for proteome analysis.[21,25,26] The analysis of protein digests can also be performed with an online combination of ESI and capillary LC.[27–29]

9.12.3.3.2 Peptide sequence tags

In essence, this approach is similar to peptide-mass fingerprinting but with a fundamental difference that instead of measuring the molecular mass, sequence ion information of the cleaved peptides is obtained using tandem mass spectrometry (MS/MS). The observed fragment ion pattern is matched with the predicted patterns of fragment ions of peptides derived from all protein entries in the database.[29–32] Because of the higher level of structure specificity of the sequence ion data, a high probability assignment of proteins is possible with sequence ion information of just one peptide. MALDI and ESI are both appropriate as ionization methods. MALDI is optimally used with gel-separated proteins. The gel spots are digested as above and are deposited onto a large-format MALDI plate. The matrix is added, and the spotted samples are analyzed on any of the tandem MS instrument (TOF/TOF, ion traps, Q-TOF, etc.).[33]

9.12.3.3.3 Peptide sequence tags with multidimensional protein identification technology

MudPIT (multidimensional protein identification technology) is a gel-free 'shotgun' proteomics approach in which upstream protein separation is not done (Figure 4).[34–36] Instead, the entire proteome is tryptic digested and analyzed directly with an online 2D-LC–ESI–MS/MS system to obtain the molecular mass and amino acid sequence ions pattern of the eluting peptides. The sequence ions data are correlated with the simulated fragment ions pattern obtained from *in silico* tryptic digest of all protein entries in a protein database or of all translated protein entries in a given nucleotide database. Usually, the peptide digest is very complex; it might contain more than 20 000 peptides; thus, for optimal detection, a high degree of fractionation of peptides is required. Such a need can be met with a powerful multidimensional LC system consisting of two orthogonal separation modes. As shown in Figure 4, the most frequently used system consists of a strong cation-exchange (SCX) capillary column, where peptides are resolved on the basis of their charge, coupled to a C-18 RPLC capillary column. In this biphasic SCX–RPLC system, separation occurs through the concerted effort of both columns: first, the digest is passed through the SCX column, and a small fraction of the trapped peptides is eluted onto the RPLC column using a step gradient of increasing salt concentration. The RPLC column resolves this fraction using an acetonitrile gradient. At the end of the gradient, the RPLC column is re-equilibrated. Next, the salt gradient is stepped up to transfer another fraction of peptides from the SCX column. This process is repeated until the entire peptide digest trapped in SCX has been analyzed. The eluent from the RPLC column enters directly into the ESI–MS system. Mass spectrometry is performed in the data-dependent mode, which involves switching the data system to the MS/MS mode when a doubly or triply charged ion is detected in the normal m/z scan. This procedure has increased the dynamic range for the detection of low-abundance proteins in complex mixtures. The other successful combination uses hydrophilic interaction chromatography in place of SCX; this mode is especially useful for resolving phosphopeptides.[37] A triphasic RPLC–SCX–RPLC system has also been used.[38] Several off-line multidimensional LC systems have been developed, including SCX–RPLC, IEF–RPLC, affinity–RPLC, and RPLC–RPLC,[39] with the objective to provide greater peak capacity, flexibility, and dynamic range of detection. The tandem mass spectrometry instrumentation includes quadrupole ion trap (QIT), linear-ion trap (LIT), or any hybrid system (Q-TOF, LIT–FTICR, etc.).

9.12.3.3.4 Accurate mass tags

The underlying basis of accurate mass tags (AMT) is when the mass of a peptide fragment is measured with <1 ppm accuracy, then the mass of a single peptide becomes a unique identifier of the protein.[40] This kind of a high degree of mass accuracy can be obtained with FT–ICR instruments. A higher level of specificity is imparted to the measurement when the retention time or the p*I* values obtained from the LC/MS or IEF–MS analysis, respectively, are also used along with the measured molecular mass.[41,42]

9.12.3.3.5 Database search

A variety of protein/DNA databases, such as GenBank, EMBL, NCBI, GenPept, Swiss-Prot, TrEMBL, PIR, OWL, IPI, and dbEST, are maintained by independent research groups for use by the public for proteome analysis. Databases have links to other databases and also provide vital information related to the identified proteins such as functions, any PTMs, domain and sites, 3D structures, homology to other proteins, associated diseases, sequence conflicts, and variants.

Several software search engines, including MS-Fit, MOWSE, Prot-ID, Expasy tools, ProFound, Mascot, and PeptideSearch, are available for use with the peptide-mass fingerprinting data. During the search process, the algorithm matches the measured molecular masses of peptides in the query against the list of theoretical mass

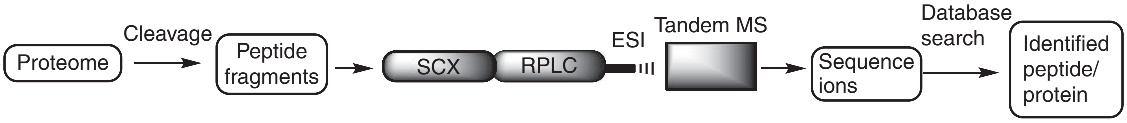


Figure 4 Depiction of shot-gun proteomics using multidimensional protein identification technology; a complex mixture of peptide fragments in the digest are resolved by a combination of ion-exchange and reversed-phase liquid chromatography.

values of peptides that are created by *in silico* digestion of each protein entry in the database with the same protease that was used in the experimental step. In a typical search algorithm, the cleavage database is first selected (it can be a protein database, gene sequence database, or both) and to limit the number of hits, the criteria of the search are defined.[19,22] The parameters entered in the search criteria are the origin of the sample (i.e., species), the mass-matching tolerance, an approximate upper value of the molecular mass of the protein, the p*I* value of the protein, a minimum number of matches, the number of cleavage sites that might have been missed by the trypsin digestion, and the type of cysteine modification. The output of the search gives a ranked list of most likely candidates (or hits); the entry that produces the best score has the highest probability of being the protein. Mismatches of the listed peptide candidates can reveal the possibility of mutation, PTMs, or coeluting/comigrating proteins. Four to six peptide masses are usually sufficient for a successful search.[19,24] The mass-matching tolerance should be below 3 ppm and the mass measurement accuracy within ±0.05 Da.

Among the plethora of search algorithms (MASCOT, SEQUEST, ProFound, MS-Tag, PeptideSearch, X!Tandem, X!Hunter, OMSSA, Parallax, DBDigger, GlobalLynx, Paragon, Spectrum Mill, SpectraST, Bonanza, BiblioSpec, and more) that have been developed for use with the sequence ion data, MASCOT, X!Tandem, and SEQUEST are the leading ones. In this search, the algorithm matches the observed fragment ion pattern with the theoretical fragment ion patterns that are expected from the database entries. Take the example of the SEQUEST algorithm, in which the search proceeds as follows: first, the algorithm selects a peptide ion and simplifies its acquired MS/MS spectrum. Next, it identifies all possible peptide sequences in the database that may fit with the measured mass of this selected peptide ion. The algorithm then predicts theoretical sequence ions patterns of all those peptides.[30,32] The spectra are subjected to fast Fourier transform, and each virtual spectrum is matched with the experimentally observed sequence ion spectrum; the peptide with a highest cross-correlation score is reported. This program can also identify PTMs by assuming that each putative modification site is modified and unmodified in one pass through the database. One must, however, remember that the computer programs for characterizing peptides using databases are not nearly 100% reliable, so the identity of the peptide must be verified through manual *de novo* sequencing as a cross-check, even when the protein is found in the database.

9.12.4 Protein Expression Profiling through Quantitative Proteomics

The expression level of proteins in cells is constantly in flux owing to changes in biochemical pathways. Differential expression of proteins is also the result of different physiological states (e.g., healthy versus diseased subjects) or stress due to external stimuli (e.g., by drugs or toxins). These changes can be monitored through differential proteomics; also called quantitative proteomics.[43] The task involves identification of proteins and comparison of their expression level in control versus test samples. This subfield of proteomics has also opened the avenues to discover biomarkers of diseases and opportunities for early diagnostic intervention and prevention of diseases. Protein–protein interactions can also be explored with quantitative proteomics.

9.12.4.1 Quantification of Proteins by 2D Gel Electrophoresis

This is an entirely gel-based procedure; no mass spectrometry is involved. Proteins from the two comparison samples are separated by a 2-DE protocol and images of the stained spots are captured after staining the gels with a suitable dye (e.g., CCB or SYPRO Ruby); next, the measured densities of the spots are compared using image analysis software to provide differential expression of proteins.[44,45] Identification of proteins is accomplished as mentioned in Section 9.12.3.

9.12.4.2 Quantification of Proteins by 2D-Differential Imaging Gel Electrophoresis

2D-differential imaging gel electrophoresis (DIGE) is also a gel-based procedure, but one that is more precise than the 2-DE approach. A key element of this procedure is labeling of proteins in the two experimental samples with different colored amine reactive fluorescent dyes.[46,47] A common procedure is to label the control sample with Cy-5 cyanine dye and the test sample with Cy-3 cyanine dye. These two dyes display fluorescence at 670 and 570 nm, respectively. The two samples are then mixed in equal concentrations and separated simultaneously on a single 2-DE gel. This eliminates inevitable variations that are encountered due to differing experimental conditions when the two samples are analyzed separately, as in the 2D-gel imaging protocol. The separated protein spots are visualized and comatched by scanning the gel at two different wavelengths.

9.12.4.3 Quantification of Proteins by Isotope-Coded Affinity Tags

This method is one of the earliest attempts for quantifying the relative levels of proteins in two samples with mass spectrometry.[48] It is applicable to cysteine-containing proteins only; fortunately, this limitation is not a big handicap because nearly 90% of proteins contain cysteine. The underlying principle of this technique is illustrated in **Figure 5**. Essentially, both samples are treated with isotope-coded affinity tag (ICAT) reagents to selectively alkylate cysteine residues; one sample is alkylated with light hydrogens (d_0) and the other with heavy hydrogens (d_8). The two samples are combined and tryptic digested. Next, the labeled cysteine peptides are affinity purified on an avidin column. Mass spectrometry is performed in the LC–ESI–MS and LC–ESI–MS/MS modes. The ratio of the ion abundances of the d_0- versus d_8-labeled peptides is a quantitative measure of the two proteomes and the MS/MS spectra identify those peptides/proteins. The ^{13}C-, ^{15}N-, and ^{18}O-labeled reagents have also been used in place of deuterium-labeled reagents.

9.12.4.4 Quantification of Proteins by iTRAQ Reagents

In this procedure, a set of four isobaric amine-specific labeling reagents (114, 115, 116, or 117), consisting of a reporter group, a balance group, and a peptide-reactive group (e.g., *N*-hydroxy succinimide) are used to label

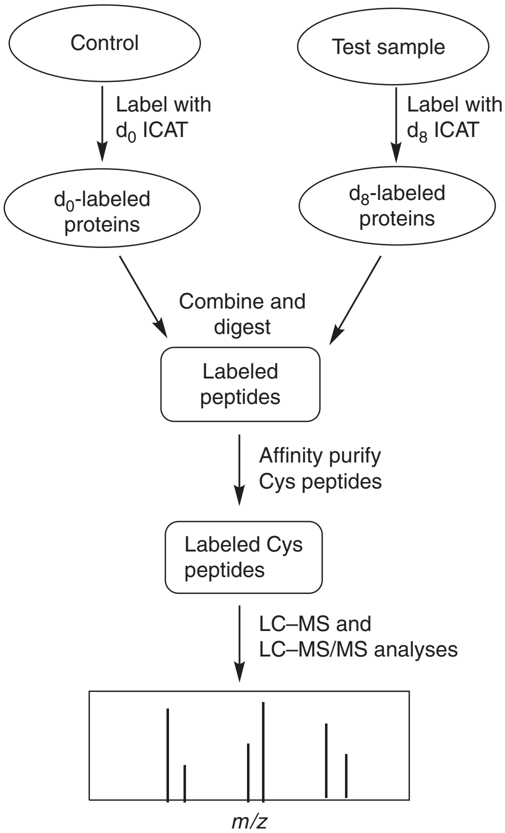


Figure 5 Illustration of ICAT quantification procedure for the analysis of differentially expressed proteins. Reproduced from C. Dass, *Fundamentals of Contemporary Mass Spectrometry*; Wiley-Interscience: Hoboken, NJ, 2007, with permission from Wiley-Interscience, Copyright 2007.

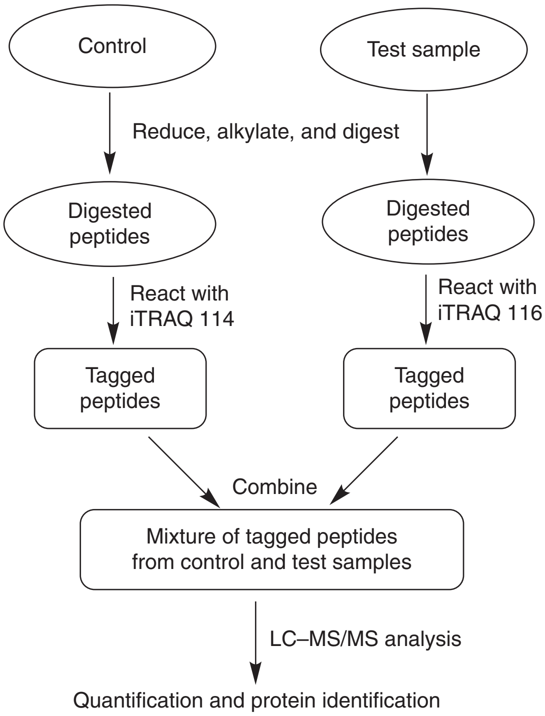


Figure 6 Principle of iTRAQ quantification. Reproduced from C. Dass, *Fundamentals of Contemporary Mass Spectrometry*; Wiley-Interscience: Hoboken, NJ, 2007, with permission from Wiley-Interscience, Copyright 2007.

peptides; their use allows multiplexing of up to four different samples in a single experiment.[49–51] Unlike ICAT reagents, labeling occurs at the peptide level. The iTRAQ protocol is illustrated in **Figure 6**. Proteins in the control and test samples are enzymatically digested, and the digests are treated with iTRAQ reagents to chemically tag the primary amino group of lysines and the N-terminus of peptides. The two labeled samples are combined and analyzed by LC–ESI–MS/MS. The spectrum will contain eight distinct reporter ions; measuring their relative abundances enables quantification of the peptides/proteins. Database searching of the sequence ion data provides identification of the proteins.

9.12.4.5 Quantification of Proteins by the Proteolytic ^{18}O-Water–Labeling Approach

In this method, the two C-terminal carbonyl oxygens of the peptides are labeled with light or heavy oxygen isotopes. Labeling occurs during trypsin digestion of the protein samples either in ordinary water or ^{18}O-water.[52] Peptides from the two digests are combined and analyzed by LC–MS; the ratio of peak heights in an extracted-ion chromatogram of the coeluting labeled and unlabeled peptides is the measure of relative levels of proteins in control and test samples. One potential advantage of this procedure is that each tryptic peptide, except the one at the C-terminal, gets labeled with ^{18}O. As a consequence, quantification can be performed with multiple targets, resulting in better precision and confidence.

9.12.4.6 Quantification of Proteins with Stable-Isotope Labeling by Amino Acids in Cell Culture

In this method, protein samples are labeled during growth of the cell culture with fortification of the culture media with the stable isotope analogs of essential amino acids.[53] In practice, cell lines are grown in a medium that is enriched with either light isotope-containing or heavy isotope-containing amino acids (e.g., Lys or d_4-Lys). The frequently used heavy isotope-labeled amino acids are ^{15}N-Lys, ^{13}C^{15}N-Lys, ^{13}C-Lys, ^{15}N-Arg, ^{13}C-Arg, and ^{13}C-Leu. As shown in **Figure 7**, the two labeled cell cultures are mixed in a 1:1 ratio and proteins

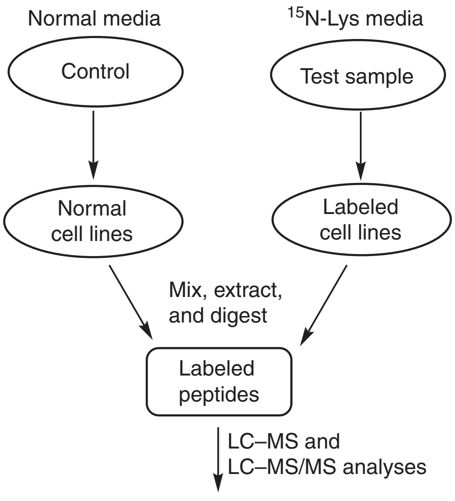


Figure 7 Protocol for quantification of proteins by the SILAC method. Reproduced from C. Dass, *Fundamentals of Contemporary Mass Spectrometry*; Wiley-Interscience: Hoboken, NJ, 2007, with permission from Wiley-Interscience, Copyright 2007.

are extracted from the mixed culture. After proteins have been purified and digested, the peptides are analyzed by LC–MS or LC–MS/MS. Signal intensities of the labeled and unlabeled peptides allow proteins to be quantified.

9.12.4.7 Quantification of Proteins by the Label-Free Approach

It is also feasible to quantify proteins without resorting to isotope labeling.[54] Signal intensities of the peptide ions in the extracted-ion chromatograms from the two test samples can provide a direct measure of the relative concentrations of proteins, if it is ensured that the ions selected in the two samples are from the same peptide. To this end, the sample handling and LC–MS analysis conditions must be strictly controlled to assure that the ions selected have precisely the same m/z, charge, and retention time.

9.12.4.8 Absolute Quantification of Proteins

Determining the absolute amount of an analyte in a sample is the ultimate goal of any quantification method. AQUA (which stands for absolute quantification) is a step in this direction, but the present state of technology allows quantification of only one or two target proteins at a time from a single sample.[55] Conceptually, the procedure is similar to that used for quantification of small molecules and requires the use of a stable isotope-labeled internal standard (IS). A ‘signature peptide’ unique to the target protein is chosen as an analyte and an analog of it that contains heavy isotope-labeled amino acids (e.g., ^{13}C and ^{15}N) is synthesized for use as an IS. The sample is spiked with a known amount of this IS and analyzed using a capillary LC–ESI–MS/MS system operating in the selected-reaction monitoring (SRM) mode. The amount of the native peptide is calculated by correlating the ratio of the areas of the target peptide peak and the IS peak with a known amount of the added IS.

9.12.5 *De Novo* Sequencing of Peptides

The amino acid sequence determination of peptides through *de novo* peptide sequencing procedure is one of the most familiar applications of mass spectrometry. A precise knowledge of the amino acid sequence of peptides is required in many situations – to understand their biological functions, to characterize components of the metabolic cycle of precursor proteins, to map changes in the metabolic profile of a peptide family caused by pathological stress or therapeutical treatment, to assess the purity of newly synthesized peptides, and to perform sequencing of newly discovered proteins or of those that are not present in the database, for example. Also, in some cases, manual *de nova* sequencing is required to evaluate the results of an automatic database search.

De novo sequencing is the term used for peptide sequencing performed without prior knowledge of its amino acid sequence and includes the following experimental steps: (1) the converting peptides into gas-phase ions, (2) inducing dissociation of those ions into amino acid sequence-specific ions, and (3) deciphering the mass spectrum into the sequence of the peptide. The last step is very tricky and requires a reasonable understanding of the peptide fragmentation rules. Guidelines for interpretation of the mass spectrum. Computer software can also be used to automate spectral interpretation.[56]

9.12.5.1 Peptide Fragmentation Rules

In the positive-ion mode, a convenient means of ionization by MALDI or ESI is the protonation of peptide molecules; MALDI yields monoprotonated ions, whereas, ESI produces multiply protonated ions, depending upon the number of basic residues in a peptide. The positive charge is initially localized at the amino group of the N-terminus or at the side chain of one of the basic residues. The basic residue proton is strongly stabilized, but the N-terminal proton easily migrates to any one of the peptide backbone amide nitrogens through a process known as 'internal solvation', resulting in a heterogeneous population of peptide ions, each component having the same sequence but differing in the site of protonation.

Energy deposited during ionization or ion activation causes cleavage of bonds all along the peptide backbone.[57–64] A recent tutorial on gas-phase fragmentation of peptides is worth reading.[64] A major driving force that determines which bond in the backbone must dissociate is the site of proton attachment.[65] The peptide backbone consists of three different types of bonds: the alkyl carbonyl bond (–CHR–CO), the peptide amide bond (–CO–NH), and the amino alkyl bond (–NH–CHR). In theory, cleaving these bonds leads to six types of fragment ions, three from each terminal of the peptide (**Figure 8**). According to the conventional nomenclature, the N-terminal charged fragments formed by cleavage of the CHR–CO, CO–NH, and NH–CHR bonds are designated by the symbols a_n, b_n, and c_n, respectively, and the corresponding C-terminal ions by x_n, y_n, and z_n, where n refers to the number of amino acid residues from the respective peptide terminus.[66,67] The *b*- and *y*-type ions are the key elements in the interpretation of a peptide spectrum as they are invariably formed during CID of peptides. Often, *a* ions are also formed by the loss of CO (28 Da) from *b* ions (see **Figure 9**). A simplistic view of the formation of *b*- and *y*-ions from a singly protonated peptide is depicted in **Figure 9** and from a doubly protonated tryptic peptide in **Figure 10**. Although the stable form of *b*-ions is the protonated oxazolone,[61,68] they are usually represented as acylium ions, for simplicity, that is, as the truncated peptide minus the OH group; both forms are shown in **Figure 9**. The *y*-ion is also represented as the truncated peptide plus a proton. Doubly protonated tryptic peptides are more useful for sequence analysis of peptides because they fragment in a more predictable manner throughout the peptide backbone; cleavage of the peptide bond leads to the formation of singly charged *b*- and *y*-ions (pathway *a*, **Figure 10**; the *b*-ion in this figure is shown as an acylium ion and not as a protonated oxazolone). In addition, because of localization of one charge at the C-terminal basic residue, these ions also fragment to produce doubly charged *y*-ions and neutral moieties (pathway *b*, **Figure 10**), but to a lesser degree.

In addition to *b*- and *y*-ions, high-energy CID of peptides might also produce the *z*- and *c*-type ions, as well as the second-generation products w_n, v_n, and d_n; the *w*-ions are formed from *z*-ions via cleavage of the β,γ-bond, *v*-ions from *y*-ions via cleavage of the α,β-bond (i.e., cleavage of the entire side chain), and *d*-ions from $(a+1)$-ions via cleavage of the β,γ-bond.[63,69] Although not typical, they have also been observed in a full-scan mass spectrum.[60,70] The *w*- and *d*-ions have been used as a probe to differentiate between Leu and Ile.[63]

Simultaneous cleavage of two bonds in the peptide backbone is also a common phenomenon, which leads to the formation of two additional types of ions, immonium ions and internal fragments, which may have either acylium- or immonium-type structure (see **Figure 11**). Immonium ions are usually observed at less than 200 m/z values. These ions might also undergo fragmentation to produce the second-generation ions.[71] Although immonium ions cannot predict the sequence of the peptide, they provide a useful clue to the presence of certain amino acids in the sequence (see **Table 2**).

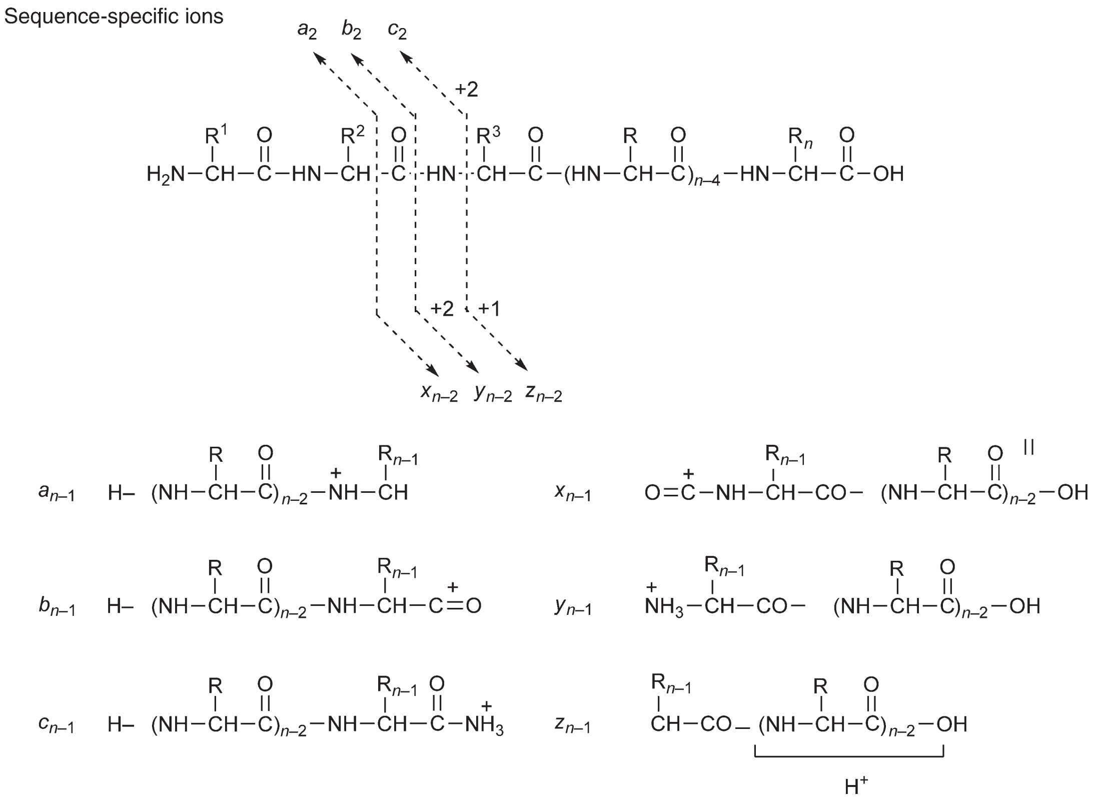


Figure 8 The nomenclature and structure of sequence-specific peptide ions. Reproduced from C. Dass, *Principles and Practice of Biological Mass Spectrometry*; Wiley-Interscience: New York, 2000, with permission from Wiley-Interscience, Copyright 2001.

Diagnostic losses of certain neutral species are commonly observed from *b*- and *y*-ions. For example, Ser, Thr, Asp, and Glu side chains exhibit prominent loss of water (–18 Da), and the Asn-, Gln-, Lys-, and Arg-containing ions similarly show abundant loss of ammonia (–17 Da). The loss of 48 Da ($HSCH_3$) is observed from Met-containing sequence ions, but if Met is in oxidized form, the mass of the expelled neutral ion becomes 64 Da. The Cys-containing ions show a loss of 34 Da (H_2S), which shifts to 92 Da ($HSCH_2COOH$) if the precursor protein is alkylated with iodoacetic acid prior to digestion. The peptides that contain a basic residue at the C-terminus are likely to expel the C-terminal residue to produce the (b_{n-1} + OH) ion.

Certain features in a peptide have an overwhelming influence on the yield of a particular type of sequence ion. One cardinal rule is that the more easily cleaved sites are those that can be easily protonated. For example, because acidic amino acid side chains – glutamate and aspartate – can easily stabilize a positive charge, the adjacent bond has a high proclivity for cleavage. The presence of a residue with relatively high basicity near either terminus of a peptide helps in charge retention by that terminal fragment. Tryptic peptides by virtue of arginine or lysine at the C-terminus produce more intense *y*-ion series than *b*-ion series. If arginine is present at the N-terminus, then *b*-ion series will be more prominent, and if it is in the middle of the sequence, there may be absence of fragmentation in the bonds around arginine. Cleaving the bond C-terminal to proline is less likely to occur, leading to a reduced abundance or absence of corresponding *b*- and *y*-ions, but cleavage of the bond N-terminal to proline is more facile.[68,72–74] Proline is an unusual naturally occurring amino acid; its side chain exists as a cyclic structure by forming a bond with its nitrogen atom. This cyclic structure prevents cleavage of the bond C-terminal to proline. In forming the oxazolone structure, the oxygen atom of the preceding carbonyl group combines with the carbonyl carbon of the amide bond undergoing cleavage (see Figure 9). As shown in Figure 12, this process in proline would lead to the formation of a strained [3,3,0] bicyclic system, which reduces the possibility of cleavage to its C-terminus but encourages cleavage to its N-terminus. Glycine also behaves in a similar manner.

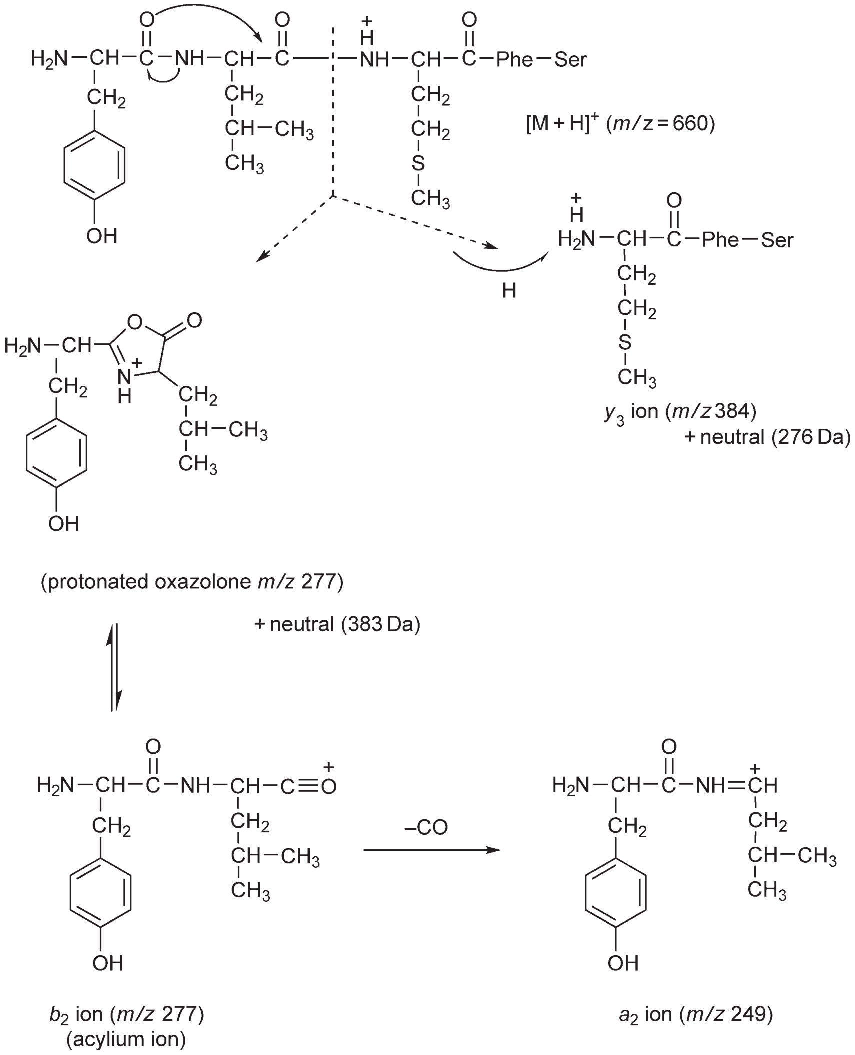


Figure 9 Formation of *b*- and *y*-ions from singly protonated peptides.

9.12.5.2 Ion-Activation and Sequence Determination of Peptides

The following ion-activation techniques have been used at one time or other to sequence peptides: (1) fast atom bombardment (FAB) ionization, (2) CID–tandem MS (MS/MS), (3) ESI in-source CID, (4) MALDI ion-source decay, (5) MALDI postsource decay (PSD), (6) electron-capture dissociation (ECD) and electron-transfer dissociation, and (7) peptide ladder sequencing. Because of the lack of space, only (2) and (4) will be discussed further.

9.12.5.2.1 Collision-induced dissociation-tandem mass spectrometry

Tandem mass spectrometry is the leading technique in sequence determination of peptides and other biomolecules.[57,59,61–65] The technique involves mass selection of the target peptide ion, fragmentation of the mass-selected ion, and *m*/*z* analysis of the fragment ions. To improve the fragment ion yield, it is advantageous to use external ion activation of the mass-selected peptide ions. A widely used and easier to implement process is CID, which involves collisions of the fast-moving peptide ions with neutral inert gas atoms. Instruments are available in which MS/MS spectra are acquired through high-energy CID, low-energy CID, or no CID. Other

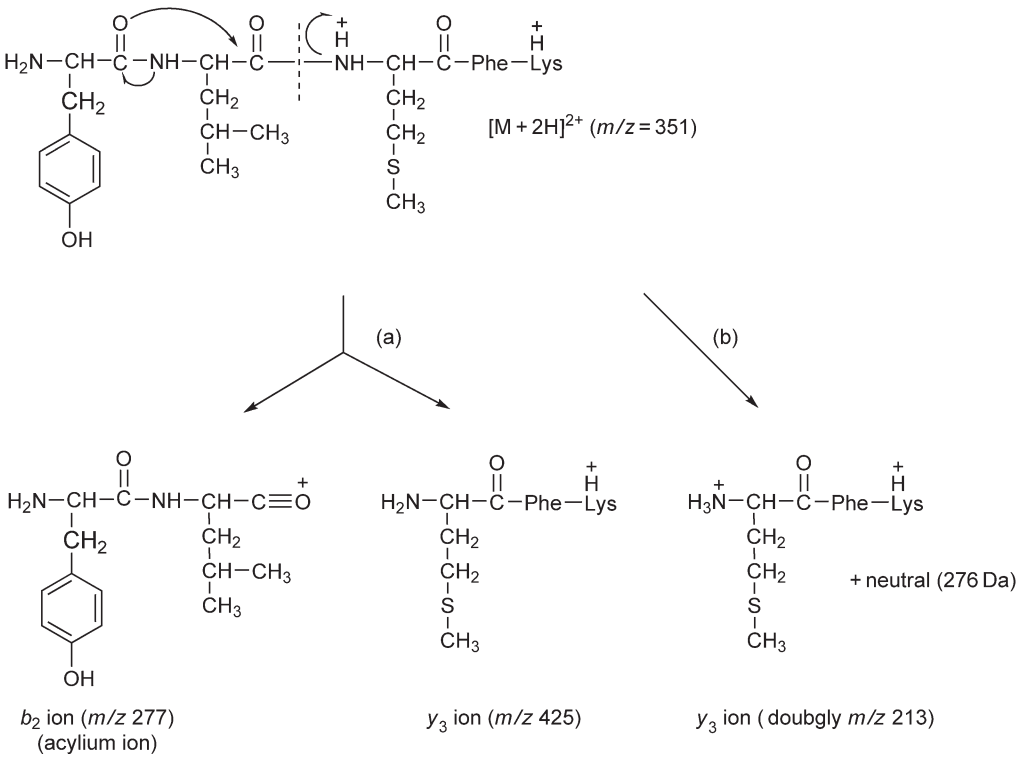


Figure 10 Formation of *b*- and *y*-ions from doubly protonated peptides; the pathway *a* yields singly charged *b*- and *y*-ions, and the pathway *b*, doubly charged *y*-ions.

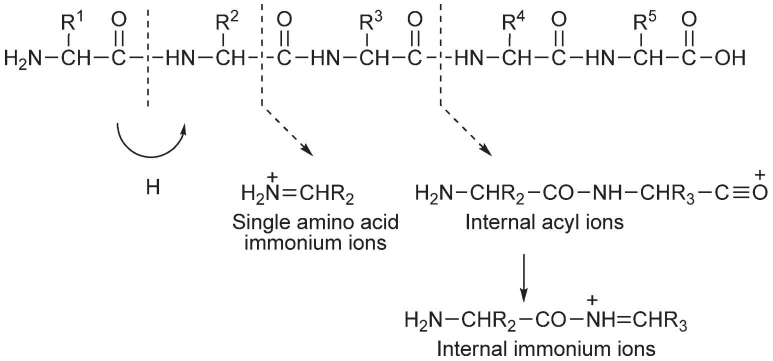


Figure 11 Formation of immonium ion and internal fragments. The immonium ions provide a clue to the composition of the peptide. Reproduced from C. Dass, *Fundamentals of Contemporary Mass Spectrometry*; Wiley-Interscience: Hoboken, NJ, 2007, with permission from Wiley-Interscience, Copyright 2007.

less common ion-activation methods are surface-induced dissociation, ultraviolet (UV) photodissociation, infrared multiphoton dissociation, and blackbody-induced radiative dissociation.[6]

Labeling the C-terminal carboxyl group of the peptide with the ^{18}O isotope simplifies the assignment of the fragment ion peaks in the spectrum as *b*- or *y*-ions.[75] Labeling is performed by digesting the protein in 1:1 $H_2^{16}O/H_2^{18}O$. Each *y*-ion from the labeled peptides is split into a doublet of peaks of nearly equal abundance and separated by 2 Da, but the *b*-ions remain unaffected.

9.12.5.2.2 Electron-capture dissociation and electron-transfer dissociation

In ECD, an ESI-produced multiply protonated peptide ion $[M + nH]^{n+}$ is first converted to an odd-electron $[M + nH]^{(n-1)+\bullet}$ ion by the capture of a thermal energy electron. Subsequent transfer of $H^\bullet$ to the backbone carbonyl group leads to the formation of *c*- and $z^\bullet$-type sequence-specific fragment ions.[76,77] The $a^\bullet$- and *y*-ions

Table 2 Masses of amino acid residues, their immonium ions, and their side chains and of neutral loss

Amino acid	*Symbol*		*Residue mass (Da)*	*Immonium ion mass (Da)*	*Side chain mass (Da)*	*Neutral loss (Da)*
Alanine	Ala	A	71.0371	44	15	–
Argnine	Arg	R	156.1011	129	100	17
Asparagine	Asn	N	114.0429	87	58	17
Aspartic acid	Asp	D	115.0269	88	59	18
Cysteine	Cys	C	103.0092	76	47	34[a]
Glutamic acid	Glu	E	129.0426	102	73	18
Glutamine	Gln	Q	128.0586	101	72	17
Glycine	Gly	G	57.0215	30	–	–
Histidine	His	H	137.0589	110	81	–
Isoleucine	Ile	I	113.0841	86	57	–
Leucine	Leu	L	113.0841	86	57	–
Lysine	Lys	K	128.0950	101	72	17
Methionine	Met	M	131.0405	104	75	48[b]
Phenylalanine	Phe	F	147.0684	120	91	–
Proline	Pro	P	97.0528	70	–	–
Serine	Ser	S	87.0320	60	31	18
Threonine	Thr	T	101.0477	74	45	18
Tryptophan	Trp	W	186.0793	159	30	–
Tyrosine	Tyr	Y	163.0633	136	107	–
Valine	Val	V	99.0684	72	43	–

[a] *S*-carboxymethyl derivative of Cys will lose 92 Da.
[b] Oxidized Met will lose 64 Da.
Reproduced from C. Dass, *Fundamentals of Contemporary Mass Spectrometry*; Wiley-Interscience: Hoboken, NJ, 2007, with permission from Wiley-Interscience, Copyright 2007.

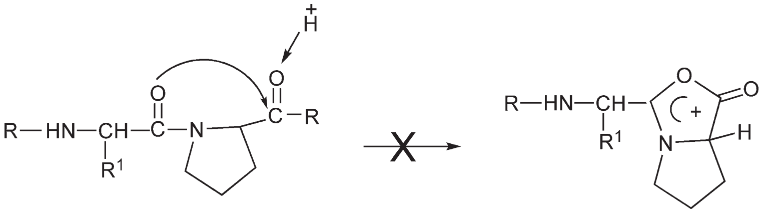


Figure 12 The effect of proline on cleavage of the peptide bond; attack of carbonyl oxygen on carbonyl carbon would form a strained [3,3,0] bicyclic system, thus preventing cleavage of the bond C-terminus to proline.

are also produced, but to a lesser extent, when H$^•$ migrates to the in-chain amide nitrogen. Both these processes induce far more backbone cleavages, allowing sequencing of much bigger peptides than that are amenable to CID. Also, because neutral losses are reduced and fragmentation is less affected by the peptide sequence, ECD is much more useful for detecting phosphorylation and glycosylation.

Electron-transfer dissociation (ETD) is a variation of ECD, in which an ion–ion reaction, for example, between an anthracene anion and a multiprotonated peptide cation, is used to transfer an electron to the peptide ion.[64,78] Similar to ECD, subsequent fragmentation of the odd-electron ions produces *c*- and *z*$^•$-type sequence-specific ions. ETD is also better suited for investigating phosphorylation and glycosylation.

9.12.5.3 How to Retrieve the Amino Acid Sequence from a Mass Spectrum

The task of retrieving the amino acid sequence from a mass spectrum involves arranging certain ions into a sequence-specific series, for example, either into the *y*-ion series or *b*-ion series. The following steps are helpful in this exercise:

- Ions below 55 Da of the $[M+H]^+$ ion are uninformative; ignore them.
- First, search for immonium ions in the low-*m*/*z* (<200) region of the spectrum; they are diagnostic of certain amino acids, and they can provide a clue to the composition of the peptide (see **Table 2**). Another clue to

the composition is gained from the ions that are formed by the loss of side-chain moieties from the $[M+H]^+$ ion (e.g., 107 Da is indicative of tyrosine and 92 Da of phenylalanine) of certain neutral molecules (e.g., ammonia and water) from *b*- and *y*-ions.

- Look for pairs of ions that are separated by 28 Da; they are the *b*/*a*-ions.
- It is helpful to identify the b_2-ion in the low-*m*/*z* region of the spectrum. This ion gives a clue to the first two amino acid residues, all possible combinations of which are listed in **Table 3**. Those *m*/*z* values are for singly protonated ions and are equal to 'the sum of residue masses + 1'. One must be careful in using this information because the combined mass of the two amino acid residues may fortuitously be equal to the residue mass of a single amino acid. These possibilities are mentioned in the footnote of **Table 3**. Once the b_2-ion is known, it is easy to recognize the y_{n-2} ion.
- Identify all members of the *b*- or *y*-sequence ion-series. For this, it is helpful to identify the $(n-1)$th ion first and then by matching the mass difference between each of the fragment ion peaks with the residue masses, the amino acids present are then recognized. Only the residue masses of 20 naturally occurring amino acids are considered (**Table 2**). This procedure is used in **Figure 13**.
- Let us try to identify the *y*-ion series first. To identify the y_{n-1} ion, subtract the mass of each of the two possible residues that make up the b_2-ion from the $[M+H]^+$ ion and match it with an ion in the high-*m*/*z* region of the spectrum. If the b_2-ion is unknown, subtract the masses of all 20 residues one at a time from the $[M+H]^+$ ion and match the result with an ion in the spectrum The next member of the *y*-ion series is recognized by subtracting the mass of all residues from the mass of the y_{n-1} ion and matching it with an ion in the spectrum. This procedure is repeated until a complete *y*-ion series has been recognized. Often, the last two residues in the sequence are difficult to recognize because the corresponding ions in the spectrum are usually absent. In that case, the list of the mass values (1 + sum of two residue masses) in **Table 3** can provide a clue. The y_1-ion is easy to identify for a tryptic peptide; it will be seen at *m*/*z* 147 or 175 because of the presence of Lys or Arg at the C-terminus, respectively.
- Once the *y*-ions are identified, search for the corresponding *b*/*y*-ion pairs through the relation 'mass of $b+y$ ions = the peptide mass + 2'.
- Recognize the remaining members of the *b*-ion series; this will confirm the sequence from the other end of the peptide.
- Tally the mass of the peptide from the amino acid sequence thus determined.
- Now, let us start fresh and search for the *b*-ion series, this procedure is especially helpful for tryptic peptides, which are recognized by the presence of *m*/*z* 147 or 175 in the spectrum. From these two mass values, we can identify the b_{n-1} ion through the relation 'the mass of which will be $= [M+H]^+ - 18 - 156$ (or 128)'. Continue to recognize the other *b*-ions in the series by subtracting the mass of all 20 residues one at a time from the mass of the last recognized *b*-ion and matching it with ions in the spectrum. This search will be easier if some *b*/*a* pairs have been identified. Once a *b*-ion is known, the corresponding complementary *y*-ion can be easily identified.

9.12.5.4 An Illustrative Example

For the sake of practice, let us try to determine the sequence of a tryptic peptide, whose MS/MS spectrum (**Figure 14**) was acquired by low-energy CID of the mass-selected doubly protonated ion at *m*/*z* 574.3. The peptide molecular mass is calculated to be = 1146.6 Da ($M = 2 \times 574.3 - 2 \times 1.0078$). The *m*/*z* 175.1 is identified as the y_1 ion and indicates that the C-terminus residue is Arg. The corresponding b_{n-1} ion is present at *m*/*z* 973.4. The ion pairs of *m*/*z* 973.4/945.4 and 844.4/816.4 differ by 28 Da, which strongly suggest *m*/*z* 973.4 and 844.4 are *b*-type ions. If this is correct, the mass difference of 129.0 between these two ions points to Glu as the next residue from the C-terminus. The *m*/*z* 715.4 is identified as the next *b*-ion because the mass difference between *m*/*z* 844.4 and 715.4 matches the residue mass of Glu. The next identified *b*-ion is *m*/*z* 628.4; the mass difference between *m*/*z* 715.4 and 628.4 is identical with the residue mass of Ser. The next *b*-ion could be *m*/*z* 571.3 because it is 57.1 Da (the residue mass of Gly), lower in mass from *m*/*z* 628.4. The next two identified *b*-ions are of *m*/*z* 472.4 and 343.2, and the corresponding amino acid residues are Val and Glu. There is a strong possibility that *m*/*z* 286.2 is the b_2-ion. The mass difference of 57.0 is between *m*/*z* 343.2 and 286.2 also supports this conjecture and implies that Gly is the next amino acid residues. The first two residues could be one of the

Table 3 Masses of b_2-ions

	G	*A*	*S*	*P*	*V*	*T*	*C*	*I/L*	*N*	*D*	*K/Q*	*E*	*M*	*H*	*F*	*R*	*Y*	*W*
G	115																	
A	129	143																
S	145	159	175															
P	155	169	185	195														
V	157	171	187	197	199													
T	159	173	189	199	201	203												
C	161	175	191	201	203	205	207											
I/L	171	185	201	211	213	215	217	227										
N	172	186	202	212	214	216	218	228	229									
D	173	187	203	213	215	217	219	229	230	231								
K/Q	186	200	216	226	228	230	232	242	243	244	257							
E	187	201	217	227	229	231	233	243	244	245	258	259						
M	189	203	219	229	231	233	235	245	246	247	260	261	263					
H	195	209	225	235	237	239	241	251	252	253	266	267	269	275				
F[a]	205	219	235	245	247	249	251	261	262	263	276	277	279	285	295			
R	214	228	244	254	256	258	260	270	271	272	285	286	288	294	304	313		
Y	221	235	251	261	263	265	267	277	278	279	292	293	295	301	311	320	327	
W	244	258	274	284	286	288	290	300	301	302	315	316	318	324	334	343	350	373

[a] Oxidized methionine will also show this combination.

GG = N = 114; GA = K/Q = 128; GV = R = 156; GE = AD = SV = W = 186.

Reproduced from C. Dass, *Fundamentals of Contemporary Mass Spectrometry*; Wiley-Interscience: Hoboken, NJ, 2007, with permission from Wiley-Interscience, Copyright 2007.

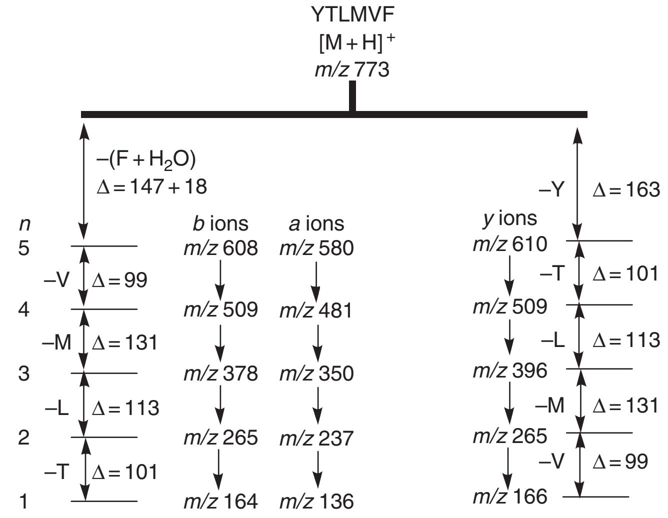


Figure 13 The *a*-, *b*-, and *y*-ions in the spectrum of YTLMVF (MW = 772.28 Da).

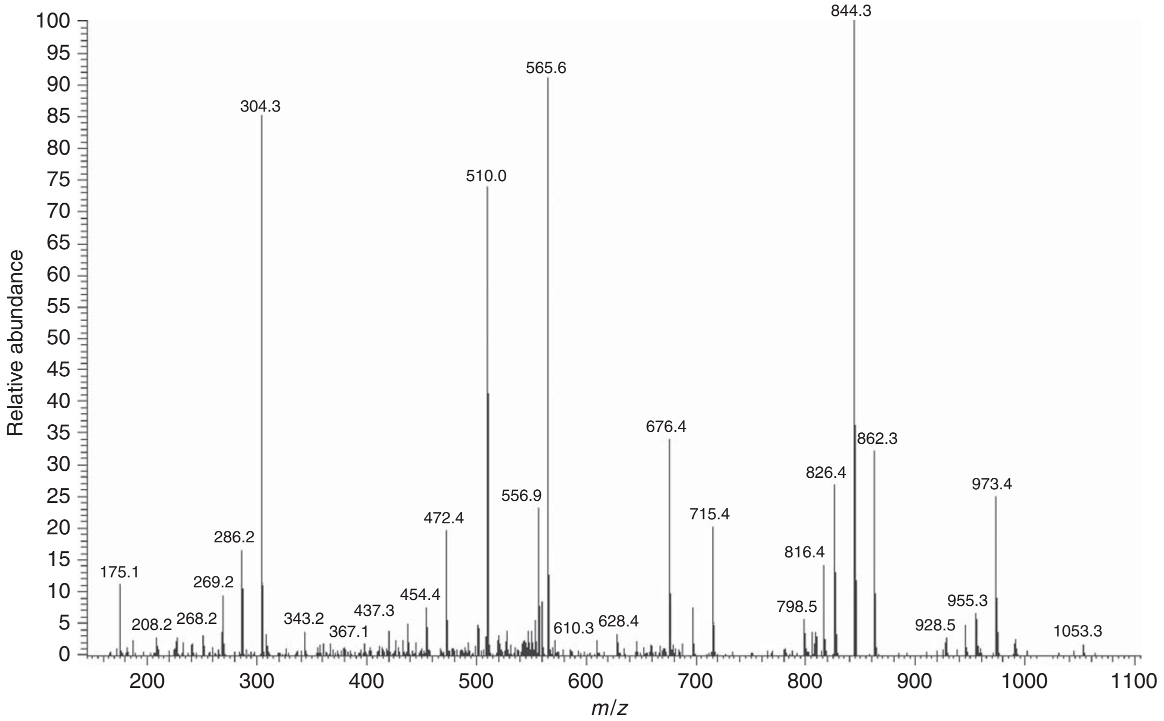


Figure 14 Low-energy CID–MS/MS spectrum of an unknown peptide. Reproduced from C. Dass, *Fundamentals of Contemporary Mass Spectrometry*; Wiley-Interscience: Hoboken, NJ, 2007, with permission from Wiley-Interscience, Copyright 2007.

combinations of Trp–Val–, Val–Trp–, Arg–Glu, or Glu–Arg– (see **Table 3**) on the basis of *m*/*z* of the b$_2$-ion. The presence of *m*/*z* 269.2 as a satellite peak of *m*/*z* 286.2 (i.e., the loss of NH$_3$) suggests that the Glu–Arg–combination is a better choice for the first two residues from the N-terminus. The corresponding y_{n-2} ion is seen at *m*/*z* 862.3. Thus, the *b*-ion series is composed of ions of *m*/*z* 1146.6, 973.4, 844.3, 715.4, 628.4, 571.3, 472.4, 343.2, 286.2, XX and the identified amino acid sequence of the peptide is Glu–Arg–Gly–Glu–Val–Gly–Ser–Glu–Glu–Arg. The complementary *y*-type ions should be present at *m*/*z* 1146.6, (1018.6), 862.3, 805.4, 676.4, (577.2), 520.2, 433.2, 304.3, and 175.1. The *y*-ions (of *m*/*z* 1018.6 and 577.2) complementary to b_1 and b_5 are missing in the spectrum. A partial sequence derived from this series is Xxx–Xxx–Gly–Glu–Xxx–Xxx–Ser–Glu–Glu–Arg.

9.12.6 Posttranslational Modifications in Proteins and Peptides

Certain active group-containing (e.g., amine, thiol, thioether, carboxylic acid, and hydroxyl) amino acid residues undergo covalent modifications while transferring the nascent protein to various intercellular locations. More than 300 PTMs are known; some common ones are acylation (acetyl, formyl, and myristyl), carboxylation, glycosylation, lipidation, amidation, oxidation, nitration, nitrosylation, phosphorylation, sulfation, acetylation, hydroxylation, and proteolysis. Several cellular functions are triggered by these modifications; for example, PTM is a common mechanism for regulating localization, function, and protein turnover. Because of this, research on modified proteins has attracted much attention. Mass spectrometry protocols are increasingly being adopted to analyze such modified proteins and peptides. These techniques rely upon the measurement of the precise change in the residue mass upon modification. Two of the most important PTMs, phosphorylation and glycosylation, are discussed further here.

9.12.6.1 Phosphoproteomics: Analysis of Phosphoproteins

Protein phosphorylation is arguably one of the most important and ubiquitous events, with 30–60% of all proteins known to exist at one time or other as phosphoproteins. Phosphorylation is a key component of many cellular processes, including proliferation, differentiation, metabolism, signal transduction, and adaptation to environmental stress and in the function of many proteins, hormones, neurotransmitters, and enzymes. An abnormal regulation of phosphorylation often results in disease. Phosphorylation most commonly occurs on serine, threonine, and tyrosine residues whereby the OH group is replaced by a phosphate group, resulting in increase in the mass of those residues by 80 Da. Phosphorylation is a reversible process, mediated by phosphotransferase enzymes, called protein kinases, and are reversed by protein phosphatases.

The phosphoproteome is a specific subset of the proteome, and phosphoproteomics is the comprehensive analysis of the entire complement of phosphoproteins in biological systems, including detection of all phosphoproteins, location of the exact site of phosphorylation, and quantification of differentially expressed phosphoproteins.

Despite the advantages of sensitivity, specificity, and faster analysis time, the greatest challenge in mass spectrometric characterization of phosphorylation is the low stoichiometry of phosphoproteins in biological samples. Suppressing ionization in the presence of other proteins/peptides also poses a problem. Prior to the integrated proteomics approach, several mass spectrometry techniques were developed to detect phosphopeptides in a protein digest and to identify the sites of phosphorylation.[79,80] The discovery of FAB ionization in 1981 enhanced the use of mass spectrometry for analyzing phosphoproteins/peptides.[70,81] The current state of development in isolating and enriching phosphopeptides and in mass spectrometric detection has made possible the proteome-wide scale analysis of phosphoproteins, including their quantification.[79,82–86]

A general protocol for mass spectrometric analysis of phosphoproteins is illustrated in **Figure 15**: various steps of this protocol are cleavage of purified phosphoproteins, isolation and preferential enrichment of phosphopeptides, selective detection of phosphopeptides in the digest, identification of the phosphorylation sites using tandem mass spectrometry, and identification of phosphopeptides/proteins through a database search.

9.12.6.1.1 Digestion of phosphoproteins

Similar to the analysis of ordinary proteins, phosphoproteins also need to be cleaved into small manageable peptide fragments; any one of the proteases listed in **Table 1** can be used. The concept of chemical tagging through β-elimination/Michael addition reactions has been adopted to facilitate the cleavage process. The β-elimination reaction converts phosphoserine and phosphothreonine to dehydroalanine and

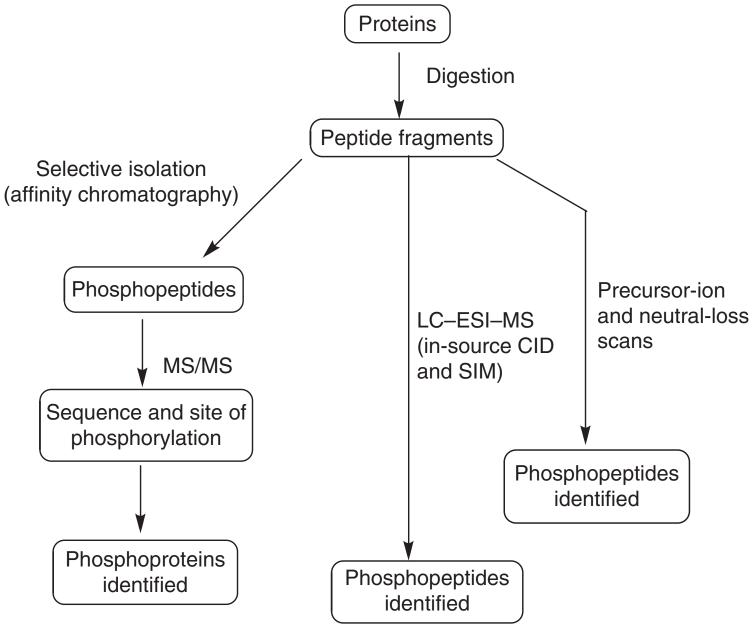

Proteins
Digestion
Peptide fragments
Selective isolation (affinity chromatography)
Phosphopeptides
MS/MS
Sequence and site of phosphorylation
Phosphoproteins identified
LC–ESI–MS (in-source CID and SIM)
Phosphopeptides identified
Precursor-ion and neutral-loss scans
Phosphopeptides identified

Figure 15 General protocol for the analysis of phosphoproteins.

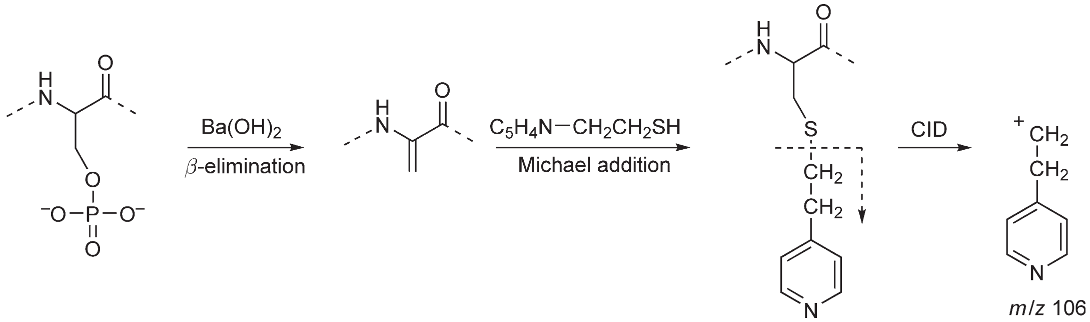


Figure 16 A reaction scheme for chemical tagging of a phosphoserine- and phosphothreonine-containing peptides for their selective detection by the positive-ion precursor-ion scanning.

β-methyldehydroalanine, respectively,[87] which after a Michael addition with cysteamine, are converted to aminoethylcysteine and β-methylaminoethylcysteine, respectively (see **Figure 16**), for a similar reaction with 2-[4-pyridyl]ethanethiol. These products being isosteric with lysine, are cleaved with lysine-specific proteases.

9.12.6.1.2 Fractionation of peptides in protein digests

A major development in phosphoproteomics is the use of affinity-based chromatography[88] with immobilized antibodies, metal ions (IMAC),[89,90] metal oxides (TiO_2),[91] and nanoparticles[92,93] for selective isolation and enrichment of phosphopeptides from an astronomically large pool of peptides in a digest. Among these methods, IMAC has been a major contributor to the success of proteome-wide scale profiling of phosphoproteins. Fe^{3+} and Ga^{3+} ions are the most frequently used metal ions. A solution of phosphopeptides at pH 3.5 is passed through the IMAC column, whereupon phosphopeptides are selectively trapped and unmodified peptides are collected in the flow-through. By passing a basic solution through the column, the bound phosphopeptides can be eluted, allowing enrichment of phosphopeptides. Nonspecific binding of acidic amino acid residues with metal ions is a major concern of IMAC that can be addressed by capping all COOH groups in the peptides by esterification prior to IMAC separation.[94]

Another revolutionary idea in the enrichment of phosphopeptides is to attach unique chemical tags to the peptides.[95,96] In one approach, applicable to phosphoserine- and phosphothreonine-containing peptides, the phosphate group is replaced with 1,2-ethanedithiol, and biotin is attached to the thiol group for selective isolation of phosphopeptides by avidin column chromatography.[95] A chemical tag that is applicable to all three phospho residues has been developed.[96] The enrichment of phosphopeptides in peptide digests has also been achieved with calcium phosphate precipitation.[97]

9.12.6.1.3 Identification of phosphopeptides

Because phosphorylation and other PTMs are not coded by genes, their location cannot be predicted; they must be analyzed directly by one of the following procedures. If the sequence of the protein is known from the knowledge of its cDNA, then the peptide mapping approach can be used to rapidly identify phosphopeptides in the protein digest; a simple molecular mass determination of the peptide fragments will show the mass increase of 80 Da in phosphopeptides. A similar concept is used in the phosphatase treatment approach; here, the peptide mixture is treated with the phosphatase enzyme to clip the phosphate group; the molecular mass of phosphopeptides will decrease by 80 Da for the loss of each phospho unit, allowing identification of phosphopeptides/proteins.

In the ESI in-source CID approach, which is implemented with an LC/ESI–MS system, the peptide ions are subjected to CID in the ESI ion source transport region by having the skimmer voltage adjusted; the phosphate group marker ion of m/z 79 is produced by operating the ESI in the negative-ion mode. The ion current due to the marker ion is recorded in the selected-ion monitoring (SIM) mode. Only phosphopeptides will show any response, thus providing rapid detection of those peptides.[98,99]

Several versions of the precursor-ion scan can be found in the literature for selective identification of phosphopeptides. In one version, the precursor-ion scan of m/z 79 is performed in the negative-ion mode.[100] In another version, termed phosphotyrosine-specific immonium-ion (PSI) scanning, only phosphotyrosine-containing peptides are detected by monitoring the CID-produced immonium ion of m/z 216.043 in the positive-ion mode.[101,102]

A derivatization strategy has been developed to generate unique marker ions that can be used for the selective detection of phosphopeptides through precursor-ion scanning. In one approach, phosphoserine and phosphothreonine residues are converted to *S*-pyridylethylcysteine and *S*-pyridylethyl-β-methylcysteine, respectively, using the base-catalyzed β-elimination of the phosphate group and a Michael addition with 2-[4-pyridyl]ethanethiol (**Figure 16**). The thioether bond of these derivatives readily cleaves under CID to produce the pyridylethyl ion ($C_5H_4N–CH_2CH_2^+$) of m/z 106; the positive-ion precursor-ion scan of this marker ion selectively detects phosphoserine- and phosphothreonine-containing peptides. In another similar strategy, derivatization is performed with 2-dimethylaminoethanethiol, followed by oxidation of the thioether derivative to sulfoxide, which upon CID undergoes facile β-elimination with the loss of m/z 122.06. Monitoring of the ion current due to this marker ion in the precursor-ion scan mode provides selective detection of phosphopeptides.[102]

Under CID, all three types of O-phosphorylated peptides exhibit the loss of 98 Da due to the expulsion of H_3PO_4 (and/or $HPO_3 + H_2O$). This reaction has been exploited to detect phosphopeptides in a protein digest with neutral-loss scan.[103]

9.12.6.1.4 Identification of phosphorylation sites

Proteome-wide scale identification of phosphoproteins can be accomplished by acquiring the amino acid sequence-specific fragment ion information of peptides in the protein digest. CID of the ESI-produced peptide ions is a highly popular approach in this respect.[104,105] Most of our understanding of peptide ion fragmentation (see Section 9.12.5) also applies to phosphopeptides. For example, *b*- and *y*-ions are the dominant sequence-specific ions in the CID spectra of phosphopeptides. The residue masses of phosphoserine, phosphothreonine, and phosphotyrosine are shifted to 167, 181, and 243 Da, respectively. Thus, on the basis of these unique mass differences, the exact site of phosphorylation can be ascertained quickly and accurately. Complications arise in those cases where loss of the phosphate group (as HPO_3 or H_3PO_4) from the precursor ion dwarfs the *b*- and *y*-ion series. Specifically, the phosphoserine- and phosphothreonine-containing peptides undergo a facile loss of H_3PO_4 via β-elimination.[104] Loss of the phosphate group from *b*- and *y*-ions also complicates the assignment of sequence ions.

One proposal to increase the visibility of the *b* and *y* sequence-specific ions and to provide complete sequence information is to derivatize phosphopeptides.[106] Sequence analysis can also be improved by using the β-elimination chemistry.[102,107] As described above in this section, phosphoserine and phosphothreonine can be converted to *S*-ethylcysteine and *S*-ethyl-β-methylcysteine, respectively, upon the base-catalyzed β-elimination of the phosphate group, followed by the reaction with ethanethiol. CID of the modified peptides results in more evenly distributed sequence-specific fragment ions.

As mentioned earlier in the chapter, ECD and ETD both are much more useful for identifying the location of a phosphate group in peptides because the unwelcome loss of the phosphate group from the precursor ion is suppressed, and the abundance of sequence-specific ions is commensurately enhanced.[108–111] Also, the sequence of the peptide has little influence on fragmentation of the peptide ion, which is solely driven by free-radical chemistry to produce *c*- and $z^{\bullet}$-type sequence-specific ions.

Another ion activation method that is well suited for identification and sequence analysis of phosphopeptides in the positive and negative polarity modes is infrared multiphoton dissociation (IRMPD).[105,112] In this technique, phosphopeptides are irradiated with 10.6 μm photons emitted from a CO_2 laser. The phosphate group behaves like a chromophore for these photons, allowing evenly distributed cleavages in the peptide chain and more sequence coverage than the CID technique.[105]

The final step of phosphoproteomics is database searching with the fragmentation pattern of the target phosphopeptide. The search algorithms are similar to those used for identification of nonmodified proteins, except for the consideration of a mass shift of 80 Da to serine, threonine, and tyrosine residues. The sequence of the identified phosphopeptide, however, must be verified through manual *de novo* sequencing procedure to confirm that the site of phosphorylation is correctly identified.[107] The search algorithm can also reveal additional information about specific kinases and phospho-specific binding interactions.

9.12.6.2 Glycoproteomics: Analysis of Glycoproteins

Glycosylation is another ubiquitous PTM of proteins that has attracted much attention; over 50% of all mammalian proteins in eukaryotic systems exist in glycosylated form at some point during their life spans. Glycosylation refers to the attachment of carbohydrate chains, called glycans, to the polypeptide either through the primary amide nitrogen atom of the asparagine side chain (N-glycosylation) or the oxygen atoms of serine and threonine side chains (O-glycosylation). The most common constituent monosaccharides of the carbohydrate side chains of mammalian proteins are D-mannose (Man or Hex), D-galactose (Gal or Hex), L-fucose (Fuc or dHex), *N*-acetylglucosamine (GlcNAc or HexNAc), *N*-acetylgalactosamine (GalNAc or HexNAc), and *N*-acetylneuraminic acid (sialic acid, Neu5Ac, or NANA). Glycosylation of proteins serves several critical functions in cell biology, including intracellular transport, cell recognition, and cell–cell interactions; it also has a profound influence in modulating the physicochemical (e.g., solubility and stability) and biological (e.g., immunological and proteolytic stability) properties of proteins.

Various steps used in characterization of glycoproteins are depicted in **Figure 17**. The first step is to measure the molecular mass of glycoproteins, which can provide a direct global assessment of glycosylation. From the known mass of the nonglycosylated protein (e.g., from its cDNA sequence information) and the measured mass of the target glycoprotein, the glycan mass can be deduced. MALDI and ESI both are well suited for the molecular mass measurement of glycoproteins. A successful MALDI–MS approach has used aminophenylboronic acid-derivatized magnetic beads to selectively capture glycoproteins;[113] the beads are spotted directly onto a MALDI sample plate along with a matrix. Immobilization also helps in selective isolation of glycoproteins from a complex mixture, thereby obviating the need for a separate isolation step.

9.12.6.2.1 Identification of glycopeptides

Because mass spectrometry analysis is conveniently performed at the peptide level, a general practice is to cleave glycoproteins into small peptide segments by digestion with trypsin or with other enzymes listed in **Table 1**. To find whether the protein is glycosylated or not, peptide maps of glycosylated and nonglycosylated forms of the target protein are compared.[114] The two sets of peptide maps are generated by proteolysis of the

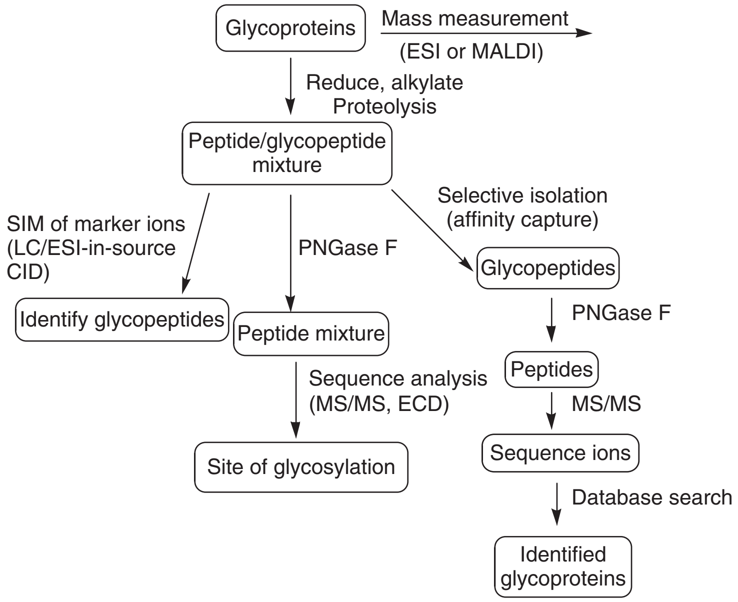


Figure 17 General protocol for the analysis of glycoproteins.

glycoprotein before and after the release of the carbohydrate chains (via PNGase treatment). The appearance of new peaks in the peptide map of the PNGase-treated fraction will provide the identity of glycopeptides. If the sequence of the protein is known from its cDNA sequence, then the m/z value of the new peptide will also pinpoint its location in the glycoprotein.

For selective detection of glycopeptides in the protein digest, an in-source CID strategy, similar to that used for phosphopeptides, is also applicable; glycopeptides are analyzed using the positive-ion LC–ESI–SIM of the carbohydrate-specific marker ions.[115] The oxonium ions of m/z 163 (Hex), 204 (HexNAc), 274 and 292 (NeuAc), and 366 (Hex–HexNAc) have served as marker ions. The precursor-ion scan of those ions (e.g., m/z 204) on a tandem mass spectrometer is another option to selectively identify glycopeptides.

9.12.6.2.2 Sites of glycosylation

For precise identification of glycosylation sites, glycopeptides are analyzed by tandem mass spectrometry.[116,117] A CID of an intact glycopeptide, however, may not provide the needed information due to poor MS response and lack of fragmentation in the peptide backbone in the presence of carbohydrate side chains; in glycopeptides the activation energy is mainly directed toward the cleavage of the labile peptide–carbohydrate bond. To circumvent these limitations, the N- and O-linked carbohydrate chains are released by treating glycopeptides with PNGase and $NaBH_4$/NaOH, respectively, prior to the MS/MS analysis.

As mentioned earlier, ECD is ideally suited for the analysis of the glycosylated sites in peptides; this 'mild' activation process does not require the removal of the carbohydrate chains. In the mass spectrum of intact glycopeptides, the carbohydrate chains remain attached to the sequence-specific ions, thereby providing direct evidence of the glycosylation sites.[76]

9.12.6.2.3 Proteome-wide analysis of glycoproteins

The proteomics-based approaches for identification of glycoproteins use the mass spectrometry procedures described in this section.[83] Various steps used in a generic approach (see **Figure 17**) are fractionation of the protein sample, proteolysis to produce a mixture of glycosylated and nonglycosylated peptides, isolation of glycosylated peptides, and MS/MS analysis of those peptides after the removal of the carbohydrate chain by treatment with PNGase F or $NaBH_4$/NaOH. Isolating glycopeptides is an important step; this task is accomplished by the selective capture of the glycopeptides using lectin-affinity chromatography, size-exclusion chromatography, or onto a solid support modified by hydrazide chemistry.[118–120]

A shotgun glycoproteomics approach similar to that discussed in Section 9.12.2.4.2 has also been described,[119] various steps of which are cleavage of glycoproteins, enrichment of glycopeptides in the digest by selective capture onto a modified solid support, release of peptides by PNGase treatment, and MS/MS analysis of the released peptide. Finally, the database search identifies the proteins.

9.12.6.2.4 Top-down sequence analysis of the intact protein via ion–ion reactions

A top-down sequence analysis approach has been developed in which sequence information on intact glycoproteins is obtained.[121] First, the charge state of the ES-ionized protein is optimized through ion–ion reactions with $[\mathrm{M}-\mathrm{F}^{-}]$ and $[\mathrm{M}-\mathrm{CF}_3{}^{-}]$ anions. The CID spectrum of that precursor ion is well endowed with abundant sequence-specific ions, which are used to identify glycosylation sites. The top-down approach has the advantage that additional time- and sample-consuming purification and proteolytic cleavage experiments can be avoided.

Once glycopeptides have been detected, the corresponding RP–HPLC fractions are collected for additional experiments aimed at characterization of the carbohydrate side chains. To this end, the collected glycopeptide fractions are treated with exoglycosidases to release specific sugar residues. The molecular mass is determined again, and the difference in the two masses provides the identity of the released carbohydrate residue. Other experiments to characterize the oligosaccharide side chain include determination of the composition of monosaccharides, sequencing and branching of each monosaccharide, stereochemistry, interglycosidic linkage position, and anomeric configuration (see Chapter 9.13).

9.12.7 Higher-Order Structures of Proteins and Peptides

Earlier in this chapter, we defined the four levels of protein structures: primary, secondary, tertiary, and quaternary. In their native states, proteins exist in highly ordered compact forms that contain surface pockets and interior cavities where specific binding of other ligands can occur. What type of roles a protein plays in living systems is largely determined on its very specific 3D structure. Changes in this structure frequently occur at the localized, as well as at the globular, level to regulate the function of enzymes and receptors. Subtle changes in the protein structure are reflected in the diversity and complexity of life. To understand fully the function of a protein, these changes must be known in terms of atomic locations, folding–unfolding dynamics, and thermodynamics.

Proteins can be denatured to unfolded random coil structures by increasing the temperature and/or changing pH of the solution or exposing the proteins to high concentrations of detergents, organic solvents, or certain chaotropic compounds such as urea and guanidinium chloride. The folding–unfolding process is usually reversible; unfolded proteins will refold to their native states once the denaturants are removed.

A suite of techniques is at the disposal of researchers to analyze a protein's higher-order structures, dynamics, and changes in its native conformation.[122] These techniques include electron microscopy, neutron diffraction, ultracentrifugation, LC, X-ray crystallography, circular dichroism (CD), nuclear magnetic resonance (NMR), FT infrared (IR), Raman, UV–visible absorption, and fluorescence spectroscopy. NMR and X-ray crystallography both have the advantage of providing high-resolution structural details, but they have their own limitations; NMR-based techniques have a limited mass range and poor sensitivity; many lesser populated transient structures may escape detection. The limitation of X-ray crystallography is that it provides static snapshots of an average structure and is accessible only to crystallized proteins. Computational algorithms can also provide details of 3D protein structures.[123]

The development of relatively softer ionization techniques of ESI and MALDI has engendered the use of mass spectrometry to determine conformations of proteins.[122,124–130] ESI has made especially tremendous strides, as it samples proteins directly from the native solvent environment. The belief is that the solution-phase structures of proteins are largely preserved during their conversion to the gas-phase ions with these methods; thus, the spectrum obtained is a reflection of features of the protein's solution chemistry.

Mass spectrometry-based techniques have distinct advantages over existing methods in terms of sensitivity, protein stability, and extended mass range. Furthermore, the solubility and purity of a protein are of lesser concern. Other techniques are limited to provide information that represents an average of entire protein ensembles, whereas mass spectrometry can reveal structural details on transient or folding intermediates.

9.12.7.1 Charge-State Distribution for Probing Conformations of Proteins

A qualitative, low-resolution picture of the conformations of a protein can be obtained in the positive-ion ESI mode through charge-state distribution (CSD).[131] Although the charge states observed in an ESI mass spectrum may not match those that exist in the bulk solution,[132] the shape of this distribution has direct correlation with the solution-phase conformation. The underlying concept is that the basic sites are open to protonation in the unfolded flexible conformation but are not accessible to the solvent in the tightly folded native state, resulting in lower charge states. Once the protein unfolds, those buried or bound basic sites become accessible to the solvent and the protein exhibits a greater degree of charging and a shift in the spectrum to lower *m/z* values. An illustrative example is the acid-induced unfolding of cellular retinoic acid binding protein I (CRABP I).[133] At pH 5 and above, only three charge states (+7, +8, and +9) are seen in the ESI spectrum of the native folded protein. Unfolding of the protein begins at pH 4.5; at a more acidic pH, the open structure dominates and CSD shifts to lower *m/z* values.

From the CSD profile, it is also feasible to identify coexisting protein structures. Those experiments are conducted by dissolving the protein in a mildly denaturing solvent environment, so that both equilibrium structures are present in the solution. The CSD profile in the ESI spectrum of this solution will be bimodal, an indication of two coexisting conformers.

CSD can also be used to study kinetics of folding–unfolding processes using time-dependent experiments.[134] For those experiments, the protein is initially brought to a denatured state, folding is initiated by slowly replacing the denaturing solvent with a refolding buffer, and the solution is analyzed at different time points. The plot of the abundance of a particular charge state versus time provides kinetic information.[135] For rapidly converting conformers, kinetic experiments are performed with an online rapid-mixing device directly coupled to ESI–MS (see Section 9.12.7.2.1).[136]

Figure 18 demonstrates that CSD can also be used to explain the structure of peptides as short as containing 13 amino acids, which, unlike proteins, exhibit fewer charge states.[135,137] This figure compares the CSD profiles of dynorphin (1–13) in 100% water and in 20% and 50% trifluoroethanol (TFE) solutions in water. TFE is a helix-inducing solvent. It is evident that the open structure that exists in water exhibits a relatively higher charge state ion profile, which shifts to a lower charge state profile as the concentration of TFE is increased. This shift is a reflection of the change in conformation to a compact secondary structure, a change induced by TFE.

9.12.7.2 Hydrogen–Deuterium Exchange for Probing Conformations of Proteins

Hydrogen–deuterium exchange (HX) is the premier method used to distinguish between folded and unfolded structures and to study folding–unfolding dynamics in proteins. In this isotope exchange process, the labile hydrogen atoms in proteins are replaced with deuterium atoms of the surrounding solvent. The HX approach, first introduced in the 1950s,[138] and subsequently adapted with mass spectrometry in 1991,[139] has now become a mature field to study conformations of proteins and peptides.[122,124–130,140,141] Grossly, there are three types of hydrogens in proteins (see **Figure 19**):[6] the nonexchanging alkyl hydrogens (red), fast-exchanging functional group labile hydrogens (green), and slow-exchanging in-chain amide hydrogens (blue); only the last ones are considered for these experiments as they exchange at a rate that can be measured by mass spectrometry. How fast amide hydrogen will exchange depends upon its accessibility to the solvent. The hydrogens that are part of the secondary structure and hydrophobic core exchange slowly, whereas the surface hydrogens exchange at a faster rate. Thus, the measured exchange rate can be used as a probe for determining the conformation of a protein; a faster exchange is indicative of a more open structure, and a slower exchange is associated with a more tightly folded compact state. Also, the unfolded conformer has more hydrogens available for exchange than the folded native structure.

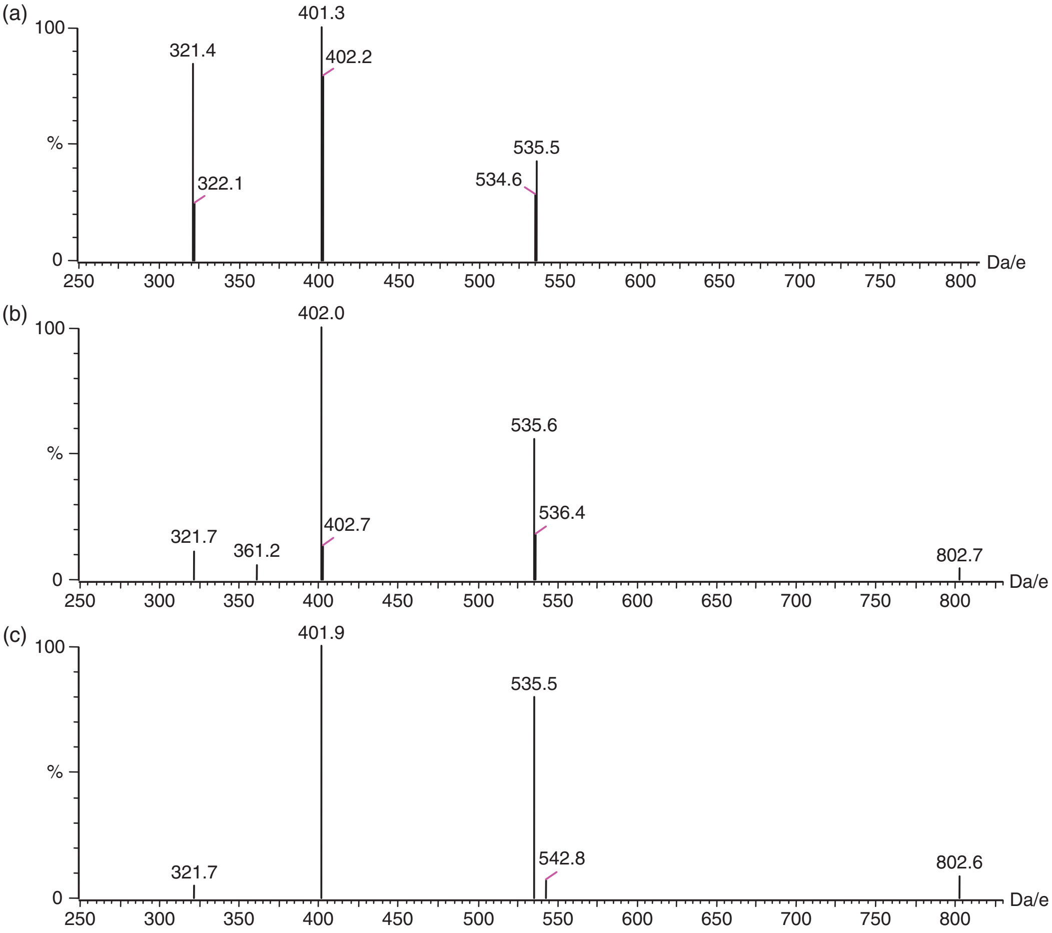


Figure 18 The CSD profile of dynorphin (1–13) in (a) water, (b) 20% TFE, and (c) 50% TFE. Reproduced from X. Cai; C. Dass, *Eur. J. Mass Spectrom.* **2007**, *13*, 2341–2346 with permission from IM publications, Copyright 2007.

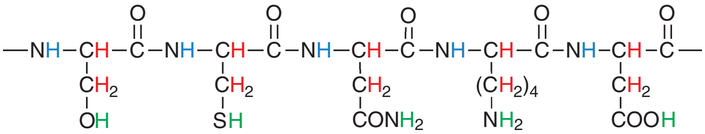


Figure 19 Three different types of hydrogens in a polypeptide chain: (1) hydrogens are nonexchanging alkyl hydrogens (red), (2) hydrogens are fast-exchanging labile hydrogens (green), and (3) hydrogens are slow-exchanging amide hydrogens (blue). Reproduced from C. Dass, *Fundamentals of Contemporary Mass Spectrometry*; Wiley-Interscience: Hoboken, NJ, 2007, with permission from Wiley-Interscience © 2007.

The H^+ and OH^- ions both can catalyze amide–hydrogen exchange; the exchange rate is the slowest at pH 2.5 and increases rapidly as the pH becomes more basic. The rate also increases as the pH becomes less than 2. Temperature also affects the exchange rate; it is 10× slower at 0 °C than at 25 °C. Because of the slowest exchange rate, the solution at pH 2.5 and 0 °C is considered at 'quench conditions'. For kinetic measurements, HX is usually conducted at physiological pH and 25 °C and after a certain time point the exchange is quenched by bringing the solution to quench conditions.

9.12.7.2.1 Experimental approaches for monitoring the amide hydrogen exchange

The global level, as well as at the small segment level, changes in the conformation of proteins can be identified through HX. For global level information, HX is performed on intact proteins using the following techniques. Such experiments provide the exchange rate averaged over all amide hydrogens.

Continuous-labeling technique: In a typical procedure, the protein is first dissolved in water at the physiological pH and denatured to the unfolded state. The protein is lyophilized and redissolved in H_2O; the exchange is initiated by diluting this solution 100-fold with D_2O. The samples are withdrawn at several fixed time points, and each is analyzed by ESI–MS or MALDI–MS. The exchange rate is calculated from the plot of number of hydrogens exchanged versus time. To study the kinetics of the unfolding process, the lyophilized folded protein is dissolved in D_2O at a pD where the folded and unfolded states both coexist. The samples are withdrawn at different time intervals for analysis by mass spectrometry.

Continuous-flow apparatus: For studying faster exchange kinetics, HX experiments are performed in a time-resolved fashion using a continuous-flow apparatus online with ESI–MS.[129,136,137] In one version of this apparatus, the exchange solution flows continuously through a capillary and a small volume of the protein/peptide solution is injected into the solvent stream to initiate the exchange.[137] In another version, the labeling and protein solutions each flow through separate channels and the two solutions are mixed together in a mixer to initiate HX (**Figure 20**).[142] The duration of the exchange (ms to s) can be varied in both versions by changing the flow rate and the dimensions of the exchange capillary. A continuous-flow apparatus can also be used to monitor the CSD profile[129] and pulse-labeling HX experiments.[142,143]

Fragmentation mass spectrometry for localized changes: To identify the peptide-level structural changes fragmentation mass spectrometry method is used.[144,145] After HX, the target protein is cleaved into small segments by pepsin digestion under quench conditions and peptide fragments are analyzed with an online LC–ESI–MS setup under conditions that minimize the back exchange.[145] The extent of the deuterium incorporation in the peptide is estimated by comparison of the mass spectrum with corresponding spectra of the same peptide when it is nondeuterated and fully deuterated.

Tandem mass spectrometry for the residue-level exchange information: Details of individual amino acid residue-level structures can be explored by acquiring the CID–MS/MS spectrum of the peptide segments formed by pepsin digestion of the exchanged protein. As mentioned earlier in this chapter, the *b*- and *y*-ion series are the primary sequence-specific ions under CID conditions. The amide hydrogen of a particular residue that is involved in HX can be identified by an increase in the mass of the sequence ion containing that residue.

The usefulness of this approach is demonstrated in identifying the structure of lipid vesicle-bound angiotensin II.[146] This study was undertaken to determine what kind of structure this hormone adopts when it is in contact with its receptor. As shown in the zoom scan of the molecular ion region, fewer hydrogens are exchanged in vesicle-bound angiotensin II (lower **Figure 21**) compared to those exchanged in water (upper **Figure 21**), an indication that angiotensin II assumes a distinct conformation when in contact with lipid vesicles. Furthermore, through HX–CID–MS/MS and other structural probes developed in this study, it was observed that a major portion of angiotensin II interacts with the phospholipid head groups on the surface of the vesicles, and it assumes a U-shaped configuration when in contact with lipid vesicles.

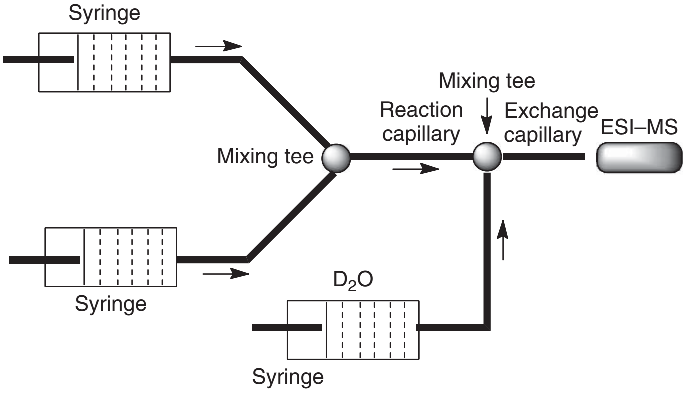


Figure 20 A sketch of the continuous-flow apparatus.

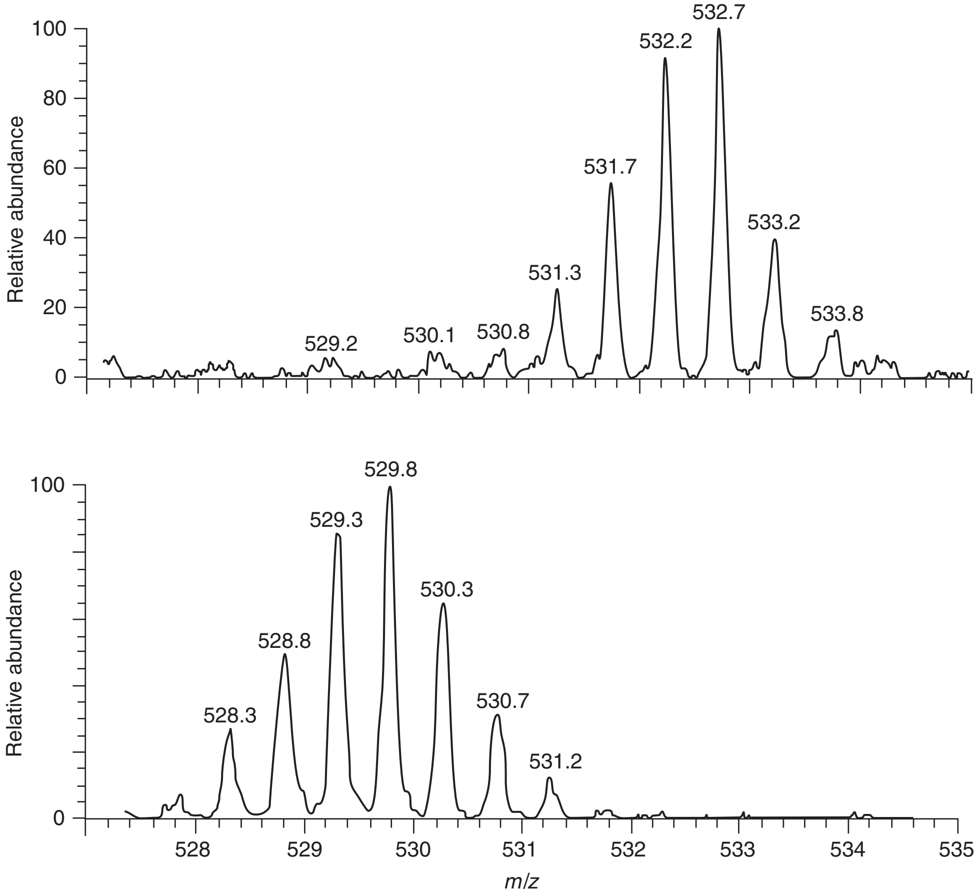


Figure 21 Zoom scan ESI mass spectra of angiotensin II after deuteration in water (upper) and in contact with lipid vesicles (lower).

9.12.7.3 Chemical Cross-Linking for Probing 3D Structures of Proteins

Another experimental approach that has achieved some success in investigating 3D structures of proteins is chemical cross-linking.[147–150] A common procedure is to covalently attach a bifunctional cross-linker to the protein and then cleave this protein with an endoprotease into smaller peptide segments; the unmodified protein is also cleaved similarly, and the resulting two sets of the peptide-mass maps are compared to reveal the identity of the cross-linked peptides. The MS/MS spectra of those peptides is also acquired to confirm their sequence and to pinpoint which two amino acid residues are cross-linked. The maximum distance between those two residues is assigned on the basis of the cross-linker arm length. This information is then used to construct the 3D structure of a protein using computational methods. Regents with varying spacer arm lengths and distinct solubility characteristics have been synthesized to cross-link different regions of proteins. For example, bis(sulfosuccinimidyl) suberate, being water soluble, is suitable for cross-linking lysines in hydrophilic or surface regions, whereas water-insoluble disuccinimidyl suberate and disuccinimidyl glutarate are both suitable for hydrophobic region lysines. Other amino acid residues that can be cross-linked are cysteine and tyrosine.

9.12.7.4 Ion Mobility Measurements for Studying Higher-Order Structures of Proteins

Ion mobility spectrometry (IMS), in which ions are separated on the basis of differences in the cross-sectional areas, can be used to determine the conformation and folding–unfolding kinetics of proteins.[151,152] The basic idea behind these measurements is that because of their distinct shapes, different conformers will travel at different rates in the IMS drift tube; the folded compact structure can be separated from a larger open structure as it travels faster. A typical IMS-based instrument, designed especially for these studies, consists of an ESI source, a mass-selecting quadrupole, an IMS drift tube, and a mass-measuring quadrupole. To identify different conformers, the solution of the protein is electrosprayed, a particular charge state is mass selected by the first quadrupole and pulsed into the IMS drift tube. The different conformers that get separated in the drift tube enter the second quadrupole where they are mass analyzed to determine their identity. Information about the number of isomers can be obtained from measured drift-time distributions.

9.12.8 Monitoring Protein–Ligand Interactions

Proteins exhibit binding affinity toward a variety of other molecular species; for example, protein–protein, enzyme–substrate, enzyme–inhibitor, antibody–antigen, receptor–ligand, protein–metal ion, and protein–DNA interactions commonly occur in biological systems. Many of these interactions are vital to body functions; they are involved in transport of small molecules, regulating biological events, controlling signaling, regulating protein function and assembly, and enhancing stability of proteins. Thus, the study of protein–ligand noncovalent interactions can aid in the fundamental understanding of many biochemical processes. Traditionally, investigators have used ultracentrifugation, light scattering, yeast two-hybrid (Y2H), immunoprecipitation, affinity chromatography, gel electrophoresis, gel-permeation chromatography, surface plasmon resonance, DNA microarrays, NMR, and X-ray crystallography. After the discovery of ESI and MALDI, mass spectrometry has become a valuable tool for the study of noncovalent protein–ligand interactions. Several years of research efforts have led to the development of a range of mass spectrometry-based methods for studying these interactions.[153–156] Any book by Downard is a good resource for this topic.

9.12.8.1 Electrospray Ionization for the Study of Noncovalent Interactions

One of the simple methods for investigating protein–ligand interactions is to acquire the ESI mass spectrum under the solvent environment that promotes complex formation.[153,155] Care must be exercised that the protein–ligand complex does not dissociate during the ESI process and its transport to the detector. To ensure the integrity of the complex, the solution pH must be close to the physiological pH, only volatile buffers must be used, the protein concentration must be low, and the interface region temperature and the skimmer voltage must also be kept low. The last two parameters must be optimized to strike a compromise; there must be sufficient heating and collision energy in the ion source region to ensure that nonspecific interactions are not observed, but not too high to cause dissociation of the protein–ligand complex. Readers are referred to two reviews on this subject for some practical examples.[153,155]

9.12.8.2 Matrix-Assisted Laser Desorption/Ionization for the Study of Noncovalent Interactions

MALDI has not been used as frequently as ESI for these studies, primarily owing to the high probability of disssociation of the protein–ligand complex during laser irradiation. The threshold energy for dissociation of intermolecular noncovalent bonds is approximately 20 kJ mol^{-1}, much less than a covalent bond dissociation energy (@ 200 kJ mol^{-1}). It is imperative that the appropriate matrix, protein concentration, pH, ionic strength, and solvent are used to detect noncovalent protein complexes via MALDI. The matrix acidity should not be too high; for example, an intact tetramer of streptovidin was observed when ferulic acid was used as a matrix but not with the more acidic 2,5-dihydroxybenzoic acid (DHB) matrix.[157] Similarly, multimers of yeast alcohol dehydrogenase and beef liver catalase were detected using dihydroxyacetophenone matrix[158] and a noncovalent complex between streptovidin and glutathione-*S*-transferase was observed at pH 3.8 with a ferulic acid matrix.[159] A lower pH and high concentration of organic solvents are counterproductive to in detecting protein complexes. Noncovalent complexes are usually stable at lower laser fluence (<10 mJ cm^{-2}). In this respect, IR lasers may have some advantages over UV lasers. One can also use the 'first-shot' phenomenon (i.e., the spectra collected following the first laser pulse) in the detection of protein complexes.[156] Another parameter that should be controlled is the source ion extraction voltage. At higher acceleration potentials, ionized protein complexes have kinetic energy high enough to cause dissociation as a result of collisons with the background gas molecules during their transport to the detector.

The use of surface-enhanced laser desorption/ionization (SELDI) is a viable alternative in the detection of noncovalent complexes.[160] In this approach, the protein-binding ligand is immobilized on a common chromatographic support; the protein sample is added onto the surface to form the complex. The surface-bound protein complex is analyzed using MALDI–MS after being washed free of unbound proteins and any interfering compounds. The interaction of S100 proteins have been studied using this appraoch.[160] Surface plasma resonance has also been coupled with MALDI–MS to detect noncovalent complexes.[161]

9.12.8.3 Hydrogen–Deuterium Exchange for the Study of Noncovalent Interactions

The hydrogen–deuterium exchange protocol discussed earlier in this chapter for studying higher-order protein structures also applies to the study of noncovalent interactions.[162,163] The premise underpinning these experiments is that the epitope of a protein participating in complex formation exchange to a lesser extent than those regions that are free and easily accessible to the solvent. Often, binding of a ligand leads to unfolding of the protein; the regions that get exposed to the solvent will show an increase in mass upon deuteration. In some complexes, only side chains are involved in the interaction; such interactions are difficult to detect by HX. Only time those complexes will be detected is when changes in the conformation of the protein also occur.

A typical HX procedure used in the study of noncovalent protein complexes is exemplified in the detection of antibody–antigen interactions; the protein is first deuterated at pH 7, passed through an antibody-packed column to form the antibody–antigen complex. The deuteriums are back exchanged and, after a fixed period, are quenched by lowering the pH to 2.5 and the temperature to 0 °C. Finally, the complex is digested with pepsin and the peptides produced are analyzed using LC–ESIMS; the epitopic peptides involved in binding will show an increase in the mass (due to retention of the D-label).

Protein–ligand interactions can also be detected by comparison method as long as there is change in the deuterium uptake.[162] The baseline HX data for individual proteins are obtained first and compared to similar data for those proteins in the complex. The location of changes due to complex formation is identfied by digesting the proteins and comparing the deuterium uptake for the two sets of peptides.

9.12.8.4 Limited Proteolysis for the Study of Noncovalent Interactions

Another simple approach for the detection of noncovalent interactions is to compare the peptide maps produced by proteolysis of the target protein and its complex with other ligands. This approach, also known as epitope mapping, relies on the fact that the two maps differ qualitatively because the contact regions of the interacting proteins are shielded from the protease's activity in the complex. An illustrative example is the detection of protein–protein interaction between a protein representing the kinase inhibitory domain of the cell cycle regulatory protein (p21-B) and cyclin-dependent kinase 2 (cdk).[164]

An offshoot of this protocol is the limited proteolysis strategy, in which proteolysis is restricted to a single event that cleaves the complex into two complementary fragments only; this identifies the most susceptible bond in the protein and hence the protease-accessible amino acid residue. To identify other accessible residues, proteolysis is performed in a time-course fashion with a suite of enzymes of diverse specificities. The potential of the limited proteolysis strategy has been demonstrated by mapping the topology of Ca^{2+}–calmodulin–melittin ternary complex.[165]

9.12.8.5 Chemical Cross-Linking for the Study of Protein–Protein Interactions

An approach similar to that used for the study of 3D structures can also be used to detect protein–protein interactions.[166] Two strategies have been developed for this purpose: bottom-up and top-down. In the former approach, the protein complex is cross-linked, as mentioned earlier, and digested proteolytically; the complete peptide map is then analyzed by mass spectrometry to provide the identity of the cross-linked peptides. In the top-down approach, the cross-linked protein is analyzed directly without undergoing proteolytic digestion. MS analysis includes measuring the accurate mass and acquiring the MS/MS spectrum of the cross-linked protein; the measured mass provides an idea of the number of incorporated cross-linkers and thus the modification sights.

9.12.8.6 Functional Proteomics and Protein–Protein Interactions

In living systems, most proteins express their function when in association with other proteins. Understanding of these associations is necessary to define cellular machinery. Coimmunoprecipitation (Co-IP), Y2H, protein microarrays, chemical cross-linking, and phage display are some of the techniques that have made a big impact in our understanding of cellular signaling networks.[167–175] Some of them have been combined with mass spectrometry-based proteomics for a system-wide study of the protein–protein interaction networks.[167] The terms ‘interactomics’ and ‘interactomes’ have been coined in parallel with proteomics and proteomes to address the expanding field of studying protein networks.

In Co-IP experiments, the target protein, along with associated proteins, is immunoprecipitated.[168] To this end, the protein whose interacting partners are being searched is attached to a solid matrix and packed in a column. The mixture of proteins is passed through the column; all proteins that form a complex with the target protein are retained by the column. The interacting proteins are eluted from the column and resolved using a 1D-SDS–PAGE system. After staining, the protein spots are picked up, tryptic digested, and analyzed by MALDI–TOFMS; a database search of the molecular mass data of peptides identifies the proteins that were involved in the interaction with the target protein. Alternately, the coimmunoprecipitated proteins are identified using shotgun proteomics with LC–ESI–MS/MS analysis of the protein digest.[169]

An I-DIRT (isotopic differentiation of interactions as random or targeted) strategy has been developed to distinguish between specific and nonspecific protein interactions.[170] For this, one set of cells is grown in a light and the other in a heavy isotopic medium in the presence of the affinity-tagged proteins; the two samples are mixed in a 1:1 ratio and the protein of interest and its complex are immunoprecipitated and analyzed by mass spectrometry; the specific interactions appear as isotopically light signal only, whereas combined light and heavy isotopes signal indicate nonspecific interactions. Co-IP has also been coupled with protein quantification techniques for comprehensive protein–protein interactome study.[171]

The Y2H is one of the most widely used technologies for studying protein–protein interactions.[172,173] In this approach, the reaction is studied inside the nucleus of yeast. A common approach is to create the plasmids containing the DNA encoding the target protein and its partner protein and to fuse the two plasmids by expressing the pair of hybrid genes in yeast cells. The signal due to the activation of the reporter gene indicates the protein–protein interaction. The DNA of the fusion protein is isolated and sequenced to provide identification of the associated proteins. For high-throughput experiments, specific test proteins are screened against large compilations of randomly cloned proteins, and interacting proteins are identified by searching a Y2H cDNA or some other library.[173]

The use of protein microarrays has also made a big impact in the study of protein–protein interactions. In this method, thousands of proteins are spotted in a microarray format at known locations on a chip, then the solution of the target protein is spread on the chip. Interacting proteins are detected by fluorescent imaging, surface plasmon resonance, or mass spectrometry.[174–177]

A comprehensive proteomics approach has been developed to identify the components of phosphoprotein complexes and includes the following experimental steps: isolation of native phosphoproteins and associated proteins by affinity chromatography, 1D PAGE separation of the affinity-isolated proteins, proteolysis of the protein spots, LC–ESI–MS/MS analysis of the cleaved peptides, and database searching for indentification of the proteins involved in the complex formation.[178]

9.12.9 Conclusions

This chapter presented contemporary mass spectrometry-based approaches that are used to investigate the structures and functions of proteins and peptides. The theme of the chapter centered around proteomics, an enabling technology for large-scale analysis of proteins. The treatment of these procedures is by no means complete. Over the years, mass spectrometry has witnessed a dramatic growth and still continues to expand;

existing methods are constantly evolving, and new techniques are emerging fast. In view of these, it is an impossible task to provide a thorough coverage of all of the existing methods. Who could imagine 20–25 years ago that it would be possible for mass spectrometry to play a major role in a systemwide investigation of biological processes and be an indispensable component of a contemporary biochemistry laboratory? The future will see an even more expanded role of mass spectrometry in biochemical research and in unraveling the secret of life.

Abbreviations

2-DE	two-dimensional gel electrophoresis
AQUA	absolute quantification
CCB	colloidal Coomassie Blue
CHAPS	3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonate
CNBr	cyanogen bromide
Co-IP	coimmunoprecipitation
CSD	charge-state distribution
CSF	cerebrospinal fluid
D	dimensional
DIGE	differential imaging gel electrophoresis
DNA	deoxyribonucleic acid
FFE	free-flow fractionation
HX	hydrogen–deuterium exchange
ICAT	isotope-coded affinity tag
I-DIRT	isotopic differentiation of interactions as random or targeted
IEF	isoelectric focusing
IMAC	immobilized metal ions affinity chromatography
IPG	immobilized pH gradient gel
iTRAQ	isobaric amine-specific labeling reagents
MASCOT	a database search engine
mRNA	messenger ribonucleic acid
NMR	nuclear magnetic resonance
PAGE	polyacrylamide gel electrophoresis
p*I*	isoelectric point
PNGase	*N*-glycosidase, an enzyme used to cleave carbohydrate chain from *N*-glycoproteins/peptides
Proteomics	systematic study of the proteome
PSD	postsource decay
PTM	post-translational modification
RNase	ribonuclease
SCX	strong cation-exchange
SDS	sodium dodecyl sulfate
SELDI	surface-enhanced laser desorption/ionization
SEQUEST	database search engine
SPE	solid-phase extraction
X!TANDEM	database search engine
Y2H	yeast two-hybrid

References

1. T. Hunter, Protein Kinase Classification. In *Methods in Enzymology*; T. Hunter, B. M. Sefton, Eds.; Academic Press: San Diego, 1991; Vol. 200, pp 3–37.
2. J. B. Fenn; M. Mann; C. K. Meng; S. F. Wong; C. M. Whitehouse, *Science* **1989**, *246*, 64–71.
3. M. Karas; F. Hillenkamp, *Anal. Chem.* **1998**, *60*, 2299–2301.
4. K. Tanaka; H. Waki; H. Ido; S. Akita; T. Yoshida, *Rapid Commun. Mass Spectrom.* **1988**, *2*, 151–153.
5. C. Dass, *Principles and Practice of Biological Mass Spectrometry*; Wiley-Interscience: Hoboken, NJ, 2000.
6. C. Dass, *Fundamentals of Contemporary Mass Spectrometry*; Wiley-Interscience: Hoboken, NJ, 2007.
7. R. Aebersold; D. Goodlett, *Chem. Rev.* **2001**, *101*, 269–295.
8. M. R. Wilkins; C. Pasqualli; R. D. Appel; K. Ou; O. Golaz; J. C. Sanchez; J. X. Yan; A. A. Gooley; G. Hughes; I. Humphery-Smith; K. L. Williams; D. F. Hochstrasser, *Bio-Technology* **1996**, *14*, 61–65.
9. K. Bjorhall; T. Miliotis; P. Davidsson, *Proteomics,* **2005**, *5*, 307–317.
10. M. Dreger, *Eur. J. Biochem.* **2003**, *270*, 589–599.
11. H. Zischka; G. Weber; P. J. A. Weber; A. Posch; R. J. Braun; D. Buhringer; U. Schneider; M. Nissum; T. Meitinger; M. Ueffing; C. Eckerskorn, *Proteomics* **2003**, *3*, 906–916.
12. X. Sun; J.-F. Chiu; Q.-Y. He, *Methods Mol. Biol.* **2008**, *424*, 205–212.
13. J. Klose; U. Kobalz, *Electrophoresis* **1995**, *16*, 1034–1059.
14. F. Van; B. Subramanian; A. Nakeff; T. Barder; S. J. Parus; D. M. Lubman, *Anal. Chem.* **2003**, *75*, 2299–2308.
15. R. W. Nelson, *Mass Spectrom. Rev.* **1997**, *16*, 353–376.
16. N. L. Kelleher, *Anal. Chem.* **2004**, *76*, 197A–203A.
17. J. S. Cottrell, *Pept. Res.* **1994**, *7*, 115–118.
18. J. R. Yates, III, *J. Mass Spectrom.* **1998**, *33*, 1–19.
19. W. J. Henzel; T. M. Billeci; J. T. Stults; S. C. Wong; C. Grimley; C. Watanbe, *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 5011–5015.
20. D. J. C. Pappin; P. Hojrup; A. J. Bleasby, *Curr. Biol.* **1993**, *3*, 327–332.
21. B. Thiede; W. Hoehenwarter; A. Krah; J. Mattow; M. Schmid; F. Schmidt; P. R. Jungblut, *Methods* **2005**, *35*, 237–247.
22. J. R. Yates; S. Speicher; P. R. Griffin; T. Hunkapillar, *Anal. Biochem.* **1993**, *214*, 397–408.
23. M. Mann; P. Hojrup; P. Roepstorff, *Biol. Mass Spectrom.* **1993**, *22*, 338–345.
24. P. James; M. Quadroni; E. Carafoli; G. Gonnet, *Biochem. Biophys. Res. Commun.* **1993**, *195*, 58–64.
25. S. P. Gygi; G. L. Corthals; Y. Zhang; Y. Rochon; R. Aebersold, *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 9390–9395.
26. A. Gorg; W. Weiss; M. J. Dunn, *Proteomics* **2004**, *4*, 3665–3685.
27. C. Dass, *Curr. Org. Chem.* **1999**, *3*, 193–209.
28. M. T. Davis; T. D. Lee, *J. Am. Soc. Mass Spectrom.* **1998**, *9*, 194–201.
29. J. R. Yates; J. K. Eng; A. L. McCormack; D. M. Schieltz, *Anal. Chem.* **1995**, *67*, 1426–1436.
30. J. K. Eng; A. L. McCormack; J. R. Yates, III, *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976–989.
31. J. R. Yates, III; A. L. McCormack, *Anal. Chem.* **1996**, *68*, 534A–540A.
32. J. R. Yates, III; S. F. Morgan; C. L. Gatlin; P. R. Griffin; J. K. Eng, *Anal. Chem.* **1998**, *70*, 3557–3565.
33. A. Qualtieri; E. Urso; M. Le Pera; M. Scornaienchi; A. Quattrone; L. Di Donna; A. Napoli; G. Sindona, *J. Am. Soc. Mass Spectrom.* **2006**, *17*, 117–123.
34. D. A. Wolters; M. P. Washburn; J. R. Yates, III, *Anal. Chem.* **2001**, *73*, 5683–5690.
35. T. Kislinger; A. O. Gramolini; D. H. MacLennan; A. Emili, *J. Am. Soc. Mass Spectrom.* **2005**, *16*, 1207–1220.
36. C. Delahunty; J. R. Yates, III, *Methods* **2005**, *35*, 248–255.
37. D. E. McNulty; R. S. Annan, *Mol. Cell. Proteomics* **2008**, *7*, 971–980.
38. A. J. Link; J. Eng; D. M. Schieltz; E. Carmack; G. J. Mize; D. R. Morris; B. M. Garvik; J. R. Yates, III, *Nat. Biotechnol.* **1999**, *17*, 676–682.
39. P. R. Jalili; C. Dass, *Rapid Commun. Mass Spectrom.* **2004**, *18*, 1877–1884.
40. Y. Shen; N. Tolic; C. Masselon; L. Pasa-Tolic; D. G. Camp, II; K. K. Hixson; R. Zhao; G. A. Anderson; R. D. Smith, *Anal. Chem.* **2004**, *77*, 144–154.
41. A. D. Norbeck; M. E. Manroe; J. N. Adkins; K. K. Anderson; D. S. Daly; R. D. Smith, *J. Am. Soc. Mass Spectrom.* **2005**, *16*, 1239–1249.
42. B. J. Cargile; J. L. Stephenson, Jr., *Anal. Chem.* **2006**, *76*, 267–275.
43. M. J. McCass, *Anal. Chem.* **2005**, *77*, 294A–302A.
44. A. Gorg; W. Weiss; M. J. Dunn, *Proteomics* **2004**, *4*, 3665–3685.
45. E. Marengo; E. Robotti; F. Antonucci; D. Cecconi; N. Campostrini; P. G. Righetti, *Proteomics* **2005**, *5*, 654–666.
46. M. Unlu; M. E. Morgan; J. S. Minden, *Electrophoresis* **1997**, *18*, 2071–2077.
47. H. Hu; J. P. Malone; A. M. Fagon; R. R. Townsend; D. M. Holtman, *Mol. Cell. Proteomics,* **2005**, *4*, 2000–2009.
48. S. P. Gygi; B. Rist; S. A. Gerber; F. Turcek; M. H. Gelb; R. Aebersold, *Nat. Biotechnol.* **1999**, *17*, 994–999.
49. P. L. Ross; Y. N. Huang; J. N. Marchese; B. Williamson; K. Parker; S. Hattan; N. Khainovski; S. Pillai; S. Dey; S. Daniels; S. Purkayastha; P. Juhasz; S. Martin; M. Bartlet-Jones; F. He; A. Jacobson; D. J. Pappin, *Mol. Cell. Proteomics* **2004**, *3*, 1154–1169.
50. Y. Ogata; M. C. Chrlesworth; L. Higgins; B. M. Keegan; S. Vernino; D. C. Muddiman, *Proteomics* **2007**, *7*, 3726–3734.
51. T. J. Griffins; H. Xei; S. Bhandakavi; J. Popko; A. Mohan; J. V. Carlis; L. Higgins, *J. Proteome Res.* **2007**, *6*, 4200–4209.
52. X. D. Yao; C. Afonso; C. Fenselau, *J. Proteome Res.* **2003**, *2*, 147–152.
53. N. Ibarrola; D. E. Klume; M. Gronborg; A. Iwahori; A. Pandey, *Anal. Chem.* **2003**, *75*, 6043–6049.
54. P. V. Bondarenko; D. Chelius; T. A. Shaler, *Anal. Chem.* **2002**, *74*, 4741–4749.
55. D. S. Kirkpatrick; S. A. Gerber; S. P. Gygi, *Methods* **2005**, *35*, 265–273.
56. M. T. Olsen; J A. Epstein; A. L. Yargey, *J. Am. Soc. Mass Spectrom.* **2006**, *17*, 1041–1049.
57. D. F. Hunt; J. R. Yates; J. Shabanowitz; S. Winston; C. R. Hauer, *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83*, 6233–6237.

58. K. Biemann; S. A. Martin, *Mass Spectrom. Rev.* **1987**, *6*, 1–75.
59. I. A. Papayannopoulos, *Mass Spectrom. Rev.* **1995**, *14*, 49–73.
60. C. Dass; D. M. Desiderio, *Anal. Biochem.* **1987**, *163*, 52–66.
61. T. Yalcin; C. Khouw; I. G. Csizmadia; M. R. Peterson; A. G. Harrison, *J. Am. Soc. Mass Spectrom.* **1996**, *7*, 233–242.
62. X.-J. Tang; R. K. Boyd, *Rapid Commun. Mass Spectrom.* **1992**, *6*, 651–657.
63. R. S. Johnson; S. A. Martin; K. Biemann, *Int. J. Mass Spectrom. Ion Process* **1988**, *86*, 137–154.
64. C. K. Barlow; R. A. J. O'Hair, *J. Mass Spectrom.* **2008**, *43*, 1301–1319.
65. V. H. Wysocki; G. Tsaprailis; L. L. Smith; L. A. Breci, *J. Mass Spectrom.* **2000**, *35*, 1399–1406.
66. P. Roepstorff; J. Fohlman, *Biomed. Mass Spectrom.* **1984**, *11*, 601.
67. K. Biemann, *Biomed. Environ. Mass. Spectrom.* **1988**, *16*, 99–111.
68. L. L. Smith; K. A. Herrmann; V. H. Wysocki, *J. Am. Soc. Mass Spectrom.* **2006**, *10*, 20–28.
69. R. S. Johnson; S. Martin; K. Biemann; J. T. Stults; J. T. Watson, *Anal. Chem.* **1987**, *59*, 2621–2625.
70. C. Dass; P. Mahalakshmi, *Rapid Commun. Mass Spectrom.* **1995**, *9*, 1148–1154.
71. C. Dass, *Rapid Commun. Mass Spectrom.* **1989**, *3*, 264–266.
72. J. A. Loo; C. G. Edmonds; R. D. Smith, *Anal. Chem.* **1993**, *65*, 425–438.
73. T. V. Vaisar; J. Urban, *J. Mass Spectrom.* **1996**, *31*, 1185–1187.
74. D. L. Tabb; L. L. Smith; L. A. Breci; V. H. Wysocki; D. Lin; J. R. Yates, *Anal. Chem.* **2003**, *75*, 1155–1163.
75. A. Shevchenko; O. N. Jensen; A. V. Podtelejnikov; F. Sagliocco; M. Wilm; O. Vorm; P. Mortensen; A. Shevchenko; H. Boucherie; M. Mann, *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 14440–14445.
76. E. Mirgorodskya; P. Roepstroff; R. A. Zubarev, *Anal. Chem.* **1999**, *71*, 4431–4436.
77. R. A. Zubarev; N. A. Kruger; E. K. Fridriksson; M. A. Lewis; D. M. Horn; B. K. Carpenter; F. W. McLafferty, *J. Am. Chem. Soc.* **1999**, *121*, 2857–2862.
78. J. E. P. Syka; J. J. Coon; M. J. Schroeder; J. Shabanowitz; D. F. Hunt, *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 9528–9533.
79. D. E. Kalume; H. Molina; A. Pandey, *Curr. Opin. Chem. Biol.* **2003**, *7*, 64–69.
80. B. A. Garcia; J. Shabanowitz; D. F. Hunt, *Methods* **2005**, *35*, 256–264.
81. B. W. Gibson; A. M. Falick; A. L. Burlingame; L. Nadasdi; A. C. Nguyen; G. L. Kenyon. *J. Am. Chem. Soc.* **1987**, *109*, 5343–5348.
82. M. O. Collins; L. Yu; J. S. Choudhary, *Proteomics* **2007**, *7*, 2751–2768.
83. C. Temporini; E. Calleri; G. Massolini; G. Caccialanza, *Mass Spectrom. Rev.* **2008**, *27*, 207–236.
84. A. Paradela; J. P. Albar, *J. Proteome Res.* **2008**, *7*, 1809–1818.
85. M. Rossignol, *Curr. Opin. Plant Biol.* **2006**, *9*, 538–543.
86. J. C. Smith; D. Figeys, *Biochem. Cell Biol.* **2008**, *86*, 137–148.
87. Z. A. Knight; B. Schilling; R. H. Row; D. M. Kenski; B. W. Gibson; K. M. Shokat, *Nat. Biotechnol.* **2003**, *21*, 1047–1054.
88. J. Rush; A. Moritz; K. A. Lee; A. Guo; V. L. Goss; E. J. Speck; H. Zhang; X. M. Zha; T. D. Polkiewicz; M. J. Comb, *Nat. Biotechnol.* **2005**, *23*, 94–101.
89. W. Zhou; B. A. Merrick; M. G. Khaledi; K. B. Tomer, *J. Am. Soc. Mass Spectrom.* **2000**, *11*, 273–282.
90. S. Li; C. Dass, *Anal. Biochem.* **1999**, *270*, 9–14.
91. M. W. H. Pinske; M. P. Uitto; M. J. Hillhorst; B. Ooms; J. R. Albert, *Anal. Chem.* **2004**, *76*, 3935–3943.
92. S. Xu; J. C. Whitin; T. T-Sang Yu; H. Zhou; D. Sun; H.-J. Sue; H. Zou; H. J. Cohen; R. N. Zare, *Anal. Chem.* **2008**, *80*, 5542–5549.
93. J. Wei; Y. Zhang; J. Wang; F. Tan; J. Liu; Y. Cai; X. Quan, *Rapid Commun. Mass Spectrom.* **2008**, *22*, 1069–1080.
94. S. B. Ficarro; M. L. McCleland; P. T. Stukenberg; D. J. Burke; M. M. Ross; J. Shabanowitz; D. F. Hunt, *Nat. Biotechnol.* **2002**, *20*, 301–305.
95. D. T. McLachlin; B. T. Chait, *Anal. Chem.* **2003**, *75*, 6826–6836.
96. H. Zhou; J. D. Watts; and R. Aebersold, *Nat. Biotechnol.* **2001**, *19*, 375–378.
97. Q. Xia; D. Cheng; D. M. Duong; M. Gearing; J. J. Lah; A. I. Levey; J. Peng, *J. Proteome Res.* **2008**, *7*, 2845–2851.
98. M. J. Huddleston; R. S. Annan; M. F. Bean; S. A. Carr, *J. Am. Soc. Mass Spectrom.* **1993**, *4*, 710–717.
99. X. Zhu; C. Dass, *J. Liq. Chromatgr. Relat. Techn.* **1999**, *22*, 1635–1647.
100. M. Wilm; G. Neubauer; M. Mann, *Anal. Chem.* **1996**, *68*, 527–533.
101. H. Steen; B. Kuster; M. Fernandez; A. Pandey; M. Mann, *Anal. Chem.* **2001**, *73*, 1440–1448.
102. H. Steen; M. Mann, *J. Am. Soc. Mass Spectrom.* **2002**, *13*, 996–1003.
103. A. Schlosser; R. Ripkorn; D. Bossemeyer; W. D. Lehmann, *Anal. Chem.* **2001**, *73*, 170–176.
104. J. P. DeGnore; J. Quin, *J. Am. Soc. Mass Spectrom.* **1998**, *9*, 1175–1188.
105. M. C. Crowe; J. S. Brodbelt, *J. Am. Soc. Mass Spectrom.* **2004**, *15*, 1581–1592.
106. N. Sadagopan; M. Malone; J. T. Watson, *J. Mass Spectrom.* **1999**, *34*, 1279–1282.
107. H. Jaffe; H. C. Pant, *Biochemistry* **1998**, *37*, 16211–16224.
108. A. Stensballe; O. N. Jensen; J. V. Olsen; K. F. Haselmann; R. A. Zubarev, *Rapid Commun. Mass Spectrom.* **2000**, *14*, 1793–1800.
109. S. D. H. Shi; M. E. Hemling; S. A. Carr; D. M. Horn; I. Lindh; F. W. McLafferty, *Anal. Chem.* **2001**, *73*, 19–22.
110. M. J. Chalmers; W. Kolch; M. R. Emmett; A. G. Marshall; H. Mischak, *J. Chromatgr. B* **2004**, *803*, 111–120.
111. J. J. Coon; J. Shabanowitz; D. F. Hunt; J. E. P. Syka, *J. Am. Soc. Mass Spectrom.* **2005**, *16*, 880–882.
112. J. W. Flora; D. C. Muddiman, *Anal. Chem.* **2001**, *73*, 3305–3311.
113. J. H. Lee; Y. Kim; M. Y. Ha; E. K. Lee; J. Choo, *J. Am. Soc. Mass Spectrom.* **2005**, *16*, 1456–1460.
114. K. F. Medzihradszky; D. A. Maltby; S. C. Hall; C. A. Settineri; A. L. Burlingame, *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 350–358.
115. S. A. Carr; M. J. Huddleston; M. F. Bean, *Protein Sci.* **1993**, *2*, 183–196.
116. M. J. Kieliszewski; M. O'Neil; J. Leykam; R. Orlando, *J. Biol. Chem.* **1995**, *270*, 2541–2549.
117. F.-G. Hanisch; B. N. Green; R. H. Bateman; J. Peter-Katalinic, *J. Mass Spectrom.* **1998**, *33*, 358–362.
118. K. Kubota; Y. Sato; Y. Suziki; N. Goto-Inaoue; T. Toda; M. Suziki; S.-I. Hisanaga; A. Suziki; T. Endo, *Anal. Chem.* **2008**, *80*, 3693–3698.
119. B. Sun; J. A. Ranish; A. G. Utleg; J. T. White; X. Yan; B. Lin; L. Hood, *Mol. Cell. Proteomics* **2007**, *6*, 141–149.
120. G. Alvarez-Manilla; J. Atwood, III; Y. Guo; N. L. Warren; R. Orlando; M. Pierce, *J. Proteome Res.* **2006**, *5*, 701–708.

121. G. E. Reid; J. L. Stepheson, Jr.; S. A. McLuckey, *Anal. Chem.* **2002**, *74*, 577–583.
122. X. Cai; C. Dass, *Curr. Org. Chem.* **2003**, *7*, 1841–1854.
123. J. Skolnick; A. Kolinski, *Comput. Sci. Eng.* **2001**, *3*, 40–50.
124. D. L. Smith; Y. Deng; Z. Zhang, *J. Mass Spectrom.* **1997**, *32*, 135–146.
125. S. D. Maleknia; K. M. Downard, *Mass Spectrom. Rev.* **2001**, *20*, 388–401.
126. A. Miranker; C. V. Robinson; S. E. Radford; C. M. Dobson, *FASEB J.* **1996**, *10*, 93–101.
127. J. R. Engin; D. L. Smith, *Anal. Chem.* **2001**, *73*, 256A–265A.
128. I. Kaltashov; S. J. Eyles, *J. Mass Spectrom.* **2002**, *37*, 557–565.
129. L. Konermann; D. A. Simmons, *Mass Spectrom. Rev.* **2003**, *22*, 1–26.
130. X. Yan; J. Watson; P. S. Ho; M. L. Deinzer, *Mol. Cell. Proteomics* **2004**, *3.1*, 10–23.
131. S. K. Chowdhury; V. Katta; B. T. Chait, *J. Am. Chem. Soc.* **1990**, *112*, 9012–9013.
132. L. Konermann; D. J. Douglas, *Rapid Commun. Mass Spectrom.* **1998**, *12*, 435–442.
133. A. Dobo; I. A. Kaltashov, *Anal. Chem.* **2001**, *73*, 4763–4773.
134. D. Fabris, *Mass Spectrom. Rev.* **2005**, *24*, 30–54.
135. X. Cai; C. Dass, *Eur. J. Mass Spectrom.* **2007**, *13*, 2341–2346.
136. L. Konermann; J. Pan; D. A. Simmons; D. J. Wilson, Time-Resolved Electrospray Ionization. In *The Encyclopedia of Mass Spectrometry*; M. L. Gross, R. M. Caprioli, Eds.; Elsevier: Amsterdam, 2007; pp 802–810.
137. H. Lin; C. Dass, *Eur. J. Mass Spectrom.* **2002**, *8*, 381–387.
138. A. Hvidt; K. Linderstrom-Lang, *Biochem. Biophys. Acta* **1954**, *14*, 574–575.
139. V. Katta; B. T. Chait, *Rapid Commun. Mass Spectrom.* **1991**, *5*, 214–217.
140. S. W. Englander, *J. Am. Soc. Mass Spectrom.* **2006**, *17*, 1481–1489.
141. D. A. Simmons; S. D. Dunn; L. Konermann, *Biochemistry* **2003**, *42*, 5896–5905.
142. D. A. Simmons; L. Konermann, *Biochemistry* **2002**, *42*, 1906–1914.
143. A. Miranker; C. V. Robinson; S. E. Radford; R. T. Aplin; C. M. Dobson, *Science* **1993**, *262*, 896–900.
144. G. Thevenon-Emeric; J. Kozlowski; Z. Zhang; D. L. Smith, *Anal. Chem.* **1992**, *64*, 2456–2458.
145. Y. Deng; Z. Zhang; D. L. Smith, *J. Am. Soc. Mass Spectrom.* **1999**, *10*, 675–684.
146. P. Jallili; C. Dass, *Rapid Commun. Mass Spectrom.* **2008 (2005)**, *22*, 1–8.
147. B. Schilling; R. H. Row; B. W. Gibson; X. Guo; M. M. Young, *J. Am. Soc. Mass Spectrom.* **2003**, *14*, 834–850.
148. A. Sinz, *J. Mass Spectrom.* **2003**, *38*, 1225–1237.
149. B. X. Huang; C. Dass; H.-Y. Kim, *J. Am. Soc. Mass Spectrom.* **2004**, *15*, 1237–1247.
150. B. X. Huang; C. Dass; H.-Y. Kim, *Biochem. J.* **2005**, *387*, 695–702.
151. K. Shelimov; M. F. Jarrold, *J. Am. Chem. Soc.* **1997**, *119*, 2987–2994.
152. R. W. Purves; D. A. Barnnett; B. Ells; R. Guevermont, *J. Am. Soc. Mass Spectrom.* **2000**, *11*, 738–745.
153. J. A. Loo, *Mass Spectrom. Rev.* **1997**, *16*, 1–23.
154. B. N. Pramanik; P. L. Bartner; U. A. Mirza; Y.-H. Liu; A. K. Ganguly, *J. Mass Spectrom.* **1998**, *33*, 911–920.
155. J. A. Loo, *Int. J. Mass Spectrom.* **2000** *200*, 175–186.
156. T. B. Farmer; R. M. Caprioli, *J. Mass Spectrom.* **1998**, *33*, 697–704.
157. L. R. H. Cohen; K. Strupat; F. Hillenkamp, *J. Am. Soc. Mass Spectrom.* **1997**, *8*, 1046–1052.
158. B. Rosinke; K. Strupat; F. Hillenkamp; J. Rosenbusch; N. Dencher; U. Krüger; H. J. Galla, *J. Mass Spectrom.* **1995**, *30*, 1462–1468.
159. S. Jespersen; W. M. A. Niessen; U. R. Tjaden; J. van der Greef, *J. Mass Spectrom.* **1998**, *33*, 1088–1093.
160. R. Lehman; C. Melle; N. Escher; F. von Eggling, *J. Proteome Res.* **2005**, *4*, 1717–1721.
161. D. Nedelkov; R. W. Nelson, *J. Mol. Recognit.* **2003**, *16*, 9–14.
162. J. Engen, *Analyst (Lond.)* **2003**, *128*, 623–628.
163. C. V. Robinson; M. Groβ; S. J. Eyles; J. J. Ewbank; M. Mayhew; F. U. Hartl; C. M. Dobson; S. E. Radford, *Nature* **1996**, *372*, 645–651.
164. R. Kriwacki; J. Wu; G. Siuzdak; P. E. Wright, *J. Am. Chem. Soc.* **1996**, *118*, 5320–5321.
165. A. Scaloni; N. Miraglia; S. Orru; P. Amodeo; A. Motta; G. Marino; P. Pucci, *J. Mol. Biol.* **1998**, *277*, 945–958.
166. A. Sinz, Chemical Cross-Linking and Mass Spectrometry for Investigation of Protein-Protein Interactions. In *Mass Spectrometry of Protein Interactions*; K. M. Downward, Ed.; Wiley-Interscience: Hoboken, NJ, 2007; pp 83–107.
167. T. Koecher; G. Superti-Furga, *Nat. Methods* **2007**, *4*, 807–815.
168. K. Markham; Y. Bai; G. Schmitt-Ulms, *Anal. Bioanal. Chem.* **2007**, *389*, 461–473.
169. R. M. Ewing; P. Chu; F. Elisma; H. Li; P. Taylor; S. Climie; L. McBroom-Cerajewski; M. D. Robinson; L. O'Connor; M. Li; R. Taylor; M. Dharsee; Y. Ho; A. Heilbut; L. Moore; S. Zhang; O. Ornatsky; Y. V. Bukhman; M. Ethier; Y. Sheng; J. Vasilescu; M. Abu-Farha; J.-P. Lambert; H. S. Duewel; I. I. Stewart; B. Kuehl; K. Hogue; K. Colwill; K. Gladwish; B. Muskat; R. Kinach; S.-L. Adams; M. F. Moran; G. B. Morin; T. Topaloglou; D. Figeys, *Mol. Syst. Biol.* **2007**, *3*, 89.
170. D. J. Lee; L. F. Westblade; B. T. Chait, *J. Bacteriol.* **2008**, *190*, 1284–1289.
171. X. Wang; L. Huang, *Mol. Cell. Proteomics* **2008**, *7*, 46–67.
172. V. Ratushny; E. A. Golemis, *BioTechniques* **2008**, *44*, 655–662.
173. J. DeGardo-Warren; M. Dufford; J. Chen; P. L. Bartel; D. Shattuck; G. C. Frech, *BioTechniques* **2008**, *44*, 265–273.
174. N. Ramachandran; H. Hainsworth; B. Bhullar; S. Eisenstein; B. Rosen; A. Y. Lau; J. C. Walter; J. LaBaer, *Science* **2004**, *306*, 86–90.
175. P. Mitchell, *Nat. Biotechnol.* **2002**, *20*, 225–229.
176. D. Nedelkov, *Anal. Chem.* **2007**, *79*, 5987–5990.
177. J. S. Yuk; K.-S. Ha, *Exp. Mol. Med.* **2005**, *37*, 1–10.
178. K. Kristansdottir; D. Wolfjeher; N. Lucius; D. S. Angulo; S. J. Kron, *J. Proteome Res.* **2008**, *7*, 2812–2824.

Biographical Sketch

Dr. Chhabil Dass currently holds the position of an associate professor in the Department of Chemistry, The University of Memphis, Memphis, Tennessee, USA. He received his B.Sc. and M.Sc. degrees from the University of Rajasthan, Jaipur, India in 1962 and 1964, respectively. After completing a 1 year training, he joined Bhabha Atomic Research Center, Mumbai in 1965 as a scientific officer. He remained at this institute until August 1979, after which he immigrated to the United States. During this period, 1 year (September 1969–August 1970) was spent at the Central Bureau for Nuclear Measurements in Geel, Belgium. After completing his Ph.D. in 1984 from the University of Nebraska, Lincoln, he joined the Department of Neurology, University of Tennessee Health Science Center as a research associate, and later rose to the ranks of assistant professor (1986) and associate professor (1994). He continues to hold the position of associate adjunct professor at this institute. He is a member of the American Society for Mass Spectrometry, American Chemical Society, and Indian Society for Mass Spectrometry. His research interests are in proteomics and applications of mass spectrometry in structure determination of proteins and peptides. He has published over 90 research articles and has authored two books on mass spectrometry: *Principles and Practice of Biological Mass Spectrometry* (2000) and *Fundamentals of Contemporary Mass Spectrometry* (2007). His hobbies include photography and playing bridge. He and his wife Asha Midha are proud parents of two sons Hemesh (wife Hema) and Yatesh (wife Suzane) and three lovely granddaughters Grace, Madeline, and Hannah.

9.13 Application of Mass Spectrometry to Rapid Analysis of Bacterial Polysaccharides

Jianjun Li and Eleonora Altman, Institute for Biological Sciences, Ottawa, ON, Canada

9.13.1 Introduction

Lipopolysaccharides (LPSs) and capsular polysaccharides (CPSs) are major components of the outer membrane of Gram-negative bacteria. These complex membrane components are believed to be essential for normal maintenance, growth, and reproduction of bacterial cells. LPS typically consists of a hydrophobic domain known as lipid A (or endotoxin), a nonrepeating core oligosaccharide region, and a distal O-chain polysaccharide (O-PS) (the O-antigen or O-chain) consisting of oligosaccharide repeating units usually containing one to five or six monosaccharides.[1,2] The majority of CPSs are heteroglycans and are generally, but not always, acidic. The acidic component is most often a uronic acid, a pyruvic acid acetal, or a phosphate diester group.[3] Many of CPSs represent attractive vaccine candidates,[4,5] while application of the LPSs to the vaccine development has been slow forthcoming.[6,7]

In order to understand the antigenic and the immunogenic properties of polysaccharides (PSs) it is important to delineate their precise chemical structure. This involves analysis of the sugar composition, the linkage analysis, determination of the sugar ring configurations (pyranose or furanose), the sequence analysis, and determination of the anomeric configuration (α or β) of the constituent monosaccharides. Additionally, PSs may contain noncarbohydrate substituents, such as *O*-acetyl and phosphate groups, pyruvate, and amino acids, which often influence immunogenicity and antigenicity of PSs and, therefore, it is important to confirm their presence. The qualitative analysis of PSs could be achieved by chemical, usually acid, or enzymatic degradation, followed by derivatization of the resultant sugars and gas–liquid chromatography–mass spectrometry (GLC–MS) of their derived alditol acetates. In order to gain information on the linkages of the constituent monosaccharides, methylation analysis is performed. The resultant partially methylated alditol acetates are analyzed by GLC–MS. Methylation analysis is an important tool in structural analysis and allows for determination of the linkages of constituent monosaccharides as well as their ring configurations. In addition, it confirms the number of amino sugars present and determines whether a PS has a linear or branched structure. Other degradation and derivatization techniques, such as partial hydrolysis and periodate oxidation provide information on the sequence of PSs. However, these traditional chemical degradation techniques require more material for analysis and have been largely replaced by nondegradable nuclear magnetic resonance (NMR) spectroscopy methods allowing complete structural analysis using milligram quantities of the purified PSs (for review see Uhrin and Brisson[8]).

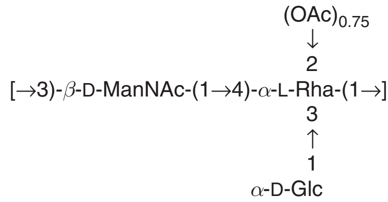
Mass spectrometry (MS) has been widely used in the analysis of biomolecules, such as MS-based approaches for proteomics (reviewed in Lee *et al.*,[9] Nesvizhskii *et al.*,[10] Ahn *et al.*,[11] Qian *et al.*,[12] Mueller *et al.*,[13] Hitchen and Dell,[14] and Nedelkov *et al.*[15]), glycomics,[16–20] lipidomics,[21–23] and metabonomics.[24,25] Of the many mass spectrometric techniques that have been applied to glycomics, electrospray ionization–mass spectrometry (ESI–MS), and matrix-assisted laser desorption/ionization mass spectrometer (MALDI-MS) are the most important. These techniques make the power and elegance of mass spectrometric analysis applicable to the large and fragile polar molecules in biological systems.[26,27] ESI technique has been coupled to different types of mass analyzers,[28–30] such as quadrupole,[31] ion trap,[32,33] Fourier transform mass spectrometer (FTMS),[34–38] hybrid quadrupole/time-of-flight (Qq-TOF),[39] and Orbitrap.[40] These instruments allow conducting tandem mass spectrometry (MS/MS) on selected ions, in order to obtain sequence information for biopolymers. The MS/MS spectra provide the sequence information through a fragmentation pattern produced by collision-induced dissociation (CID) of selected ions with a neutral collision gas, such as nitrogen or argon. ESI–MS, coupled with either liquid chromatography (LC) or capillary electrophoresis (CE), has been widely used for the analysis of proteins/peptides, oligonucleotides, oligosaccharides, and lipooligosaccharides. However, it is difficult to obtain ESI–MS on quadrupole instruments for a wide range of carbohydrate polymers (e.g., MW > 50 kDa) that give predominantly singly charged molecular ions.[31,41]

Partial hydrolysis of PSs is one of the common techniques for the structural elucidation of PSs. The generated oligosaccharides are isolated and structurally analyzed by a variety of chemical and physical methods. This strategy has been extensively used by our group and others.[31,42–46] To degrade PSs, two common approaches can be employed, chemical and enzymatic. In both cases, concentrations of different reagents and reaction conditions have to be optimized depending on the nature of biopolymers. For neutral sugars, the most commonly used reagents for hydrolysis include mineral acids, such as sulfuric acid (H_2SO_4), trifluoroacetic acid (TFA), and hydrochloric acid (HCl). In some instances, particularly when dealing with acid-labile sugars, solvolysis reagents are utilized, such as sulfuric acid in absolute methanol, methanolic HCl, and hydrofluoric acid (see Garna *et al.*[47] and the references therein). Undoubtedly, this strategy is feasible but is often time consuming and is not universally applicable in cases of acid-labile sugar residues or noncarbohydrate components, such as *O*-acetyl groups. In addition, it is often hard to find hydrolysis conditions suitable for generation of the intact repeating unit.

9.13.2 In-Source Fragmentation and Analysis of Polysaccharides by Capillary Electrophoresis–Mass Spectrometry

9.13.2.1 Method Development and Application to the Analysis of the O-Chain Polysaccharide of *Aeromonas salmonicida*

The O-PS of *A. salmonicida* was analyzed by chemical and NMR methods and found to be a branched polymer consisting of trisaccharide repeating units containing L-rhamnose (Rha), D-glucose (Glc), 2-acetamido-2-deoxy-mannose (ManNAc), and *O*-acetyl (OAc) group and having the following structure[43] (**Scheme 1**).



Scheme 1 Structure of the O-PS from *Aeromonas salmonicida* strain A449.

In order to validate our new method, we have used TFA as a chemical reagent for PS hydrolysis. In order to determine optimal conditions for generation of an intact repeating unit, the reaction temperature was kept constant at 100 °C, while effects of TFA concentrations (e.g., 0.5, 1.0, and 2.0 mol l^{-1}) and reaction times (e.g., 0.5, 1, and 2 h) were investigated for O-PS of *A. salmonicida* strain A449. CE–MS analysis results for nine TFA-hydrolyzed O-PS samples are presented in **Figure 1**. It was noted that the intensity of observed ions increased with the increase in TFA concentration or prolonged reaction time. It implied that optimal conditions for complete hydrolysis of this polymer were achieved using 2.0 mol l^{-1} TFA at 100 °C for 2 h. Although predominant ions were observed under all conditions examined (e.g., *m/z* 326.3, 488.3, 530.3, and 837.8), intensities of the most abundant ions (e.g., *m/z* 368.3) in the O-PS samples subjected to hydrolysis decreased with an increase in TFA concentration, and eventually disappeared following hydrolysis with 2.0 mol l^{-1} TFA at 100 °C for 1 h. Based on the structure of O-PS, one can assign ions at *m/z* 368.3 to a disaccharide containing ManNAc (203 Da) and Rha (146 Da). Fragment ions corresponding to one O-deacetylated repeating unit were observed at *m/z* 530.3. Consequently, consecutive additions of Rha and ManNAc residues gave rise to ions at

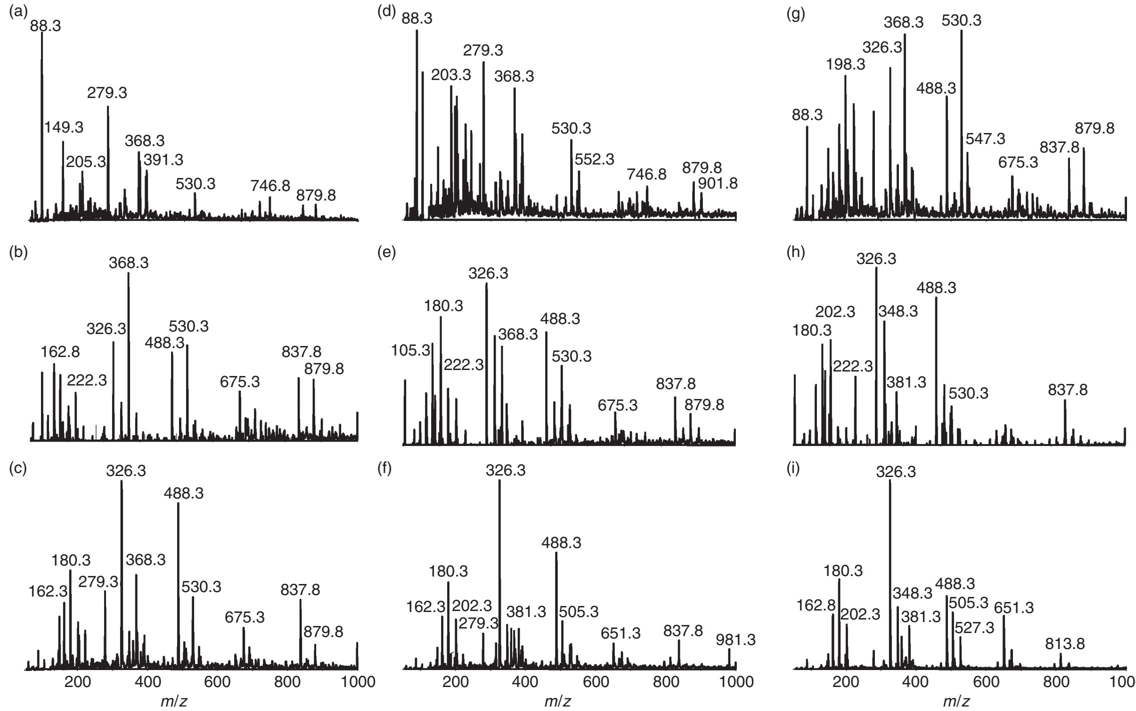


Figure 1 CE–MS analysis of the TFA-treated O-PS (1.0 μg μl^{-1}) from *Aeromonas salmonicida.* All reactions were performed at 100 °C. Other conditions are as follows: (a) 0.5 mol l^{-1} TFA, 0.5 h; (b) 0.5 mol l^{-1} TFA, 1 h; (c) 0.5 mol l^{-1} TFA, 2 h; (d) 1.0 mol l^{-1} TFA, 0.5 h; (e) 1.0 mol l^{-1} TFA, 1 h; (f) 1.0 mol l^{-1} TFA, 2 h; (g) 2.0 mol l^{-1} TFA, 0.5 h; (h) 2.0 mol l^{-1} TFA, 1 h; (i) 2.0 mol l^{-1} TFA, 2 h. Separation conditions are as follows: bare fused-silica (90 cm × 50 μm i.d., 185 μm o.d.), 15 mmol l^{-1} ammonium acetate, pH 7.0, +15 kV, 300 mbar. Orifice voltage: +30 V. Reproduced from J. Li; Z. Wang; E. Altman, *Rapid Commun. Mass Spectrom.* **2005**, *19*, 1305. © Crown in the right of Canada.

m/z 879.8. Although NMR data suggested the existence of OAc group on Rha, no corresponding ions were detected. It indicated that TFA might have caused the removal of the OAc group even under very mild experimental conditions (e.g., Figure 1(e), 0.5 mol l^{-1} TFA, 1 h). More importantly, when TFA concentration and/or reaction time were increased, the N-linked acetyl (NAc) residues were removed from ManNAc as shown in Figure 1. The relative ratios of ion intensities of *m/z* 368.3 versus *m/z* 326.2 (difference by 42 Da) were dramatically decreased and eventually disappeared as shown in Figures 1(f), 1(**h**), and 1(**i**), respectively. The ions observed at *m/z* 326.3, *m/z* 488.3, and *m/z* 837.8 could be attributed to N-deacetylation. These observations suggested that it was possible to generate intact repeating units of the PS backbone when hydrolysis was conducted with 1.0 mol l^{-1} TFA at 100 °C for 1 h. Based on this experimental evidence, all further experiments involving TFA hydrolysis were conducted under these conditions.

It is well known that ESI is able to generate multiply charged ions and the recorded spectrum can then be deconvoluted for the molecular mass of intact biopolymers. Furthermore, the multiply charged ions can be fragmented using conventional CID to obtain structural information. Initially, the intact O-PS samples were analyzed using conditions already established for the analysis of the TFA-treated O-PS samples, typically with an orifice voltage of 30 V. However, no multiply protonated (positive ion mode) or deprotonated (negative ion mode) ions were detected for polymers of interest. Figure 2(a) represents the

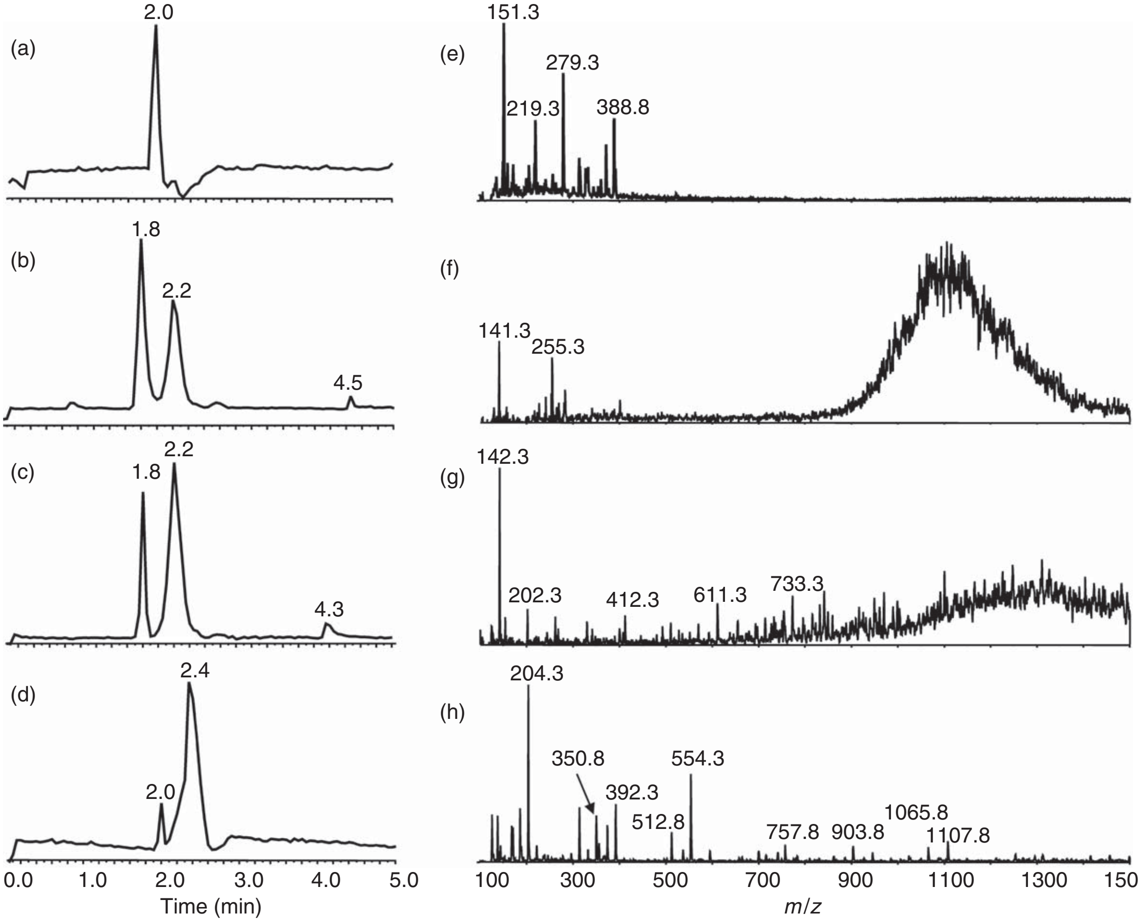


Figure 2 CE–MS analysis of the O-PS (1.0 μg μl^{-1}) from *Aeromonas salmonicida* A449. (a) TIE (positive ion mode, *m/z* 100–1500) and extracted spectrum at 2.4 min, orifice voltage: +30 V; (b) TIE (negative ion mode, *m/z* 100–1500) and extracted spectrum at 2.2 min, orifice voltage: −30 V; (c) TIE (negative ion mode, *m/z* 100–1500) and extracted spectrum at 2.2 min, orifice voltage: −200 V; (d) TIE (positive ion mode, *m/z* 100–1500) and extracted spectrum at 2.4 min, orifice voltage: +200 V. Reproduced from J. Li; Z. Wang; E. Altman, *Rapid Commun. Mass Spectrom.* **2005**, *19*, 1305. © Crown in the right of Canada.

CE–MS analysis of the intact O-PS in a positive ion mode with a normal orifice voltage of +30 V. The peak at 2.0 min corresponds to electroosmotic flow (EOF) and neutral products. The negative peaks in the total ion electropherogram (TIE) resulted from the biopolymers. However, the extracted mass spectrum revealed that the observed ions do not correspond to PSs or oligosaccharides. For comparison, two orifice voltages, for example, −30 and −200 V, were used for the negative ion mode. Similar results were obtained, as shown in **Figures 2(b)** and **2(c)**, respectively. Again, the peak at 1.8 min corresponds to EOF and neutral components in the sample and the peak at 2.2 min corresponds to the PS. The peaks at 4.5 min (**Figure 2(b)**) and 4.3 min (**Figure 2(c)**) are unknown compounds. The broad peaks in the extracted mass spectra revealed the heterogeneity of the sample (**Figure 2(b)**). Even with an orifice voltage of −200 V, no repeating units were detected (**Figure 2(c)**).

It was reported that a heavier curtain gas is not necessarily an effective means for increasing the degree of ion fragmentation in the IS-CID.[48] Therefore, nitrogen can be used as the curtain gas for all studies. The effect of orifice voltages, from 100 to 200 V, on the degree of ion fragmentation was investigated for the analysis of O-PS. It was found that the degree of ion fragmentation increased with the increase in orifice voltage. When orifice voltages were lower than 150 V, almost no repeating units were detected. Since the maximum orifice voltage of API 3000 in this lab is 200 V, all the IS-CID mass spectra were obtained using this voltage. The IS-CID mass spectrum (IS-CID-MS) of the 1.0 $\mu g\, \mu l^{-1}$ O-PS is presented in **Figure 2(d)**. As shown in the TIE, the total analysis time was less than 5 min, with two baseline resolved peaks. The extracted mass spectrum at 2.4 min (**Figure 2(d)**) suggested that the observed ions were generated from PSs; whereas the peak at 2.0 min corresponded to EOF and neutral products in the sample. It is noteworthy to mention that the O-PS was well separated from other background ions even when a high pressure (300 mbar) was applied. The obtained spectrum was consistent with the proposed structure. The ions at *m/z* 204.3 suggested the presence of ManNAc and the ions at *m/z* 350.3 indicated that ManNAc was attached to a Rha residue. Presence of fragment ions at *m/z* 392.3 confirmed that ManNAc was attached to a RhaOAc moiety. Taken together the fragmentation pattern of CE-IS-CID-MS was consistent with the reported structure of O-PS from *A. salmonicida* composed of a trisaccharide repeating unit containing Glc (162 Da), ManNAc (203 Da), and RhaOAc (188 Da). The CE-IS-CID-MS/MS of precursor ions at *m/z* 554.5 and *m/z* 757.8 was consistent with the presence of one repeating unit and one repeating unit plus one additional ManNAc residue, respectively. The fragmentation of ions at *m/z* 757.7 has led to the observation of ions at *m/z* 554.3 corresponding to the loss of ManNAc (data not shown). Consequently, ions at *m/z* 554.3 have lost Glc and RhaOAc to give rise to fragment ions at *m/z* 203.8. The observed ions at *m/z* 314.3 probably resulted from a disaccharide ManNAc-Rha losing two H_2O moieties. Depending on the reaction being carried out in liquid or gas phase, the generated repeating unit might differ by 18 Da or a water moiety.

These observations demonstrate that the IS-CID-MS is a powerful technique to obtain structural information on the composition of PS repeating units. To evaluate the sensitivity of this technique, serial dilutions of O-PS were analyzed. As expected, the signal-to-noise ratio degraded with the decrease in concentrations, 0.5, 0.1, 0.05 $\mu g\, \mu l^{-1}$. It was found that the concentration limit of detection (CLOD) for the analysis of O-PS of *A. salmonicida* strain A449 is about 0.05 $\mu g\, \mu l^{-1}$. More importantly, the obtained fragment pattern was conserved for all O-PS concentrations used, making it possible to obtain sequence information. By comparing the mass spectra between TFA-hydrolyzed and in-source fragmentation samples, we can assign fragment ions in **Figure 2(d)** to the corresponding ions in **Figure 1**. For example, the fragment ion at *m/z* 350.3 (**Figure 2(d)**) can be assigned to the ion at *m/z* 368.3 (**Figure 1**). Once again, the difference of 18 Da resulted from a different reaction mechanism (gas- or liquid-phase reaction). These results also suggested that most abundant ions were different by OAc (42 Da) for correlated ion pairs, since OAc could not survive under TFA treatment even under mild conditions. In order to confirm that the OAc group is not due to an artifact of IS-CID (gas-phase reaction), the O-deacetylation of O-PS samples was carried out using acetic acid. The de-OAc PSs then underwent chemical hydrolysis and in-source fragmentation. In addition, the ions detected in both spectra could be easily correlated with an addition or subtraction of water (18 Da). For example, the ions at *m/z* 350 in **Figure 2(d)** could be attributed to the same disaccharide as the ions at *m/z* 368.3 in **Figure 1**.

9.13.2.2 Carbohydrate Analysis and Serological Classification of Typical and Atypical Isolates of *Aeromonas salmonicida*: A Rationale for the Lipopolysaccharide-Based Classification of *A. salmonicida*

To detect structural variations in the O-PS structure of *A. salmonicida* from typical and atypical isolates we have used a microscale CE–MS-based method for analysis of *A. salmonicida* LPS directly on the bacterial cells. This method involves pretreatment of bacterial cells ($\sim 10^8$–10^{10} cells) with proteinase K, RNase and DNase followed by delipidation with mild acetic acid and subsequent MS analysis. Using this method, we have analyzed LPS structure of 39 typical and atypical isolates of *A. salmonicida* and related species (**Table 1**). All *A. salmonicida* strains examined could be divided into three structural patterns, type A, type B, and type C, according to the structure of their corresponding O-PSs (**Scheme 2**). Majority of typical *A. salmonicida* isolates belonged to type A displayed the

Table 1 Strains of *Aeromonas salmonicida* examined and their O-chain polysaccharide structural type

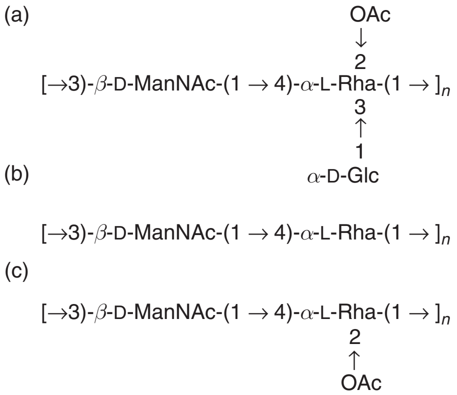
Strain	*Species*	*O-chain type*
A449	Brown trout, *Salmo trutta*, France, *A. salmonicida* subsp. *salmonicida*	Type A[a]
A450	Brown trout, *Salmo trutta*, France, *A. salmonicida* subsp. *salmonicida*	Type A
80204	Atlantic salmon, *Salmo salar*, Canada, *A. salmonicida* subsp. *salmonicida*	Type A
80204-1	*A. salmonicida* subsp. *salmonicida*, laboratory derived an Et-Bromide mutant	Type A
33658	Atlantic salmon, *Salmo salar*, ATCC, *A. salmonicida* subsp. *salmonicida*	Type A
51413	Brown trout, *Salmo trutta*, ATCC, *A. salmonicida* subsp. *salmonicida*	Type A
BC5085	Atlantic salmon, *Salmo salar*, Canada, *A. salmonicida* subsp. *salmonicida*	Type A
BC6129	Atlantic salmon, *Salmo salar*, Canada, *A. salmonicida* subsp. *salmonicida*	Type A
N2758	Atlantic salmon, *Salmo salar*, Norway, *A. salmonicida* subsp. *salmonicida*	Type A
N3395	Atlantic salmon, *Salmo salar*, Norway, *A. salmonicida* subsp. *salmonicida*	Type A
97-13	Coho salmon, *Oncorhynchus kisutch*, Canada, *A. salmonicida* subsp. *salmonicida*	Type A
4043	Char, *Salvelinus alpinus*, Finland, *A. salmonicida* subsp. *achromogenes*	Type A
33659	Brown trout, *Salmo trutta*, ATCC, *A. salmonicida* subsp. *achromogenes*	Type B[b]
4078	Atlantic cod, *Gadus morhua*, Iceland, *A. salmonicida* subsp. *achromogenes*	Type B
4101	Atlantic cod, *Gadus morhua*, Iceland, *A. salmonicida* subsp. *achromogenes*	Type B
4102	Atlantic cod, *Gadus morhua*, Canada, *A. salmonicida* subsp. *achromogenes*	Type B
4000	Atlantic salmon, *Salmo salar*, Iceland, *A. salmonicida* subsp. *achromogenes*	Type B
4035	Masou salmon, *Oncorhynchus masou*, Japan, *A. salmonicida* subsp. *Masoucida*	Type A
27013	Masou salmon, *Oncorhynchus masou*, *A. salmonicida* subsp. *Masoucida*	Type A
N2461	Turbot, *Scophthalmus maximus*, Norway, *A. salmonicida* atypical	Type A
4117	Turbot, *Scophthalmus maximus*, Norway, *A. salmonicida* atypical	Type A
4082	Atlantic salmon, *Salmo salar*, Norway, *A. salmonicida* atypical	Type A
4133	Atlantic salmon, *Salmo salar*, Chile, *A. salmonicida* atypical	Type A
N2517	Atlantic halibut, *Hippoglossus hippoglossus*, Norway, *A. salmonicida* atypical	Type B
4099	Atlantic cod, *Gadus morhua*, Norway, *A. salmonicida* atypical	Type B
4143	Atlantic cod, *Gadus morhua*, Norway, *A. salmonicida* atypical	Type B
4002	Char, *Salvelinus alpinu*s, Norway, *A. salmonicida* atypical	Type B
4123	Carp, *Cyprinus carpio*, Yugoslavia, *A. salmonicida* atypical	Type B
4050	Halibut, *Hippoglossus hippoglossus*, Norway, *A. salmonicida* atypical	Type B
4153	Halibut, *Hippoglossus hippoglossus*, Norway, *A. salmonicida* atypical	Type B
4128	Spotted wolffish, *Anarhichas minor*, Iceland, *A. salmonicida* atypical	Type B
4137	Spotted wolffish, *Anarhichas minor*, Norway, *A. salmonicida* atypical	Type B
4089	Turbot, *Scophthalmus maximus*, Norway, *A. salmonicida* atypical	Type B
N4705	Turbot, *Scophthalmus maximus*, Norway, *A. salmonicida* atypical	Type C[c]
4108	Halibut, *Hippoglossus hippoglossus*, Norway, *A. salmonicida* atypical	Type C
4059	Spotted wolffish, *Anarhichas minor*, Norway, *A. salmonicida* atypical	Type C
4067	Spotted wolffish, *Anarhichas minor*, Norway, *A. salmonicida* atypical	Type C
4129	Spotted wolffish, *Anarhichas minor*, Norway, *A. salmonicida* atypical	Type C
4141	Spotted wolffish, *Anarhichas minor*, Norway, *A. salmonicida* atypical	Type C

[a] Rha-ManNAc backbone with Glc and OAc substituents (**Scheme 2**(**a**)).

[b] Rha-ManNAc backbone (**Scheme 2**(**b**)).

[c] Rha-ManNAc backbone and OAc substituent (**Scheme 2**(**c**)).

Reproduced from Z. Wang; X. Liu; A. Dacanay; B. A. Harrison; M. Fast; D. J. Colquhoun; V. Lund; L. L. Brown; J. Li; E. Altman, *Fish Shellfish Immunol*. **2007**, *23*, 1095. © Elsevier Ltd.


(a)
OAc
↓
2
[→3)-β-D-ManNAc-(1 → 4)-α-L-Rha-(1 →]n
3
↑
1
α-D-Glc
(b)
[→3)-β-D-ManNAc-(1 → 4)-α-L-Rha-(1 →]n
(c)
[→3)-β-D-ManNAc-(1 → 4)-α-L-Rha-(1 →]n
2
↑
OAc

Scheme 2 Structural variants of the O-chain polysaccharide of *Aeromonas salmonicida* LPS: (a) type A, representative strain A449; (b) type B, representative strain 33659; (c) type C, representative strain N4705. Reproduced from Z. Wang; X. Liu; A. Dacanay; B. A. Harrison; M. Fast; D. J. Colquhoun; V. Lund; L. L. Brown; J. Li; E. Altman, *Fish Shellfish Immunol*. **2007**, *23*, 1095. © Elsevier Ltd.

complete O-PS structure previously determined for *A. salmonicida* subsp. *salmonicida* strains A449, 80204, and 80204-1 (**Figure 3**(**a**)) and composed of the trisaccharide repeating unit containing L-rhamnose (L-Rha), *N*-acetyl-D-mannosamine (D-ManNAc) and D-glucose (D-Glc), and noncarbohydrate O-acetyl substituent (OAc). Identical fragmentation patterns in the CE–MS spectrum were obtained for 18 *A. salmonicida* strains examined belonging to this structural group. Characteristic fragment ions corresponding to one O-acetylated repeating unit were observed at *m*/*z* 554.5. Fragment ions at *m*/*z* 392.4 were consistent with the loss of Glc residues. Other characteristic fragment ions included ions at *m*/*z* 757.5 and *m*/*z* 945.7, which were consistent with consecutive additions of ManNAc and RhaOAc residues, respectively. The second group, type B, was represented by the O-PS of *A. salmonicida* strain 33659 and consisted of a backbone structure containing L-Rha and D-ManNAc only (**Scheme 2(b)**, **Figure 3**). Identical fragmentation patterns in CE–MS spectra were obtained for 15 *A. salmonicida* strains examined belonging to type B. Characteristic fragment ions corresponding to one repeating unit were observed at *m*/*z* 350.4. Additional fragment ions corresponding to consecutive additions of ManNAc (*m*/*z* 553.6) and Rha residues (*m*/*z* 699.7) were also observed. In order to verify the composition and substitution pattern of sugars the purified O-PS of *A. salmonicida* subsp. *achromogenes* strain 33659 was subjected to compositional and linkage analysis and the results compared with structural data obtained for the O-PS of *A. salmonicida* subsp. *salmonicida* strain A449.[42] Finally, six atypical *A. salmonicida* strains belonging to the third structural group, type C, were represented by the O-PS of atypical strain N4705 consisting of the backbone PS structure containing L-Rha and D-ManNAc and *O*-acetyl group. Three chemical repeating units were illustrated for each PS and fragment ions were assigned according to the Domon and Costello nomenclature.[49] Since these tandem mass spectra were obtained with in-source CID, they represented a mixture of glycoforms resulted from the cleavage of acid-labile Rha residues. The compositional and linkage analysis was performed on the purified O-PS of *A. salmonicida* strain N4705,[42] confirming its structure (**Figure 3(c)**). The CE–MS fragmentation pattern of *A. salmonicida* strains belonging to type C contained two characteristic fragment ions at *m*/*z* 392.4, corresponding to the O-acetylated repeating unit, and at *m*/*z* 350.4, corresponding to the repeating unit without the *O*-acetyl group. Other characteristic fragment ions included ions at *m*/*z* 553.5 and at *m*/*z* 699.7, which were consistent with additions of ManNAc and Rha residues, respectively. Based on relative intensities of fragment ions at *m*/*z* 392.4 (O-acetylated molecular species) and at *m*/*z* 350.4 (non-O-acetylated molecular species), the degree of O-acetylation of Rha residues was approximately 30%.

9.13.2.3 Structural Studies of the CPS and O-PS of *A. salmonicida*: Comparison between *In Vitro* and *In Vivo* Growth Conditions

Formation of CPS covering the A-layer has been reported to be produced during the *in vivo* culture of *A. salmonicida* in surgically implanted intraperitoneal culture chambers.[50] Moreover, Merino *et al.*[51] have reported that when grown under conditions promoting capsule formation, strains of *A. salmonicida* exhibited

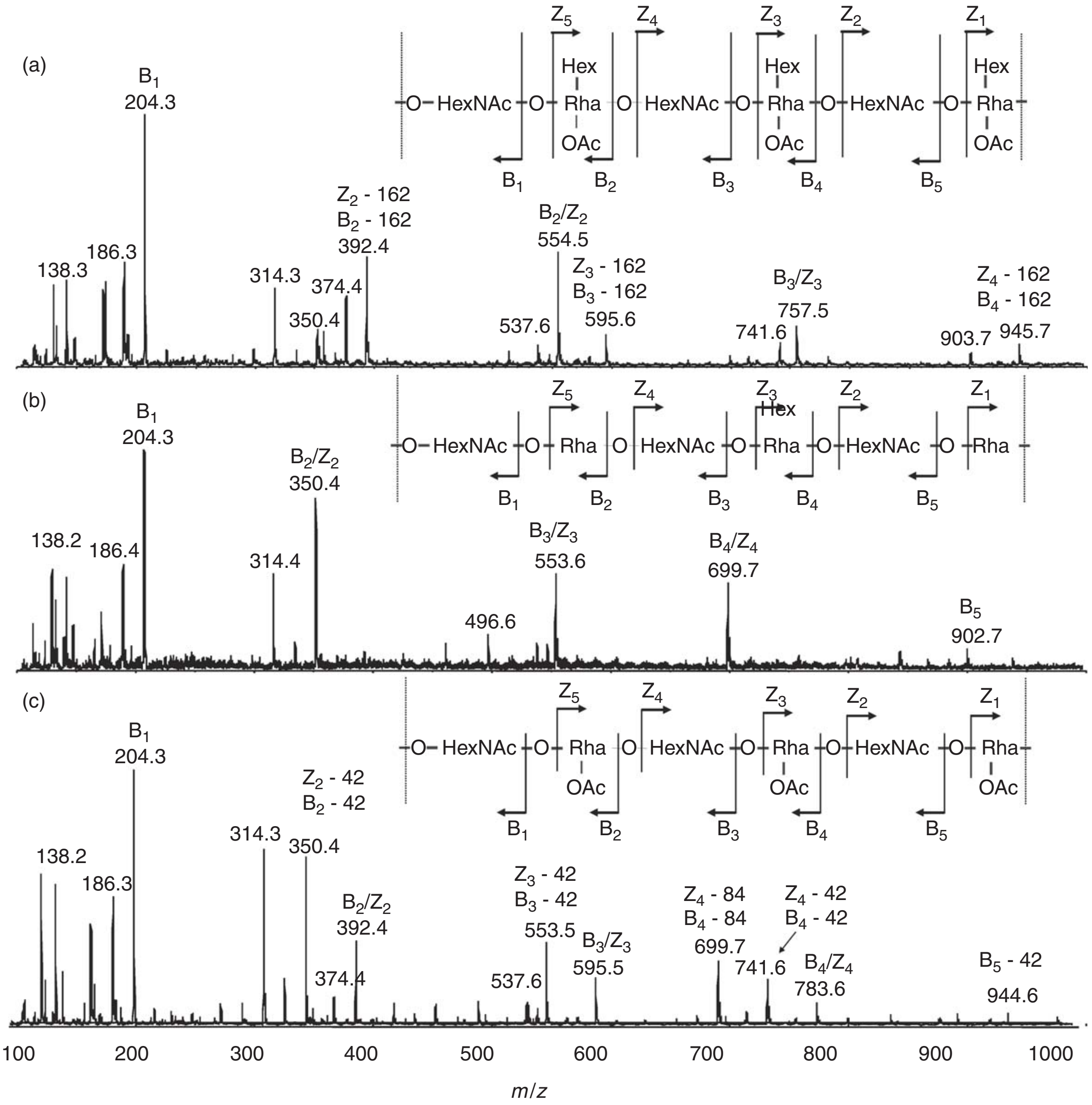
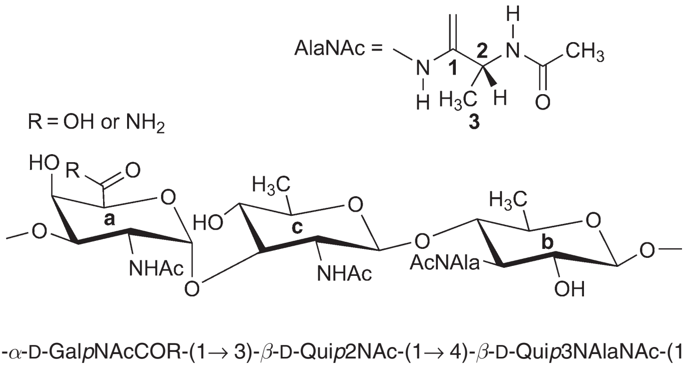


Figure 3 CE–MS analysis of the whole cell lysates of *Aeromonas salmonicida* in the positive ion mode: (a) type A, representative strain A449; (b) type B, representative strain 33659; (c) type C, representative strain N4705. Separation conditions are as follows: bare fused-silica (90 cm × 50 μm i.d., 185 μm o.d.), 15 mmol l^{-1} ammonium acetate, pH 7.0, +15 kV, 300 mbar. Orifice voltage: +200 V. Adopted from Z. Wang; X. Liu; A. Dacanay; B. A. Harrison; M. Fast; D. J. Colquhoun; V. Lund; L. L. Brown; J. Li; E. Altman, *Fish Shellfish Immunol*. **2007**, *23*, 1095. © Elsevier Ltd. Three chemical repeating units, based on the acid-labile feature of Rha residue, were illustrated for each polysaccharide and fragment ions were assigned according to the Domon and Costello nomenclature.[49]

significantly higher ability to invade fish cell lines. It suggests that, as with the A-layer and LPS, CPS is an important virulence factor, essential for host cell invasion and bacterial survival. Previous studies have determined the structure of the O-PS of *A. salmonicida* strain SJ-15.[50] Partial structure of the core oligosaccharide from the same strain of *A. salmonicida* was also determined.[52] In both instances, *A. salmonicida* strain SJ-15 was cultured in tryptic soy broth (TSB) at 25 °C. In addition, other reports describe capsular material isolated from cells grown on yeast extract–peptone–glucose–mineral salts.[53] The relevance of these structures to *in vivo*-cultured bacteria and their role in pathogenesis has not been established. We have isolated and characterized the cell-surface carbohydrate antigens of *A. salmonicida* strain 80204-1 produced under *in vitro* growth conditions on tryptic soy agar (TSA) and *in vivo* have demonstrated that their structures are chemically and antigenically distinct from the previously described O-PS[50] and capsule.[53] Cells of the A-layer-avirulent strain of *A. salmonicida*, 80204-1, were grown on TSA, harvested, washed with 2.5% saline, and subjected to the phenol/water extraction followed by purification of aqueous- and phenol-phase soluble LPS by ultracentrifugation. Crude CPS was recovered from the initial 2.5% saline wash of bacterial cells and from 1% saline wash of the aqueous- and phenol-phase soluble LPS and purified by gel permeation chromatography on a Sephadex G-100 column. CPS eluted as a broad peak in a void volume of a Sephadex G-100 column. Fractions were collected and analyzed colorimetrically for aldose. Three fractions, designated fraction #1–fraction #3 (not shown), were pooled and analyzed by NMR. Fraction #1 was contaminated by an α1,6-glucan, while fraction #2 and fraction #3 were homogeneous yielding a glucan-free CPS that was used for further analysis. Methylation analysis performed on *in vivo* cultured *A. salmonicida* strain 80204-1 cells confirmed these findings and showed the presence of 2,6-dideoxy-4-*O*-methyl-2-(*N*-methylacetamido)-glucose and 3,6-dideoxy-2-*O*-methyl-3-(*N*-methylacetamido)-glucose, in accordance with the proposed structure of CPS (**Scheme 3**). The linkage of GalNAcA could not be verified as the methylation analysis was carried out without the carboxyl group reduction step.

To confirm the sequence of this newly formed PS detected in the *in vivo* cultured cells, CE–MS analysis was carried out on both the concentrated bacterial cell saline wash containing CPS and on the bacterial pellet following mild acid hydrolysis to release delipidated LPS. Initially, the CPS and LPS samples were subjected to CE–MS analysis with the typically used orifice of 45 V. The mass spectra indicated no individually separated ion peaks and showed only a broad peak (data not shown) corresponding to a range of molecular masses, which could be attributed to CPS or LPS O-PS repeating unit fragments of different length. As validated in Section 9.13.2.1, we have successfully applied this approach to the partial degradation of the PS into shorter oligosaccharide units due to front-end CID (**Figure 4**). The presence of *m*/*z* 662.3 could not be unambiguously confirmed due to a high background, which also resulted in the loss of other characteristic fragment ions at *m*/*z* 1067.4 and 325.5. However, CE–MS/MS analysis of the precursor ion at *m*/*z* 662.3 was consistent with its previously determined composition of a trisaccharide repeating unit consisting of Qui3NAlaNAc, GalNAcAN, and QuiNAc and confirming its presence in the *in vivo* cultured cells. Similarly, based on the CE–MS/MS analysis of the precursor ion at *m*/*z* 662.3, the same trisaccharide repeating unit could also be detected in 2.5% saline wash of the bacterial cells grown *in vivo*. It is noteworthy that the fragment ion at *m*/*z* 663.3 corresponding to a trisaccharide-repeating unit containing GalNAcA was not detected in the *in vivo* cultured cells, suggesting a complete amidation of GalNAcA. We have confirmed by direct CE–ES–MS analysis that both CPS and O-PS were also present in the *in vivo*-grown cells of *A. salmonicida* strain 80204-1 harvested at 72 h postimplant surgery. These PSs were not detected in the *in vitro*-grown bacterial inoculum TSB culture used for the implants.



[→3)-α-D-Gal*p*NAcCOR-(1→ 3)-β-D-Qui*p*2NAc-(1→ 4)-β-D-Qui*p*3NAlaNAc-(1 →]$_n$

Scheme 3 The proposed structure of the CPS and O-chain polysaccharide of *Aeromonas salmonicida* strain 80204-1. Reproduced from Z. Wang; S. Larocque; E. Vinogradov; J.-R. Brisson; A. Dacanay; M. Greenwell; L. L. Brown; J. Li; E. Altman, *Eur. J. Biochem.* **2004**, *271*, 4507. © Blackwell Publishing Ltd.

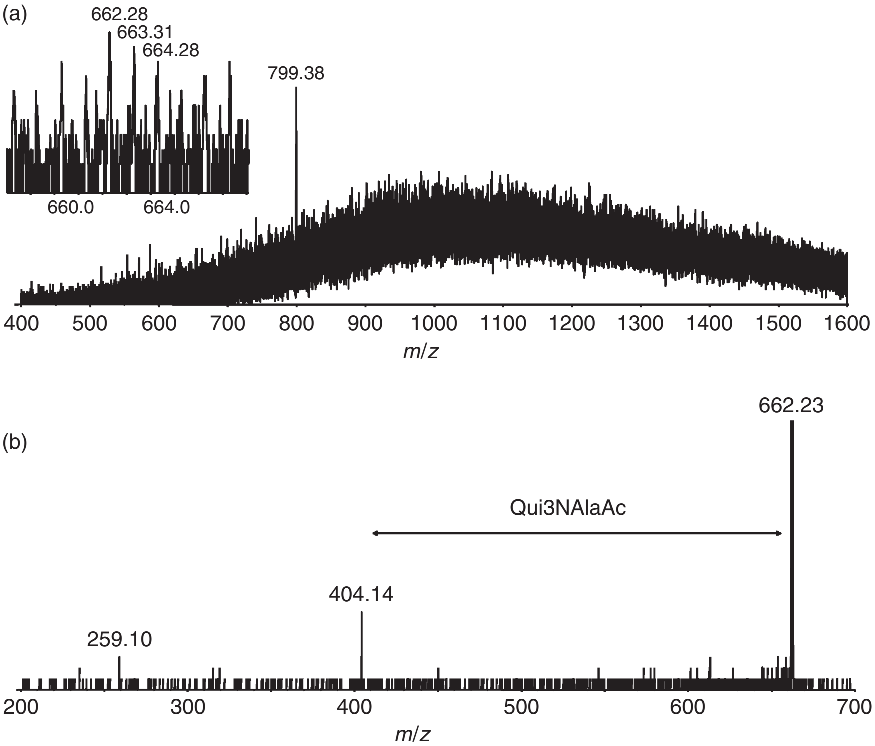


Figure 4 CE–MS and CE–MS/MS analysis (+ion mode) of LPS from *in vivo Aeromonas salmonicida* strain 80204-1 cells. Extracted mass spectrum and extracted MS/MS spectrum of precursor ion *m/z* 662.3 promoted using an orifice voltage of 120 V. Reproduced from Z. Wang; S. Larocque; E. Vinogradov; J.-R. Brisson; A. Dacanay; M. Greenwell; L. L. Brown; J. Li; E. Altman, *Eur. J. Biochem.* **2004**, *271*, 4507. © Blackwell Publishing Ltd.

9.13.3 Materials and Methods

9.13.3.1 Bacterial Culture and Isolation of Lipopolysaccharide

Aeromonas salmonicida strains were obtained from the Institute for Marine Biosciences, National Research Council of Canada, Halifax, NS, Canada. The bacteria were cultured on TSA or TSB without glucose at 18 or 25 °C for 48–72 h. The cells were killed with 1% (w/v) phenol solution (22 °C, 4 h), washed with 0.01 mol l^{-1} phosphate-buffered saline pH 7.4, and harvested by low-speed centrifugation (3000 *g*, 4 °C, 25 min). The cells were washed with 2.5% saline solution (w/v), digested enzymatically and LPS extracted by the hot phenol–water extraction method. Phenol and water layers were collected separately, dialyzed against tap water, and lyophilized. The lyophilizates were then dissolved in 1% saline solution (w/v), subjected to ultracentrifugation (105 000 *g*, 4 °C, 16 h), and the LPS pellets were redissolved in water and lyophilized. The 1% saline supernatant containing CPS was dialyzed against dH_2O water until salt free, lyophilized and used for the isolation of CPS according to the procedure described below.

9.13.3.2 Preparation of the O-Chain Polysaccharide and Its O-Deacetylated and Periodate-Oxidized Derivatives

9.13.3.2.1 Mild acid hydrolysis of lipopolysaccharide

LPS (60 mg) was hydrolyzed with 0.2 mol l^{-1} AcOH (100 °C, 2 h). The reaction mixture was cooled down on ice and the insoluble lipid A was removed by centrifugation. The water-soluble part was lyophilized and purified by gel chromatography on a Bio-Gel P-2 column (Bio-Rad). The fraction containing crude O-PS was further purified on a Bio-Gel P-10 column (Bio-Rad).

9.13.3.2.2 O-Deacetylation of the O-chain polysaccharide

O-PS (10 mg) was dissolved in 5% ammonium hydroxide (2 ml) and incubated at 37 °C overnight. The solution was lyophilized, and the product purified by gel permeation chromatography using a Bio-Gel P-10 column, resulting in a yield of 7 mg.

9.13.3.2.3 Periodate oxidation and Smith degradation

The O-PS (10 mg) was oxidized with 0.05 mol l^{-1} sodium metaperiodate (1.6 ml) at 4 °C in the dark for 6 days. Then ethylene glycol (1 ml) was added to destroy the excess of the metaperiodate and stop the reaction. The mixture was stirred at 22 °C for 30 min, followed by reduction with sodium borohydride at 22 °C for 18 h and neutralization with 10% (v/v) AcOH. The solution was dialyzed and lyophilized. Smith-type hydrolysis of the periodate-oxidized and reduced polymer was affected with 0.5 mol l^{-1} TFA at 22 °C for 48 h, the sample was purified on a Bio-Gel P-2 column, and the sugar composition of the product was determined.

9.13.3.3 Isolation and Purification of Capsular Polysaccharide

Bacterial cells were washed with 2.5% (w/v) sodium chloride (saline) solution and following low-speed centrifugation, the saline supernatant was dialyzed against dH_2O water until salt free.[54] The dialyzate was lyophilized, redissolved in dH_2O water, and digested with trypsin, ribonuclease, and deoxyribonuclease, and, following dialysis, the crude PS was dissolved in a minimal volume of dH_2O and precipitated with five volumes of 95% ethanol. It was further purified by precipitation with 1% (w/v) cetyltrimethylammonium bromide. After keeping at 4 °C overnight, the precipitated complex was collected by low-speed centrifugation and redissolved in a minimal volume of 10% (w/v). The CPS was recovered by precipitation with five volumes of 95% ethanol, dialyzed extensively against dH_2O water until salt free, and lyophilized. Pure CPS was obtained by gel filtration on a column of Sephadex G-50 (Pharmacia).

9.13.3.4 Carboxyl Group Reduction

Carboxyl group reduction of the CPS and LPS samples was performed as previously described.[55] Briefly, LPS (10 mg) was dissolved in distilled water (10 ml) and following the addition of 1-cyclohexyl-3-(2-morpholinoethyl) carbodiimide metho-*p*-toluenesulfonate (113 mg), the stirred mixture was maintained at pH 4.7 by titration with 0.1 mol l^{-1} HCl for 3 h. Following completion of the reaction a 2 mol l^{-1} solution of sodium borohydride (12.5 ml) was added slowly and the reaction mixture was maintained at pH 7 by titration with 4 mol l^{-1} HCl. The reaction was allowed to proceed for 2 h at 22 °C, and the solution was dialyzed and lyophilized. The product was purified by gel permeation chromatography on Sephadex G-100 and lyophilized (yield 6 mg).

9.13.3.5 Acid Hydrolysis with Trifluoroacetic Acid

The sample of O-PS of *A. salmonicida* strain A449 (100 μg) was subjected to acid hydrolysis using three concentrations of TFA, 0.5, 1.0, and 2.0 mol l^{-1}. Each solution was subjected to a time-course hydrolysis at 100 °C for 0.5, 1, or 2 h. The reaction mixture was evaporated to dryness under the stream of nitrogen at 22 °C. Hydrolyzed PS samples were concentrated, redissolved in dH_2O and analyzed by CE–MS without any further purification.

9.13.3.6 Compositional Analysis

LPS samples (0.5 mg) were hydrolyzed with 2 mol l^{-1} TFA at 100 °C for 18 h followed by reduction in dH_2O with $NaBH_4$ and subsequent acetylation with acetic acid/pyridine (1:1, v/v) at 100 °C for 1 h. The resultant alditol acetates were extracted thrice with dichloromethane, the combined organic layer concentrated to dryness and analyzed by GLC using a Hewlett–Packard chromatograph equipped with a 30 m DB-17 capillary column (190 °C (32 min), 16 °C min^{-1} to 270 °C (32 min)) and by GLC–MS in the electron impact mode (EI) recorded using a Varian Saturn 2000 mass spectrometer. For compositional analysis of acid-labile deoxy aminosugars, PS or LPS samples were converted into their methyl ester derivatives by sealed-tube hydrolysis with 3% (w/v) methanolic hydrogen chloride at 100 °C for 16 h and then neutralized with silver carbonate (Aldrich). The resultant methyl esters were directly acetylated with acetic acid/pyridine (1:1, v/v) mixture at 100 °C for 1 h. The resultant acetates were extracted thrice with dichloromethane, the combined organic layer concentrated to dryness and subjected to GLC–MS analysis.

9.13.3.7 Methylation Analysis

The O-PS and CPS samples were methylated according to the method of Ciucanu and Kerek.[56] Briefly, the PS (3–5 mg) was dissolved in 0.5 ml of dry DMSO and stirred overnight at 22 °C. Twenty-five milligrams of freshly finely powdered sodium hydroxide was added and the solution was stirred for 2 h at 22 °C. Methyl iodide (1 ml) was added and the mixture was stirred for 4 h at 22 °C. The reaction was stopped by addition of 0.5 ml dH_2O. The product was extracted with chloroform (3 × 0.5 ml), the organic layers combined, dried by addition of a few crystals of anhydrous sodium sulfate and concentrated to dryness. The permethylated PS was subjected to hydrolysis with 4 mol l^{-1} TFA for 1 h at 125 °C (or for 4 h at 125 °C for heptose derivatives) for aminosugar-containing PSs and analyzed according to the previously reported conditions for partially methylated alditol acetates.[57]

9.13.3.8 Capillary Electrophoresis–Mass Spectrometry

All experiments were performed as described previously in detail.[41] Briefly, a CE instrument was coupled to a mass spectrometer, API 3000 or Q-Star (Applied Biosystems/Sciex, Concord, Canada) through a microlon spray interface. Sheath solution (isopropanol–MeOH, 2:1) was delivered at a flow rate of 1 μl min^{-1}. An electrospray stainless-steel needle (27-gauge) was butted against the low dead volume tee and enabled the delivery of the sheath solution to the end of the capillary column. The separations were obtained on approximately 90 cm length bare fused-silica capillary using 10 mmol l^{-1} ammonium acetate in deionized water, pH 9.0, containing 5% MeOH. A voltage of 25 kV was typically applied at the injection. The outlet of the capillary was tapered to approximately 15 μm i.d. using a laser puller (Sutter Instruments, Novato, CA, USA). Mass spectra were acquired with an orifice voltage of +200 V for API 3000 or +120 V for Q-Star.

9.13.3.9 Direct Capillary Electrophoresis–Mass Spectrometry Analysis of *Aeromonas salmonicida* Cells

Bacterial cells, 2.5×10^{11} cfu, were washed with 2.5% (w/v) saline, and the pellet recovered by low-speed centrifugation (3000 *g*, 4 °C, 10 min) and lyophilized. The lyophilized pellet was treated with RNase and DNase to release LPS (final concentration 10 μg ml^{-1} in 0.02 mol l^{-1} ammonium acetate, pH 7.5, 37 °C, 2 h) and lyophilized following low-speed centrifugation (yield, 27 mg dry weight). It was treated with proteinase K as described above and the product was recovered by low-speed centrifugation. Lyophilized sample was treated with 1% AcOH (100 °C, 1 h), desalted using a centrifugal filter device (Pall Corporation, Novato, CA, USA) and analyzed directly by a Crystal Model 310 CE instrument coupled to an API 3000 mass spectrometer. In addition, lyophilized cells were subjected to sugar composition and methylation analyses as described above for purified PS samples.

9.13.4 Conclusions

We have shown that MS provides a rapid and reliable method for screening bacterial PSs. Fragmentation of PSs to repeating units having lower molecular mass makes it possible to obtain their sequence information. In-source fragmentation was applied to promote the formation of structurally relevant repeating units of heterogeneous CPS and O-PS that would remain undetected using conventional ESI conditions. This approach was proven particularly useful for probing the subtle structural differences in monosaccharide composition and functionalities arising across bacterial serotypes; however, the complete structure analysis is impossible without NMR analysis.

Abbreviations

A-layer	*Aeromonas* surface layer
CE	capillary electrophoresis
CID	collision-induced dissociation
CPS	capsular polysaccharide
dHexOAc	O-acetylated deoxy hexose
EOF	electroosmotic flow
GalNAcA	2-acetamido-2-deoxy-D-galacturonic acid
GalNAcAN	2-acetamido-2-deoxy-D-galacturonamide
GLC	gas–liquid chromatography
Glc	glucose
HexNAc	2-acetamido-2-deoxy-hexose
IS-CID	in-source collision-induced dissociation
LPS	lipopolysaccharide
ManNAc	2-acetamido-2-deoxy-mannose
MS	mass spectrometry
NMR	nuclear magnetic resonance
OAc	*O*-acetyl
O-PS	O-chain polysaccharide
Qq-TOF	hybrid quadrupole/time-of-flight
Rha	rhamnose
TFA	trifluoroacetic acid
TIE	total ion electropherogram
TSA	tryptic soy agar
TSB	tryptic soy broth

References

1. R. Raetz; C. Whitfield, *Annu. Rev. Biochem.* **2002**, *71*, 635.
2. M. Caroff; D. Karibian, *Carbohydr. Res.* **2003**, *338*, 2431.
3. A. Weintraub, *Carbohydr. Res.* **2003**, *338*, 2539.
4. H. J. Jennings; A. Gamian; F. E. Ashton, *J. Exp. Med.* **1987**, *165*, 1207.
5. H. J. Jennings, *Curr. Top. Microbiol. Immunol.* **1990**, *150*, 97.
6. A. D. Cox; W. Zou; M. A. Gidney; S. Lacelle; J. S. Plested; K. Makepeace; J. C. Wright; P. A. Coull; E. R. Moxon; J. C. Richards, *Vaccine* **2005**, *23*, 5045.
7. J. H. Passwell; E. Harlev; S. Ashkenazi; C. Chu; D. Miron; R. Ramon; N. Farzan; J. Shiloach; D. A. Bryla; F. Majadly; R. Roberson; J. B. Robbins; R. Schneerson, *Infect. Immun.* **2001**, *69*, 1351.
8. D. Uhrin; J.-R. Brisson, Structure Determination of Microbial Polysaccharides by High Resolution NMR Spectroscopy. In *NMR in Microbiology:Theory and Applications*; J. N. Barotin, J. C. Portais, Eds.; Horizon Scientific Press: Wymonden, 2000; 165–190.
9. E. Y. Lee; D. S. Choi; K. P. Kim; Y. S. Gho, *Mass Spectrom. Rev.* **2008**, *27*, 535–555.
10. A. I. Nesvizhskii; O. Vitek; R. Aebersold, *Nat. Methods* **2007**, *4*, 787.
11. N. G. Ahn; J. B. Shabb; W. M. Old; K. A. Resing, *ACS Chem. Biol.* **2007**, *2*, 39.
12. W. J. Qian; J. M. Jacobs; T. Liu; D. G. Camp; R. D. Smith, *Mol. Cell Proteomics* **2006**, *5*, 1727.
13. L. N. Mueller; M. Y. Brusniak; D. R. Mani; R. Aebersold, *J. Proteome Res.* **2008**, *7*, 51.
14. P. G. Hitchen; A. Dell, *Microbiology* **2006**, *152*, 1575.
15. D. Nedelkov; U. A. Kiernan; E. E. Niederkofler; K. A. Tubbs; R. W. Nelson, *Mol. Cell Proteomics* **2006**, *5*, 1811.
16. M. V. Novotny; H. A. Soini; Y. Mechref, *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* **2008**, *866*, 26.
17. L. D. Roberts; G. McCombie; C. M. Titman; J. L. Griffin, *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* **2008**, *871*, 174.
18. R. Raman; S. Raguram; G. Venkataraman; J. C. Paulson; R. Sasisekharan, *Nat. Methods* **2005**, *2*, 817.
19. Z. Shriver; S. Raguram; R. Sasisekharan, *Nat. Rev. Drug Discov.* **2004**, *3*, 863.
20. J. Zaia, *Mass Spectrom. Rev.* **2004**, *23*, 161.
21. A. D. Watson, *J. Lipid Res.* **2006**, *47*, 2101.
22. X. Han; R. W. Gross, *Mass Spectrom. Rev.* **2005**, *24*, 367.
23. X. Han; R. W. Gross, *J. Lipid Res.* **2003**, *44*, 1071.
24. D. G. Robertson; M. D. Reily; J. D. Baker, *J. Proteome Res.* **2007**, *6*, 526.

25. X. Lu; X. Zhao; C. Bai; C. Zhao; G. Lu; G. Xu, *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* **2008**, *866*, 64.
26. D. Harvey, *Mass Spectrom. Rev.* **2006**, *25*, 595.
27. D. Harvey, *Mass Spectrom. Rev.* **2008**, *27*, 125.
28. G. L. Glish; D. J. Burinsky, *J. Am. Soc. Mass Spectrom.* **2008**, *19*, 161.
29. C. Campa; A. Coslovi; A. Flamigni; M. Rossi, *Electrophoresis* **2006**, *27*, 2027.
30. W. J. Griffiths; A. P. Jonsson; S. Liu; D. K. Rai; Y. Wang, *Biochem. J.* **2001**, *355*, 545.
31. J. Li; Z. Wang; E. Altman, *Rapid Commun. Mass Spectrom.* **2005**, *19*, 1305.
32. F. F. Hsu; J. Turk; R. M. Owens; E. R. Rhoades; D. G. Russell, *J. Am. Soc. Mass Spectrom.* **2007**, *18*, 479.
33. H. L. Cheng; G. R. Her, *J. Am. Soc. Mass Spectrom.* **2002**, *13*, 1322.
34. P. B. O'Connor; E. Mirgorodskaya; C. E. Costello, *J. Am. Soc. Mass Spectrom.* **2002**, *13*, 402.
35. P. B. O'Connor; J. L. Pittman; B. A. Thomson; B. A; Budnik; J. C. Cournoyer; J. Jebanathirajah; C. Lin; S. Moyer; C. Zhao, *Rapid Commun. Mass Spectrom.* **2006**, *20*, 259.
36. H. J. An; T. R. Peavy; J. L. Hedrick; C. B. Lebrilla, *Anal. Chem.* **2003**, *75*, 5628.
37. Y. Xie; J. Liu; J. Zhang; J. L. Hedrick; C. B. Lebrilla, *Anal. Chem.* **2004**, *76*, 5186.
38. J. Zhang; L. LaMotte; E. D. Dodds; C. B. Lebrilla, *Anal. Chem.* **2005**, *77*, 4429.
39. Z. Wang; S. Larocque; E. Vinogradov; J.-R. Brisson; A. Dacanay; M. Greenwell; L. L. Brown; J. Li; E. Altman, *Eur. J. Biochem.* **2004**, *271*, 4507.
40. M. Hardman; A. A. Makarov, *Anal. Chem.* **2003**, *75*, 1699.
41. J. Li; J. C. Richards, *Mass Spectrom. Rev.* **2007**, *26*, 35.
42. Z. Wang; X. Liu; A. Dacanay; B. A. Harrison; M. Fast; D. J. Colquhoun; V. Lund; L. L. Brown; J. Li; E. Altman, *Fish Shellfish Immunol.* **2007**, *23*, 1095.
43. Z. Wang; E. Vinogradov; S. Larocque; B. A. Harrison; J. Li; E. Altman, *Carbohydr. Res.* **2005**, *340*, 693.
44. D. J. McNally; M. P. Lamoureux; A. V. Karlyshev; L. M. Fiori; J. Li; G. Thacker; R. A. Coleman; N. H. Khieu; B. W. Wren; J.-R. Brisson; H. C. Jarrell; C. M. Szymanski, *J. Biol. Chem.* **2007**, *282*, 28566.
45. D. J. McNally; H. C. Jarrell; N. H. Khieu; J. Li; E. Vinogradov; D. M. Whitfield; C. M. Szymanski; J.-R. Brisson, *FEBS J.* **2006**, *273*, 3975.
46. D. J. McNally; H. C. Jarrell; J. Li; N. H. Khieu; E. Vinogradov; C. M. Szymanski; J.-R. Brisson, *FEBS J.* **2005**, *272*, 4407.
47. H. Garna; N. Mabon; B. Wathelet; M. Paquot, *J. Agric. Food Chem.* **2004**, *52*, 4652.
48. B. B. Schneider; D. D. Chen, *Anal. Chem.* **2000**, *72*, 791.
49. B. Domon; C. E. Costello, *Glycoconj. J.* **1988**, *5*, 397.
50. D. H. Shaw; Y. Z. Lee; M. J. Squires; O. Luderitz, *Eur. J. Biochem.* **1983**, *131*, 633.
51. S. Merino; S. Alberti; J. M. Tomas, *Infect. Immun.* **1994**, *62*, 5483.
52. D. H. Shaw; M. J. Hart; O. Luderitz, *Carbohydr. Res.* **1992**, *231*, 83.
53. A. Garrote; R. Bonet; S. Merino; M. D. Simon-Pujol; F. Congregado, *FEMS Microbiol. Lett.* **1992**, *74*, 127.
54. E. Altman; J.-R. Brisson; M. B. Perry, *Biochem. Cell Biol.* **1986**, *64*, 707.
55. R. L. Taylor; H. E. Conrad, *Biochemistry* **1972**, *11*, 1383.
56. I. Ciucanu; F. Kerek, *Carbohydr. Res.* **1984**, *131*, 209.
57. K. Leontein; B. Lindberg; J. Lonngren, *Carbohydr. Res.* **1978**, *62*, 359.

Biographical Sketches

Jianjun Li is a senior research officer and the Head of the Glycoanalysis-MS Facility in the Glycobiology program at the Institute for Biological Sciences (NRC-IBS). He received a B.Sc. in chemistry and a Ph.D. in analytical chemistry from Wuhan University, China. Dr. Li's research involves the development of high-resolution separation techniques coupled with MS and their applications to the structural characterization of various biological molecules. From 1997 to 2001, he focused on the development of chip-based capillary electrophoresis–mass spectrometry (Chip-CE–MS) systems. He has successfully integrated different functional materials, such as C_{18}, MAbs, and IMAC beads onto a chip channel to perform preconcentration chip-CE–MS with applications to functional proteomics. Since 2001, Dr. Li's research projects extended to the glycomics, an emerging field of glycobiology that combines the disciplines of both carbohydrate biochemistry and molecular biology to study and understand the structure, biosynthesis, and biological function of glycans. He has authored or coauthored over 130 original peer-reviewed scientific papers and three book chapters.

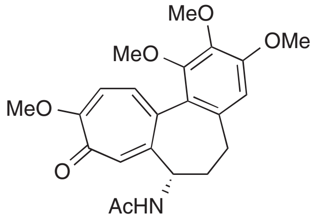
Eleonora Altman is a senior research officer at the Institute for Biological Sciences, National Research Council of Canada. She joined NRC-IBS in 1984 after receiving her Ph.D. degree in carbohydrate chemistry from the University of British Columbia. She has an extensive expertise in structural characterization of CPS and LPS of medically important bacteria and the application of high-resolution NMR spectroscopy and MS to structural analysis. In the early 1990s her interests shifted to studies related to antibody–carbohydrate interactions and specific modifications of carbohydrates. Over the past 10 years, Dr. Altman has been involved in studies related to bacterial adhesion and cell–cell interactions. More recently, this work has expanded to include investigation of the role of LPS in the pathogenesis of *Helicobacter pylori*, associated with chronic gastritis, peptic ulcers, and gastric cancer. Dr. Altman has published over 75 refereed papers in international journals within the area of structural analysis and glycobiology.

9.14 Modern Methods for the Isolation of Natural Product Receptors

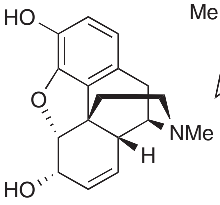
Peter Karuso, Macquarie University, Sydney, NSW, Australia

9.14.1 Introduction

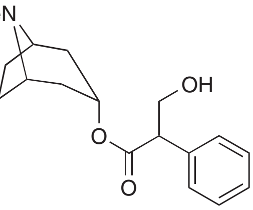
Humans have probably made use of natural products for tens of thousands of years. Certainly, by the time writing was invented the human pharmacopeia already contained thousands of plants, oils, and extracts. The Mesopotamians (2600 BC) recorded approximately 1000 plant substances on clay tablets in cuneiform. The Egyptian use of medicinal plants dates back to 2900 BC and the Ebers papyrus (1500 BC) records over 700 medicines derived from natural sources (mostly plants). Similarly, the Chinese *Materia Medica* records prescriptions from at least 1100 BC and the Ayurvedic system in India dates back to the same period (1000 BC).[1] These systems were based around supernatural beliefs and in many ancient societies, such as the Australian aboriginals, medicine and religion are closely bound.[2] In ancient Greek society, medicines and religion began to separate as the action of medicines began to be understood. Hippocrates (~400 BC) believed that health was based on the balance of black bile, yellow bile, phlegm, and blood (the four humors) and that natural products could be used to restore a balance once lost. With the loss of traditional knowledge to the Europeans in the Dark Ages, the Greco-Roman tradition was expanded to include the Indian and Chinese traditions by the Arabs, who documented these traditions in such works as *Canon Medicinae* and *Corpus of Simples*.[1] By the sixteenth century, Paracelsus had expanded and challenged the Hippocritean view, which was largely accepted at the time, with his belief that sickness was caused by attack from outside agents on the body, and that these agents can be neutralized with specific chemicals. This basic concept inspired the development of the pharmaceutical industry.[3] By the early nineteenth century many natural products had been purified, including colchicine, morphine, atropine, and strychnine, which are in use today. Interestingly, while the Mesopotamians had identified opium as a medicine in their writings of 2600 BC, it took until AD 1826 for E. Merck to commercially produce morphine as a chemical product. However, it was not long (1899) before Bayer produced the first synthetic natural product analog (aspirin) as a drug.[1] Aspirin is an analog of salicin, found in willow bark (and other plants) that were described by Hippocrates (400 BC) as having analgesic properties. In more recent times, even with the maturation of organic synthesis and the advent of combinatorial chemistry, natural products play a preeminent role with about 50% of new small-molecule drugs (1981–2006) being either natural products, natural product derivatives, or synthetic compounds based on a natural product.[4]



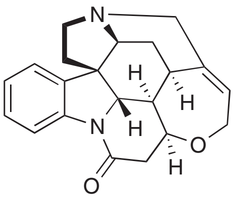
Colchicine



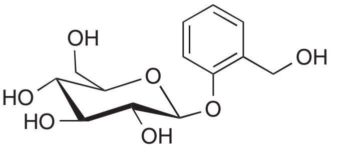
Morphine



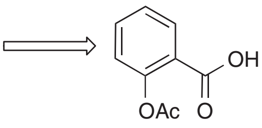
Atropine



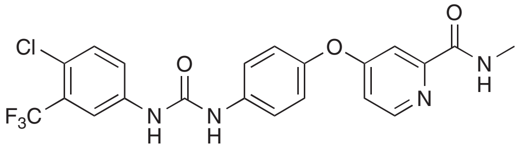
Strychnine



Salicin

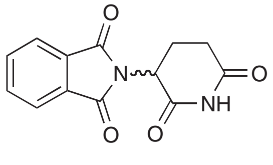


Aspirin

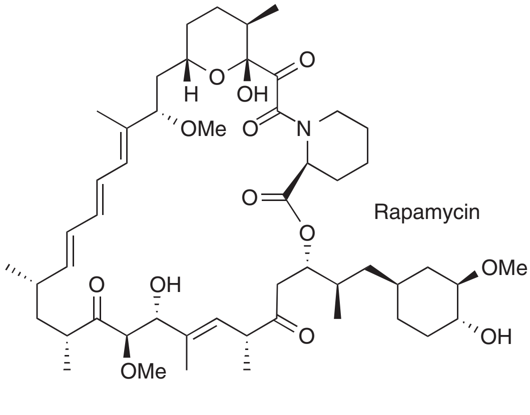


Sorafenib

Compared to combinatorial compounds, natural products are more ‘drug-like’, have higher molecular weights, incorporate fewer nitrogen, halogens (especially fluorine), or sulfur atoms but more oxygen atoms, and are sterically more complex, with more bridgehead atoms, rings, and chiral centers.[5] Despite these desirable properties and proven track record, emphasis on natural products research has declined in recent years as pharmaceutical companies embrace new technologies, such as combinatorial chemistry, which are not only faster and cheaper, but also have the advantage of clear delineation of intellectual property issues. While this led to a surge in activity, the expected productivity has not materialized, with the number of new active substances (NCE) falling to a 20-year low.[6] To our knowledge, only one drug has been developed *de novo* from combinatorial chemistry. That compound is sorafenib, a kinase inhibitor developed by Merck for the treatment of clear-cell renal cancer (and more recently, advanced liver and lung cancer) that was widely released in 2006.[7]



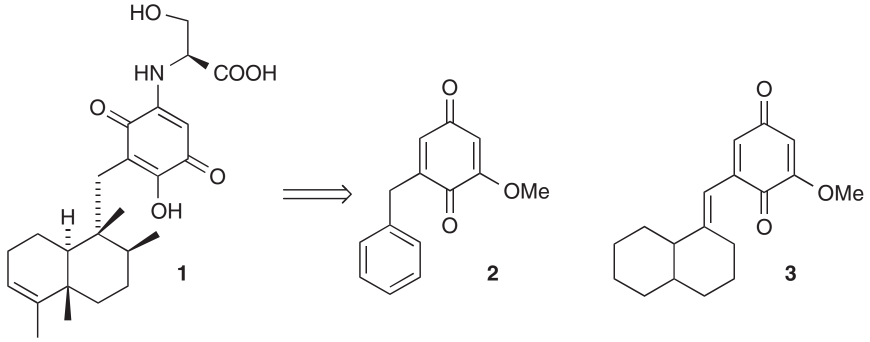
Thalidomide



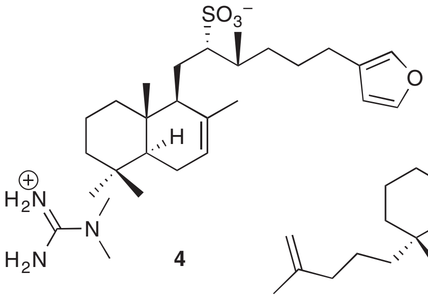
Rapamycin

As reactions in nature are heavily biased toward function, it follows that every natural product should have a biological target (or ‘receptor’) and that receptor will be a potentially drugable target. Only if one accepts this premise can the biochemical expense of producing natural products be rationalized. Therefore, we should consider all natural products as privileged structures and potential leads in drug discovery. The concept of ‘privileged structures’ has been widely used in medicinal chemistry, but was first espoused by Hirshfield in relation to the benzodiazepine structure, which seemed to produce a disproportionate number of active drugs.[8] This concept was extended to indole alkaloids by Labaudiniere and coworkers[9] and all biologically active natural products by Newman.[4] The topic of why inscrutable secondary metabolism produces many related structures (diversity orientated), while primary metabolism is more target orientated has had some robust recent discussion[10–17] and regardless if one accepts the ‘screening hypothesis’, the ‘waste metabolite hypothesis’, or the ‘all-active hypothesis’ the real question remains; can the structures of natural products be useful to us?

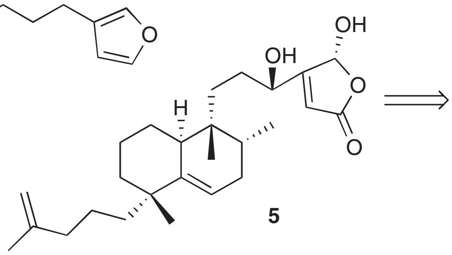
Humans and natural products generally have not coevolved, so it is unlikely that the true molecular target of a natural product is a human protein. Therefore, it can be argued that the importance of natural products is lost if they are applied to human diseases.[18] This apparent contradiction can be reconciled by the belief that natural products have emerged in nature to interact with biomolecules and that nature is inherently conservative as exemplified by the astonishingly similar genomes displayed by apparently divergent species (e.g., bananas and humans share 50% of their genes).[19] Thus, most proteins in lower eukaryotes and even prokaryotes have functional counterparts with sequence and fold similarities to human proteins. Jerrold Meinwald put this the best when he said: “Natural products have evolved to interact with something, and that something may not be so different from human proteins.”[20] There is thus ample evidence that natural products are important structures and that many, if not most, have protein-binding partners.



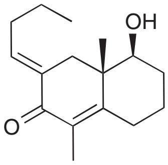
Although combinatorial chemistry has underperformed as an engine for drug discovery, the technique is effective for developing and optimizing natural product pharmacophores. This was demonstrated by Waldmann and coworkers,[21] who prepared a 74-compound library based on the natural product nakijiquinone C (1), which was originally isolated from a marine sponge.[22] Nakijiquinone C inhibits the c-erbB2 kinase but the library did not yield any compounds that were more potent inhibitors of c-erbB2. Two compounds (2, 3) were, however, found to be selective Tie-2 kinase inhibitors. Similarly, the same group have recently constructed a 483-member combinatorial library around the decalin core of the marine natural products sulfircin (4)[23] and dysidiolide (5)[24] that are known to be Cdc25A phosphatase inhibitors. Again, none of the library members displayed the desired phosphatase activity, although two (6, 7) showed low micromolar inhibition of acetyl choline esterase (AchE), which is a protein with a very similar protein-fold topology and ligand-binding site. These examples also demonstrate that a natural product’s shape is recognized not only by the biosynthetic enzyme that produced it but also by other (therapeutic) proteins that share a similar protein fold, which goes a long way toward explaining why natural products can be useful in systems that they did not coevolve with.[25]



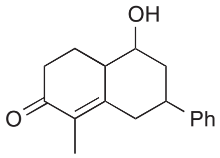
4



5



6 5.3 µmol l^{-1}



7 3.7 µmol l^{-1}

There is a general lack of information concerning cellular targets of most natural products but, without an understanding of the 'receptors' of natural products, efforts to generate more potent analogs through structure–activity relationships, structure-based design, or combinatorial chemistry may prove futile. Similarly important, but even less well known, are the identities of 'off-targets' that are responsible for toxicity or side effects. It would be advantageous to determine protein targets responsible for toxicity early in drug development and off-targets may be important in identifying other uses of a natural product. For example, the drug thalidomide, originally prescribed as a hypnotic to alleviate morning sickness in pregnant women (with disastrous consequences), is now used as an anti-inflammatory drug for treating Hansen's disease (leprosy) and as an anticancer agent against multiple myeloma (Thalomid). Similarly, the natural product rapamycin was first described as an antifungal agent but is now registered as an immunosuppressant (Rapamune). Rapamycin is also currently being developed as an anticancer drug, a side effect noticed in transplant patients using the drug.[26] The immunosuppressive and anticancer activity of rapamycin are mediated through its interaction with the protein mTOR and this fact has led to advances on multiple fronts. The development of rapamycin from antifungal to immunosuppressant and anticancer drug has taken over three decades and, it could be argued that, this process could have been dramatically shortened if there were methods to quickly isolate the protein-binding partners for the natural product. Substantial progress has been made toward achieving this goal over the past decade with the development of several new technologies but new tools are desperately needed for this daunting challenge.[27]

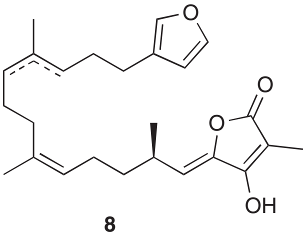
In this chapter, we review various methods that have been used to isolate the binding proteins for natural products but also look critically at methods that might be used in the near future to achieve this goal. Each has advantages and disadvantages and there is no one perfect method, but many are complementary. The review is not meant to be exhaustive but examples are chosen to highlight particular advantages or disadvantages of each approach. There are also certainly other techniques that may be adapted for the isolation of natural product-binding proteins but in the interests of brevity they all cannot be detailed.

9.14.2 Traditional Approaches

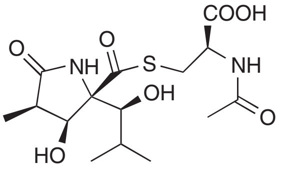
The 'Golden Age of Antibiotics' (1940–70), can be considered to have its rather unpretentious beginnings in an obscure Belgian journal that reported the serendipitous discovery of 'penicillin' by a Scottish microbiologist by the name of Alexander Fleming in 1928.[28] It is fittingly ironic that the impetus to take this obscure finding further came a decade later due to the allied need to increase troop survival rates after injury in World War II. It took another decade to elucidate the structure of penicillin and required the 'new' technique of X-ray crystallography.[29] The structure was new to science, resisting all traditional degradative approaches. With even more irony, penicillin was heralded as a wonder drug and that man's fight against disease was finally over. Naturally, bacterial resistance was already noted in 1945, soon after the drug was introduced into hospitals. Life was just not meant to be that easy. It took another decade to find that penicillin targeted the biosynthesis of the bacterial cell wall,[30,31] and a further decade to determine that the actual protein target was the transpeptidase enzyme.[32] The discovery of the cellular target for penicillin consequently led to the rapid development of many new transpeptidase and β-lactamase inhibitors (e.g., penems and monobactams). Interestingly, it has been estimated that over 10 000 penicillin analogs have been synthesized and tested but some of the most active compounds (cephalosporins, clavulanates, sulbactams, monobactams) are natural products.[1] Once a protein-binding partner of a drug is discovered, that protein is considered a therapeutic target and the search for new, more potent analogs can begin.

The important point of this story is that it took two decades of trial and error to discover the proteins that bind to penicillin but once found, a plethora of analogs was synthesized and discovered in nature in the 1970s. Today progress is faster, but an example from the modern literature will serve to demonstrate the classical approach to finding the protein target of a natural product, and its shortcomings.

In 2004, Jee H. Jung isolated a series of new and known sesterterpenes from the sponge *Psammocinia* sp. by bioassay-guided isolation.[33] The assay (brine shrimp lethality) suggested the authors that the compounds may act through some form of cytotoxicity. A panel of five cancer cell lines was thus tested and selective activity for some of the compounds was found for human skin cancer cells (SK-MEL-2) with low micromolar ED_{50} values. For whatever reason, many cytotoxic marine natural products tend to show selectivity for melanoma cancer cell lines. A common mode of anticancer activity involves the inhibition of DNA replication, so the authors tested their compounds in a ^{31}P-based DNA replication assay and found that the compounds inhibited SV40 DNA replication *in vivo*. DNA replication can be slowed by inhibition of topoisomerase or DNA polymerase, so the authors tested their compounds against two enzymes (topoisomerase I and DNA polymerase R) by previously established assays and found activity (IC_{50}) of a 1:1 mixture of strobilinin and felixinin (8) against both enzymes at 5 and 10 μmol l^{-1}, respectively. The results have since appeared in several reviews, which state that 8 is a ligand for topoisomerase and DNA polymerase[34–38] but the series of linear steps highlights some of the problems of a classical approach to the determination of natural product-binding proteins. At each step, a series of assumptions must be made about the underlying mode of action that depends intrinsically upon the current state of knowledge: Only activity for which a test exists can be used to support or refute each assumption, initiating a somewhat circular argument. For all other activities, as Wittgenstein put it "Wovon man nicht sprechen kann, darüber muß man schweigen" (what we cannot speak of we must pass over in silence).[39] The point is that in the traditional approach, we can only find activity in places where it can be found, which means in systems that we know about and with assays that are established.



8



9

Other than as leads in drug discovery, natural products are also effectively used as probes in chemical genetics.[40] To understand the function and role of a particular gene, a method is required to modulate the gene product. This is classically done indirectly through a genetic approach, by knocking out that gene. Unfortunately, the cascade of compensatory responses during development often conspire to mask the expected outcome. In addition, the gene product is often redundantly coded, resulting in no effect or is required at some stage of development and its loss is thus fatal. The deletion of the gene is mostly irreversible so cannot be simply regulated. In contrast, chemical genetics can be used to regulate gene products through inhibition or binding to proteins in cells giving substantial spatial and temporal regulation. Most natural products exert their biological effects through binding to proteins and many have found uses in understanding the cellular role of their binding protein. For example, lactacystin (9) was isolated from a species of *Streptomyces* and was found to induce neurite outgrowth.[41] Tritium-labeled lactacystin was used to identify the highly conserved N-terminal threonine residue of the mammalian proteasome subunit X as the protein-binding partner. The ability of lactacystin to inhibit cell-cycle progression and induce neurite outgrowth correlated with its ability to inhibit the proteasome. This demonstrated, for the first time, that the proteasome is a major regulatory complex required for cell-cycle progression. When lactacystin was added to proliferating cells, the oscillation of cyclin protein levels stopped, resulting in accumulation of the cyclin-dependent kinase (CDK) inhibitor p27, showing that p27 was required for cells to get from G_0/G_1 to S phase.[42] Lactacystin has also helped elucidate the role of the proteasome in inflammation. Inhibition of the proteasome with lactacystin stabilizes IκB and phosphorylated IκB by blocking the degradation of these proteins, which triggers the inflammatory response.[43] Other natural product proteasome inhibitors have also been used to gain a better understanding of the diverse roles of the proteasome in cell biology.[44]

To avoid the pitfall of assumptions based on current knowledge and to facilitate rapid drug development, it is important to develop unbiased approaches to finding the cellular partners for natural products, or any small molecule.

9.14.3 Genome-Wide Approaches

Genome-wide approaches can be differentiated at two basic levels. The first is forward chemical proteomics, which uses a cell's proteome as a library from which the natural product's binding partner is isolated. The second is reverse chemical proteomics, which uses the cell's genome. If the aim of chemical genomics is to discover a small-molecule modulator for every gene product (protein)[40] in a cell then the aim of chemical proteomics can be articulated as the discovery of the protein target of every biologically active small molecule.[45] This can be done in two ways, starting from a cell's total protein set (proteome), in which case we can call the process forward chemical proteomics or by starting from the cell's genome (or transcriptome),[46] in which case we can call the process reverse chemical proteomics.[47] In both cases, selection is based on an interaction between the natural product and a protein library (proteome) but in the latter, the protein is attached in some way to its encoding gene.

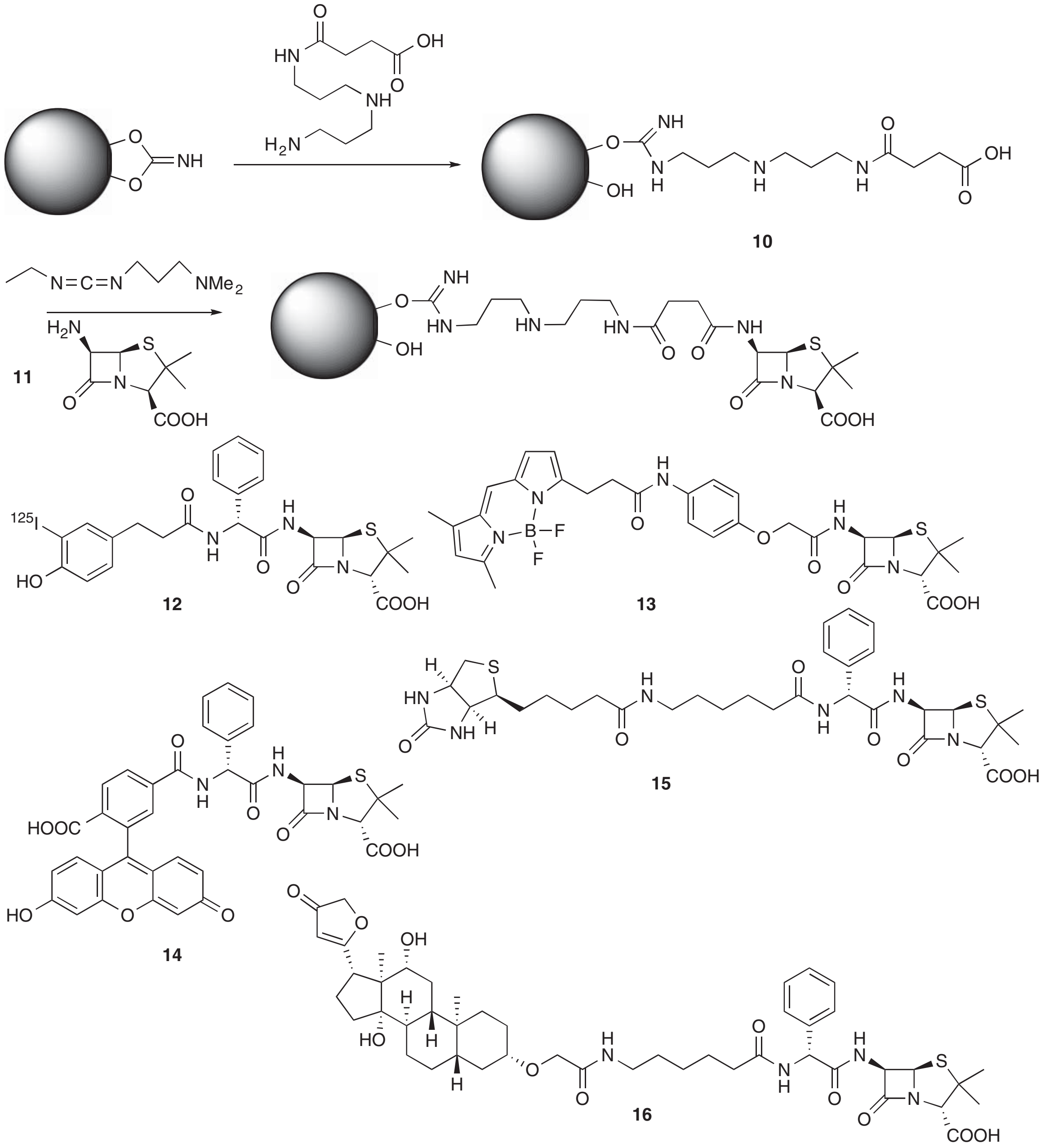
Growing interest, particularly from the pharmaceutical industry, in protein–protein interactions[48,49] has increased the demand for new methods for the rapid identification of interaction partners. These methods can (at least potentially) be modified and optimized by chemists to identify natural product-binding proteins though many of these technologies have not yet been widely used for small molecules, let alone natural products.

Knowledge of the cellular target for a natural product automatically identifies that target as 'druggable' and allows the full repertoire of medicinal chemistry, chemical genetics, and combinatorial chemistry to be applied to structure-based design and lead optimization. Thus, the determination of natural product-binding proteins is also a key step in increasing the diversity of known drug targets. It has been estimated that 50% of all approved drugs target just four families of proteins, and that the human genome (29 679 genes) currently includes only 207 known drug targets.[50] One of the reasons for this low number is that the pharmaceutical industry has relied for many years on the maxim that the best way to develop a new drug is to start with an old one.[50] This has led, as noted above, to a plethora of penicillin analogs such that over 4% of all known drugs target penicillin-binding proteins but also led to a critical lack of new antibiotics in recent years to fill the gap left by the development of antibiotic resistance to β-lactam-based antibiotics.

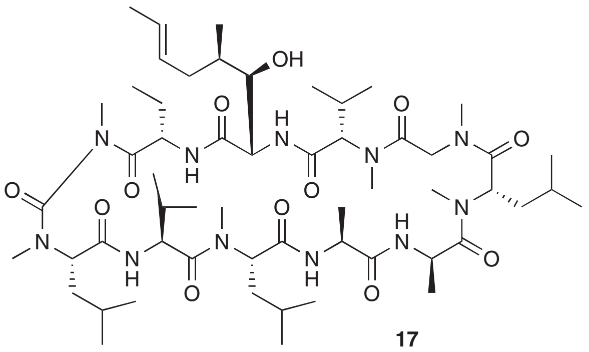
The full potential of high-throughput/high-content screening of new small molecules with pharmaceutical effect (NCEs) will not be realized until the supporting fields that allow the rapid isolation of small-molecule-binding proteins mature.

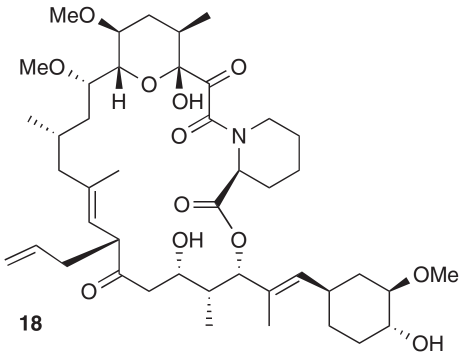
9.14.3.1 Chemical Proteomics

Chemical proteomics offers a convenient method of unbiased isolation of cellular protein-binding partners of natural products that does not rely on a circular argument but requires some way of tagging or immobilizing the natural product or the proteome. The natural product is commonly biotinylated or covalently attached to a chromatography medium but can also be tagged with a fluorophore, radioactive label, or photoaffinity probe. For example, Strominger isolated five penicillin-binding proteins from the membrane of *Bacillus subtilis* in the early 1970s using a penicillin-derivatized sepharose-affinity resin.[51] He started by activating Sepharose with cyanogen bromide and coupling succinyldiaminodipropylamine to prepare the carboxy-derivatized sepharose (**10**), which was coupled to 6-aminopenicillanic acid (**11**) with the aid of 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide. As penicillins make covalent bonds to their binding proteins (see Section 9.14.3.3.2) extensive washing of the column can reduce background binding and the captured proteins can be eluted with neutral NH_2OH, which cleaves the penicilloylenzyme bond. SDS–PAGE was used to separate the proteins and one was identified as D-alanine carboxypeptidase.[51] Several groups have also used radioactively labeled penicillin probes (e.g., **12**) and gel electrophoresis to identify penicillin-binding proteins, taking advantage of the covalent linkage between tagged penicillin and binding proteins.[52–54] More recently, fluorescent and biotinylated penicillin derivatives (e.g., 13–15) have replaced radioactive probes.[55–59] In the case of biotinylated penicillins, labeled proteins can be isolated by chromatography on streptavidin sepharose, eluting with detergent then separated by electrophoresis, electroblotted, and then visualized with fluorescently labeled streptavidin or by chemiluminescence through horseradish peroxidase (HRP)-linked streptavidin. With fluorescent penicillins, cell lysates or membrane proteins were incubated with tagged penicillin and then the proteins separated by 1D or 2D gel electrophoresis.[57] The fluorescently labeled spots can then be manually picked and identified by mass spectrometry. Digoxigenin-labeled penicillin (16) has also been used to show that chemiluminescence is more sensitive and safer than ^{125}I labeling.[60] Digoxigenin's small size and relative ease with which it can be attached to biomolecules, along with readily available antibodies, make it a standard chemical label and immunohistochemical marker for *in situ* hybridization.



The covalent binding of labeled penicillins to penicillin-binding proteins has been used in numerous studies including drug discovery, antibiotic mechanisms of action and resistance, and cell wall physiology.[61]


17

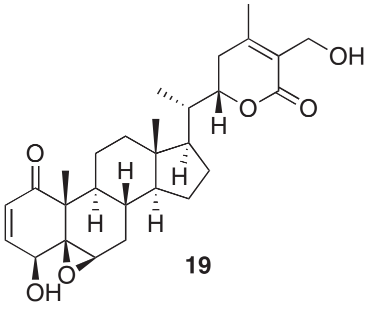

18

One of the first examples of chemical proteomics was the isolation of tubulin as the protein-binding partner of colchicine by Taylor and coworkers in 1965.[62–66] In this work, pig-brain proteome was incubated with a tritiated analog of colchicine and radioactive fractions were purified by column chromatography, analyzed by gel electrophoresis that led eventually to the discovery of tubulin. Similarly, Handschumacher and coworkers used a tritiated analog of the fungal (*Tolypocladium inflatum*)[67,68] natural product cyclosporin A (**17**) to isolate two cytosolic proteins from bovine thymus homogenates.[69,70] These were new proteins dubbed 'cyclophilins' and binding of cyclosporin A to human cyclophilin is responsible for suppression of the immune system. There are many other examples of radiolabeled natural products being used to isolate their cellular-binding partners. This technique works well for natural products that bind covalently to their partner, such as penicillins[53] (see above), lactacystin (**9**) (see below), and for compounds that have very high affinity for their binding protein (e.g., FK506; **18**); however, this would not be expected to work for natural products with moderate or high affinity as the interaction would not be expected to survive the purification steps (e.g., electrophoresis, ion exchange chromatography, and HPLC). The solution to this problem was to adapt the methods biochemists use to isolate proteins that interact with other proteins.

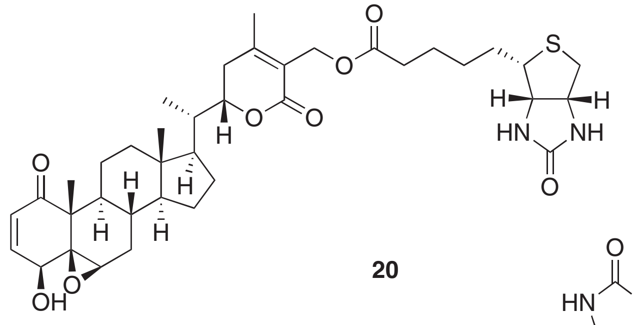
9.14.3.1.1 Affinity chromatography

One of the simplest techniques for identifying noncovalent natural product-binding partners involves tethering the natural product to an affinity resin (typically agarose, sepharose, or less commonly polystyrene) that has been functionalized to react with common groups such as amines, carboxylic acids, thiols, or alcohols. Agarose/sepharose-affinity matrices have the tendency to interact nonspecifically with proteins. As a result, many other affinity matrices have been adapted for chemical proteomics including glycidyl methacrylate, polymethacrylate, polyacrylamide, polystyrene, and polyethyleneglycol-based resins. A recent study of the suitability of a large range of gels used in peptide synthesis found that polyethylene glycol dimethylacrylamide copolymer (PEGA resin) was the best for the affinity isolation of the protein-binding partner for the FK506 (**18**) analog AP1497.[71] In a typical experiment (**Figure 1**), a cellular lysate or tissue homogenate is incubated at 4 °C in the presence of protease inhibitors with an affinity matrix. Typically, an hour to overnight is required to maximize contact of the proteome with the displayed natural product. The resin is then washed with buffer, usually containing a low concentration of a detergent such as Tween 20, to remove nonspecific-binding proteins. Binding proteins are then eluted either specifically (with free natural product) or, more conventionally, under denaturing conditions (e.g., 2% SDS) removed under denaturing conditions. The eluted proteins can then be resolved by 1D or 2D gel electrophoresis (SDS–PAGE) and identified through sequencing or more commonly by mass spectrometry.[72] The most popular technique is peptide mass fingerprinting (PMF), which involves excision of the putative protein band from the gel, tryptic digestion of the protein and then identification of that protein by its characteristic pattern of peptides. One of the great challenges of affinity chromatography is the minimization of nonspecific interactions. This can be achieved using extensive washes but risks elution of moderate or even high-affinity proteins. Other methods involve serial-affinity chromatography, where the cell lysate is first passed through a resin that is derivatized with an inactive analog of the natural product to remove nonspecific binders. This can also be done in parallel, where the eluate from the inactive analog resin is compared to the eluate from the natural product resin. In the competition variant, two sets of resin derivatized with the natural product are used where in one experiment, excess soluble natural product (soluble control) is added to the cell lysate and just solvent to the other. Comparison of the two eluates reveals the specific binding proteins missing from the soluble control experiment.

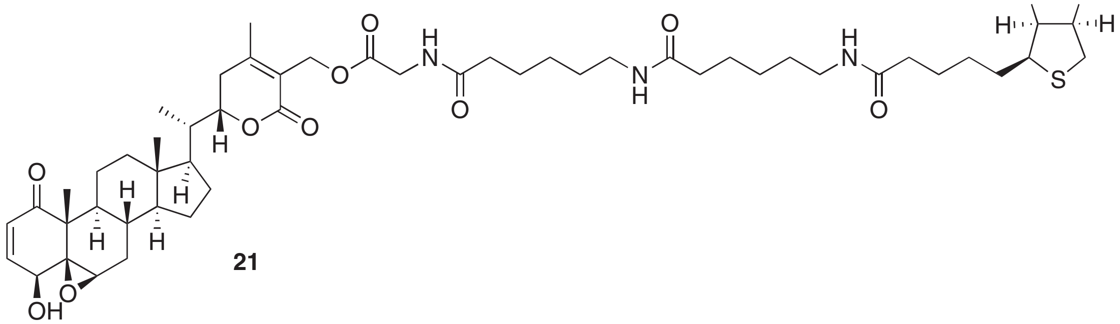
Affinity chromatography is by far the most popular method for the isolation of natural product-binding proteins. For example, withaferin A (**19**) a member of the withanolide family of C28 steroidal lactones occurs in many solanaceous plants, especially *Withania somnifera*.[73] It exhibits antiangiogenic and antitumor activity *in vivo*, which result from this drug's potent growth inhibitory activities.[74,75] Whitesell and coworkers isolated the putative human target for this natural product through a classical forward chemical proteomics approach.[76] To achieve this withaferin A was modified at the C27 position with commercially available pentafluorophenyl-biotin to yield biotinyl-withaferin (**20**), as available structure–activity relationship data suggested that this position was noncritical.[77] Cytotoxicity testing of biotinylated withaferin A confirmed preservation of its bioactivity. This is an important step often not done to confirm that the affinity probe retains its biological activity. The biotinylated withaferin was immobilized onto streptavidin-coated agarose beads and incubated with MCF-10A breast epithelial cell extract that had been previously cleared of nonspecific-binding proteins with underivatized streptavidin-sepharose (serial-affinity chromatography). Analysis of the proteins that bound to the underivatized beads revealed a surprisingly large amount of protein, highlighting the need to remove background binders that could conspire to mask any real binding partners. The derivatized streptavidin-sepharose was preincubated with either free withaferin A or solvent (dimethyl sulfoxide – DMSO). Each set of beads was then incubated with the precleared cell lysate and washed with buffer to retain both tight and loosely bound proteins. Elution of the bound proteins (It is difficult to break the biotin–avidin noncovalent interaction but it can be done with, for example, boiling SDS or 8 mol l^{-1} guanidine (pH 1.5) or by addition of excess free biotin. The authors did not state specifically how the protein elution was achieved.) was followed by electrophoresis (sodium dodecyl sulfate–polyacrylamide gel electrophoresis – SDS–PAGE under reducing conditions) and staining with SYPRO ruby. A single band at approximately 36 kDa was observed in the DMSO-treated lysate that did not appear in the withaferin A-treated lysate. The conclusion was that this band was effectively displaced by prior addition of unmodified withaferin A (competition variant). Tryptic digestion of the excised protein gave a series of peptides that were analyzed by liquid chromatography–tandem mass spectrometry (LC–MS/MS) to identify human annexin II as the protein-binding partner.[76] Further investigation of the peptides generated from tryptic digestion of pure annexin II in the presence of withaferin A suggested that the withaferin A reacted with Cys133. The authors assumed this was a nucleophilic attack of the thiol or thiolate of the protein on the ring B epoxide of the withaferin A but it is possible that conjugate addition of the thiol on the ring A enone occurs.



19



20



21

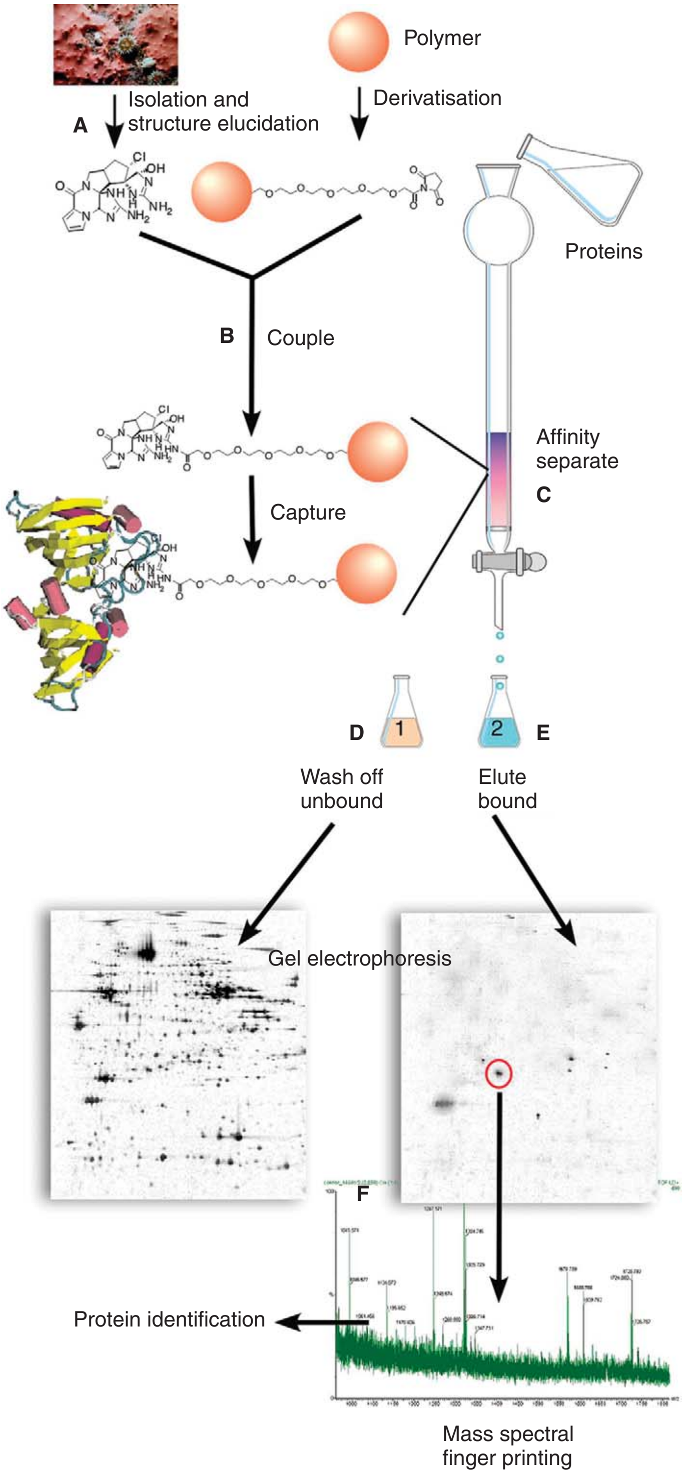
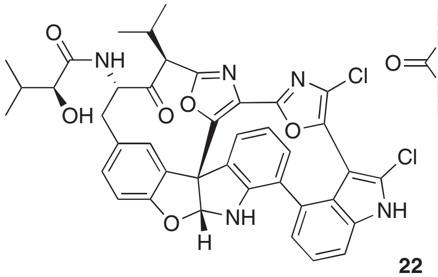


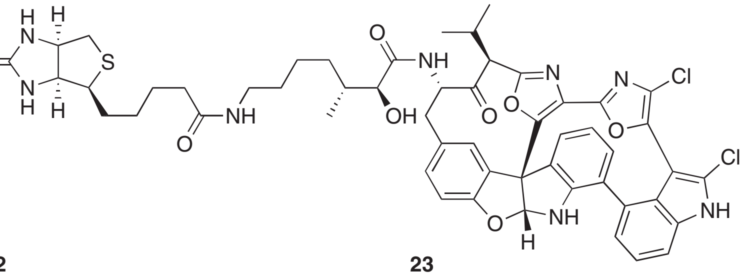
Figure 1 Chemical proteomics starts with a pure natural product (A) that is tethered to a solid support through some appropriate chemistry (B), covalently or through biotin–avidin, using a suitable chemical linker. It is often necessary to synthesize several analogs with requisite functional groups for coupling to a linker and to test these to ensure they retain the biological activity of the natural product. The resulting affinity matrix, which could be a classical affinity column, for example, (C) is incubated with a cell lysate to capture specific-binding proteins. After washing (D), the binding proteins are eluted (E) with underivatized natural product (specific elution) or under denaturing conditions (e.g., 2% SDS; nonspecific elution) and identified by either LC–MS or SDS–PAGE–PMF analysis (F). Experimental parameters, such as the immobilization chemistry, stringency of washes, and elutions need to be optimized to achieve an acceptable level of background, nonspecific binding.

Interestingly, a year later, Mohan and coworkers showed that withaferin A binds to the 56 kDa intermediate filament protein, vimentin, by covalently modifying its cysteine residue, which is present in the highly conserved α-helical 2B domain. The natural product induces vimentin filaments to aggregate *in vivo*, seen as punctate cytoplasmic aggregates, that colocalize vimentin and F-actin.[78] The methods used by Mohan were almost identical to those of Whitesell except that they used human umbilical vein endothelial cells and the withaferin A probe (**21**) contained a long linker between the natural product and the biotin. The authors could find no trace of annexin II and concluded that the linker allowed the isolation of an alternative binding protein. It is possible that withaferin A binds to more than one protein with high affinity and that the longer linker allowed the Mohan probe to bind to vimentin, which was excluded by the lack of a linker in the Whitesell probe. Such examples highlight the requirement for long linkers that are able to effectively present the natural product to proteins that may have deep-binding pockets, not accessible to probes with a short linker.

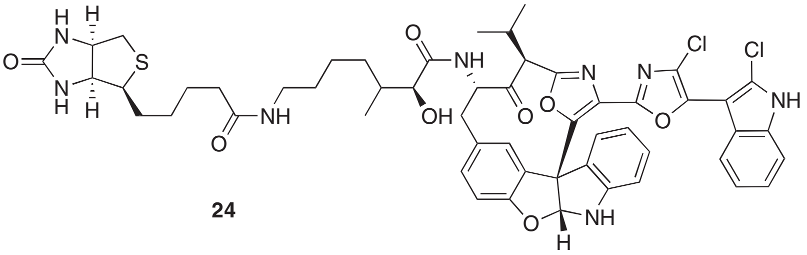
Diazonamide (**22**) is a marine natural product isolated from the tunicate *Diazona chinensis*, though the originally reported structure was wrong.[79] Diazonamide inhibits human cancer cell growth and, at low nanomolar concentrations, induces an M-phase growth arrest.[80] Wang *et al.* used affinity chromatography to isolate the binding protein for diazonamide immobilized on avidin–agarose resin. They used two affinity resins, one derivatized with a biotinylated diazonamide (**23**) and the other with a very close structural analog (**24**) that was missing just one bond.[81] Incubation (4 °C, overnight) of each resin with partially purified protein extracts from HeLa cells revealed two proteins of approximately 50 kDa in the SDS–PAGE gels that were not present in the eluate from resin derivatized with the control (**24**). In addition, they used a soluble control (free diazonamide) to show that purification of the two 50 kDa proteins was inhibited in the presence of **22**. By mass spectrometry, both proteins were found to be ornithine δ-amino transferase (OAT), a hitherto uncharacterized mitochondrial enzyme. They confirmed that inhibition of OAT was responsible for diazonamide's antimitotic effects by knocking down endogenous OAT with siRNA. Interestingly, diazonamide does not affect OAT's enzymatic activity suggesting that OAT has a role in mitotic spindle assembly unrelated to its catalytic activity.



22



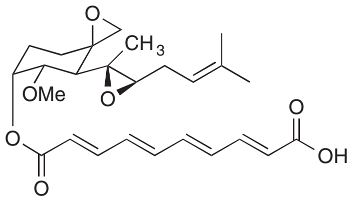
23



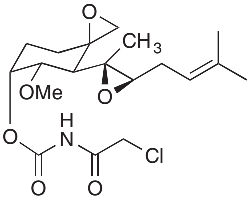
24

The Crews lab at Yale specializes in the isolation of protein receptors for epoxide containing natural products such as fumagillin,[82] eponemycin, and epoxomicin.[83] These examples are notable because the probes were incubated with live cells, followed by lysis and affinity purification. Fumagillin (**25**) and TNP-470 (**26**), a semisynthetic analog of fumagillin, are inhibitors of angiogenesis, causing late G_1 phase endothelial cell-cycle arrest.[84] Despite the fact that TNP-470 had undergone numerous pharmacological studies and clinical trials, little was known about its cellular target or mode of action. Sin *et al.*[82] used a biotinylated analog of fumagillin (**27**) to identify its cellular target from human umbilical venous endothelial cells. This led to the isolation of human methionyl aminopeptidase (MetAP-2) – a cobalt-dependent metalloprotease. Mass spectrometry and

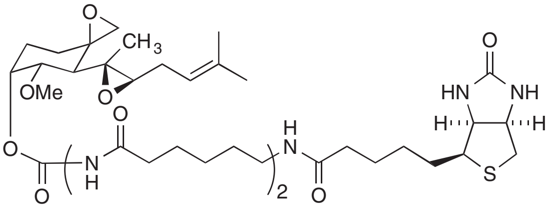
X-ray crystallography revealed the imidazole nitrogen of His231 in the active site of MetAP-2 that forms a covalent bond with the carbon of the spirocyclic epoxide of fumagillin.[85,86] Metalloproteases are known to be involved in angiogenesis by helping break down of the intercellular matrices required for new blood vessel growth.



25

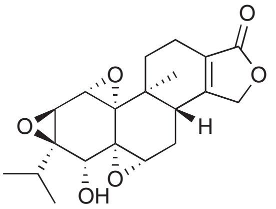


26

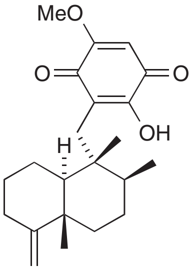


27

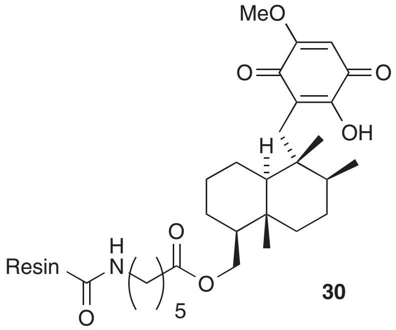
A similar strategy was employed to identify the cellular receptor for triptolide (**28**), a terpenoid from the Traditional Chinese Medicine (TCM), the plant *lei gong teng* (*Trypterygium wilfordii*). Triptolide is a triepoxide that induces rapid apoptosis in many cancer cell lines by inhibiting NF-κB transactivation.[87] The natural product was nonspecifically tritiated and incubated with growing HeLa cells. After lysis and anion exchange chromatography, the radioactive fraction was analyzed by SDS–PAGE. Radioactivity was specifically associated with a 110 kDa protein that was subsequently identified as a calcium channel protein (PC2).[88] Knowledge of the protein-binding partner for triptolide allowed the compound to be developed for treatment of polycystic kidney disease.[89] In this case, as most from the Crews lab, identification of the cellular partner was possible only because a covalent bond is formed.



28



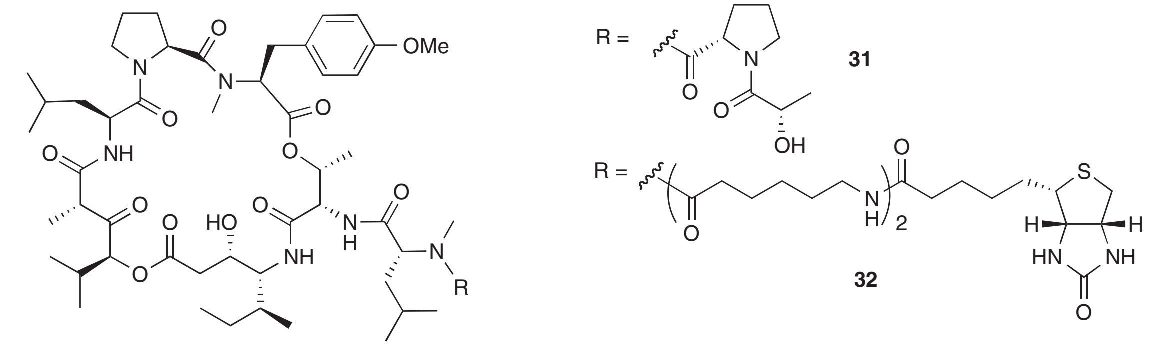
29



30

Ilimaquinone (**29**) is a marine sesquiterpene quinone first isolated in 1979 from the sponge *Hippospongia metachromia*.[90] It has been reported to have mild antibacterial, antiviral, antifungal, and anti-inflammatory activities, but more interestingly, ilimaquinone has been found to break down the Golgi apparatus into small vesicles, thereby blocking cellular secretion.[91,92]

In a classic paper, Snapper synthesized an ilimaquinone–agarose-affinity resin (**30**), which was incubated with homogenized bovine liver and then washed extensively.[93] Proteins retained by the resin were separated by gel electrophoresis, yielding six main protein bands. Amino acid sequencing of these bands revealed three proteins involved in the activated methyl cycle – SAHase, *S*-adenosylmethionine synthetase (SAM synthetase), and catechol-*O*-methyltransferase (COMT) – as well as three unrelated proteins. Subsequent enzymatic assays established that ilimaquinone is a competitive inhibitor of SAHase, but has little effect on the activity of SAM synthetase or COMT. The authors noted that a consequence of SAHase inhibition would be the intracellular accumulation of SAH, which is a potent feedback inhibitor of methyltransferases. These results support the assertion that methylation events play an important role in cellular secretory events and vesicle-mediated processes. The study also highlighted the problem of nonspecific interactions as only one of the six isolated proteins was shown to interact in any way with the natural product.

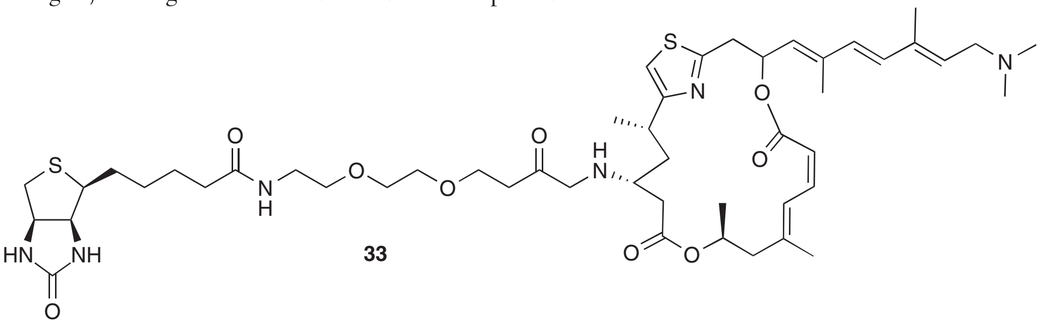


31

32

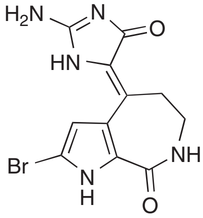
Didemnin B (**31**) is a cyclic depsipeptide, first isolated from a Caribbean tunicate, *Trididemnum solidum*, in 1997.[94] Didemnin B has potent anticancer activity, inducing G_1 cell-cycle arrest at nanomolar concentrations.[95] In addition, didemnin B also exhibits antiviral[96] and immunosuppressive[97] activities and has been shown to inhibit protein synthesis.[98] Schreiber and coworkers used affinity chromatography to identify cellular receptors for didemnin B.[99] In their study, they used bovine brain tissue homogenate and passed through a streptavidin-agarose column to remove endogenous streptavidin- and agarose-binding proteins. The column eluate was then incubated with a biotinylated analog of didemnin B (**32**) and the resulting mixture was passed through a fresh streptavidin–agarose column (serial-affinity chromatography). Proteins retained by the column were eluted with SDS, separated by gel electrophoresis, and identified by amino acid sequencing.[99] In addition to a lot of background, a 49-kDa protein with 95% homology to human translation elongation factor 1α (EF-1α) was retained by the column. Further investigation revealed didemnin B binds to EF-1α only in the presence of GTP. EF-1α is an abundant guanine nucleoside-binding protein that transports amino-acyl tRNAs to the ribosomal A site in a GTP-dependent manner.[100] The authors hypothesized that didemnin B blocks protein synthesis by binding to GTP:EF-1α and preventing the GTP:EF-1α-amino-acyl tRNA complex from interacting with the ribosome. However, this interaction does not explain didemnin B-induced late G_1 cell-cycle arrest at nanomolar concentrations as binding to EF-1α was in the millimolar range. The significance of Schreiber's findings was quickly questioned based on the apparent low binding to EF-1α but also because the characteristic side chain of didemnin B was missing from the probe (**32**).[101]

Schreiber responded by removing the majority of EF-1α from the bovine brain tissue using cation-exchange chromatography.[102] The EF-1α-depleted brain lysate was then passed through a column of didemnin-B immobilized on Affigel and proteins retained by the affinity matrix were separated by electrophoresis. Two new didemnin-B-binding proteins (of 34 and 36 kDa) were observed. Amino acid sequencing and subsequent cloning of the cDNA encoding these proteins revealed them to have similarity to human palmitoyl protein thioesterase (PPT), an enzyme that removes palmitate from H-Ras and the $G_{\alpha s}$ subunits of heterotrimeric GTP-binding proteins *in vitro*. Further investigation by Crews revealed that didemnin B uncompetitively inhibits the enzymatic activity of recombinant human PPT.[103] However, there is speculation that the didemnin B receptors isolated may not be related to the biological activity of the molecule.[104] This is because the affinity probes employed had the linker attached to didemnin B in place of the Pro^{8}Lac^9 side chain, which previous studies have confirmed to be important for cytotoxicity and immunosuppression.[105] Joullié and coworkers have synthesized a range of fluorescent,[106] photoaffinity,[107] and radioactive[108] analogs of didemnin B with intact side chains that could be useful for identifying the cellular targets, although no such studies have been reported.

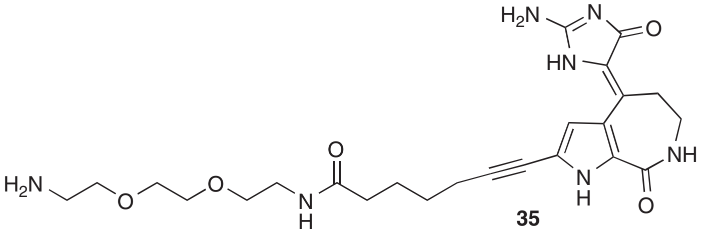


33

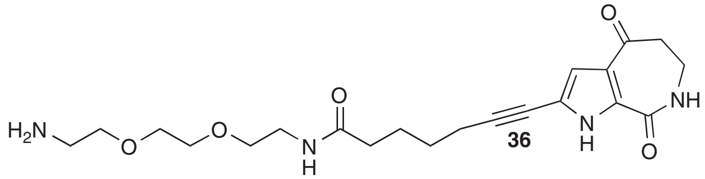
Pateamine A is a dilactone macrocyclic natural product first isolated in New Zealand from the sponge *Mycale* sp.[109] The compound was described as possessing possible selective, potent cytotoxicity (P388; IC$_{50}$ 150 pmol l^{-1}). Romo employed his total synthesis of pateamine A to synthesize a biotinylated version (**33**) that was used in a streptavidin sepharose pull-down experiment to isolate the threonine/serine kinase eIF4A from RKO cells.[110] It was later shown that binding of pateamine A to eIF4A disrupts binding to eIF4E and eIF4G and, in a rare case, actually increases the enzymatic activity of eIF4A thereby inhibiting eukaryotic translation.[111]



34



35

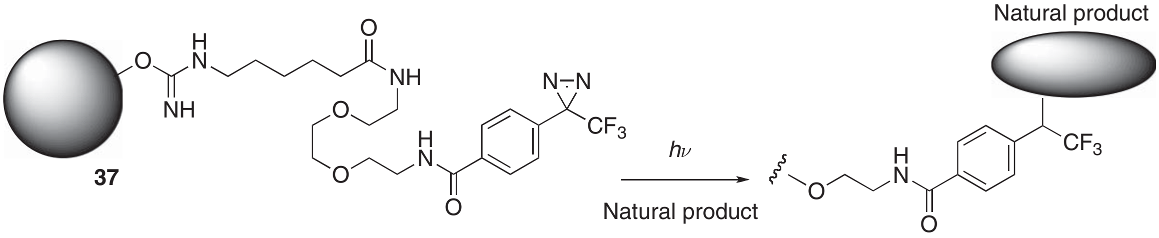


36

Hymenialdesine (**34**) is a cyclic oroidin alkaloid that has been isolated, with several close analogs, from several axinellid sponges[112] and been shown to be a potent inhibitor of CDKs.[113] The crystal structure of hymenialdesine with CDK2 revealed that the molecule bound to the ATP-binding site of the kinase and that the bromo-group pointed out of the pocket. Meijer and Gray then used a Sonogashira coupling to attach a seven-carbon linker to the C5-position (**35**), which was then covalently coupled to an affinity matrix through an ethylene glycol linker. As a control resin, an aldesine analog (**38**) was synthesized. Mouse brain extracts were incubated with the affinity resins, washed, and nonspecifically eluted with Laemmli buffer (95 °C). Electrophoresis revealed many bands from the resin loaded with hymenialdisine and very little from the aldesine-resin. In-gel tryptic digestion of the major bands followed by MALDI–MS analysis revealed three kinases known to be strongly inhibited by hymenialdisine (GSK3α, GSKβ, and Mek1), two new kinases that have not been previously known to be targets of hymenialdisine (p90RSK and an isoform of BIKe), and β-tubulin.[114] Unlike other examples of affinity chromatography, this example shows that it is possible to isolate multiple-binding partners in one experiment.

As illustrated by the few examples above, affinity chromatography, using natural products as bait, is an effective method for the isolation of natural product-binding proteins. This facilitates the identification of cellular processes that would be difficult to study without the aid of the natural product and provides invaluable information for rational drug development. Despite its success, the examples above also highlight the limitations of the techniques. First, chemical proteomics will usually isolate the most abundant binding protein over the most avid as highlighted by the didemnin B work of Schreiber and Crews. Second, despite the first point, binding interactions need to be strong so as to survive the washing conditions required to reduce background binding to an acceptable level. Even so, one can expect many false positives as exemplified by the ilimaquinone example. Using a precleared extract (against an affinity matrix that is very close to the natural product derivatized resin) is a popular strategy to avoid this problem. Third, the natural product needs to be derivatized without affecting its biological activity. This can be a real challenge, not only from a chemical perspective but also from a supply perspective as many natural products are rare or difficult to make. Seemingly small changes can affect what protein-binding partners are isolated as exemplified by the withaferin example above.

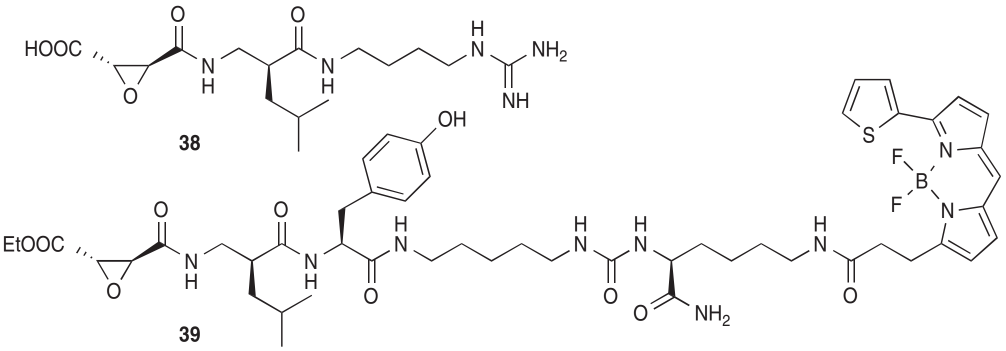
One enterprising solution to the problem of derivatization, by Osada and coworkers, is to use random photo cross-linking of the natural product to a surface or resin. Thus, aryl diazirine groups (see Section 9.14.3.4) covalently attached to glass slides have been used to create small-molecule microarrays.[115] In a proof of principle, FK506, cyclosporine, digitoxin, and a number of other small molecules were photo-cross-linked to the glass and shown to bind the respective fluorescently labeled antibodies. The aryl diazirine group is transformed, with UV light, into a highly reactive carbene, which then reacts irreversibly with small molecules in a manner that is independent of the functional groups.[116–118] Osada has recently reported a similar approach to derivatize an affinity gel, such as agarose beads, with natural products and shown that they retain their binding capability by rescuing their binding proteins from cell lysates.[119] In this paper, the group attached a photoactive linker to CH agarose 4B beads (**37**) and then suspended the underivatized natural product ($10\,\mathrm{mmol\,l^{-1}}$) in methanol with the beads. After drying, the beads were irradiated with UV light (365 nm) and then washed thoroughly with methanol and water. In this case, cyclosporine (**17**) and FK506 (**18**) were used. Incubating the cyclosporine beads with Jurkat cell lysate, washing, and elution allowed the isolation of CsA, the cellular receptor for cyclosporine. Similarly, FKBP12 was isolated from beads derivatized with FK506.



A similar approach has been reported by Angela Koehler except that the glass slide is derivatized with a diethyleneglycol linker terminated with an isocyanate group instead of a photoreactive group. After the slides are printed with natural products, they are exposed to pyridine vapor to catalyze the covalent attachment with amines, alcohols, and thiols. The slides, printed with AP1497, FK506, or rapamycin were found to bind fluorescently labeled, recombinant FKBP1a and corticosterone and digoxin to their respective fluorescently labeled antibodies.[120] The method was then trialed on a natural product extract from *Streptomyces hygroscopicus*, the organism that produced rapamycin.[121] Crude extracts under various growth conditions were screened with fluorescently labeled FKBP1a and it was found that rapamycin could be detected in the presence of quite complex mixtures.[122] However, the obvious extension to printing a nanoLC gradient elution of a crude extract to identify fractions containing specific natural products was not done. Presumably this would be difficult to achieve as the solvent (or impurities) used for HPLC could react with the isocyanate, and it would be difficult to get the volumes small enough to fit onto one glass slide. Other disadvantages are that the natural product is not randomly derivatized but through a possibly critical functional group and that different functional groups react at widely different rates with isocyanates that could bias binding. The method also relies on the availability of fluorescently labeled proteins of interest and can therefore not discover new proteins that bind to natural products. However, if the adsorbed proteins could be analyzed by mass spectrometry (see Section 9.14.3.1.5), then photochemical immobilization could become a very useful method for the identification of natural product-binding proteins. Finally, it should be possible to use a reverse affinity chromatography approach, where a protein of interest is covalently linked to a support and a crude extract of natural products passed through the column, to isolate compounds that bind specifically to that protein (reverse chemical genetics). This approach is used for combinatorial libraries but there are no examples from the field of natural products chemistry.

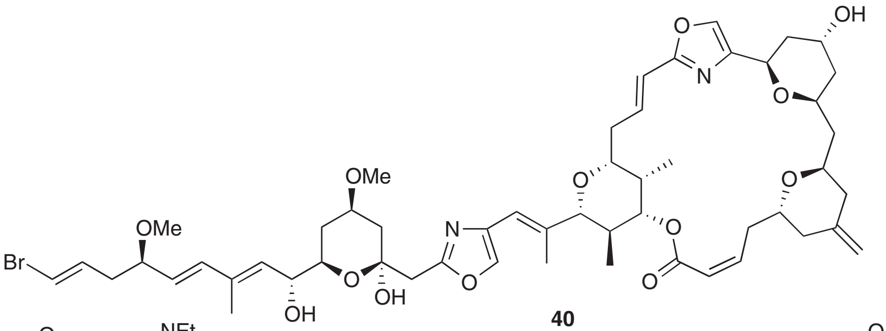
9.14.3.1.2 Fluorophore tags

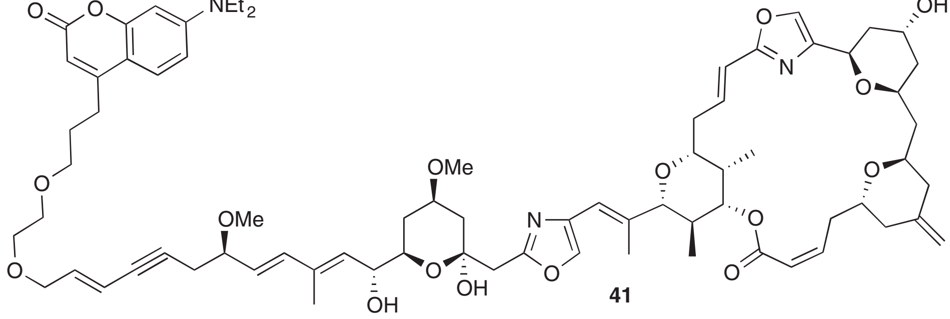
Fluorescently tagging natural product analogs has been used for activity-based proteomics to identify putative classes of enzymes in SDS–PAGE gels.[45,123] For example, the natural product, E-64 (**38**), first isolated and characterized from *Aspergillus japonicus* by Hanada *et al.*[124–126] has been used as an activity probe based on the natural product's ability to covalently bind cysteine proteases.[127] For example, yellow DCG 04 (**39**) was incubated with a rat kidney homogenate and separated by gel electrophoresis. Because E-64 is specific to papain-like proteases and derivatizes the active site cysteine covalently, the protein–probe adduct is stable to the denaturing conditions of gel electrophoresis (SDS–PAGE). This allowed Bogyo and coworkers to identify a number of cathepsins in the rat kidney proteome under various conditions.[127]


38

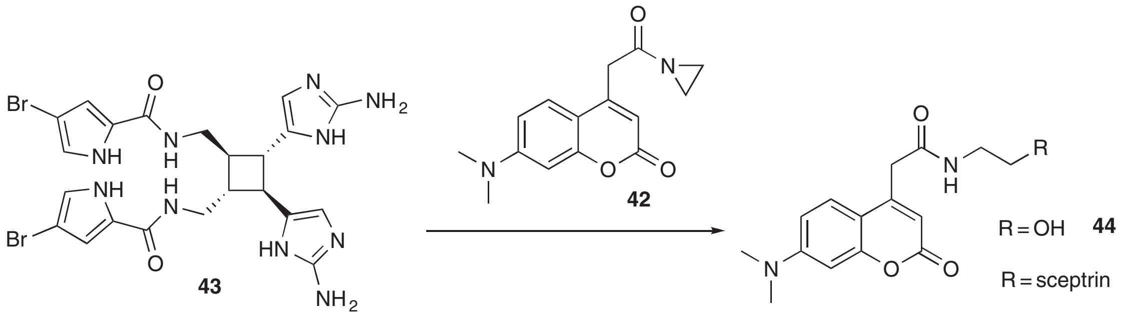
39

Phorboxazole A (**40**) is a very potent cytostatic agent from the Western Australian sponge, *Phorbas* sp. first isolated by Ted Molinski in 1995.[128] It causes S-phase cell-cycle arrest at picomolar concentrations (average cytostatic activity in NIH's 60-cell panel is ∼1 nmol l^{-1}) but the target of phorboxazole A is somewhat mysterious. The total synthesis of phorboxazole A[129] allowed the synthesis of a fluorescent analog (**41**).[130] As it was previously determined that modification of the C45,46-vinyl bromide of phorboxazole A did not substantially diminish cytostatic activity against cancer cells, this site was used to introduce a fluorescent *N*,*N*-dialkyl-7-aminocoumarin group through a Sonogashira coupling between the corresponding vinyl iodides and the C46 terminal alkyne. In this example, the fluorescent analog was incubated with HeLa cells and the lysate fractionated into nuclear, cytosolic, endoplasmic reticulum, Golgi, and membrane partitions with only the cytosolic fraction displaying significant fluorescence. Native gel electrophoresis yielded two protein bands at 32 and 54 kDa, with only the 54 kDa band persisting after purification on an anti-*N*,*N*-dialkyl-7-aminocoumarin antibody-affinity column. The protein was identified through tryptic digestion, nanoLC-MS/MS as human cytokeratins KRT1, KRT9, or KRT10; cytoskeletal proteins. HeLa cells grown in the presence of phorboxazole A were lysed and run down a KRT10 antibody column. Elution and SDS–PAGE isolated the 32 kDa protein and MS analysis revealed this to be the kinase cdk4. Further studies showed that only in the presence of phorboxazole A, or the fluorescent analog, did cdk4 associate with KRT10 and that it was most likely that this association was responsible for cell-cycle arrest.[130] As the association between phorboxazole and cytokeratins is in the picomolar range, it allowed the isolation of its putative cellular receptor by native gel (where the proteins are not denatured) electrophoresis. This would not have been possible if the affinity was lower due to equilibrium dissociation during electrophoresis.


40


41

Le Clair has also recently proposed a system to identify protein targets in parallel with natural product isolation.[131] In a proof of principle, they used a crude extract of the sponge *Agelas conifera* to simultaneously isolate new antibiotics and their protein-binding partners. First, the crude sponge extract was passed through a series of affinity columns that had the entire *Escherichia coli* proteome immobilized in different ways. The aim was to isolate any natural product from the sponge that bound specifically to any bacterial protein. The columns were washed with buffer and eluted with ethanol (95%; 50 °C). The crude protein-binding extract of the sponge was dried and reacted with a fluorescent tag (**42**), designed to react randomly with nucleophiles in the natural product. The tagged library of natural products was then removed by an antibody (to the fluorescent tag) affinity column equilibrated and washed with RIPA buffer and then eluted again with ethanol. The crude fluorescently tagged protein-binding sponge extract was then incubated with a crude *E. coli* protein extract and captured on a tag-antibody column, eluted under denaturing conditions and analyzed by SDS–PAGE to reveal a single band that appeared at approximately 40 kDa. Tryptic digestion of the excised band followed by LC–MS–MS indicated the protein was MreB, a protein that has been identified as a drug target in bacteria because of its critical role in cell wall assembly.[132] The tagged small molecule was partially identified by HRMS ($C_{37}H_{40}Br_2N_{12}O_5$) and the natural product was re-isolated from more crude sponge extract and identified as sceptrin (**43**), a natural product first isolated from *Agelas sceptrum*,[133] that has been shown to disrupt membranes and cause spheroplast formation of the cell walls of bacteria.[134] Although interesting, this method will likely be limited to natural products that bind to abundant proteins (such as MreB) with high affinity. While it may be possible to isolate enough protein to identify, the method is unlikely to yield enough natural product for identification (unless the compound is well known with a characteristic mass spectrum as in this example) because the moles of protein isolatable dictate the amount of small molecule isolatable. The authors suggested that it may be possible to prepare affinity columns containing a single recombinant protein to increase the amount of natural product isolatable but this would be logistically difficult for every single protein, even in a small proteome and adds another step, requiring isolation of the gene and overexpression of the putative natural product-binding protein in another host. The method would also require high concentrations of the natural product in the extract and that the natural product had a suitable nucleophile that reacted efficiently with the tag and that in so doing did not abrogate its biological activity. There are also a number of technical problems to do with knowing how much of the reactive fluorescent tag (**42**) to add to the natural product extract as excess needs to be decomposed and would form a large background (**44**) if the tag is used for purification (as in this example).

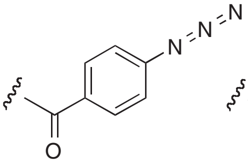


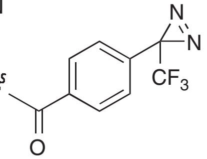
9.14.3.1.3 Protein arrays

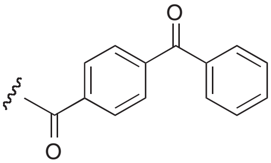
Printing of proteins (proteome chips) is an emerging technology that holds promise if a number of fundamental technical hurdles can be overcome.[135] One advantage of this technique is that it can overcome the limitations of sample often experienced with affinity chromatography techniques, but protein arrays are in the early stages of development.[136] The method is analogous to gene chips except that it relies on the printing of proteins onto slides or membranes. A major disadvantage of protein chips over gene chips is the fragility of proteins, many of which can be easily denatured. It is also not trivial to create a representative normalized proteome arrayed and have the proteins functional. So far, there have only been proof of concept studies for finding small-molecule receptors using this technique.[137]

9.14.3.1.4 Photoaffinity labeling

Photoactivated probes are widely used in structure–function studies of biomolecules such as proteins, peptides, DNA, and glycolipids.[138] The technique has been available since the 1970s but is little used in natural products research.[139] However, it should be popular among chemists as it requires quite a bit of chemical manipulation of the probe but not much biological manipulation. Unlike other methods, photoaffinity labeling can be used at two levels. First, it can be used to screen binding proteins for a particular small molecule and has the advantage that low-affinity interactions are easily detected. Second, if the protein partner is known, it can be used to determine exactly where on the protein the small molecule binds. The second mode is also possible with small molecules that form stable covalent bonds to their target proteins (e.g., the penicillins; see Section 9.14.3.3.2) but is applicable to all interactions in photoaffinity labeling. The field has been extensively reviewed.[138,140–144]

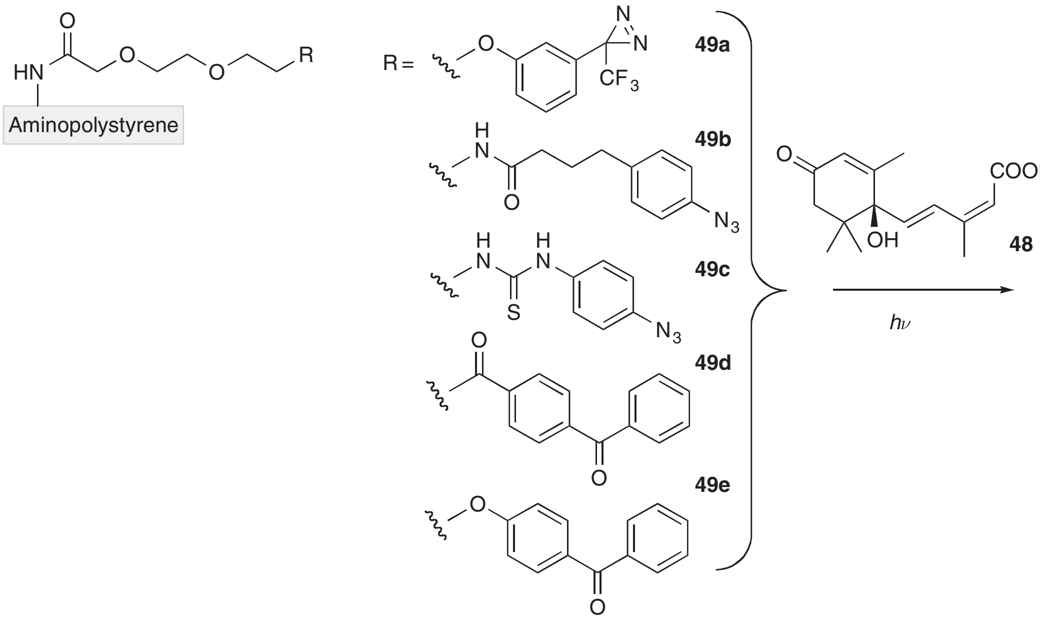

45


46


47

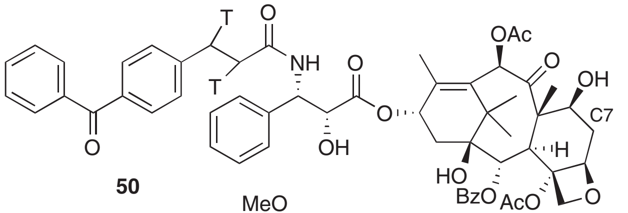
Even though the technique was proposed in the 1960s[145] it was not until recently that it has become popular due to the advent of more advanced photophores and the tools and techniques of proteomics.[140] Ideal photophores need to be stable in the absence of UV light, be activated under mild conditions that do not adversely affect the biological system, not affect the biological activity of the probe it is attached to, react irreversibly and nonspecifically with the nearest molecule, and have half-lives shorter than the k_{off} of the ligand–receptor dissociation. In addition, the concentration should be kept below the concentration of the putative protein partner to avoid labeling of nontarget biomolecules. This has, in the past, been a technical difficulty as the amount of protein isolated was often too small to identify. For gel analysis, fluorescent protein stains, such as Deep Purple, have superseded silver stains because they can detect picograms of protein, are MS compatible, and do not produce heavy metal waste.[146] In combination with new MS techniques, proteins can be readily identified at six orders of magnitude lower levels than the natural products chemists need to identify a small molecule.

The photophores used today are generally phenylazides (**45**), phenyldiazirines (**46**), or benzophenones (**47**) that produce nitrenes, carbenes, or diradicals, respectively.[143] While azides are the easiest to synthesize, the resulting nitrenes suffer from long half-lives, rearrangements into ketenimines and the formation of unstable *N*-heteroatom bonds. Benzophenone-based probes have a very low reactivity toward water but require prolonged UV irradiation causing considerable damage to other molecules. The trifluoromethylphenyldiazirines are the most difficult to make but are quickly activated to carbenes that form carbon-based bonds but suffer from high reactivity toward water.[138] There is thus no perfect photophore but the method has been used to isolate the binding partners for many biomolecules and been found especially useful for uncovering very weak-binding interactions such as those within lipid membranes.[140] There have been several comparative studies of using different photophores producing conflicting results. The Hatanaka group generally find the phenyldiazirines the best. For example, they found that only diazirine-modified tetrodotoxin (**48**) was capable of identifying the toxin-binding region at the mouth of the sodium channel.[147] In contrast, the Prestwich group have found that the benzophenone group is the best overall because they are activated in a reversible manner through excitation–relaxation cycles, only C–H bonds within 3.1 Å of the carbonyl oxygen are modified and are stable to common protic solvents.[148] However, more recently, Taylor has taken a slightly different approach by photoimmobilizing abscisic acid onto polystyrene derivatized with six different photophores. Recognition of immobilized abscisic acid was monitored by two fluorescent antibodies; one recognizing the carboxy chain and the other the cyclohexyl ring. They found that **49a** showed a significant response to the antibody that recognizes the side chain but nothing for the other antibody. Compounds **49b** and **49c** were equally well recognized by both antibodies but at a lower intensity than **49a**, whereas benzophenone **49e** showed no reactivity with either antibody. Corning's Universal-BIND plate showed recognition only by the cyclohexyl-recognizing antibody and was most similar in response to compound **49d**. These results show that the right choice of chemistry can strongly affect the efficiency and orientation of labeling and that the factors that are involved are not yet well understood.

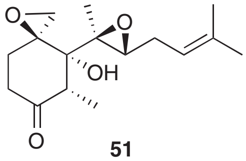


In a typical experiment, a protein extract is prepared from a tissue homogenate and subjected to initial protein purification by ion exchange or size-exclusion chromatography.[149] Fractions containing the desired activity are incubated with the photoprobe to establish a noncovalent interaction and then irradiated to establish a covalent bond between the ligand and the nearest other molecule. Low efficiency of cross-linking is unavoidable due to inevitable side reactions leading to high nonspecific background labeling. The mixtures can then be separated by SDS–PAGE. As the covalent linkages are generally stable, this can be done under denaturing conditions without loss of label. The gel can then be imaged on a phosphorimager (^{125}I) or by fluorography on X-ray film (^{3}H). Competition assays can be done with underivatized probe to establish if the interaction is specific. Finally, labeled proteins can be isolated by HPLC or CE and subjected to tryptic digestion and PMF to establish the location of the probe on the protein backbone. Alternatively, the most-radioactive fragments can be purified by HPLC and identified by MALDI–TOF mass spectroscopy.[144]

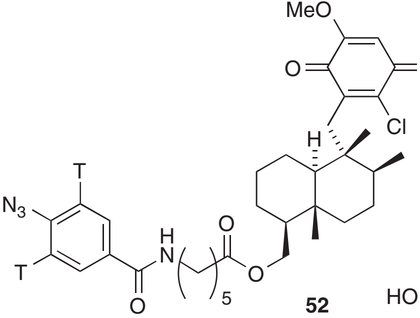
As mentioned earlier tetrodotoxin has been used as a probe to determine the binding site of the toxin in sodium channels. Other natural products that have been used with photophores include taxol. In a study by Prestwich, photoaffinity labeling with benzophenone/tritium-labeled taxol (**50**) led to incorporation into tubulin, as expected, but also P-glycoprotein, a drug efflux pump.[150,151] In another early example, the antiangiogenic natural product ovalicin (**51**) (isolated from *Pseudeurotium ovalis*)[152] was derivatized with a phenylazide photoaffinity label and a radioactive iodine.[153] Cell extracts were incubated with the probe in darkness and then irradiated with UV light. SDS–PAGE revealed a number of radioactive bands but pre-incubation of the cell extract with unlabeled ovalicine resulted in the disappearance of one band at 67 kDa. That protein was identified as methionine aminopeptidase (MetAP). Ironically, it was discovered, in the same year, that ovalicin (and fumagillin) probably covalently modify a conserved (across all species) histidine residue in the active site of the MetAP from *E. coli*.[85] It has recently been proved that the histidine attacks one of the epoxides using single-crystal X-ray analysis of the MetAP–ovalicin adduct.[154] MetAP–1 and MetAP-2 have recently been shown to be involved in cancer, confirming a clear link between ovalicin (**51**) and fumagillin (**25**), their protein target and their biological activity – a link that would not have been made so easily without the use of chemical proteomics.[155]



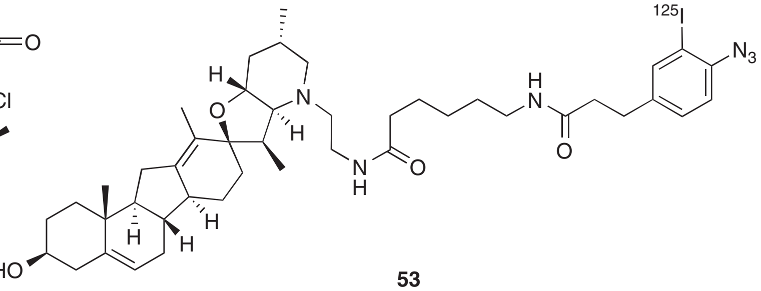
50



51



52

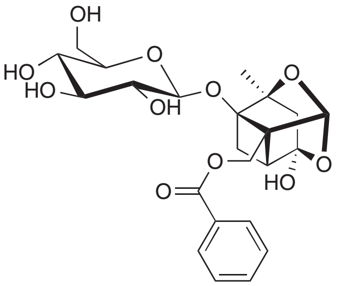


53

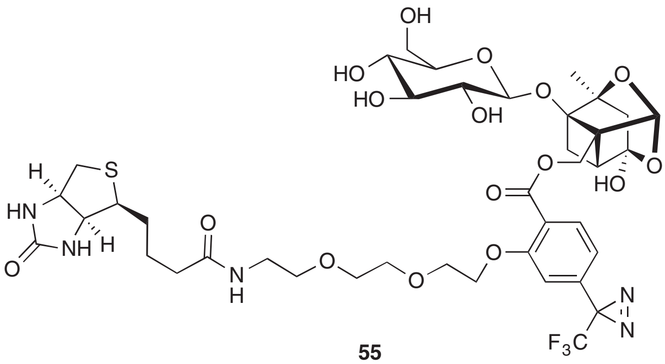
Radeke and Snapper[156] used affinity chromatography to isolate the binding proteins for ilimaquinone (see above) but, in a related experiment, used a photoaffinity reagent to identify the cellular receptors for ilimaquinone. A tritiated azidobenzene moiety was attached to a synthetic chloroquinone analog of ilimaquinone to form a photoaffinity reagent (**52**). Bovine liver extract was incubated with this reagent in the presence of ultraviolet light. The crude mixture was then separated by ion-exchange chromatography and fractions containing high levels of radioactivity were further purified by size exclusion, yielding two major radioactive bands (48 and 55 kDa). Amino acid sequencing of the 48 kDa band revealed it to be the enzyme *S*-adenosylhomocysteinase (SAHase). SAHase plays a key role in cellular methylation chemistry by catalyzing the breakdown of *S*-adenosylhomocysteine (SAH) to homocysteine and adenosine and was shown to bind ilimaquinone. Interestingly, in this case, photoaffinity labeling worked much better than affinity chromatography, the latter producing far more background.

Cyclopamine is a terpene alkaloid from the American wild corn lily *Veratrum californicum* that was found to induce congenital birth defects in lambs, including cyclopia.[157] It was known that the teratogenic effects of cyclopamine, and the related jervine, are due to inhibition of cellular responses to the Hedgehog family of growth factors.[158] However, how cyclopamine specifically inhibits Hedgehog pathway activation remained unclear. To help elucidate the mechanism of Hedgehog activation, Beachy derivatized cyclopamine with a bifunctional probe (**53**), containing a phenyl azide photophore and ^{125}I radionucleotide for detection.[159] They found that cyclopamine binds directly to the heptahelical bundle of the protein smoothened thereby regulating its activation and providing a molecular basis for cyclopamine action.

Paeoniflorin (**54**) is a monoterpene glycoside from the TCM *chi chao yao* (*Paeonia lactiflora*) that was first isolated in 1963.[160] The compound has been reported to exert anticonvulsive, antithrombotic, antihypertensive and antihyperglycemic effects, and much more.[161] It has also recently been reported that paeoniflorin has neuroprotective effects by activation of the adenosine A1 receptor.[162] Taking into account the fact that the benzyl group of paeoniflorin has little effect on its activity,[163] the group designed and synthesized a trifunctional probe (**55**) that was used to photoaffinity label the natural product's binding protein. The probe was incubated with a homogenate of rat brain (4 °C, 8 h) and then irradiated with UV light. The concentrated homogenate was separated by SDS–PAGE and then electroblotted onto PVDF. The membrane was blocked with BSA/PBS/Tween, washed and incubated with streptavidin–horse radish peroxidase, washed again and then incubated with Enhanced ChemiLuminescence reagent (GE Healthcare), and visualized by exposure to X-ray film. Photolabeled and control preparations all had bands at 73 and 120 kDa for naturally biotinylated proteins but only preparations that contained the photophore were exposed to UV light showing a new band at 55 kDa. Unfortunately, this new band was not identified and there have been no subsequent papers from the group following up these results.[161]

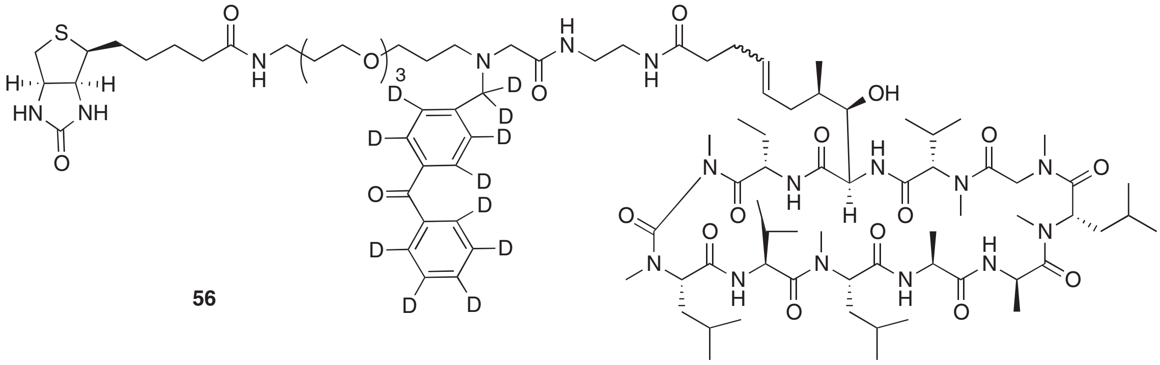


54



55

A new approach that is now becoming popular are the tetrafunctional probes that, in addition to the biotin, photophore, and natural product also contain an isotope tag (see Section 9.14.3.1.5) to facilitate mass spectral identification. Borrowing directly from modern proteomics (ICAT)[164] allows the cross-linked proteins to be readily identified by mass spectrometry. In a proof of principle, Belshaw and coworkers[165] tagged cyclosporin A (**17**), a natural nonribosomal peptide from *Trichoderma polysporum*,[67] with a long biotinylated linker that also contained a perdeuterated benzophenone photophore (**56**). The tag was coupled to the natural product through a cross metathesis reaction. Cyclosporin A is known to bind to cyclophilin *in vivo*,[70] which in turn binds to calcineurin to mediate its immunosuppressive effects.[166] Belshaw used a 1:1 mixture of **56** and the undeuterated analog in the labeling experiments resulting in a heavy and light version of any labeled protein, separated by 11 Da. To demonstrate this approach the researchers made up a mixture of four proteins, contain cyclosporine, FKBP, carbonic anhydrase, and ovalbumin. The photophore reagent was preincubated with protein for 1 h and then frozen in liquid nitrogen and subjected to UV radiation in a photoreactor (350 nm, 15 min), with intermittent freezing of the sample. The proteins were purified by size-exclusion chromatography then adsorbed onto monomeric avidin resin, washed, and then eluted with biotin. The eluate was reduced (dithiothreitol – DTT) and alkylated (iodoacetamide) and subject to trypsin digestion and the digest analyses be LC–MS/MS. Only two peptides were found to contain the unique isotopic signature of the probe and corresponded to the cyclophilin peptides 92–118 and 56–82. Both peptides are on the exposed cyclosporine-binding side of cyclophilin. This method overcomes a major disadvantage of other photoaffinity labeling techniques by simplifying the purification and identification step.

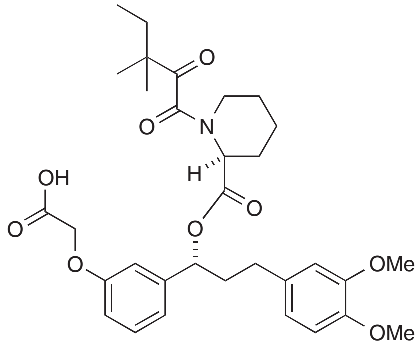


56

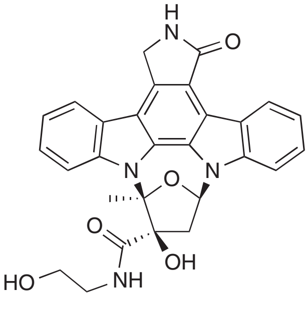
General disadvantages of photoaffinity tagging are that the probes are rather complex and reactive and there are relatively few generally useful photoreactive groups. Low efficiency of cross-linking is unavoidable due to inevitable side reactions leading to high nonspecific background labeling and UV damage. This generally requires some form of prefractionation of the proteome or tissue homogenate, which must be assayed for the desired activity before photoaffinity labeling.[144] The method is also noniterative but with the rapidly increasing power of mass spectrometry it can be predicted that photoaffinity labeling will become increasingly popular with natural products chemists to determine the protein-binding partners. Unlike display technologies, the isolation of DNA, lipid, and very weak interactions is already possible and likely to improve with further research into improved photoactivated functional groups.

9.14.3.1.5 Isotope labeling

There are a number of emerging techniques related to the previous section that have done away with the photoaffinity labeling by incorporation of the isotope tag into a proteome. Stable isotope labeling techniques (e.g., SILAC, iTRAQ) are currently the best methods in quantitative proteomics. In SILAC,[167,168] cells are cultured in a medium that contains either light (natural isotopes) or heavy (^{15}N and ^{13}C) versions of lysine and arginine. After about five population doublings, $^{13}C_6{}^{15}N_2$–Lys and $^{13}C_6$–Arg are fully incorporated into the proteome. Tryptic digestion of proteome, which yields peptides with just one Lys or Arg, shows a characteristic +6 or +8 Da shift in the mass spectrum. Mixing equal quantities of cells from the heavy or light cultures, followed by tryptic digestion and then HPLC-yields peptides that have an equal abundance of the light and heavy isotope as indicated by MS peaks of equal height. This method relies on the lack of any primary isotope effect on the production of proteins by cells. iTRAQ[169,170] does not make this assumption because peptides are labeled with a reactive *N*-hydroxysuccinimide ester after tryptic digestion. Four isobaric (equal molecular weight) versions of the iTRAQ reagent are available that, when reacted with a tryptic digestion, label all lysine-containing peptides. In the mass spectrometer (LC–MS), peptides labeled with any of the reagents appear at the same mass but MS–MS leads to intense fragments of different mass for the reporter group (variously ^{13}C- and ^{15}N-labeled *N*-methylpiperizine; 114–117 amu), where the isobaric tag is split into its light and heavy components. This technique allows cells to be grown under four different conditions, labeled with the iTRAQ reagent and then mixed and analyzed by mass spectrometry. Theoretically, all proteins that are differentially expressed can be detected if unique peptides can be identified. Although the iTRAQ method has not been used to identify the protein-binding partner for a natural product, it has been used to reveal the mechanisms of action of clinical ABL kinase inhibitors.[171] The research group of Cellzome AG in Heidelberg developed ‘kinobeads’, an affinity resin derivatized with a mixture of pan-kinase inhibitors (such as staurosporine and bisindolmaleimide) to bind all kinases in a particular sample. By adding three different concentrations of Gleevec to K562 cell lysates, the authors showed that, in addition to the known target (kinase domain of BCR–ABL), the receptor tyrosine kinase DDR1 binding to the kinobeads was specifically inhibited by increasing concentrations of Gleevec. This system could be used immediately for the detection of kinase-binding natural products and to determine selectivity of known kinase inhibitors.



AP1497 **57**



K252a **58**

Another elegant example uses SILAC to achieve a similar result. Here, the Carr and Schreiber groups derivatized affigel 102 (BioRad) with an analog of FK506 (AP1497; **57**).[172] HeLa S3 cells grown in light SILAC media were lysed in the presence of the FK506 analog and those grown in the heavy SILAC medium without the addition of the analog. The lysates were incubated with the affigel derivatized with the same compound (AP1497), washed, and then combined. The affigel was then boiled in SDS and the isolated proteins reduced and alkylated before tryptic digestions and analysis by mass spectrometry. In this way, the authors were able to show that addition of soluble AP1497 was able to reduce the binding of FKBP1a, FKBP2, FKBP4, FKBP5, FKBP9, and FKBP10 to affigel derivatized with AP1497. Similarly, the staurosporine analog K252a reduced the binding of approximately 50 kinases to affigel derivatized with **58**. Importantly, the authors showed that the soluble competitive-binding method worked much better than using control beads that contained everything except the natural product analog. In the latter case, the beads were less able to discriminate real from background binding. This is a common problem in affinity-based methods. For example, Oda *et al.*[173] identified 377 proteins as potential binding partners for the anticancer drug E7070 using two affinity matrices, one labeled with the drug and the other with an inactive close analog. The eluted proteins from each resin were labeled with a cysteine-reactive fluorophore (Cy-3 or Cy-5) for DIGE or ICAT reagents for MS analysis but no convincing protein partner could be identified.

In these types of experiments, the magnitude of the SILAC or iTRAQ isotope ratios is dependent on the protein abundance in the lysate and its affinity to the probe. For example, a very low abundance protein with very high affinity could have the same isotope ratio as a high-abundance protein with low affinity, stressing the need for appropriate controls. Another disadvantage of both methods is that the stable isotope reagents are very expensive. A clear advantage of this technique is that the natural product (or drug) does not have to be derivatized but relies on the availability of resins that can effectively bind particular classes of protein, which is currently a limitation.

9.14.3.1.6 Drug western

In this method, bacteria are transfected with a human cDNA plasmid library (e.g., λTriplEx) and plated onto an agar plate. A nitrocellulose membrane, soaked in isopropyl β-thiogalactopyranoside (IPTG) to induce protein production, is placed over the top and the plate/membrane incubated for 4–5 h at 37 °C. The membrane was removed, washed, and blocked with gelatin before incubation with a small molecule that had been chemically conjugated to the protein BSA. After 6 h at room temperature, excess drug–BSA conjugate is removed and the membrane treated with anti-BSA antibody conjugated to HRP and the HRP detected by chemiluminescence. Using this method, Tanaka *et al.* were able to screen approximately 10^6 transformants against HMN-154 (a synthetic anticancer lead), and identify 10 plaques with significant chemiluminescence. Six clones contained DNA with no homology to any known sequence, the remaining four encoded NF-YB, thermosin β-10, human growth hormone, and gonadotropin releasing hormone. The latter two were considered promiscuous binders to BSA and not investigated further but NF-YB and thermosin β-10 clones were isolated and the expressed protein purified and shown to bind HMN-154. Considering the small number of transformants ($\sim 2 \times 10^4$) that can be produced in one experiment and the large amount of manipulation required it is unlikely that this method will have particular advantages over other better established methods. The results are also not convincing as only one clone of each protein was detected among 80% background.

9.14.3.2 Reverse Chemical Proteomics

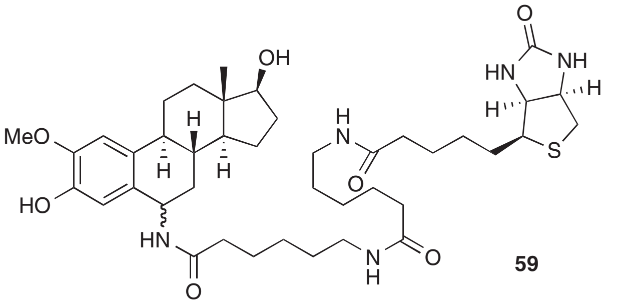
Over the past few years, a number of strategies have been developed to screen random peptide libraries for sequences targeting specific proteins. Peptides with desired binding properties can be serially selected and amplified based on a physical link with the encoding cDNA. These technologies include mRNA/DNA, viral, ribosomal, and cell displays with phage display being by far the most commonly used and probably the best understood.

In contrast, the display of alien proteins on the surface of cells, organelles, and viruses has been far less common but has provided biologists with important tools for the directed evolution of proteins (e.g., enzymes and antibodies) and analysis of protein–protein interactions.

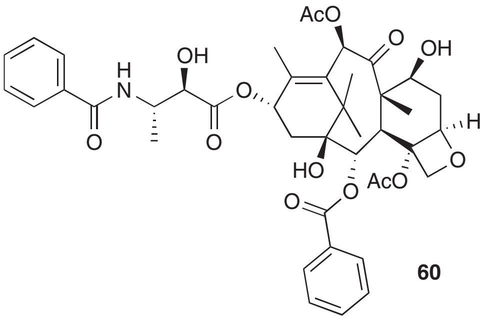
This section deals with display technologies and their possible use in the discovery of natural product receptors.

9.14.3.2.1 Phage display

Phage-displayed peptides, antibodies, and cDNA libraries have proven invaluable in mapping protein–protein interactions, protein–drug interactions, and the molecular evolution of enzymes and antibodies and has been well reviewed.[174–177]



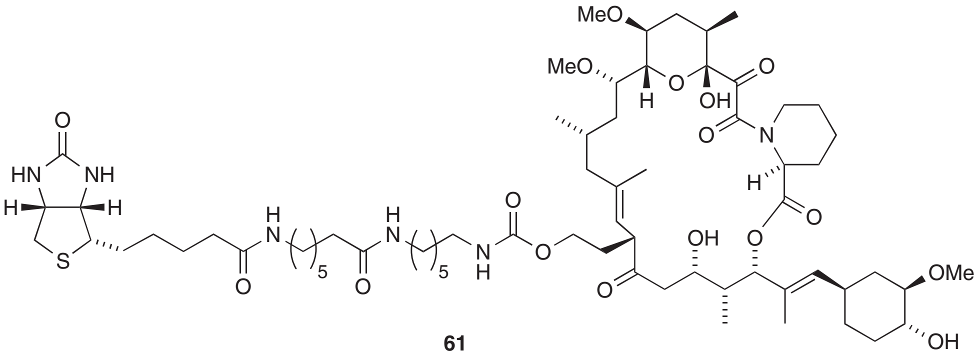
59



60

Phage display was first developed with the *E. coli*-specific, filamentous bacteriophage M13.[178] Filamentous phages replicate and assemble without killing their host by assembly in the bacterial membrane periplasm and secretion through the outer membrane. The M13 phage particles consist of single-stranded DNA (ssDNA) surrounded by a protein coat composed of primarily protein pVIII ($2.5–3 \times 10^3$ copies). The ends of the filament are capped by five molecules of the minor coat protein pIII and PVI, which are involved in bacterial binding and in termination of the phage particle assembly process. The other end of the phage is capped by five complexes of minor coat proteins pVII and pIX, which are required for initiation and maintenance of phage assembly in the host bacteria. The most widely used format at present is display on the surface of M13 bacteriophages through fusion to the pIII or pVIII coat proteins. However, these proteins can only take N-terminal display as the derivatization of the C-terminus is not tolerated. Unfortunately, this means that the bacterial ribosome first translates the alien protein and terminates when it reaches the 3′-stop codon of a cDNA fusion. The only C-terminal display system where the coat protein is translated first followed by the alien protein is with coat protein pVI.[179] There are limitations on the size of the chimera and this system has not proved popular. In an alternative strategy, Crameri and Suter designed a pIII-based cDNA-display system that is utilized for leucine–zipper interaction of c-Fos and c-Jun. Here the cDNA library is fused to c-Fos and incorporated into the M13 genome. The pIII protein is uniformly fused to c-Jun and in the cytoplasm, the two protein associate and form a disulfide bond, covalently linking the alien protein to pIII through c-Fos and c-Jun.[180] As the pIII protein is also required for bacteriophage assembly and infectivity, there are some limits to the size of the protein that can be successfully expressed and displayed without causing an unacceptable bias toward phage with no alien gene or a very small insert during bacteriophage replication. Because of these and other difficulties, there are few reports of successful cDNA library display with M13 phages and no reports of using this system to isolate small-molecule target proteins. However, there are two papers from Makowski and coworkers using pIII, M13 random peptide libraries (commercial 7-mer and 22-mer) to look for estradiol- (**59**) and taxol(**60**)-binding motifs. For taxol, they did not find any peptide sequence with homology to tubulin but some with homology to bcl-2.[181] For 2-methoxyestradiol, after two rounds of selection, they found some peptides with homology to β-tubulin, COMT, and 17-β-hydroxysteroid dehydrogenase.[177] However, the results were ambiguous and unconvincing and, as a result, the technique has not been taken up. The fact that M13 phage particles must be secreted through the bacterial membrane acts as a barrier to some clones that can potentially bias selections. These problems have led to the search for alternate display systems that are lytic and facilitate C-terminal display to overcome the stop codon problem.

The T7-phage-display system invented by Rosenberg *et al.*[182] and commercialized by Novagen avoids the stop codon problem by C-terminal display on the cp10 protein. T7 phages are also lytic, so display and reproduction are not dependent on secretability through the bacterial membrane. The size of the alien protein can be quite large with functional enzymes of 1200 amino acids having been successfully displayed. The number of copies per phage particle is also adjustable from 0.1 to 15 out of 415 copies of cp10 in the phage capsid. These phages also have a rapid life cycle and are resistant to severe conditions such as 1% SDS, $4\,\mathrm{mol\,l^{-1}}$ urea, or $2\,\mathrm{mol\,l^{-1}}$ guanidine-HCl making them perfect for application in a chemistry laboratory. Other bacteriophages have also been used for display systems, including lytic lambda phage[183] and T4 phage.[184]

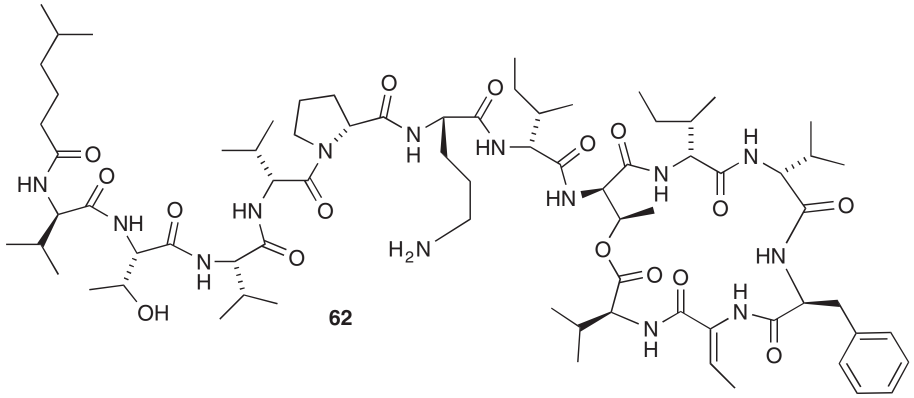


61

The first example and clear indication of the potential of phage display to isolate protein-binding partners for small molecules was demonstrated by Austin, who used the immunosuppressant drug FK506 (**18**) to isolate its known receptor, FKBP, from a T7-phage-display library.[185] Initially, FK506 was biotinylated at the allyl group through hydroboration of the double bond. This site was chosen because it was already well known that the allyl group was exposed to solvent in the FK506–FKBP crystal structure so that derivatization should not interfere with binding to the known protein target.[186] The biotinylated analog (**61**) was immobilized onto monomeric avidin–agarose to produce an affinity column. Lysate from T7 phages displaying a human brain cDNA library was first passed through columns containing underivatized avidin–agarose and a biotin-derivatized avidin–agarose to remove background binding to the column, agarose, avidin, and biotin. The pretreated lysate was then passed through the FK506-derivatized affinity column, and the column was washed thoroughly with buffer. Phage particles retained by the column were eluted with free biotin and transfected into fresh *E. coli* (BLT5615). After lysis, the amplified phages were precipitated and washed to remove biotin from the previous elution step. Finally, the phages were resuspended in buffer and absorbed onto a fresh FK506-derivatized affinity column for the next round of selection. Following each round of selection (biopanning), random phage plaques were picked and their DNA amplified by PCR. After the second round of selection, all clones picked gave rise to a DNA band of 450 bp, suggesting that the library had converged. DNA sequencing of these bands revealed identical copies of a full-length, in-frame gene coding for human FKBP. Therefore, it appeared that after only two rounds of selection with the FK506-affinity resin, phage particles displaying FKBP on their surface had been amplified selectively to become the dominant members of the library. However, it was later discovered that the FKBP clone isolated was actually a synthetic positive control FKBP clone, rather than from the brain cDNA library.[187] As mentioned above, T7 bacteriophages are small and hardy and can survive on surfaces for many years. Consequently, cross-contamination is a common problem associated with all phage-display systems. However, the risk of contamination can be reduced by using disposable plasticware, minimizing the formation of aerosols, and disinfecting exposed surfaces with bleach or UV light. Austin repeated the experiment, performing seven rounds of selection and taking particular care to avoid contamination. Sixteen random clones were picked after the sixth round of selection and their DNA inserts were amplified by PCR. Five of these clones were found to be identical and contained the entire coding sequence of FKBP1a. This number increased to 11 out of 16 clones in round 7. The results of this experiment show that, contrary to expectations, it is possible to isolate the receptor for a small epitope such as a natural product using a cDNA library displayed on T7 bacteriophages, thereby providing an important proof of concept.

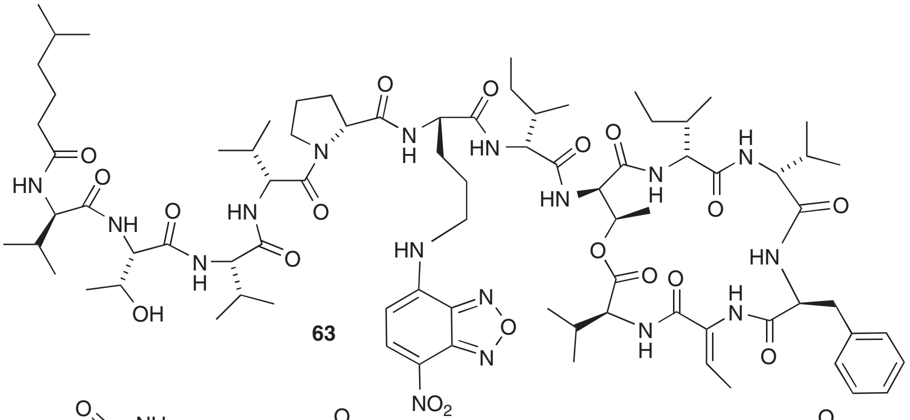
Soon after, Belasco reported the application of T7-phage display for the isolation of RNA-binding proteins.[188] Using hairpin II of U1 small nuclear RNA (U1hpII) or the 3′ stem loop of histone mRNA as bait, they were able to selectively amplify T7 phage that display either the spliceosomal protein U1A or the histone stem loop-binding protein from a lung cDNA library containing more than 10^7 independent clones. While this is not relevant to natural products and the protein targets were already known, it did show that it was possible to rapidly isolate protein-binding partners that have K_D values in the micromolar range and that there is no need to elute phage from an affinity matrix as the matrix itself can be used to infect fresh *E. coli.*

The Austin group went on to use phage display to isolate the cellular receptors for AP1497 (**57**) – a synthetic analog of FK506.[189] Initially, biotinylated AP1497 was immobilized on a streptavidin-coated polystyrene microtitre plate, and the resulting affinity support was used to probe a human brain T7-phage cDNA library. After performing three rounds of selection, 96 random plaques were selected and were tested individually for their ability to bind to an AP1497-derivatized plate. The clones displaying the highest level of affinity were submitted for DNA sequencing, revealing that 22 of the top 30 clones contained either an FKBP1a, FKBP1b, or an FKBP2 gene, all of which are known receptors for FK506. This experiment confirmed that the level of derivatization attainable on the surface of a microtiter plate well is sufficient to isolate phages displaying the target protein from the plethora of other phages in a cDNA library. These experiments also proved that not only the most avid binder but multiple protein targets could be isolated using phage display.

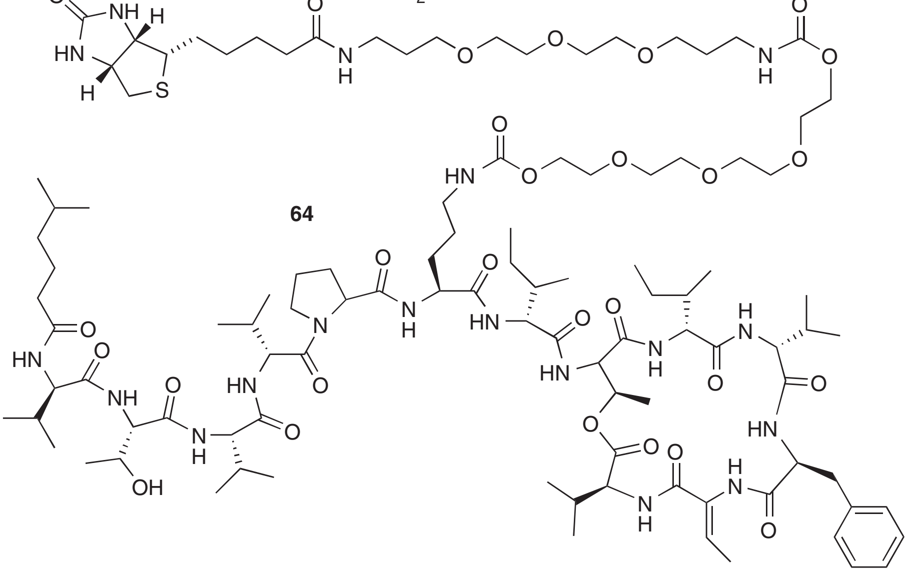


62

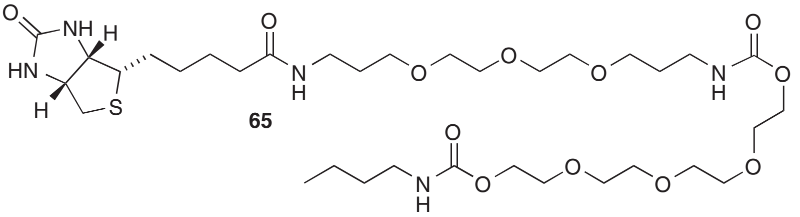
Kahalalide F (**62**) is currently in phase II clinical trials as an anticancer drug against a range of difficult-to-treat solid tumors and has relatively low toxicity to nontumor cells[190,191] and low toxicity in humans.[192,193] Kahalalide F was originally isolated by Hamann and Scheuer from the sacoglossan marine mollusk, *Elysia rufescens*, and subsequently from the sacoglossan's food source, the green alga *Bryopsis* sp.[194,195] The transmission of natural products down the food chain is a common feature of marine natural products, especially in nudibranchs and related sea slugs.[196] Kahalalide F appears to disrupt lysosomes, with treated cells swelling dramatically, forming large vacuoles before lysing. The natural product induces sub-G_1 cell-cycle arrest and cytotoxicity independent of MDR, HER2, p53, and blc-2 signaling.[197] Piggott and Karuso[198] constructed two kahalalide F probes, one tagged with a fluorophore (**63**) and the other with biotin (**64**). The former was used for fluorescence microscopy to show the localization of kahalalide F in specific features of the cytoplasm and both probes were shown to retain their cytotoxicity toward cancer cell lines, confirming that conjugation to the ornithine side chain did not affect the biological activity of the probe. In addition, they constructed a control probe that contained the biotin, linker, and ornithine side chain (**65**). This was used to preclear three human T7-phage-display libraries of nonspecific-binding phages and those that bind specifically to plastic, avidin, biotin of the linker. The precleared libraries were incubated separately in avidin-coated microtiter plates, derivatized with biotinylated kahalalide F. The nonbinding phages were aspirated and the microtiter well briefly washed with cold PBS buffer (containing 0.01% Tween 20). The binding phages were eluted with 1% SDS and used to infect log-phase *E. coli* to generate the first sublibrary, which was again precleared with control probe and then transferred to a well containing biotinylated kahalalide F to absorb specific binders. This process was repeated and after nine rounds of selection, random clones were picked and the DNA inserts sequenced. It was revealed that the same protein (human ribosomal protein S25; RPS25) was the dominant clone in all three libraries and that at least five different versions of this gene were isolated. Surface plasmon resonance (SPR) was used to show that phage displaying RPS25 specifically bond to kahalalide F. Based on these results, the authors postulated that RPS25 was involved in a hitherto unidentified signaling pathway that triggers a series of events resulting in cell lysis and oncosis.[198]



63

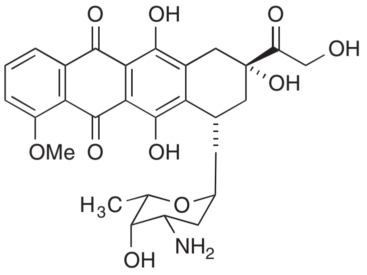


64

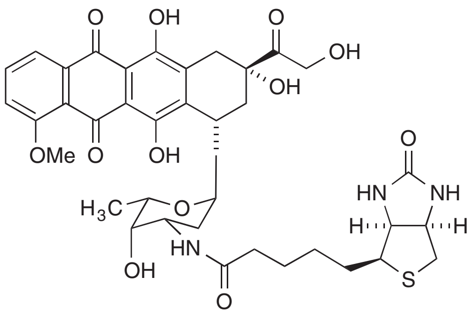


65

Despite its well-documented tubulin-stabilizing effect, taxol (**60**) has many side effects that cannot be explained by its binding to tubulin. In a study by Austin, they found that nanomolar concentrations of taxol, mimicking that found in cancer patients, induced cytosolic Ca^{2+} oscillations in a human neuronal cell line.[199] These oscillations were independent of extracellular and mitochondrial calcium but dependent on intact signaling through the phosphoinositide signaling pathway. Phage display using a human brain cDNA library and a C7-biotinylated taxol (structure not disclosed) identified a taxol-binding protein, neuronal Ca^{2+} sensor 1 (NCS-1). Free taxol increased binding of NCS-1 to the inositol 1,4,5-trisphosphate receptor and short hairpin RNA-mediated knockdown of NCS-1 in the same cell line abrogated the response to taxol but not to other agonists stimulating the phosphoinositide signaling pathway. The group also showed that other biotinylated natural products did not bind to NCS-1. These data showed that phage display can be used to identify off-target interactions that may help explain side reactions of drugs. The identification of off-targets is also important in the design of selective inhibitors.

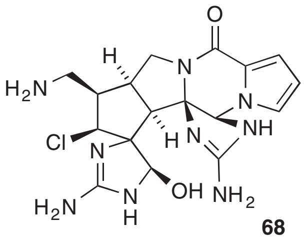


66

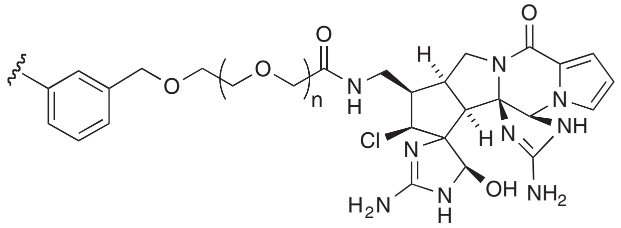


67

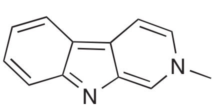
Doxorubicin (**66**), also known as adriamycin, is a cytotoxic anthracycline antibiotic, first isolated in 1963 from cultures of the actinomycete *Streptomyces peucetius* var. *caesius*.[200] Doxorubicin is an antineoplastic agent commonly used for the treatment of a wide variety of tumors, although it often produces severe side effects. The primary mode of action of doxorubicin is thought to be intercalation into cellular DNA, which results in the inhibition of DNA and RNA synthesis.[201] However, doxorubicin has also been shown to inhibit the enzymes RNA polymerase[202] and DNA topoisomerase II,[203] undoubtedly adding to the cytotoxicity of the drug. Finally, doxorubicin reacts with NADPH-cytochrome P-450 reductase to a semiquinone radical that reacts with oxygen to produce highly active superoxides, hydroxy radicals, and hydrogen peroxide. In addition, it is possible that doxorubicin interacts with other cellular components that have yet to be identified. Yu and coworkers used phage display to identify proteins that are capable of binding to doxorubicin.[204] Initially, a linkerless biotinylated analog of doxorubicin (**67**) was immobilized on a streptavidin-coated polystyrene microtiter plate. One well of the plate was then incubated with lysate from T7 phages displaying a human liver cDNA library. After washing the well thoroughly, bound phages were eluted with 1% SDS and transfected into *E. coli* to produce a sublibrary. This procedure was repeated until a total of four rounds of selection had been performed. Finally, 20 individual phage plaques were picked from the final round of selection and their DNAs were amplified by PCR and sequenced to determine the identity of the gene inserts and hence the proteins being displayed on the surface of the phages. The majority (90%) of DNA sequences obtained from the final round of selection consisted of gene fragments coding for peptides of less than 20 amino acids. However, two clones were found to contain a gene fragment encoding a polypeptide identical to the C-terminal region of human nucleolar phosphoprotein 140 (hNopp140). hNopp140 is a highly phosphorylated protein involved in the biogenesis of the nucleolus.[205] The primary sequence of the protein consists of long runs of serine residues and a very high content of charged amino acids, including a 10-fold repeat of positively and negatively charged regions.[206,207] There are 82 potential phosphorylation sites on the protein, but the degree of phosphorylation changes during the cell cycle.[208] The authors cloned the phage-displayed gene fragment into an *E. coli* expression vector to examine the interactions between this protein fragment and doxorubicin using fluorescence spectroscopy and SPR. It was found that addition of nonphosphorylated recombinant hNopp140 to doxorubicin increased the fluorescence intensity of the drug in a dose-dependent manner, suggesting that the two species do indeed interact. However, no increase in fluorescence intensity occurred when a phosphorylated form of hNopp140 was used. A similar trend was observed when phosphorylated and nonphosphorylated forms of hNopp140 were applied to doxorubicin immobilized onto the surface of a SPR chip.



68



69

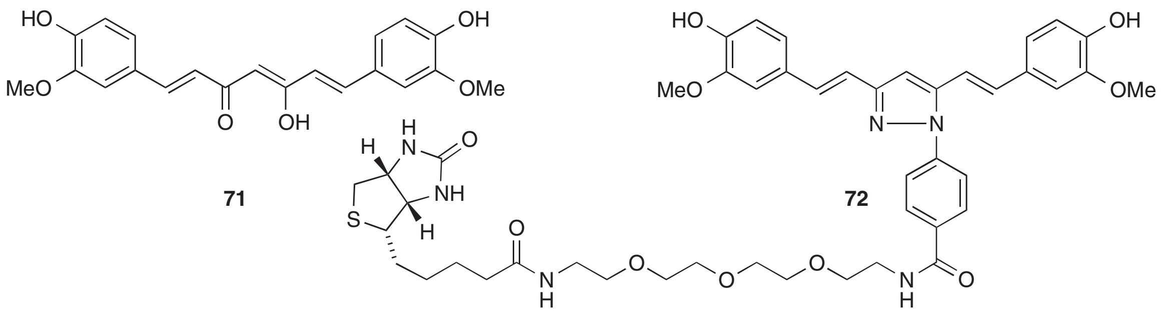


70

Interestingly, Piggott and Karuso[209] also isolated hNopp140 while screening a T7-phage-display library for binding proteins to the marine natural product, palau'amine (**68**). Palau'amine was covalently immobilized onto polystyrene microtiter plates (**69**). One well of the plate was then incubated with lysate from T7 phages displaying a human liver cDNA library and five rounds of selection were performed in an analogous fashion to that used by Yu and coworkers for doxorubicin.[204] After the final round, random plaques were picked and their cDNA inserts amplified by PCR and sequenced. The majority of the clones were unique but contained DNA inserts that were either not in the correct reading frame, were in backwards, outside the coding region of the gene from which they originated or expressed only short peptide sequences. However, 6% of the clones contained DNA inserts that were in the correct reading frame and within the coding region of their corresponding gene. One of these was identical to the hNopp140 clone isolated by Yu with doxorubicin, expressing amino acids 388–603 of the 699 residue protein. The two other clones expressed human nucleolar phosphoprotein p130 (hNopp130), which has a DNA sequence almost identical to that of hNopp140. The high percentage of charged amino acids and long runs of serine residues make hNopp140 a promiscuous nonspecific binder, which can become the dominant member of a phage-display library very quickly in the absence of a competing real protein–ligand interaction. It is thus likely that the appearance of hNopp140 in a phage-display experiment is indicative that the selection has been unsuccessful. In addition, if after several rounds of biopanning, the library has not increased dramatically in titer and has not converged onto one or two full-length, in-frame members, it would seem that the phage-display selection has not worked.

Gearhart *et al.*[210] used a simple phage-display approach in an attempt to identify proteins that bind to 2-methylnorharman (**70**). Nitrocellulose discs were soaked in a solution of 2-methylnorharman hydrochloride and then incubated with lysate from T7 phages displaying a human brain cDNA library. The discs were then washed thoroughly and the retained phage particles were transfected into *E. coli* directly from the disc to produce a sublibrary. This procedure was repeated until a total of six rounds of biopanning had been performed. Finally, 43 individual phage plaques were picked from the final round of biopanning and their DNA was amplified by PCR, sequenced, and searched against various online databases. Nearly half of the DNA sequences obtained from the final round of biopanning matched sequences coding for fragments of human proteins with unknown functions or identities. The remaining sequences encoded fragments of six known human proteins: human nucleolar phosphoprotein p130, dorfin, α-tubulin, paraoxonase 2, fatty acid-binding protein 5, and platelet-activating factor acetylhydrolase 1B1. The authors noted that all six proteins isolated contained high percentages of glutamic and aspartic acids. While it is possible that 2-methylnorharman shows a preference for highly charged proteins, it is also possible that these charged protein fragments are interacting nonspecifically with the nitrocellulose discs. The appearance of hNopp130 in the experiment is particularly noteworthy. Unfortunately, the authors did not perform any validation to ascertain whether the particular phages isolated were indeed binding to 2-methylnorharman and not just to the nitrocellulose.

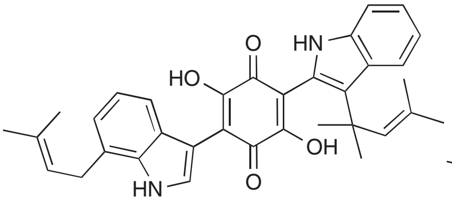
The noncovalent adsorption of 2-methylnorharman directly onto nitrocellulose discs is unlikely to produce a useful affinity support for phage-display experiments. This is because the capsid of a T7-phage particle has a diameter of 60 nm and is comprised of 415 individual proteins.[211] Therefore, the absence of a long linker molecule between the solid phase and the target molecule may make it sterically impossible for a protein attached to a T7-phage particle to bind. Even if a phage-displayed protein succeeded in binding to an adsorbed ligand, the complex could then desorb from the surface and be lost during the next wash. As there are few phage particles in the initial library displaying any cellular receptor, loss of these in an early round of selection would guarantee a failed experiment. While it is possible for phages to infect *E. coli* when bound to a solid support,[212] it is more efficient to release the phage particles into solution. This is usually achieved by treating the solid phase with a detergent, such as SDS, to disrupt protein–ligand interactions. However, the selection process can be greatly enhanced by only eluting the phage particles that are specifically bound to the target molecule, leaving behind the nonspecific binders. This can be achieved by treating the solid phase with free natural product to outcompete the immobilized natural product, or by employing a cleavable linker.[213]



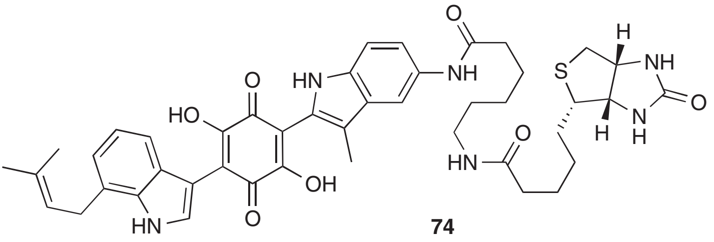
71

72

Kwon and coworkers[214] used reverse chemical proteomics to identify the cellular receptor for HBC – a synthetic curcumin (71) derivative that inhibits the proliferation of several tumor cell lines. A biotinylated analog of HBC (72) was immobilized on a streptavidin-coated polystyrene microtiter plate and the resulting affinity support was used to probe a mixture of five different T7-phage-displayed human cDNA libraries. After four rounds of selection had been performed, 17 random plaques were selected and their DNA inserts were sequenced, revealing that 12 of the clones displayed the C-terminal (aa 86–149) of calmodulin. The authors used SPR to show that the HBC–calmodulin interaction exhibited a clear dose response, with the dissociation constant for the interaction calculated to be 8 μmol l^{-1}. A flexible docking study also indicated that HBC is compatible with the binding cavity occupied by the known calmodulin inhibitor, W7. However, given that only a fragment of calmodulin was isolated, and that this fragment is 38% charged amino acid residues the result is not convincing. In addition, Baek *et al.*[215] used a yeast drug-induced haploinsufficiency screen to identify eight target genes that, when deleted, lead to HBC sensitivity but none of these was identified as calmodulin. Haploinsufficiency refers to the fact that lowering the gene dosage of a drug target increases the susceptibility of an organism to that drug (see Section 9.14.3.4.2).



73



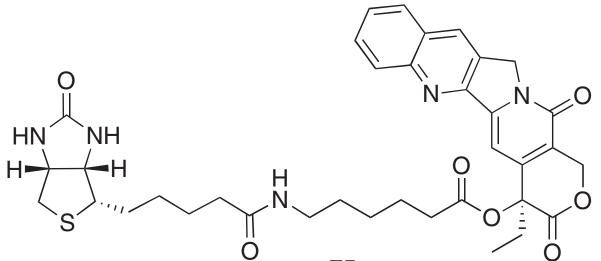
74

Demethylasterriquinone (73) is a natural product from *Aspergillus terreus*[216,217] that has been found to be a mimic for insulin. Identification of its cellular targets (other than the insulin receptor) would be useful in developing drugs that act only on the insulin receptor and not on other off-target proteins. This knowledge would allow the design and synthesis of orally active small molecules that could eventually replace injected insulin in diabetes treatment. Through a process of methyl scanning, Pirrung determined the site that least affected the biological activity of asterriquinone and derivatized this with a short-linked biotin (74). Biopanning a phage-display library against biotinylated demethyasterriquine led to the isolation of glyceraldehyde 3-phosphate dehydrogenase (GAPDH).[218] These results are supported by the findings of Min *et al.*,[219] who used a chemical genetics approach with *Caenorhabditis elegans* that GAPDH has an important role in insulin signaling. Binding of small molecules, such as demethylasterriquinone B1, to GAPDH could therefore disrupt phosphatase(s) acting upon phosphatidylinositol lipids and thereby potentiate insulin signaling through the PI3kinase pathway.

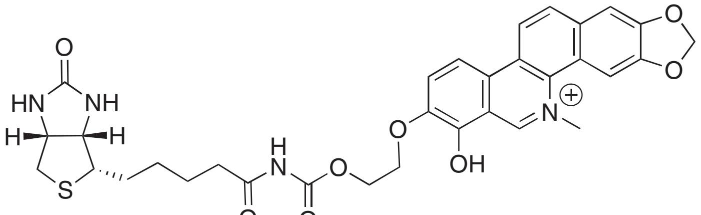
Despite the power of phage display as exemplified by these examples, technologies that require an *in vivo* step, such as phage, yeast, or bacterial displays have limitations. In phage display, libraries must be transformed into *E. coli*, limiting the number of possible independent sequences to the number of actual cells that can be grown in a convenient volume (10^9–10^{10}), of which two-thirds will be out of frame, leading to proteins or peptides that are not encoded by the attached gene and up to half are in backwards, if cloning is nondirectional. The total number of sequences represented can be further decreased by proteolysis of unfolded molecules, poor expression in the bacterial host due to differences in codon usage, failure in correct processing of the phage capsule, failure of the alien gene product to fold correctly without the mammalian endoplasmic reticulum, and toxicity of the expressed protein to the host. Further complications arise from the inability to display entire membrane proteins, oligomeric proteins, the total lack of posttranslational modifications, which means critical saccharides or phosphates are missing or N- or C-terminal domains are not removed, possibly masking activity of the gene product. However, the Austin group at Yale has proved that membrane protein domains can be isolated using phage display. They found the α-domain of F1-ATP synthase, which is a mitochondrial membrane ATP protein, was the binding partner for a synthetic ATP mimic.[220] Phage display may also select for higher infectivity rates or reduced host toxicity over increased affinity.

Despite these limitations, phage display is an important tool that has a proven track record in the isolation of protein targets for natural products. The clear advantages are that it is relatively unbiased, extremely fast and iterative in nature, ensuring that the most avid (therefore biologically most important) rather than the most common binding partner is isolated.

Optimization of the technique is required before it is globally applicable but recent advances are encouraging. For example, by employing a covalent bond between the natural product and affinity matrix in combination with preclearing of the phage lysate against an underivatized affinity matrix and direct infection of *E. coli* with the affinity matrix allowed Tang and coworkers to isolate CypA and CypB genes from a human brain cDNA library after six rounds of selection against cyclosporine A (17).[221]

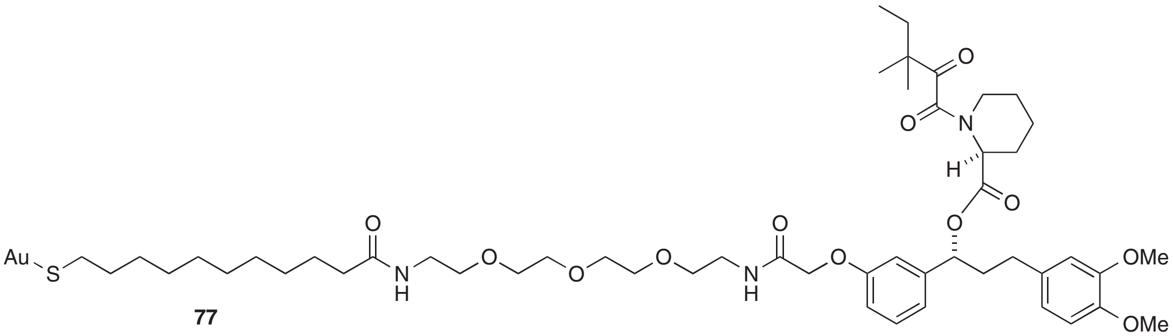


75



76

Sugawara and coworkers have recently published a series of papers exploring new ways to use T7-phage display to discover the protein-binding partners for natural products and drugs. They started in 2005 using a cDNA library constructed from human leukocytes (white blood cells) from mRNA to look for peptide sequences that bind to biotinylated camptothecin (75). After four rounds of selection, they picked 40 random plaques and sequenced each to determine what peptide was displayed on the surface of each phage; 25% of the clones displayed NSSQSARR. This peptide was synthesized, attached to a SPR chip and shown to bind to underivatized camptothecin.[222,223] This group has also recently proposed a high-throughput version of the technique by using an avidin–agarose-affinity column. A random peptide (12-mer) phage library was constructed and incubated with biotinylated NK109 (76) and then pumped through the column, washed with buffer, and eluted with excess biotin. The eluate was collected in 96-well PCR plates and amplified using real-time PCR (SYBR green) to determine which fractions contained high concentrations of phage. These can then be sequenced or subject to another round of selection after amplification in *E. coli*.[224,225] The major advances is that there is no need to optimize the elution step and that the method can be used in high-throughput mode. Another method to avoid the need to determine wash and elution conditions and to avoid large amounts of background, noted earlier as a perennial problem with phage display, involved use of quartz crystal microbalance (QCM)-type biosensor. A QCM is a very sensitive and accurate mass measuring device that works by measuring changes in resonant vibrational frequency caused by a mass increase. Binding of biotinylated camptothecin (75) to an avidin-coated QCM device allowed Sugawara and coworkers to measure absolute binding of a phage library to the probe.[226] By watching the decrease in frequency, the researchers were able to stop when phage absorption was optimal and then washed the QCM chip and used the surface of the chip directly to infect *E. coli* to produce the next phage sublibrary. After just three rounds of selection, they were able to isolate the same peptide as previously identified as binding to camptothecin.[222] The method has recently been improved by replacing the avidin–biotin link with a self-assembled monolayer where the natural product can be covalently linked to the gold surface of the QCM through a thiol. This improved link allowed the recovery of FKBP1a from a cDNA phage library after only one round of selection, obviating the need to washing, elution, and amplification, which, it was noted, biases selection toward fast-growing phage. Applying this technique to AP1497, Sugawara was able to derivatize the FK506 analog AP1497 (77) with a linker terminating with a thiol and bind this to the QCM chip.[227] A phage library was incubated with the chip for 10 min, briefly washed and the surface used directly to infect *E. coli*. Sixteen random plaques were sequenced and 10 were found to code for FKBP1a and another three expressed peptides that resembled the FK506-binding domain of FKBP1a. The system has also been used to find the protein-binding partners for methotrexate and trimanoside.[228,229]



77

9.14.3.2.2 Retroviral display

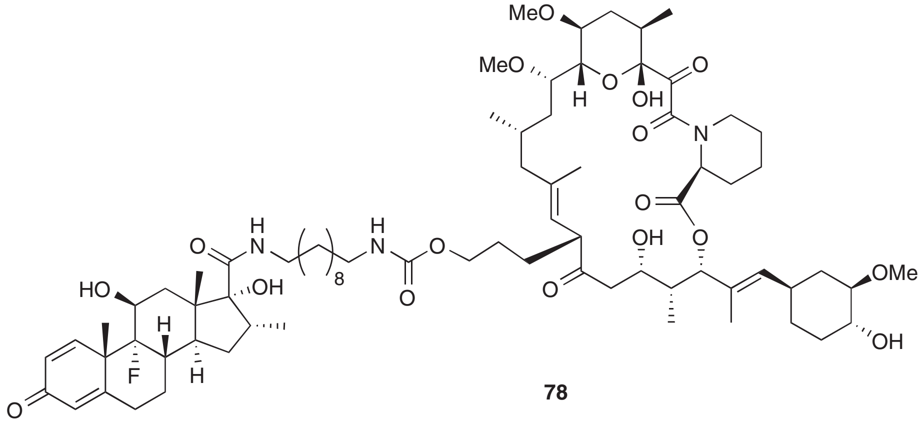
Retroviruses are interesting viral vectors that incorporate their RNA (through the cDNA) into the host's DNA. The viral particles are composed of a core of viral RNA, surrounded by replication enzymes, host membrane, and an outer viral glycoprotein coat.[230] There are seven genera and all could conceivably be used for the display of alien proteins. Retroviral display has been recently reviewed.[231]

There are several options for the display of alien proteins using this technique. Proteins can be displayed on the envelope spike protein (Env), in which case selective pressure is placed on the population of viral particles before reinfection into fresh mammalian cells, much like phage display. So far, the potential of retroviral display for the generation and screening of eukaryotic expression libraries has only been demonstrated for small peptides. For example, avian leukemia virus (ALV) has been used to display a random peptide library (8-mer) and used to select specific sequences that bind to monoclonal antibodies that recognize short peptides (FLAG and HA tag antibodies). A 100× enrichment of binding sequences was observed per round of selection.[232]

Alternatively, viral vectors (replication deficient) can be used to incorporate the alien RNA into the host's genome. Thus, the vector is used to generate a mammalian cell surface display library. Clearly, the advantage of retroviral display in mammalian cells is that the library would have access to a eukaryote environment for proper protein processing and posttranslational modifications. On the down side, the system is very new and little is known about the factors influencing expression and replication. There are no examples of using these techniques for the discovery of protein–protein-binding interactions or for small protein molecules. However, as Buchholz *et al.*[231] recently stated: "It is well conceivable that further sophisticated screening strategies based on retroviral display libraries will be established in the near future."

9.14.3.2.3 Yeast three-hybrid screening

The yeast two-hybrid system was originally developed as a genetic assay by Fields and Song[233] in the 1980s. It arises from the observation that many transcription factors (proteins that regulate the transcription of DNA to mRNA) can be broken down into two components that mediate DNA binding and transcription activation. In the yeast two-hybrid experiment, the bait is constructed by fusing the DNA-binding domain of a reporter gene with a protein of interest. A prey protein is then fused to the activation domain of the same transcription factor and only when the prey and bait come together the transcription factor is formed in a functional way. When this happens the reporter gene is transcribed and the mRNA translated to a protein that can be detected in some way. For example, the LacZ protein can be measured colorimetrically. To generate protein–protein interaction maps, one tags the protein of interest (at the gene level) with the DNA-binding domain of a reporter gene as above. Next, the entire transcriptome (cDNA) of an organism is tagged with the activation domain of the reporter gene and transfected into yeast. Each yeast cell receives only one new gene and this forms the library. The yeast strain expressing the DNA-binding domain of the reporter gene is now mated with the library and plated out. Each colony that expresses the reporter gene is then PCR amplified to reveal what gene it was carrying and this then allows the researcher to say that this isolated prey protein interacts with the bait. The process can be repeated for every prey protein discovered, which is now the bait to discover its interacting partners.[234] This is essentially how the first yeast and *C. elegans* interaction map was developed.[49] The disadvantages of this system is that there are false positives from protein that contain their own transcription factors. They are also time consuming and expensive requiring a high degree of manipulation. False negatives can also be observed if the fusion proteins are toxic, poorly folded, or easily degraded. In addition, all interactions must take place in the nucleus, which may not be possible for proteins that contain a specific location tag.



78

The yeast three-hybrid system is an extension of the two-hybrid system, in which a third component (RNA or a small molecule) is required to reconstitute the transcription factor. This was pioneered by Licitra and Liu.[235] The authors synthesized a hybrid natural product; dexamethasone linked to FK506 (78). The drug-binding domain of the glucocorticoid receptor (the target for dexamethasone) was fused to LexA, transcription factor, DNA-binding domain. The LaxA activation domain was fused with the cDNA library from Jurkat cells. Only when an FK506-binding protein is fused with the activation domain of LaxA is the gene transcribed and the LaxA protein synthesized. Positive clones were selected, sequenced, and shown to contain FKBP1a.

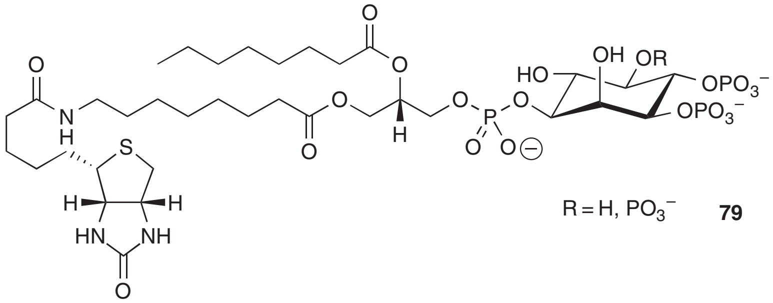
Even though this field has been heavily reviewed, there are very few examples of using the yeast three-hybrid system for uncovering the binding partners for natural products where the binding protein was not already known. It has been shown that an interaction (K_D) of $<50\,\mathrm{nmol\,l^{-1}}$ is required to achieve detectable activation.[236] This lack of sensitivity is one reason why there have been so few reports using this technique with natural products. In addition, to the same disadvantages noted for the two-hybrid system, the small-molecule probe must also be able to permeate live cells and get into the nucleus without killing the cell. This would be problematic if the natural product displayed high cytotoxicity or antifungal activity.

In spite of these difficulties, the technique has successfully been used to isolate known and new kinases that bind to purvalanol B, a potent kinase inhibitor and the authors suggest that the improved methodology is generally useful for the discovery of new drug candidates but, to date, there have been few further examples.[237]

The ability to detect and quantify fluorescent tags with commercial flow cytometers has facilitated the use of cell-surface display methods.

9.14.3.2.4 Yeast display

Yeast (*Saccharomyces cerevisiae*), a eukaryotic single-celled organism, has also been used for surface display, in much the same way as bacterial display (see Section 9.14.3.2.6). Yeast surface display was first demonstrated in 1993 as a method of immobilizing pathogen-derived proteins for vaccine development.[238] Schreuder *et al.* used a cell-wall-anchored protein α-agglutinin as the carrier for the alien protein. Many other proteins have been examined for their utility in surface display but the commonest is the Aga2p–Aga1p system.[239,240] In this system, the alien protein can be fused to either the N- or C-terminus of the Aga2p protein. Two disulfide bonds form between Aga2p and Aga1p, a GPI-anchored protein. Coexpression of the Aga2p-fusion and Aga1p leads to display of the alien protein on the surface of the yeast. The alien protein is linked by an amide bond to Aga2p, which in turn is linked through two disulfides to Aga1p that is covalently linked to the cell wall through glucan linkages and lodged, noncovalently in the cell membrane by the fatty acid tails of the GPI anchor. This system has been successfully used to clone cDNA-based domain libraries into yeast.[241,242]



R = H, PO_3^- **79**

Although no natural product-binding proteins have been identified with this technique, it has been used to identify the binding partners for a small-molecule phosphatidylinositide. Bin Liu first constructed a yeast surface-displayed human cDNA libraries containing 2.4–8.9 × 10^6 primary clones derived from testes, brain, fetal liver, and breast tumor tissue (Invitrogen) by digestion with *Eco*RI; ligation into *Eco*RI-digested and dephosphorylated pYD1 (and +1 and −1 frame shifted) yeast-display vectors.[242] As the library was composed of cDNA fragments, it is better described as a domain-display library. This was done to increase the representation of in-frame open-reading frames (ORFs) for larger proteins or membrane proteins but also because the display system used was a C-terminal system that cannot be used to display a normal cDNA library. Two biotinylated homologs (**79**) were used as bait to capture yeast cells displaying a specific-binding partner, after which the cells were incubated with (fluorescent) phycoerythrin-labeled streptavidin and subjected to three rounds of selection using fluorescence-activated cell sorting (FACS). About 8% of the third sublibrary was highly fluorescent and collected for analysis. Recovery of the plasmid from 55 clones yielded 11 unique cDNA inserts from seven different genes. Nine of the 11 inserts contained a full-length pleckstrin homology (PH) domain that is known to code for phosphatidylinositide binding.[242] Liu's results suggest that large libraries of functional human proteins (or pieces of proteins) can be efficiently displayed on the yeast surface and can theoretically be used to identify protein fragments with affinity for any soluble molecule that can be fluorescently detected. Using modern FACS, up to 10^7 library members can be screened in an hour.[242] For natural products, this could be done in a traditional way through biotinylation of the natural product and then detecting the yeast that bind the natural product through fluorescently labeled streptavidin. Alternatively, one could attach an immunogenic tag such as a FLAG tag (octapeptide; DYKDDDDK) that can be recognized by commercially available fluorescent antibodies or the fluorophore can be bonded, or through a linker, to the natural product of interest. In each case, the FLAG peptide or fluorophore alone can be used as a control to remove yeast that bind to just the tag.

Yeast display of cDNA libraries for the discovery of protein–protein or protein–natural product interactions has a number of important advantages. These include the incorporation of posttranslational modifications, such as phosphorylation, disulfides, and glycosylations, though the glycosylation pattern is unlikely to be human-like. The protein folding and secretory machineries are also similar to those of mammalian cells, thus more suited to the functional display of mammalian proteins that require endoplasmic reticulum-specific posttranslational processing for efficient folding and activity than bacterial systems. Large numbers (10^4–10^5) of identical copies of the alien protein can be displayed on each yeast cell making selection and analysis easier than with phage or *in vitro* methods and opening the possibility of functional homodimers or homotrimers forming on the surface of the yeast. There also seems to be no size limitation for the fusion proteins, unlike phage, which are presently limited to about 100 kDa.[239] However, the single largest advantage is the ability to screen using flow cytometry. Yeast cells are large, tough and because of the large number of displayed protein can be heavily labeled with fluorescent reporters and thus easily sorted. They are also generally considered safe to work with and easy to grow, though they grow much more slowly than bacteria or phage.

A direct comparison between phage and yeast displays of the same HIV-1 immune scFv cDNA library and using the same selecting antigen (HIV-1 gp120) revealed that yeast display sampled the immune antibody repertoire considerably more fully than phage display, selecting all the scFv identified by phage display and twice as many novel antibodies.[243] In general, N-terminal display is used with the ag2p protein but C-terminal is possible as well, allowing the easy construction of cDNA libraries. Indeed it has been shown that anti-CD3e scFvs exhibited 1–2 orders of magnitude greater affinity for their target when C-terminal fusions were used compared to N-terminal.[244]

Disadvantages include the lower transformation rates (typically 10^5–10^7) compared to *in vitro* methods ($\sim 10^{10}$–10^{14}) and are 10× lower than for bacterial display, which mean libraries may not cover completely an entire genome. There are questions about processing limitations, as some proteins may inhibit secretion of the carrier protein yielding an expression bias. There are also questions about biased codon usage in yeast compared to humans in the case of human library construction,[245] which may result in an expression bias for codons rarely used in yeast and mutation of some leucine residues (CUG) to serine in the displayed proteins.

Yeast display has come a long way since proof of concept in 1993 to the successful expression of a human cDNA library in 2005 and the demonstration that small-molecule-binding proteins can be isolated in 2007. Considering the rapid progress over the last two years and the inherent advantages over *in vitro* and other *in vivo* methods, it is clear that yeast display will soon overcome the current limitations and possibly become the method of choice for isolation of natural product-binding proteins.

9.14.3.2.5 Bacterial display

Bacterial display was first shown to be possible by Charbit *et al.*,[246] who displayed polio virus epitopes on the surface of *E. coli* through the LamB protein. The system was developed as an alternative to phage display in an attempt to increase the size of displayed proteins, which was limited at the time, and retains some of the advantages of phage display such as high-transformation rates and ease of handling. Most of the fusion systems developed so far are insertional, meaning the alien protein is inserted into an exposed loop in the carrier protein. This is generally necessary because the host requires both the N- and C-terminal domains for proper function. This limits the size of insertions to <100 amino acids and makes display of real proteins and random libraries problematic because of stop codons, which would truncate the fusion protein. This problem would also be manifested in N-terminal display methods (e.g., autotransporters AIDA-I, IgA protease, and CPX).[247] The only two C-terminal display methods reported so far, which have been used for library screening, are with LppOmpA and invasin.[247] Ice nucleation protein (*Pseudomonas syringae*) might also be able to be used for C-terminal display libraries but it has been used only for enzyme screening.[248]

Microbial cell-surface display has many potential applications, including live vaccine development, peptide library screening, bioconversion using whole-cell biocatalyst and bioadsorption.[249] However, it has, to our knowledge, never been used to display a cDNA library. However, the use of cell display has a number of advantages over *in vitro* and phage-display technologies. Most importantly, cells can be sorted by high-throughput flow cytometers (FACS), which also allows ready tuning of the selection protocols to achieve the desired stringency (proportional to K_D) and selectivity (% yield). After selection, the binding characteristics of individual clones can be quickly determined (on-cell) using flow cytometry without the need for subcloning and protein purification. Working with *E. coli* is also advantageous because the system is easy to handle and well understood.

Affinity screening of bacterial display libraries can be carried out using either magnetic beads, FACS or a combination of both. If the target of interest can be biotinylated then bacterial cells expressing the binding target are multiply biotinylated which can be captured on a streptavidin-coated magnetic bead. This method was used to discover streptavidin-binding motifs.[250] However, magnetic beads can be envisaged to suffer from avidity interactions arising from multiple display of the library on the cell surface and streptavidin on the bead's surface. In most cases, FACS is preferred because the selection can be visualized and easily adjusted. In this method, the target can be directly labeled with a fluorophore or an epitope that binds to a fluorescent antibody or fluorescent-labeled streptavidin for example. This results in fluorescent cells that display library members that bind to the target and can be effectively enriched by FACS. Culturing of the sorted cells into a sublibrary can then be reselected and so on with increasing stringency to isolate the best-binding partners for the target. This method has been used for antibody epitope mapping, discovering protein-binding ligands (peptides), peptide substrates for enzymes, cell-binding peptides and antibody screening. If cDNA libraries can be effectively screened through C-terminal display then this would open the door to the screening of these libraries for natural product-binding proteins. This would form a complementary method to phage display that may be very useful due to the inherent advantages of cell display. However, the techniques required further basic research before they can compete with phage display. For example, Lunder *et al.* tested phage display and bacterial display in selecting streptavidin-binding peptides from two commercially available libraries. Under similar conditions, selection of phage-displayed peptides for streptavidin binding proved convincingly better.[251]

Limitations include some of those already mentioned for phage display, such as the lack of any posttranslational modifications, difficulty in displaying (nonbacterial)-membrane proteins and multiprotein complexes, and expression biases or mistranslation errors caused by mismatched human/bacterial codon usage.[252] Other unique limitations of bacterial cell-surface display include the size and folding efficiency of the alien protein as well as its disulfide content which affect its ability to be secreted and localized on the outside of the cell. Bacterial display will also inherently select for proteins that optimally fold and process in a prokaryote cell. Unfortunately, the plethora of host strains, conditions, and display carrier proteins makes comparison of expression levels difficult.[253] Expression of alien proteins on the surface of bacteria is also known to induce extracytoplasmic stress resulting in growth arrest and poor display performance. Coexpression of Skp, a chaperone protein involved with transport and insertion of outer membrane proteins, was demonstrated to be effective to restore cell physiology.[254]

An interesting recent example shows a proof of principle for the selection of affinity proteins from large combinatorial libraries displayed on the surface of the Gram-positive bacterium *Staphylococcus carnosus*. In contrast to Gram-negative bacteria, Gram-positive bacteria are more robust because of the cell wall and contain only one cell membrane. Surface proteins do not need to be membrane spanning, but are covalently linked to the peptidoglycan cell wall through their C-terminus, resulting in a high tolerance to recombinant fusions. An affibody library of 3×10^9 58-mers was screened for binding to TNF-α yielding three different high-affinity binders.[255] Thus, development of other bacterial display systems that circumvent some of the current disadvantages of Gram-negative bacteria could see this technique become useful in the isolation of natural product-binding proteins from expressed cDNA libraries, though it is noted that Gram-positive bacteria are known to secrete large amounts of protease. Considering the number of very recent improvements, the development of a functional cDNA expression system may not be that far into the future.

9.14.3.2.6 Human cell display

Mammalian cells have also been used to display single proteins and a random peptide library fused to the CCR5 chemokine receptor.[256] However, human cell display has not yet been used to display any sort of protein library let alone isolate natural product-binding receptors but there is one recent example that indicated that this is possible.

Pastan and coworkers show that human embryonic kidney 293T cells (HEK 293T) that are widely used for transient protein expression can be used for cell surface display of single-chain Fv (scFv) antibodies.[257,258] In this example, Pastan fused the anti-CD22 scFv antibody to the transmembrane domain of the platelet-derived growth factor receptor (PDGFR). They produced two versions, one wild type and a mutant with 2× higher affinity for CD22. With a 400× excess of the wild-type antibody, the mixture was incubated with biotinylated CD22-Fc, washed and then incubated with phycoerythrin-labeled streptavidin, washed again, and sorted by FACS. After just one round of selection, a 240× enrichment of the mutant antibody displaying cells was observed. This proof-of-principle clearly demonstrates that human cell display has potential but it is yet to be seen if it can be used to display cDNA libraries. To do this a C-terminal display system needs to be found. Certainly, there are no concerns over protein folding, expression, codon usage, or posttranslational modifications as with other display technologies, but obvious limitations are the size of the libraries that can be generated with human cells, which is limited to the number that can be conveniently cultured and the length of time required for each round of selection. Whereas T7-phage display can produce libraries of 10^9 or greater and be run to produce two rounds of selection per day, mammalian cells typically take at least a week for one round of selection. Growing more than 10^7 cells per round can also become difficult. Finally, cell sorting at low speed (required for delicate human cells) can take a whole day to sort enough cells for analysis.

9.14.3.3 *In Vitro* Display Technologies

These techniques couple genotype and phenotype by creating a physical link between mRNA (genome) and the encoded protein (phenome). While phage display, cell display, and the yeast two-/three-hybrid systems are limited by the involvement of living cells for library generation and biopanning, this is not the case with *in vitro* methods where the amplification step is taken over by a chemical process; the polymerase chain reaction (PCR or RT-PCR). An associated advantage is that *in vitro* methods are not limited by the number of cells that can be grown in a given volume so very large libraries ($>10^{14}$) can be constructed because they do not require transformation in a living cell. Although, compared to the size of sequence space, for example, there are 10^{104} possible 80-mer peptides (which is more atoms than exist in the known universe), all man-made libraries have to be considered small. An advantage of cDNA libraries is that they are always small (10^4–10^5 members) but are selected for function.

9.14.3.3.1 Ribosome display

The first *in vitro* library technique for affinity selection called ‘Selective Evolution of Ligands by EXponential enrichment’ (SELEX) was introduced in 1990 by Tuerk and Gold.[259] It starts with the synthesis of a pool of DNA or RNA, containing a region of randomized nucleotides (~10–100 bases) flanked by a conserved region (priming sites for the PCR). For RNA, the conserved region usually contains a T7 RNA polymerase promoter for transcription of the DNA library to RNA. Immobilization of a ligand (protein or small molecule) allows the selection of DNA/RNA based on the ability to bind ligand molecules with high specificity and affinity. Amplification of the selected DNA (or RNA) yields a sublibrary that can be used for a second round of selection and so on until just one member (or family) of the original library remains – these are called aptamers. SELEX can also be easily performed to select for ssDNA[260] or double-stranded DNA (dsDNA).[261] In the original paper, ribosome display for *in vitro* selection of peptides (as opposed to oligonucleotides) was also proposed. In SELEX, the idea is to find oligonucleotides that bind strongly to a given ligand, thus the phenotype and genotype are on the same molecule. The technical problem with ribosome display was how to provide a physical link between the genotype (mRNA) and the phenotype (translated peptide or protein).

The first success was demonstration in 1994, with the report of a large, diverse library of decapeptides displayed and selected while associated with *E. coli* S30 polysomes and RNA.[262] The key to Dower’s success was the application of natural product antibiotics that were known to interfere with protein synthesis by stabilizing the ribosome–mRNA–protein complex. Thus, rifampicin and chloramphenicol (for prokaryotic system) or cycloheximide (for eukaryotic system) were used.[262,263] Because these antibiotics halt the translation at random locations, the ensuing libraries were composed of mostly truncated peptides and thus not really suitable for the generation of cDNA libraries. Later, removal of the stop codon from mRNA was used to stall the translation at the end of the mRNA.[264,265] Several improvements have been made more recently to stabilize the mRNA–ribosome–protein complex, to prevent nuclease (RNAase) degradation, eliminate production of truncated proteins, and improve displaying efficiency.[266–270] Further improvements to include an *in situ* reverse transcription (RT-PCR) in combination with single-primer PCR technologies have been recently published.[271]

Ribosome display (**Figure 2**) has been used to select protein-binding partners but, to our knowledge, never been used to find the unknown protein-binding partner for a small molecule.[268,272–275] However, a recent paper by Plückthun clearly indicates that this is possible. In this report, the authors show, for the first time, that it is possible to use ribosome display to select for catalytic activity based on catalytic turnover (directed evolution). In their experiment, they displayed variants of RTEM-β-lactamase and used biotinylated ampicillin sulfone (**80**) to select for catalytically active variants (**Figure 3**).[276]

Although the early versions of ribosome display were unstable and the libraries of low quality, recent improvements have produced much more robust methodology (**Figure 4**). However, from a chemist's perspective further improvements are required to make it more user friendly.[176]

The advantage of ribosome display is that it is a relatively mature technology that has been improved tremendously over the past decade. Like all *in vitro* display methods, very large libraries can be constructed but

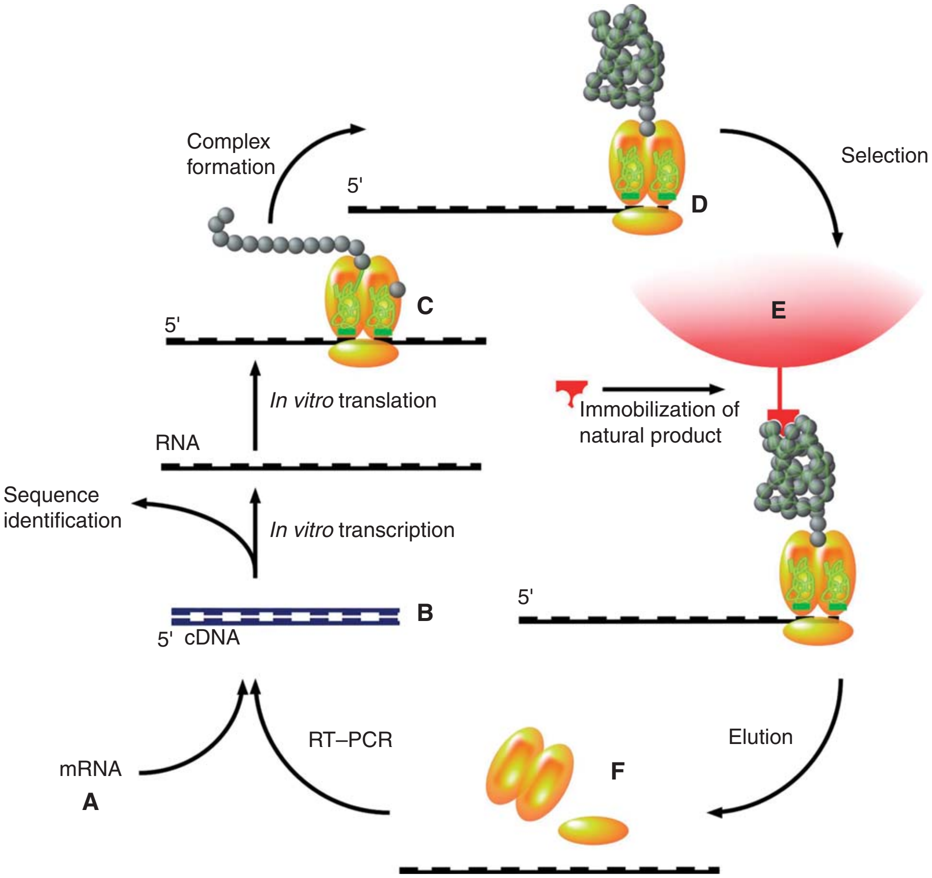


Figure 2 In ribosome display, mRNA (A) extracted from a cell is converted into a cDNA library (B) is transcribed back into mRNA with no stop codons. Prokaryotic or eukaryotic proteosomes are added and the ribosome then travels down the mRNA (C) translating until it reaches the end of the mRNA molecule (D), where the ribosome halts. With no stop codon, the release factor proteins cannot bind and so the protein, ribosome, and mRNA are physically associated and can be stabilized by high Mg^{2+} and low temperatures. This complex could then be bound directly to an immobilized natural product (E), the nonbinding library members washed away and the bound members eluted with EDTA (F), which destabilizes the ribosomal complexes by removing Mg^{2+}. The purified sublibrary is converted into cDNA by reverse transcription (RT-PCR) and amplified by regular PCR (B). The *in vitro* transcription and translation can be repeated for another round of selection or the cDNA can be analyzed by agarose electrophoresis and/or sequencing.

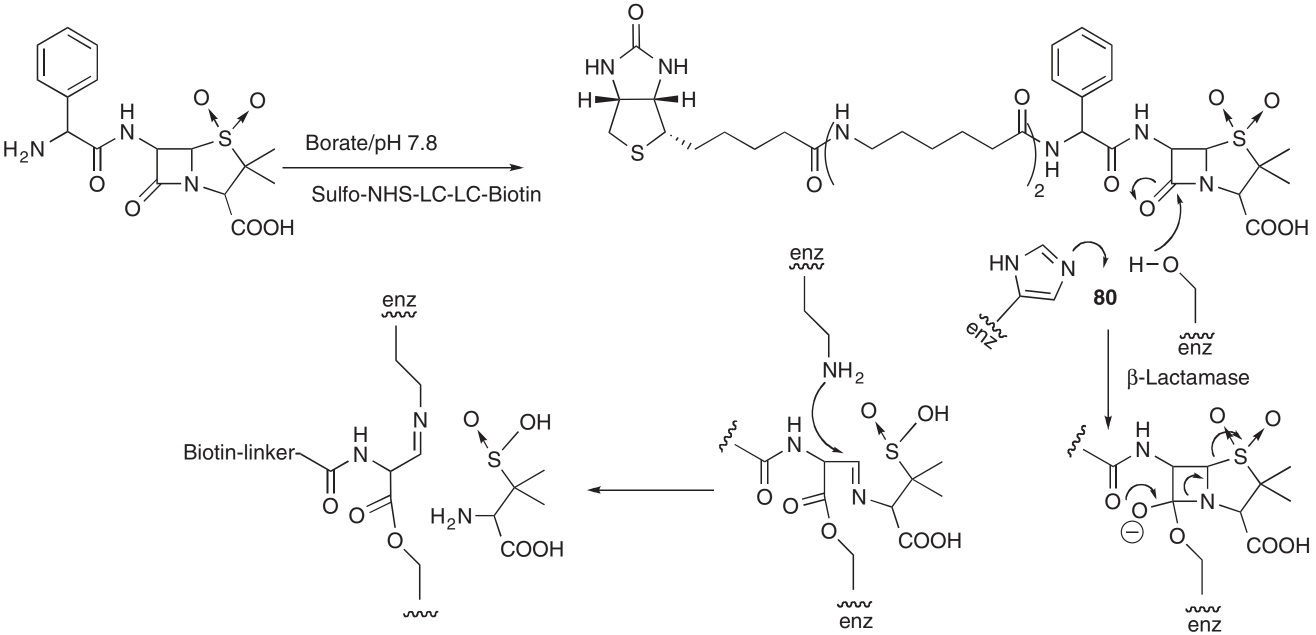


Figure 3 Construction of a covalent activity probe for β-lactamase enzymes was achieved by biotinylation of ampicillin sulfone with EZ-Link Sulfo-NHS-LC-LC-Biotin (Pierce) by Amshutz *et al.*[276] The action of the enzyme on the substrate produced a suicide inhibitor that selectively biotinylated the most active enzymes. By appropriate selection criteria, ribosome display was used to select the fastest enzymes.

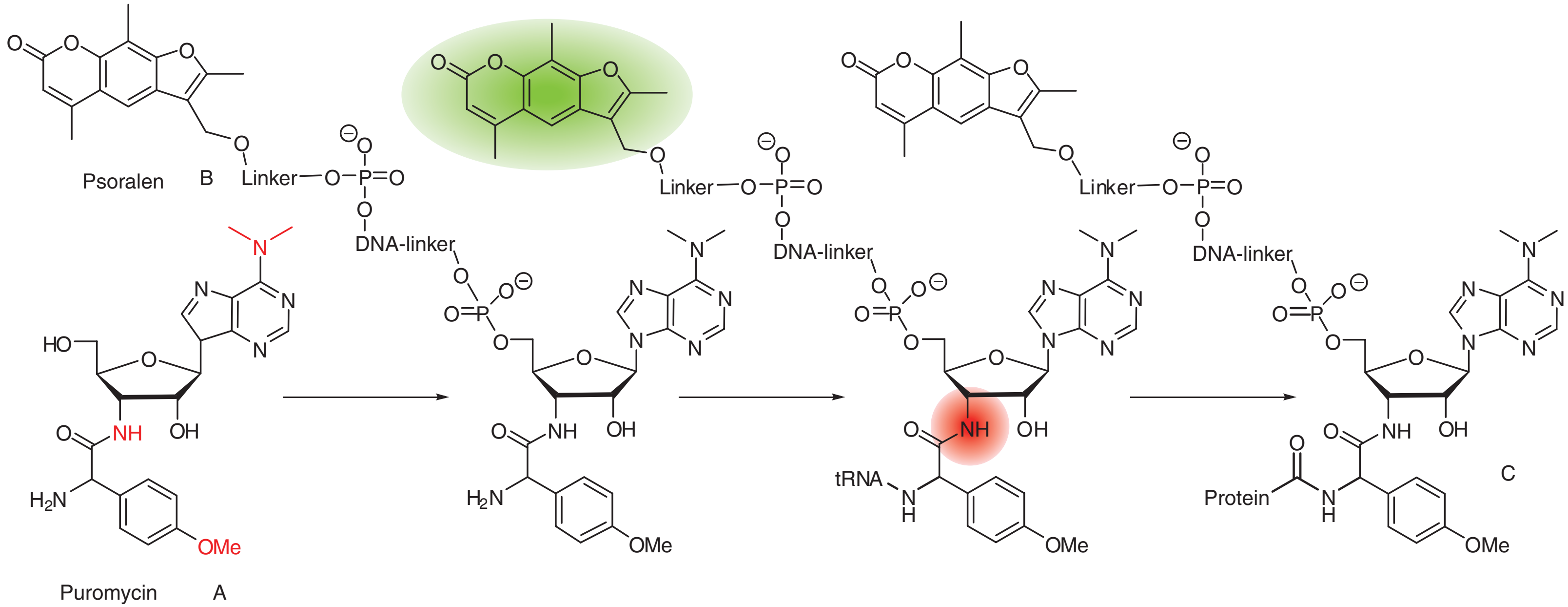


Figure 4 Puromycin (**A**) is an antibiotic analog of tyrosyl tRNA that differs by the groups in red. Once joined to a DNA linker at the 3′-end and a psoralen (B) at the 5′-end it is ready to covalently link the mRNA to nascent peptide. Photoactivation of the psoralen (green), cross-links the 5′-end of the linker and the mRNA. Once the ribosome stalls at the end of the mRNA, the puramycin enters the ribosome A-site and is transferred to the end of the newly formed protein. However, the ribosome is unable to hydrolyze the ribamine (red) amide bond thus forming a permanent link between mRNA and the encoded protein (C).

in this particular application of finding protein targets for natural products, the advantage is moot because genomes are relatively small. Disadvantages include PCR bias and artifact formation that tend to occur in multitemplate PCR, such as cDNA library generation, and provide incorrect information on the abundance and diversity of genes.[278] Other disadvantages shared by most display methods include questions about proper folding of proteins in an *in vitro* system, the lack of any sort of posttranslational modifications, the unlikely occurrence of natural protein–protein complexes, and the difficulty of expressing large or membrane proteins. Indeed, it has been reported that the ligand-binding domain of the Nogo receptor, which aggregates in phage display, does not aggregate in ribosome display.[277] As the ribosome is a large complex (>3 MDa) of proteins and rRNA and displayed peptides or proteins in typical libraries are relatively small (<10 kDa), it is possible that many displayed peptides may be lost through unpredictable interactions with the ribosome. This was referred to as the 'large display object problem' by Gold[279] but is less likely to be a problem with displayed cDNA libraries.

Unique disadvantages to ribosome display relate to the inherent instability of RNA. This is a disadvantage shared by mRNA display but an advantage of ribosome display over mRNA display is that the *in vitro* transcription and translation can be combined into one step. This renders the mRNA less vulnerable to degradation.

9.14.3.3.2 mRNA display

mRNA display was developed from a desire to display large numbers of random peptides and was thus a response to the limitations of phage display. Initially, random peptides were displayed on polysomes (ribosomes), which contained the mRNA for the displayed peptide.[262] This was quickly developed into an mRNA display system simultaneously by two groups, Roberts and Szostak in the United States and Yanagawa in Japan. The techniques have been recently reviewed by both groups.[175,280]

The key feature of mRNA display is that the gene (mRNA) and the encoded protein are covalently attached to each other by a puromycin–DNA linker (**Figure 4**). This is a key advantage over ribosome display, where the link between the mRNA and the protein is noncovalent and mediated by the ribosome. Thus mRNA display allows more stringent selection criteria to be employed that would result in the dissociation of ribosome and mRNA.

Puromycin is a natural product isolated from *Streptomyces alboniger* and a structural analog of tyrosyl tRNA but when incorporated into the ribosomal machinery causes premature termination of translation by being nonspecifically linked to the growing protein.[281,282] This information, and the recent development of ribosome display, allowed two groups to simultaneously and independently develop mRNA display based on this chemistry (**Figure 5**).[283,284] However, the technique was far from perfect because the mRNA encoding the protein had to have the stop codon removed, a DNA spacer, P-acceptor (21-mer DNA molecule with 4-mer RNA), and puromycin laboriously installed on each strand of mRNA.[284] Puromycin works by entering the A-site of the ribosome in the absence of a release factor (stop codon) and the peptidyl transferase subunit catalyzes amide bond formation between the amine group on the puromycin and the C-terminal carboxylate of the full-length protein. Because the new linkage is an amide instead of the usual ester, hydrolysis is no longer possible and protein synthesis stalls. The mRNA–protein conjugate is then released and can be purified. The ligation method was improved by using another natural product, psoralen, to induce cross-linking of the puromycin-containing oligonucleotide to the 3′-end of an mRNA template.[285] Psoralen, a furanocoumarin, occurs in many plants but was first isolated from the seeds of the Ayuvedic plant *Psoralea corylifolia*.[286] The compound had long been known to be a specific, bifunctional photo cross-linking agent that can join dsDNA after irradiation at 365 nm.[287] Another approach is to use an enzymatic ligation of a polyethylene glycol (PEG) spacer.[288]

mRNA-display libraries made from cellular cDNA have been used to identify the protein partners that bind to antiapoptotic protein Bcl-X_L.[289] Hammond *et al.* used random primers to construct the library that contained members of various lengths and in all three reading frames (i.e., two-thirds of the library displayed peptides that are not encoded by the associated gene). This approach also allows the inclusion of tissue-specific primer tags so that multiple libraries can be screened in one experiment. After immobilization of the Bcl-X_L protein, the mRNA library was screened for binding partners and the eluted sublibrary amplified by PCR and transcription. After four rounds of selection, tissue-specific primers were used to identify the selected clones by sequencing of

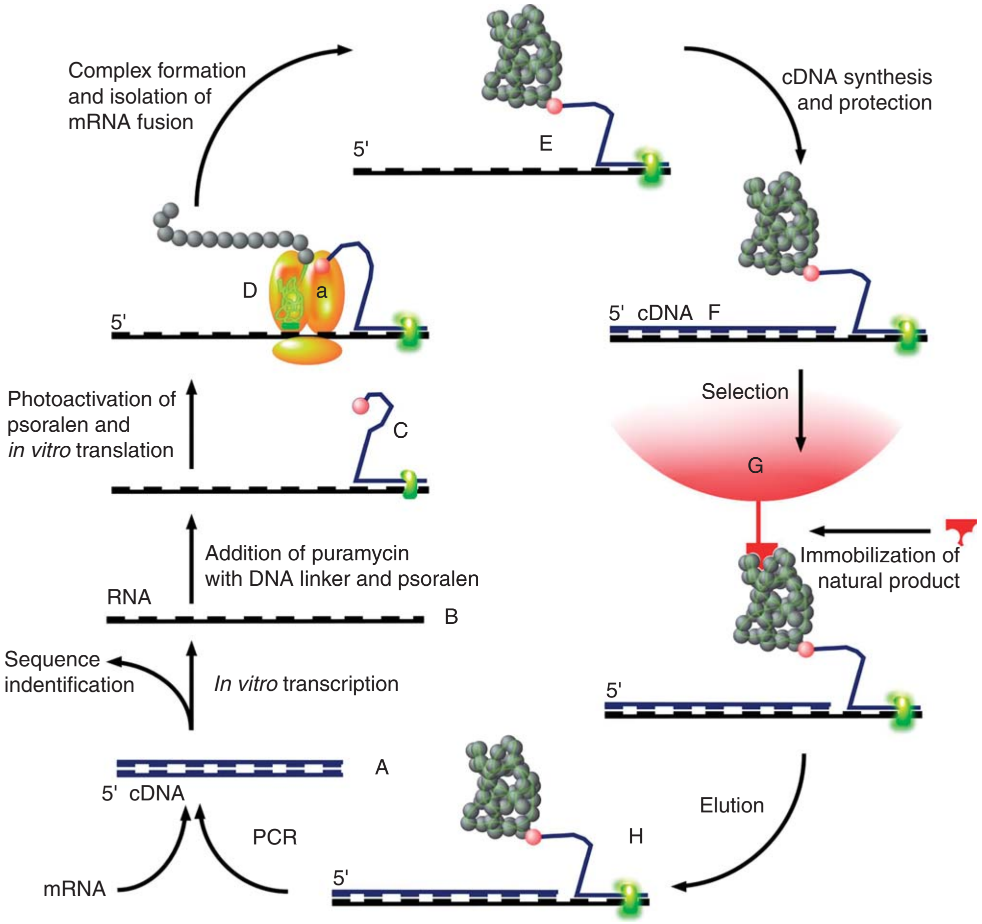
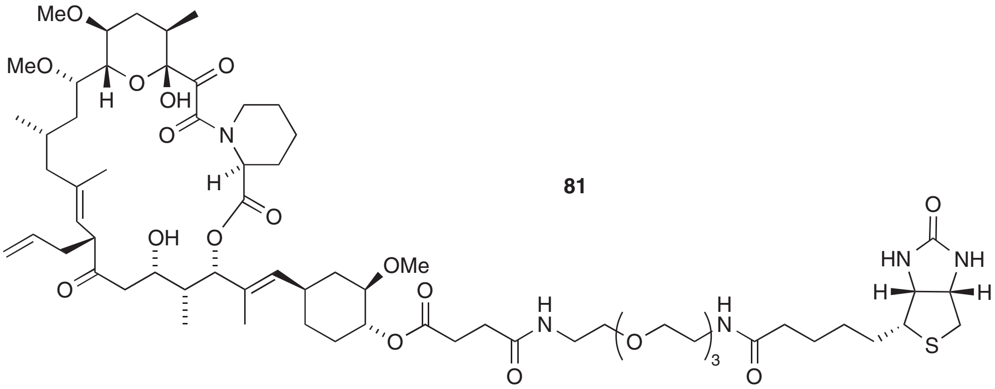


Figure 5 Starting from natural mRNA, a cDNA library (A; blue) is produced and like ribosomal display, the cDNA is transcribed into mRNA (B) with no stop codons. The 3′-end of each mRNA molecule is ligated to a short synthetic DNA linker (C) and sometimes a polyethyleneglycol spacer, which terminates with a puramycin molecule (small red sphere). The ligation is stabilized by the addition of psoralen (green clamp), which is photoactivated to covalently join both strands. Addition of crude polysomes or purified ribosomes (D) results in translation of the mRNA into protein, but the ribosome stalls at the mRNA–DNA junction. Since there are no stop codons, release factors cannot function and instead the puromycin enters the A-site of the ribosome (A). Because puramycin is an analog of tyrosyl-tRNA, the peptidyl transferase subunit catalyzes amide bond formation between the puromycin amine and the peptide carboxyl terminus, but is unable to hydrolyze the amide link (which should be an ester in tyrosyl-tRNA) to release the dimethyladenosine. The ribosome is dissociated to release the mRNA–protein fusion (E), which is protected with complementary cDNA using RT-PCR (F). The mRNA library can then be selected against an immobilized natural product probe (G), nonbinding library members washed away and the bound mRNA (H) released with SDS. PCR amplification of the cDNA provides a sublibrary (A) for another round of selection or for analysis/sequencing.

the isolated cDNA. This revealed three known binding partners for Bcl-X_L but also another 17 proteins. An advantage shared by the *in vitro* techniques and phage display is the physical association between the gene and the protein. This allows easy isolation of the cDNA and subsequent cloning into any expression vector for biochemical studies. This allowed Hammond to determine the relative binding affinities of the 20 proteins and peptides to Bcl-X_L. Similarly, protein-binding partners have been isolated for the transcription factors c-Jun[290] and c-Fos,[291] and calmodulin,[292] using mRNA display. DNA-binding proteins have been isolated by Yanagawa by using a 12-*O*-tetradecanoylphorbol-13-acetate (TPA)-responsive element (TRE) as a bait DNA.[293] These results suggest that mRNA display could also be fruitfully used to explore the binding proteins of natural products by simply substituting the Bcl-X_L protein in the Hammond example for a natural product. Surprisingly though, there has only been two publications, reporting proof of principle, using the ubiquitous

FK506–FKBP system. The first, by McPherson *et al.*,[294] biotinylated FK506 (81) through the secondary hydroxyl in 3% overall yield and used streptavidin-coated magnetic beads for immobilization and after just three rounds of selection, the library converged onto a single clone for FKBP1a.



81

More recently, Yanagawa and coworkers has further improved the mRNA display method by incorporation of a photocleavable 2-nitrobenzyl linker between genotype (mRNA) and phenotype (protein) to facilitate isolation of the mRNA for amplification or analysis.[295] They demonstrated the utility of the photocleavable linker by also isolating FKBP1a from an mRNA library using biotinylated FK506.

These two examples, using FK506–FKBP, clearly demonstrate the potential of mRNA display for the isolation of natural product receptors and, like phage display, could be used to determine the target proteins for natural products with no known target or even no known biological activity.

However, mRNA display has some of the disadvantages of phage display in that the natural variability in concentration of mRNA in cells (over 7–8 orders of magnitude)[296] means that high-abundance mRNA is amplified repeatedly and would tend to dominate in the absence of very strong selection pressure. Recently, cDNA tiling arrays have been used to address this problem, and been shown to double the coverage of Jun-associated proteins without reducing accuracy.[297] Normalization of cDNA libraries would also aid the discovery of novel protein-binding partners from rare transcripts, but because *in vitro* display systems do not have any transformation steps in selection rounds, there is no bias related to transformation efficiency (unlike phage and cell display). PCR bias and artifact formation can also occur in multitemplate PCR, such as cDNA library generation, and provide incorrect information on the abundance and diversity of genes.[278] Other disadvantages shared by most display methods include questions about proper folding of proteins in an *in vitro* system, the lack of any sort of posttranslational modifications, the unlikely occurrence of natural protein–protein complexes and the difficulty in expressing large or membrane proteins.

Unique disadvantages to mRNA display relate to the RNA itself and the stop codon in natural mRNA. Because the mRNA is linked to puromycin at the 3′-end (Figure 5), *in vitro* transcription and translation must be carried out as separate reactions (in ribosome display, *in vitro* transcription, and translation can be combined). This separation of transcription and translation renders the mRNA more vulnerable to degradation by hydrolysis or RNAase activity. A second problem for *in vitro* selection of cDNA libraries is that full-length cDNA constructed by an oligo(dT)-priming strategy always contain 3′- and 5′-UTR plus a stop codon that would disengage the ribosome before incorporation of puramycin. To circumvent this, cDNA-derived libraries for mRNA display can be constructed by means of random hexamer priming but this makes the DNA nondirectional. This can be solved through degenerate primers and directional linkers.[292]

9.14.3.3.3 DNA display

Cellular compartmentalization has inspired a range of new, and potentially useful, *in vitro* systems that treat droplets of water as microreactors. These have been recently reviewed by Griffiths and coworkers.[298] The first DNA-display experiments, using this idea, were reported by Tawfik and Griffiths, who developed an *in vitro* compartmentalization (IVC) method using water-in-oil emulsions, which can be cheaply manufactured from oil, emulsifier, and detergent. By carefully selecting the conditions, a DNA library can be spatially arrayed by partitioning into the water droplets such that no more than one DNA molecule resides in any one compartment. Multiple copies of the encoded protein can then be made in the droplet by an *E. coli*-coupled cell-free system.[299] This is interesting but not particularly useful for the selection of binding partners unless the DNA and encoded protein can be physically linked. This has been achieved through an adaptation of IVC by Yanagawa using the near-covalent interaction between streptavidin and biotin.[300] Put simply, a biotinylated DNA molecule that encodes a protein library with an N-terminal streptavidin tag is diluted into a water/oil emulsion. Now, when the water droplets that contain one DNA molecule are transcribed and translated, the expressed proteins bind to the DNA through the biotin forming a physical link. The expressed proteins can now be affinity purified against any binding epitope (such as a natural product) and the DNA amplified by PCR and the processes repeated for as many rounds as required for positive selection. However, the original method suffered from poor efficiency (<1% DNA–protein complex formation) but an adapted version has increased efficiency and been used to search for peptide ligands.[301] The method has also been demonstrated to be able to display active proteins up to 1000 amino acids[302] so there is potential to use the technique for display of cDNA libraries and then to pan for natural product receptors, although no examples exist at this time. Thus one of the major disadvantages of the system is that, while one can manipulate a range of reaction conditions that are incompatible with *in vivo* display, the cell-free transcription and translation must be carried out under defined Mg, salt, and pH conditions.

A very similar method has been reported by Tawfik and Griffiths, whereby a single piece of DNA is linked to streptavidin-coated beads through covalently attached biotin. The DNA encodes a FLAG tag at the 5′-end and biotinylated anti-FLAG antibody is used then to couple the encoded protein to the same bead.[303] It remains to be seen if this complex and demanding strategy can be adapted to cDNA library display and the isolation of natural product-binding proteins.

A covalent link between genotype (DNA) and phenotype (protein) in DNA display has been achieved through enzyme activity. Thus, a DNA library, also encoding a DNA methyltransferase (M.*Hae*III) can form a covalent bond with a 5-fluorodeoxycytidine (suicide inhibitor) that is installed at the end of the DNA fragment. The resulting library of DNA–protein fusions is extracted from the oil/water emulsion, and DNA displaying protein with desired binding properties selected from the pool of DNA–protein fusions by affinity selection.[304] The method has also been used for affinity maturation of binding domains but it is not clear if the method can be used for the expression and selection of binding proteins from cDNA libraries.[305]

More recently, a covalently linked system exploits the replication initiator of bacteriophage P2, which covalently attaches the P2A protein to its own DNA phosphate backbone.[306] This method is droplet independent, theoretically allowing libraries as large as those used in mRNA display but at present the system suffers from low complex formation efficiency (~3%) that needs to be improved.

DNA display systems have possible advantages over *in vivo* display technologies, although there are few published examples. First, DNA is much more stable than RNA allowing more varied biopanning strategies (e.g., where RNAase enzymes might be present). For DNA display, removal of the stop codon from full-length cDNA libraries is also not required. The third advantage results from possible multivalency effects. In mRNA display, one mRNA yields one protein whereas multiple proteins can be displayed from a single template DNA in an oil droplet or on a microbead. This advantage is similar to phage display where it is possible to display many copies of the alien protein on the surface, thus aiding selection and coverage.

A disadvantage of the compartmentalization-based DNA display systems is that the potential library size is restricted by the number of compartments (or microspheres) that can be handled. The microdroplets are also always of variable size, which makes it difficult to ensure that each droplet contains only one gene. A Poisson distribution would result in large droplets containing many genes. To avoid this, the DNA–droplet ratio has to be dropped such that most of the droplets are empty, decreasing the size of the library that can be used. Although it is possible to deliver substrates and ligands, either directly or indirectly (through separate microdroplets, micelles, or nanodroplets) there is no universal technique and the efficiency and rates of reactions cannot be directly measured.[298] However, developments in microfluidics and the generation of physically defined, spatially arrayed, and homogeneous droplets are now being developed, for example, in the lab of George Whitesides.[307–309] The ability to fuse, incubate, and manipulate these tiny droplets when combined with IVC DNA display should provide a powerful discovery platform for combinatorial chemistry and natural products chemistry that should be realized in the not-too-distant future.[310]

Taken together, *in vitro* display of cDNA libraries are an integral part of modern reverse chemical proteomics applications, where the goal is to functionally display all proteins and to minimize selection bias in the identification of protein ligands for natural products.

9.14.3.3.4 Plasmid display

Plasmid display is conceptually simple, avoids the potential difficulties encountered with other display systems such as the cDNA size, *in vitro* translation, mRNA stability, and is easily amenable to high-throughput versions.[311] The fusion proteins, including the DNA-binding domain, are expressed *in vivo*, and the proteins bind to the specific DNA sequence on the encoding plasmids through a sequence-specific DNA–protein interaction. Complexes of fusion proteins and the encoding plasmid DNA can be used for the *in vitro* selection from a protein library following cell lysis. This technique was first validated experimentally using the *lac* repressor protein.[312] However, the success of the plasmid display is critically linked to the correct folding of the fusion proteins in addition to the maintenance of the protein–DNA-binding interaction.[313] Plasmid display may be amenable to the microdroplet format used for DNA-display methods but this has not yet been tried. There are no examples of using plasmid display to isolate natural product-binding proteins but it is theoretically possible.

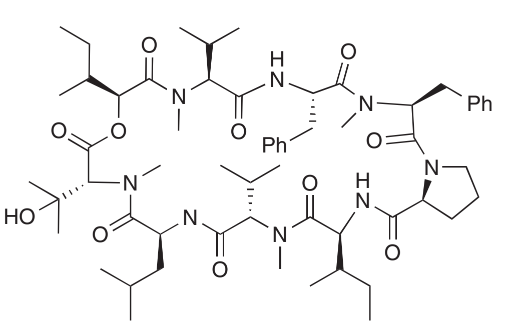
9.14.3.4 Reverse Genetics

Classical genetics cannot be used to find the binding partners for natural products. This method does, however, link a phenotype with a genotype. Generally, it involves inducing a genetic mutation in a cell or organism, selecting mutants displaying a phenotype of interest and then identifying the mutant gene that was responsible for the observed phenotype. This was first shown in 1927 through the genetic mutation of fruit flies with X-rays.[314] Subsequently, chemical mutagens such as alkylating agents (e.g., *N*-nitroso compounds), DNA intercalators, and nucleobase analogs were found to produce mutations in higher yields than those obtainable through ionizing radiation. Regardless of the method used to induce mutations, the classical genetic approach is irreproducible, inefficient, and time consuming. The random, nonspecific nature of mutations produced from mutagenic chemicals or ionizing radiation mean that it is difficult, if not impossible, to reproduce a particular mutant.

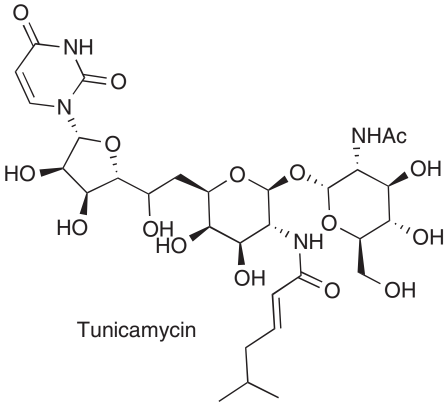
Reverse genetics involves targeting a gene of interest with a knockout mutation (deletion) or site-specific mutation (to alter the function of the gene product) and then observing the phenotypic consequences of the loss of that gene product. These mutations were traditionally introduced using homologous recombination in an embryonic stem cell. Unlike classical genetics, the reverse genetic approach is both specific and reproducible. However, a major drawback with the technique is that each mutant is forced to adapt to the loss of a particular gene, and the resulting cascade of secondary and compensatory events can obscure the true effect of the original gene knockout. Recently, there have been some innovations in these general methods that may be utilized for the determination of a natural product's binding target.

9.14.3.4.1 Small interfering RNA

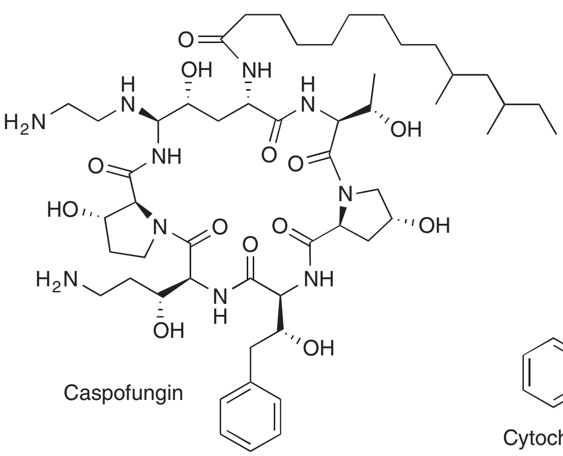
The use of posttranscriptional gene silencing by RNA interference (RNAi) has become popular.[315–317] This technique involves inserting small pieces of double-stranded RNA (siRNA), identical to small sections of the gene that needs to be suppressed. These small pieces of siRNA take advantage of the RISC (RNA-induced silencing complex) that forms a natural defense against certain viruses and is used for endogenous gene silencing. RISC binds the siRNA, unzips it into single-stranded RNA that is now able to recognize the target gene (mRNA). The RISC–RNAi complex binds to the target gene and cleaves it at the recognition site. Other proteins then degrade the mRNA pieces preventing gene translation into protein. Thus, a single gene can be downregulated for about 48 h. The limited temporal control and problematic delivery of the siRNA into cells are major concerns. Recently, it has also been found that siRNA can produce off-target gene regulation that can mask the expected outcome.[318] However, if it is possible to produce a specific siRNA for every human gene then it should be possible to set up a high-throughput screen for drug-binding targets in certain circumstances. For example, if an increased or decreased cytotoxicity of a drug can be related to the downregulation of a particular gene then this would provide *prima facie* evidence that the drug (natural product) binds to, or interferes with the gene product of the interfered with mRNA.[319,320]



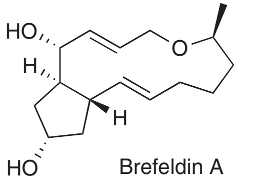
Aureobasidin A



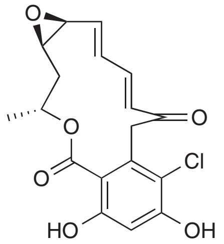
Tunicamycin



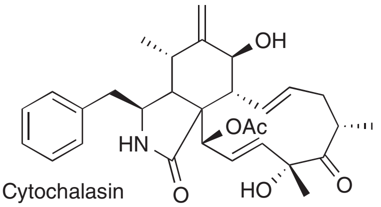
Caspofungin



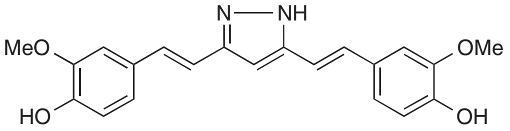
Brefeldin A



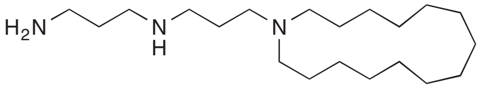
Radicicol



Cytochalasin



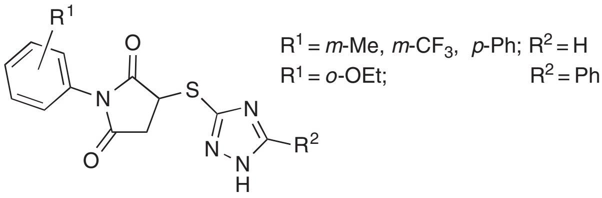
Hydrazinocurcumin



Dihydromotuporamine C

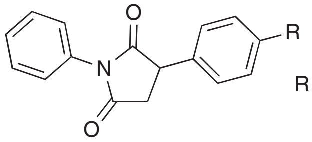
9.14.3.4.2 Drug-induced haploinsufficiency

Another reverse genetic screen that may become useful as a genome-wide screen for natural product-binding proteins is the drug-induced haploinsufficiency screen. This relies on the fact that yeast are diploid, having two copies of every gene. Deletion of one copy rarely leads to any obvious phenotypic change but it was discovered that deletion of one copy of a drug target gene led to a hypersensitivity to that drug.[321] The drug targets of several natural products have been confirmed in this way. For example, Roemer used a genome-wide fitness test to determine the drug-binding proteins for several antifungal drugs, including several natural products.[322] A *Candida albicans* library containing 2868 mutants, representing 45% of the *C. albicans* genome, each with the deletion of one copy of one gene (heterozygotes), was constructed using PCR techniques. Aliquots of the library (equal quantities of all 2868 mutants) were treated with an inhibitory compound (at different concentrations) or a control over 20 population doublings. The relative growth of each strain was monitored by DNA microarrays competitively hybridized with amplified and labeled tags (using the common primer pairs) from the drug/no drug treatments. Statistical analyses enable identification of strains significantly affected in growth rate.[322] This resulted in the identification of phosphatidylinositol/ceramide phosphoinositol transferase (IPC synthase), required for sphingolipid synthesis, as the target for aureobasidin A. Tunicamycin targeted UDP-*N*-acetyl-glucosamine-1-P transferase, which transfers GlcNAc-P from UDP-GlcNAc to Dol-P in the ER in the first step of the dolichol pathway of protein asparagine-linked glycosylation. The catalytic subunit of 1,3-beta-D-glucan synthase, that is involved in cell wall synthesis and maintenance, as the target of caspofungin, a GTP-binding protein of the rho subfamily of Ras-like proteins, involved in the establishment of cell polarity that regulates protein kinase C (Pkc1p). Ergokonin A was found to target and cell wall synthesizing enzyme – 1,3-beta-glucan synthases. Brefeldin A was found to interact with two proteins, guanine nucleotide exchange factor (GEF), involved in proliferation of the Golgi, intra-Golgi and ER to Golgi transport, and a GTPase of the Ras superfamily, involved in intracellular trafficking within the Golgi. The drug is known to bind to the interface between these two proteins. Cytochalasin D did not display induced hypersensitivity to any one mutant but a rich mixture of genes that did not allow a positive identification. Radicicol provided another example of the limitations of the haploinsufficiency screen, in that its well-characterized target (Hsp90p) was unresponsive to drug treatment as a heterozygote. In all cases, the tested inhibitors already had well-characterized targets and these were nearly always identified in their screen. They also explored the assay as an approach to predicting the binding partner of novel antifungal compounds (82–83) with no known mode of action. Their profiles were similar to known microtubule inhibitors and highlighted by marked hypersensitivity of the α-tubulin mutants to all drugs.



R^1 = *m*-Me, *m*-CF_3, *p*-Ph; R^2 = H
R^1 = *o*-OEt; R^2 = Ph

82



R = F, OCF_2Cl

83

These results not only proved that the method could isolate known targets for natural products, but also identified modes of action for new compounds, in this case as binding to α-tubulin. The discovery of this interaction suggests that the compounds should also be screened as anticancer agents.

Similarly, Roberge showed that the marine natural product dihydromotuporamine C targeted sphingolipid metabolism by producing hypersensitivity to some 21 different genes.[323] In this paper, the authors measured the growth rate of individual mutants rather than using a DNA-array method. This was much more time consuming but gave far fewer false positives.

Hoe and coworkers have used a 4158 mutant library of *Schizosaccharomyces pombe* to screen for the targets of hydrazinocurcumin, a synthetic derivative of the turmeric natural product curcumin.[215] An initial screen provided 178 hits of which 165 were determined to be false positives by growing each mutant with varying concentrations of the natural product derivative. After a second screen, they found eight genes that resulted in hypersensitivity to hydrazinocurcumin. These related to septum formation and the general transcription processes, which may be related to histone acetyltransferase, a previously reported target for curcumin.[324] No definitive binding partner was found, highlighting some of the possible difficulties of this technique. It is also notable that there was no evidence for binding of hydrazinocurcumin to calmodulin, the protein target proposed by Kwon using phage display.[214]

Reverse genetics, through either siRNA gene silencing or specific gene deletion in diploid species, shows promise as a new method for natural product or drug target elucidation. There are several limitations, as discussed above, but these may be overcome in the near future to produce a technique that would be quite powerful for the isolation of natural product receptors. Though we are not aware of any examples using siRNA, this method seems to be globally applicable whereas the drug-induced haploinsufficiency assay would not be applicable to any monoploid species and may be difficult to apply to human cells because of the large number of genes that need to be mutated. Human cells are also extremely complex because one gene could code for hundreds of proteins through alternative splicing, RNA editing, and posttranslational modifications making results difficult to interpret. One major advantage of all the reverse genetics approaches is that the natural product is not derivatized.

9.14.4 Future Prospects

Chemical proteomics approaches, especially affinity chromatography, dominate the methods used to identify natural product-binding proteins. The only other technique that is used to any great extent is phage display. Both methods have in common the need to synthesize a tagged natural product, usually with biotin. It is always challenging to prepare natural product affinity reagents without affecting their biological activity, not to mention the challenge of semi- or total synthesis. In addition, it is often difficult to minimize contamination due to nonspecific binding of proteins to an affinity matrix. Recently, interesting new techniques have been reported that could, potentially, be used to isolate natural product-binding proteins without the need for any sort of chemical derivatization. These MS methods involve SILAC, ICAT, or iTRAQ quantitative proteomics. However, with some development, it could be possible to use these systems to screen for natural products that inhibit particular types of enzymes using activity-based chemical proteomics, for example.[123] Currently, activity-based proteomics can capture serine, cysteine, and metallo-proteases, tyrosine phosphatases and kinases, glycosidases, or penicillin-binding proteins. If the activity probes could be combined with SILAC or iTRAQ, this could be a powerful discovery platform for natural products that can bind to these classes of enzymes. Reverse genetics approaches also have potential if haploid yeast can be made for every gene as, in this method, there is no need to derivatize the natural product and the assays are easy to carry out. The near future will also bring major advances in display technologies. Those outlined here, with the possible exception of mRNA display and ribosome display, are in their infancy and these could become very important techniques in discovering natural product-binding proteins. Phage display is relatively mature but there are many alternative bacteriophage-display systems that will, no doubt, surface in the next few years.[325]

Glossary

chemical genetics Chemical genetics strategies start with libraries of chemical compounds, which are screened to find compounds that produce a phenotype of interest. Once a phenotype-modifying compound is found, it is used to identify the particular target protein to which it binds in the cells. Sometimes used interchangeably with ‘chemical genomics’.

chemical proteomics Makes use of small molecules that can bind specifically to a class of enzyme (activity-based chemical proteomics) or a protein (affinity-based chemical proteomics) either covalently or noncovalently to allow their purification and/or identification as valid drug targets. The process is analogous to interaction proteomics but instead of identifying protein–protein interaction, small-molecule–protein interactions are identified.

reverse chemical genetics To discover a new drug or chemical probe, it is possible to start with a protein of interest and screen for small molecules that affect its activity, then find out whether the small molecule causes a phenotypic change in an organism or cell. This approach is analogous to reverse genetics, in which a gene is deliberately mutated or knocked out in order to study the resulting phenotype. Reverse chemical genetics is extensively used in the pharmaceutical industry to screen for new drug leads in combinatorial libraries.

reverse chemical proteomics The proteome is expressed on the surface of an amplifiable vector, such as a virus or cell, and then probed with a tagged small molecule. The starting point is the transcriptome (mRNA), which is physically linked to its gene product (protein) to facilitate identification of small-molecule–protein interactions.

reverse genetics Seeks to determine the phenotype that derives from a specific gene (obtained by DNA sequencing). It is an approach to discovering the function of a gene that proceeds in the opposite direction to classical genetics, which uses random mutations to work back from a phenotype of interest to the gene responsible for that phenotype. In reverse genetics, a gene is removed, mutated or suppressed, and the phenotype produced used to determine the function of the gene.

reverse proteomics In reverse proteomics, the starting point is the DNA sequence of the genome of an organism. First, the transcriptome (complete set of transcripts) and proteome (complete set of proteins) are predicted *in silico* and subsequently this information is used to generate reagents for their analysis.

References

1. D. J. Newman; G. M. Cragg; K. M. Snader, *Nat. Prod. Rep.* **2000**, *17*, 215.
2. C. H. Berndt, The Role of Native Doctors in Aboriginal Australia. In *Magic, Faith and Healing*; A. Kiev, Ed.; Free Press of Glencoe: London, 1964; p 264.
3. J. M. Rollinger; T. Langer; H. Stuppner, *Curr. Med. Chem.* **2006**, *13*, 1491.
4. D. J. Newman, *J. Med. Chem.* **2008**, *51*, 2589.
5. M. Feher; J. M. Schmidt, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 218.
6. S. Class, *Chem. Eng. News* **2002**, *80*, 39.
7. S. Wilhelm; C. Carter; M. Lynch; T. Lowinger; J. Dumas; R. A. Smith; B. Schwartz; R. Simantov; S. Kelley, *Nat. Rev. Drug Discov.* **2006**, *5*, 835.
8. B. E. Evans; K. E. Rittle; M. G. Bock; R. M. DiPardo; R. M. Freidinger; W. L. Whitter; G. F. Lundell; D. F. Veber; P. S. Anderson; R. S. L. Chang; V. J. Lotte; D. J. Cerino; T. B. Chen; P. J. Kling; K. A. Kunkel; J. P. Springer; J. Hirshfieldt, *J. Med. Chem.* **1988**, *31*, 2235.
9. J. S. Mason; I. Morize; P. R. Menard; D. L. Cheney; C. Hulme; R. F. Labaudiniere, *J. Med. Chem.* **1999**, *42*, 3251.
10. M. J. Stone; D. H. Williams, *Mol. Microbiol.* **1992**, *6*, 29.
11. R. D. Firn; C. G. Jones, *Mol. Microbiol.* **2000**, *37*, 989.
12. R. D. Firn; C. G. Jones, *Nat. Prod. Rep.* **2003**, *20*, 382.
13. M. A. Fischbach; J. Clardy, *Nat. Chem. Biol.* **2007**, *3*, 353.
14. S. M. Owen; J. Penuelas, *Trends Plant Sci.* **2006**, *11*, 423.
15. R. D. Firn; C. G. Jones, *Trends Plant Sci.* **2006**, *11*, 422.
16. E. Pichersky; T. D. Sharkey; J. Gershenzon, *Trends Plant Sci.* **2006**, *11*, 421.
17. M. B. Austin; P. E. O'Maille; J. P. Noel, *Nat. Chem. Biol.* **2008**, *4*, 217.
18. S. L. Schreiber, *Chem. Eng. News* **2003**, *81*, 51.
19. A. Coghlan; N. Boyce, *New Sci.* **2000**, *167*, 4.
20. A. M. Rouhi, *Chem. Eng. News* **2003**, *81*, 77.
21. L. Kissau; P. Stahl; R. Mazitschek; A. Giannis; H. Waldmann, *J. Med. Chem.* **2003**, *46*, 2917.
22. J. Kobayashi; T. Madono; H. Shigemori, *Tetrahedron* **1995**, *51*, 10867.
23. A. E. Wright; P. J. McCarthy; G. K. Schulte, *J. Org. Chem.* **1989**, *54*, 3472.
24. S. P. Gunasekera; P. J. McCarthy; M. Kelly-Borges; E. Lobkovsky; J. Clardy, *J. Am. Chem. Soc.* **1996**, *118*, 8759.
25. B. M. McArdle; R. J. Quinn, *ChemBioChem* **2007**, *8*, 788.
26. B. K. Law, *Crit. Rev. Oncol. Hematol.* **2005**, *56*, 47.
27. T. W. Corson; C. M. Crews, *Cell* **2007**, *130*, 769.
28. A. Fleming, *Bull. Soc. Chim. Belg.* **1929**, *10*, 226.
29. D. C. Hodgkin, *Adv. Sci.* **1949**, *6*, 85.
30. J. T. Park; J. L. Strominger, *Science* **1957**, *125*, 99.
31. J. Lederberg, *J. Bacteriol.* **1957**, *73*, 144.
32. K. Izaki; M. Matsuhashi; J. L. Strominger, *Proc. Natl. Acad. Sci. U.S.A.* **1966**, *55*, 656.
33. K. Choi; J. Hong; C.-O. Lee; D.-k. Kim; C. J. Sim; K. S. Im; J. H. Jung, *J. Nat. Prod.* **2004**, *67*, 1186.
34. X.-L. Hou; Z. Yang; K.-S. Yeung; H. N. C. Wong, *Prog. Heterocycl. Chem.* **2005**, *17*, 142.
35. J. W. Blunt; B. R. Copp; M. H. G. Munro; P. T. Northcote; M. R. Prinsep, *Nat. Prod. Rep.* **2006**, *23*, 26.
36. Y. Liu; S. Zhang; P. J. M. Abreu, *Nat. Prod. Rep.* **2006**, *23*, 630.
37. N. Dixon; L. S. Wong; T. H. Geerlings; J. Micklefield, *Nat. Prod. Rep.* **2007**, *24*, 1288.
38. Y. Liu; L. Wang; J. H. Jung; S. Zhang, *Nat. Prod. Rep.* **2007**, *24*, 1401.
39. L. Wittgenstein, *Ann. Naturphilos.* **1921**, *14*, 185.
40. S. L. Schreiber, *Bioorg. Med. Chem.* **1998**, *6*, 1127.
41. S. Omura; T. Fujimoto; K. Otoguro; K. Matsuzaki; R. Moriguchi; H. Tanaka; Y. Sasaki, *J. Antibiot.* **1991**, *44*, 113.
42. M. Pagano; S. W. Tam; A. M. Theodoras; P. Beer-Romero; G. Del Sal; V. Chau; P. R. Yew; G. F. Draetta; M. Rolfe, *Science* **1995**, *269*, 682.
43. M. Karin; Y. Ben-Neriah, *Annu. Rev. Immunol.* **2000**, *18*, 621.
44. B. K. Kim; C. M. Crewes, *J. Med. Chem.* **2008**, *51*, 2600.
45. D. A. Jeffery; M. Bogyo, *Curr. Opin. Biotechnol.* **2003**, *14*, 87.
46. V. E. Velculescu; L. Zhang; W. Zhou; J. Vogelstein; M. A. Basrai; D. E. Bassett, Jr.; P. Hieter; B. Vogelstein; K. W. Kinzler, *Cell* **1997**, *88*, 243.
47. A. M. Piggott; P. Karuso, *Comb. Chem. High Throughput Screen.* **2004**, *7*, 607.
48. Y. Ho; A. Gruhler; A. Hellbut; G. D. Bader; L. Moore; S.-U. Adams; A. Millar; P. Taylor; K. Bennett; K. Boutiller; L. Yang; C. Wolting; I. Donaldson; S. Schandorff; J. Shewnarane; M. Vo; J. Taggartt; M. Goudreault; B. Muskat; C. Alfarano; D. Dewar; Z. Lin; K. Michallckova; A. R. Willems; H. Sassi; P. A. Nielsen; K. J. Rasmussen; J. R. Andersen; L. E. Johansen; L. H. Hansen; H. Jespersen; A. Podtelejnikov; E. Nielsen; J. Crawford; V. Poulsen; B. D. Sorensen; J. Matthlesen; R. C. Hendrickson; F. Gleeson; T. Paweson; M. F. Moran; D. Durocher; M. Mann; C. W. V. Hogue; D. Figeys; M. Tyers, *Nature* **2002**, *415*, 180.
49. A. J. M. Walhout; M. Vidal, *Nat. Rev. Mol. Cell Biol.* **2001**, *2*, 55.
50. J. P. Overington; B. Al-Lazikani; A. L. Hopkins, *Nat. Rev. Drug Discov.* **2006**, *5*, 993.
51. P. M. Blumberg; J. L. Strominger, *Proc. Natl. Acad. Sci. U.S.A.* **1972**, *69*, 3751.
52. T. J. Dougherty; A. E. Koller; A. Tomasz, *Antimicrob. Agents Chemother.* **1980**, *18*, 730.
53. U. Schwarz; K. Seeger; F. Wengenmayer; H. Strecker, *FEMS Microbiol. Lett.* **1981**, *10*, 107.
54. B. G. Spratt; A. B. Pardee, *Nature* **1975**, *254*, 516.
55. M. Dargis; F. Malouin, *Antimicrob. Agents Chemother.* **1994**, *38*, 973.
56. C. R. DeLoney; N. L. Schiller, *Antimicrob. Agents Chemother.* **1999**, *43*, 2702.

57. K. R. Gee; H. C. Kang; T. I. Meier; G. Zhao; L. C. Blaszcak, *Electrophoresis* **2001**, *22*, 960.
58. B. Lakaye; C. Damblon; M. Jamin; M. Galleni; S. Lepage; B. Joris; J. Marchand-Brynaert; C. Frydrych; J.-M. Frere, *Biochem. J.* **1994**, *300*, 141.
59. M. Galleni; B. Lakaye; S. Lepage; M. Jamin; I. Thamm; B. Joris; J. M. Frere, *Biochem. J.* **1993**, *291*, 19.
60. L. M. Weigel; J. T. Belisle; J. D. Radolf; M. V. Norgard, *Antimicrob. Agents Chemother.* **1994**, *38*, 330.
61. M. J. Pucci; T. J. Dougherty, *Methods Mol. Med.* **2008**, *142*, 131.
62. E. W. Taylor, *J. Cell Biol.* **1965**, *25*, 145.
63. G. G. Borisy; E. W. Taylor, *J. Cell Biol.* **1967**, *34*, 525.
64. G. G. Borisy; E. W. Taylor, *J. Cell Biol.* **1967**, *34*, 535.
65. M. L. Shelanski; E. W. Taylor, *J. Cell Biol.* **1967**, *34*, 549.
66. R. C. Weisenberg; G. G. Broisy; E. W. Taylor, *Biochemistry* **1968**, *7*, 4466.
67. A. Ruegger; M. Kuhn; H. Lichti; H. R. Loosli; R. Huguenin; C. Quiquerez; A. Von Wartburg, *Helv. Chim. Acta* **1976**, *59*, 1075.
68. T. J. Petcher; H. P. Weber; A. Ruegger, *Helv. Chim. Acta* **1976**, *59*, 1480.
69. M. M. Merker; R. E. Handschumacher, *J. Immunol.* **1984**, *132*, 3064.
70. R. E. Handschumacher; M. W. Harding; J. Rice; R. J. Drugge; D. W. Speicher, *Science* **1984**, *226*, 544.
71. K. Kuramochi; Y. Miyano; Y. Enomoto; R. Takeuchi; K. Ishi; Y. Takakusagi; T. Saitoh; K. Fukudome; D. Manita; Y. Takeda; S. Kobayashi; K. Sakaguchi; F. Sugawara, *Bioconjug. Chem.* **2008**, *19*, 2417.
72. B. J. Leslie; P. J. Hergenrother, *Chem. Soc. Rev.* **2008**, *37*, 1347.
73. L. C. Mishra; B. B. Singh; S. Dagenais, *Altern. Med. Rev.* **2000**, *5*, 334.
74. R. Mohan; H. J. Hammers; P. Bargagna-Mohan; X. H. Zhan; C. J. Herbstritt; A. Ruiz; L. Zhang; A. D. Hanson; B. P. Conner; J. Rougas; V. S. Pribluda, *Angiogenesis* **2004**, *7*, 115.
75. B. Jayaprakasam; Y. Zhang; N. P. Seeram; M. G. Nair, *Life Sci.* **2003**, *74*, 125.
76. R. R. Falsey; M. T. Marron; G. M. K. B. Gunaherath; N. Shirahatti; D. Mahadevan; A. A. L. Gunatilaka; L. Whitesell, *Nat. Chem. Biol.* **2006**, *2*, 33.
77. Y. Yokota; P. Bargagna-Mohan; P. P. Ravindranath; K. B. Kim; R. Mohan, *Bioorg. Med. Chem. Lett.* **2006**, *16*, 2603.
78. P. Bargagna-Mohan; A. Hamza; Y.-e. Kim; Y. Khuan Ho; N. Mor-Vaknin; N. Wendschlag; J. Liu; R. M. Evans; D. M. Markovitz; C.-G. Zhan; K. B. Kim; R. Mohan, *Chem. Biol.* **2007**, *14*, 623.
79. N. Lindquist; W. Fenical; G. D. Van Duyne; J. Clardy, *J. Am. Chem. Soc.* **1991**, *113*, 2303.
80. N. S. Williams; A. W. G. Burgett; A. S. Atkins; X. Wang; P. G. Harran; S. L. McKnight, *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 2074.
81. G. Wang; L. Shang; A. W. G. Burgett; P. G. Harran; X. Wang, *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 2068.
82. N. Sin; L. Meng; M. Q. W. Wang; J. J. Wen; W. G. Bornmann; C. M. Crews, *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 6099.
83. N. Sin; L. Meng; H. Auth; C. M. Crews, *Bioorg. Med. Chem.* **1998**, *6*, 1209.
84. J. Abe; W. Zhou; N. Takuwa; J. Taguchi; K. Kurokawa; M. Kumada; Y. Takuwa, *Cancer Res.* **1994**, *54*, 3407.
85. W. T. Lowther; D. A. McMillen; A. M. Orville; B. W. Matthews, *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 12153.
86. S. Liu; J. Widom; C. W. Kemp; C. M. Crews; J. Clardy, *Science* **1998**, *282*, 1324.
87. S. J. Leuenroth; C. M. Crews, *Chem. Biol.* **2005**, *12*, 1259.
88. S. J. Leuenroth; D. Okuhara; J. D. Shotwell; G. S. Markowitz; Z. Yu; S. Somlo; C. M. Crews, *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 4389.
89. S. J. Leuenroth; N. Bencivenga; P. Igarashi; S. Somlo; C. M. Crews, *J. Am. Soc. Nephrol.* **2008**, *19*, 1659.
90. R. T. Luibrand; T. R. Erdman; J. J. Vollmer; P. J. Scheuer; J. Finer; J. Clardy, *Tetrahedron* **1979**, *35*, 609.
91. P. A. Takizawa; J. K. Yucel; B. Veit; D. J. Faulkner; T. Deerinck; G. Soto; M. Ellisman; V. Malhotra, *Cell* **1993**, *73*, 1079.
92. B. Veit; J. K. Yucel; V. Malhotra, *J. Cell Biol.* **1993**, *122*, 1197.
93. H. S. Radeke; C. A. Digits; R. L. Casaubon; M. L. Snapper, *Chem. Biol.* **1999**, *6*, 639.
94. K. L. Rinehart, Jr.; J. B. Gloer; J. C. Cook, Jr.; S. A. Mizsak; T. A. Scahill, *J. Am. Chem. Soc.* **1981**, *103*, 1857.
95. S. L. Crampton; E. G. Adams; S. L. Kuentzel; L. H. Li; G. Badiner; B. K. Bhuyan, *Cancer Res.* **1984**, *44*, 1796.
96. P. G. Canonico; W. L. Pannier; J. W. Huggins; K. L. Rienehart, *Antimicrob. Agents Chemother.* **1982**, *22*, 696.
97. D. W. Montgomery; C. F. Zukoski, *Transplantation* **1985**, *40*, 49.
98. L. H. Li; L. G. Timmins; T. L. Wallace; W. C. Krueger; M. D. Prairie; W. B. Im, *Cancer Lett.* **1984**, *23*, 279.
99. C. M. Crews; J. L. Collins; W. S. Lane; M. L. Snapper; S. L. Schreiber, *J. Biol. Chem.* **1994**, *269*, 15411.
100. B. Riis; S. I. S. Rattan; B. F. C. Clark, *Trends Biochem. Sci.* **1990**, *15*, 420.
101. B. V. SirDeshpande; P. L. Toogood, *Biochemistry* **1995**, *34*, 9177.
102. C. M. Crews; W. S. Lane; S. L. Schreiber, *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 4316.
103. L. Meng; N. Sin; C. M. Crews, *Biochemistry* **1998**, *37*, 10488.
104. M. D. Vera; M. M. Joullié, *Med. Res. Rev.* **2002**, *22*, 102.
105. S. C. Mayer; J. Ramanjulu; M. D. Vera; A. J. Pfizenmayer; M. M. Joullié, *J. Org. Chem.* **1994**, *59*, 5192.
106. P. Portonovo; X. Ding; M. S. Leonard; M. M. Joullié, *Tetrahedron* **2000**, *56*, 3687.
107. M. D. Vera; A. J. Pfizenmayer; X. Ding; D. Ahuja; P. L. Toogood; M. M. Joullié, *Bioorg. Med. Chem. Lett.* **2001**, *11*, 1871.
108. M. D. Vera; A. J. Pfizenmayer; X. Ding; D. Xiao; M. M. Joullié, *Bioorg. Med. Chem. Lett.* **2001**, *11*, 13.
109. P. T. Northcote; J. W. Blunt; M. H. G. Munro, *Tetrahedron Lett.* **1991**, *32*, 6411.
110. W.-K. Low; Y. Dang; T. Schneider-Poetsch; Z. Shi; N. S. Choi; W. C. Merrick; D. Romo; J. O. Liu, *Mol. Cell* **2005**, *20*, 709.
111. W.-K. Low; Y. Dang; S. Bhat; D. Romo; J. O. Liu, *Chem. Biol.* **2007**, *14*, 715.
112. E. Hao; J. Fromont; D. Jardine; P. Karuso, *Molecules* **2001**, *6*, 130.
113. L. Meijer; A. M. Thunnissen; A. W. White; M. Garnier; M. Nikolic; L. H. Tsai; J. Walter; K. E. Cleverley; P. C. Salinas; Y. Z. Wu; J. Biernat; E. M. Mandelkow; S. H. Kim; G. R. Pettit, *Chem. Biol.* **2000**, *7*, 51.
114. Y. Wan; W. Hur; C. Y. Cho; Y. Liu; F. J. Adrian; O. Lozach; S. Bach; T. Mayer; D. Fabbro; L. Meijer; N. S. Gray, *Chem. Biol.* **2004**, *11*, 247.
115. N. Kanoh; S. Kumashiro; S. Simizu; Y. Kondoh; S. Hatakeyama; H. Tashiro; H. Osada, *Angew. Chem. Int. Ed. Engl.* **2003**, *42*, 5584.

116. M. Platz; A. S. Admasu; S. Kwiatkowski; P. J. Crocker; N. Imai; D. S. Watt, *Bioconjug. Chem.* **1991**, *2*, 337.
117. M. Nassal, *J. Am. Chem. Soc.* **1984**, *106*, 7540.
118. J. Brunner; H. Senn; F. M. Richards, *J. Biol. Chem.* **1980**, *255*, 3313.
119. N. Kanoh; K. Honda; S. Simizu; M. Muroi; H. Osada, *Angew. Chem. Int. Ed. Engl.* **2005**, *44*, 3559.
120. J. E. Bradner; O. M. McPherson; A. N. Koehler, *Nat. Protoc.* **2006**, *1*, 2344.
121. S. N. Sehgal; H. Baker; C. Vézina, *J. Antibiotics* **1975**, *28*, 727.
122. K. Schmitz; S. J. Haggarty; O. M. McPherson; J. Clardy; A. N. Koehler, *J. Am. Chem. Soc.* **2007**, *129*, 11346.
123. B. F. Cravatt; A. T. Wright; J. W. Kozarich, *Ann. Rev. Biochem.* **2008**, *77*, 383.
124. K. Hanada; M. Tamai; S. Morimoto; T. Adachi; S. Ohmura; J. Sawada; I. Tanaka, *Agric. Biol. Chem.* **1978**, *42*, 537.
125. K. Hanada; M. Tamai; S. Ohmura; J. Sawada; T. Seki; I. Tanaka, *Agric. Biol. Chem.* **1978**, *42*, 529.
126. K. Hanada; M. Tamai; M. Yamagishi; S. Ohmura; J. Sawada; I. Tanaka, *Agric. Biol. Chem.* **1978**, *42*, 523.
127. D. Greenbaum; A. Baruch; L. Hayrapetian; Z. Darula; A. Burlingame; K. F. Medzihradszky; M. Bogyo, *Mol. Cell Proteomics* **2002**, *1*, 60.
128. P. A. Searle; T. F. Molinski, *J. Am. Chem. Soc.* **1995**, *117*, 8126.
129. C. J. Forsyth; F. Ahmed; R. D. Cink; C. S. Lee, *J. Am. Chem. Soc.* **1998**, *120*, 5597.
130. C. J. Forsyth; Y. Lu; J. Chen; J. J. La Clair, *J. Am. Chem. Soc.* **2006**, *128*, 3858.
131. A. D. Rodriguez; M. J. Lear; J. J. La Clair, *J. Am. Chem. Soc.* **2008**, *130*, 7256.
132. Z. Gitai; N. A. Dye; A. Reisenauer; M. Wachi; L. Shapiro, *Cell* **2005**, *12*, 329.
133. R. P. Walker; D. J. Faulkner; D. Van Engen; J. Clardy, *J. Am. Chem. Soc.* **1981**, *103*, 6772.
134. V. S. Bernan; D. M. Roll; C. M. Ireland; M. Greenstein; W. M. Maiese; D. A. Steinberg, *J. Antimicrob. Chemother.* **1993**, *32*, 539.
135. F. X. Zhou; J. Bonin; P. F. Predki, *Comb. Chem. High Throughput Screen.* **2004**, *7*, 539.
136. J. Ball; B. Schweitzer; P. Predki; M. Snyder, Development and Applications of Functional Protein Microarrays. In *Protein Microarrays*; M. Schena, Ed.; Jones and Bartlett Publishers: Sudbury, MA, 2005; p 421.
137. P. F. Predki, Ed., *Functional Protein Microarrays in Drug Discovery*; CRC Press: Boca Raton, FL, 2007.
138. Y. Sadakane; Y. Hatanaka, *Anal. Sci.* **2006**, *22*, 209.
139. J. R. Knowles, *Acc. Chem. Res.* **1972**, *5*, 155.
140. E. L. Vodovozova, *Biochemistry (Mosc.)* **2007**, *72*, 1.
141. A. Sinz, *ChemMedChem* **2007**, *2*, 425.
142. T. Tomohiro; M. Hashimoto; Y. Hatanaka, *Chem. Rec.* **2005**, *5*, 385.
143. Y. Hatanaka; Y. Sadakane, *Curr. Top. Med. Chem.* **2002**, *2*, 271.
144. G. Dormán; G. D. Prestwich, *Trends Biotech.* **2000**, *18*, 64.
145. A. Singh; E. R. Thornton; F. H. Westheimer, *J. Biol. Chem.* **1962**, *237*, 3006.
146. J. A. Mackintosh; H.-Y. Choi; S.-H. Bae; D. A. Veal; P. J. Bell; B. C. Ferrari; D. D. Van Dyk; N. M. Verrills; Y.-K. Paik; P. Karuso, *Proteomics* **2003**, *3*, 2273.
147. H. Nakayama; Y. Hatanaka; M. Taki; E. Yoshida; Y. Kanaoka, *Ann. N. Y. Acad. Sci.* **1993**, *707*, 349.
148. G. D. Prestwich; G. Dorman; J. T. Elliott; D. M. Marecak; A. Chaudhary, *Photochem. Photobiol.* **1997**, *65*, 222.
149. A. B. Theibert; G. D. Prestwich; T. R. Jackson; L. P. Hammonds-Odie, The Purification and Assay of Inositide Binding Proteins. In *Signalling by Inositides*; S. B. Shears, Ed.; Oxford University Press: London, 1997; p 117.
150. I. Ojima; O. Duclos; G. Dorman; B. Simonot; G. D. Prestwich; S. Rao; K. A. Lerro; S. B. Horwitz, *J. Med. Chem.* **1995**, *38*, 3891.
151. Q. Wu; P.-Y. Bounaud; S. D. Kuduk; C.-P. H. Yang; I. Ojima; S. B. Horwitz; G. A. Orr, *Biochemistry* **1998**, *37*, 11272.
152. H. P. Sigg; H. P. Weber, *Helv. Chim. Acta* **1968**, *51*, 1395.
153. E. C. Griffith; Z. Su; B. E. Turk; S. Chen; Y.-H. Chang; Z. Wu; K. Biemann; J. O. Liu, *Chem. Biol.* **1997**, *4*, 461.
154. A. Addlagatta; B. W. Matthews, *Protein Sci.* **2006**, *15*, 1842.
155. P. Selvakumar; A. Lakshmikuttyamma; J. R. Dimmock; R. K. Sharma, *Biochim. Biophys. Acta – Rev. Cancer* **2006**, *1765*, 148.
156. H. S. Radeke; M. L. Snapper, *Bioorg. Med. Chem.* **1998**, *6*, 1227.
157. R. F. Keeler; W. Binns, *Teratology* **1968**, *1*, 5.
158. R. G. Cooper; C. J. Etheridge; L. Stewart; J. Marshall; S. Rudginsky; S. H. Cheng; A. D. Miller, *Chem. Eur. J.* **1998**, *4*, 137.
159. J. K. Chen; J. Taipale; M. K. Cooper; P. A. Beachy, *Genes Dev.* **2002**, *16*, 2743.
160. S. Shibata; M. Nakahara, *Chem. Pharm. Bull.* **1963**, *11*, 372.
161. W.-W. Qiu; J. Xu; D.-Z. Liu; J.-Y. Li; Y. Ye; X.-Z. Zhu; J. Li; F.-J. Nan, *Bioorg. Med. Chem. Lett.* **2006**, *16*, 3306.
162. D.-Z. Liu; K.-Q. Xie; X.-Q. Ji; Y. Ye; C.-L. Jiang; X.-Z. Zhu, *Br. J. Pharmacol.* **2005**, *146*, 604.
163. F. L. Hsu; C. W. Lai; J. T. Cheng, *Planta Med.* **1997**, *63*, 323.
164. S. P. Gygi; B. Rist; S. A. Gerber; F. Turecek; M. H. Gelb; R. Aebersold, *Nat. Biotechnol.* **1999**, *17*, 994.
165. S. M. Lamos; C. J. Krusemark; C. J. McGee; M. Scalf; L. M. Smith; P. J. Belshaw, *Angew. Chem. Int. Ed. Engl.* **2006**, *45*, 4329.
166. J. Liu; J. D. Farmer, Jr.; W. S. Lane; J. Friedman; I. Weissman; S. L. Schreiber, *Cell* **1991**, *66*, 807.
167. S. E. Ong; B. Blagoev; I. Kratchmarova; D. B. Kristensen; H. Steen; A. Pandey; M. Mann, *Mol. Cell Proteomics* **2002**, *1*, 376.
168. S. E. Ong; M. Mann, *Nat. Protoc.* **2006**, *1*, 2650.
169. P. L. Ross; Y. N. Huang; J. N. Marchese; B. Williamson; K. Parker; S. Hattan; N. Khainovski; S. Pillai; S. Dey; S. Daniels; S. Purkayastha; P. Juhasz; S. Martin; M. Bartlet-Jones; F. He; A. Jacobson; D. J. Pappin, *Mol. Cell Proteomics* **2004**, *3*, 1154.
170. L. R. Zieske, *J. Exp. Bot.* **2006**, *57*, 1501.
171. M. Bantscheff; D. Eberhard; Y. Abraham; S. Bastuck; M. Boesche; S. Hobson; T. Mathieson; J. Perrin; M. Raida; C. Rau; V. Reader; G. Sweetman; A. Bauer; T. Bouwmeester; C. Hopf; U. Kruse; G. Neubauer; N. Ramsden; J. Rick; B. Kuster; G. Drewes, *Nat. Biotechnol.* **2007**, *25*, 1035.
172. S. E. Ong; M. Schenone; A. A. Margolin; X. Li; K. Do; M. K. Doud; D. R. Mani; L. Kuai; X. Wang; J. L. Wood; N. J. Tolliday; A. N. Koehler; L. A. Marcaurelle; T. R. Golub; R. J. Gould; S. L. Schreiber; S. A. Carr, *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 4617.
173. Y. Oda; T. Owa; T. Sato; B. Boucher; S. Daniels; H. Yamanaka; Y. Shinohara; A. Yokoi; J. Kuromitsu; T. Nagasu, *Anal. Chem.* **2003**, *75*, 2159.

174. R. A. Williamson; G. J. Silverman, Gene Fragment Libraries and Genomic and cDNA Expression Cloning. In *Phage Display: A Laboratory Manual*; C. F. Barbas, III, D. R. Burton, J. K. Scott, G. J. Silverman, Eds.; Cold Spring Harbor Laboratory Press: New York, 2001; p 6.1.
175. N. Matsumura; N. Doi; H. Yanagawa, *Curr. Proteomics* **2006**, *3*, 199.
176. A. Sergeeva; M. G. Kolonin; J. J. Molldrem; R. Pasqualini; W. Arap, Gene Fragment Libraries and Genomic and CDNA Expression Cloning. *Adv. Drug Deliv. Rev.* **2006**, *58*, 1622.
177. D. J. Rodi; G. E. Agoston; R. Manon; R. Lapcevich; S. J. Green; L. Makowski, *Comb. Chem. High Throughput Screen.* **2001**, *4*, 553.
178. G. P. Smith, *Science* **1985**, *228*, 1315.
179. L. S. Jespers; J. H. Messens; A. De Keyser; D. Eeckhout; I. Van Den Brande; Y. G. Gansemans; M. J. Lauwereys; G. P. Vlasuk; P. E. Stanssens, *Biotechnology (N.Y.)* **1995**, *13*, 378.
180. R. Crameri; M. Suter, *Gene* **1993**, *137*, 69.
181. D. J. Rodi; R. W. Janes; H. J. Sanganee; R. A. Holton; B. A. Wallace; L. Makowski, *J. Mol. Biol.* **1999**, *285*, 197.
182. A. Rosenberg; K. Griffin; F. W. Studier; M. McCormick; J. Berg; R. Novy; R. Mierendorf, *Innovations* **1996**, *6*, 1.
183. I. N. Maruyama; H. I. Maruyama; S. Brenner, *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 8273.
184. V. P. Efimov; I. V. Nepluev; V. V. Mesyanzhinov, *Virus Genes* **1995**, *10*, 173.
185. P. P. Sche; K. M. McKenzie; J. D. White; D. J. Austin, *Chem. Biol.* **1999**, *6*, 707.
186. K. P. Wilson; M. M. Yamashita; M. D. Sintchak; S. H. Rotstein; M. A. Murcko; J. Boger; J. A. Thomson; M. J. Fitzgibbon; J. R. Black; M. A. Navia, *Acta Crystallogr. D Biol. Crystallogr.* **1995**, *51*, 511.
187. P. P. Sche; K. M. McKenzie; J. D. White; D. J. Austin, *Chem. Biol.* **2001**, *8*, 399.
188. S. Danner; J. G. Belasco, *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 12954.
189. K. M. McKenzie; E. J. Videlock; U. Splittgerber; D. J. Austin, *Angew. Chem. Int. Ed. Engl.* **2004**, *43*, 4052.
190. S. G. Gómez; J. A. Bueren; G. T. Faircloth; J. Jimeno; B. Albella, *Exp. Hematol.* **2003**, *31*, 1104.
191. A. P. Brown; R. L. Morrissey; G. T. Faircloth; B. S. Levine, *Cancer Chemother. Pharmacol.* **2002**, *50*, 333.
192. S. Martin-Algarra; E. Espinosa; J. Rubio; J. J. L. Lopez; J. L. Manzano; L. A. Carrion; A. Plazaola; A. Tanovic; L. Paz-Ares, *Eur. J. Cancer* **2009**, *45*, 732.
193. B. Pardo; L. Paz-Ares; J. Tabernero; E. Ciruelos; M. Garcia; R. Salazar; A. Lopez; M. Blanco; A. Nieto; J. Jimeno; M. A. Izquierdo; J. M. Trigo, *Clin. Cancer Res.* **2008**, *14*, 1116.
194. M. T. Hamann; P. J. Scheuer, *J. Am. Chem. Soc.* **1993**, *115*, 5825.
195. M. T. Hamann; C. S. Otto; P. J. Scheuer, *J. Org. Chem.* **1996**, *61*, 6594.
196. P. Karuso, *Bioorg. Mar. Chem.* **1987**, *1*, 31.
197. K. Shiba, *J. Drug Target.* **2006**, *14*, 512.
198. A. M. Piggott; P. Karuso, *ChemBioChem* **2008**, *9*, 524.
199. W. Boehmerle; U. Splittgerber; M. B. Lazarus; K. M. McKenzie; D. G. Johnston; D. J. Austin; B. E. Ehrlich, *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 18356.
200. A. Di Marco; M. Gaetani; B. Scarpinato, *Cancer Chemother. Rep.* **1969**, *53*, 33.
201. F. Zunino; R. Gambetta; A. Di Marco; A. Zaccara, *Biochim. Biophys. Acta* **1972**, *277*, 489.
202. R. Y. Chuang; L. F. Chuang, *Biochemistry* **1979**, *18*, 2069.
203. K. M. Tewey; T. C. Rowe; L. Yang; B. D. Halligan; L. F. Liu, *Science* **1984**, *226*, 466.
204. Y. Jin; J. Yu; Y. G. Yu, *Chem. Biol.* **2002**, *9*, 157.
205. U. T. Meier; G. Blobel, *J. Cell Biol.* **1990**, *111*, 2235.
206. U. T. Meier; G. Blobel, *Cell* **1992**, *70*, 127.
207. C.-Y. Pai; H.-K. Chen; H.-L. Sheu; N.-H. Yeh, *J. Cell Sci.* **1995**, *108*, 1911.
208. Y.-K. Kim; Y. Jin; K. M. Vukoti; J. K. Park; E. E. Kim; K.-J. Lee; Y. G. Yu, *Protein Expr. Purif.* **2003**, *31*, 260.
209. A. M. Piggott; P. Karuso, *Mar. Drugs* **2005**, *3*, 36.
210. D. A. T. Gearhart; F. Patricia; J. Warren Beach, *Neurosci. Res.* **2002**, *44*, 255.
211. G. Rontó; M. M. Agamalyan; G. M. Drabkin; L. A. Feigin; Y. M. Lvov, *Biophys. J.* **1983**, *43*, 309.
212. E. J. Videlock; V. K. Chung; M. A. Mohan; T. M. Strok; D. J. Austin, *J. Am. Chem. Soc.* **2004**, *126*, 3730.
213. A. M. Piggott; P. Karuso, *Tetrahedron Lett.* **2005**, *46*, 8241.
214. J. S. Shim; J. Lee; H.-J. Park; S.-J. Park; H. J. Kwon, *Chem. Biol.* **2004**, *11*, 1455.
215. S. T. Baek; D.-U. Kim; S. Han; I. S. Woo; M. Nam; L. Kim; K.-S. Heo; H. Lee; H.-R. Hwang; S.-J. Choi; M. Won; M. Lee; S.-K. Park; S. Lee; H.-J. Kwon; P. J. Maeng; H.-M. Park; Y. Park; D. Kim; K.-L. Hoe, *J. Microbiol. Biotechnol.* **2008**, *18*, 263.
216. K. Arai; S. Shimizu; Y. Taguchi; Y. Yamamoto, *Chem. Pharm. Bull.* **1981**, *29*, 991.
217. Y. Yamamoto; K. Nishimura; N. Kiriyama, *Chem. Pharm. Bull.* **1976**, *24*, 1853.
218. H. Kim; L. Deng; X. Xiong; W. D. Hunter; M. C. Long; M. C. Pirrung, *J. Med. Chem.* **2007**, *50*, 3423.
219. J. Min; Y. Kyung Kim; P. G. Cipriani; M. Kang; S. M. Khersonsky; D. P. Walsh; J. Y. Lee; S. Niessen; J. R. Yates, III; K. Gunsalus; F. Piano; Y. T. Chang, *Nat. Chem. Biol.* **2007**, *3*, 55.
220. S. N. Savinov; D. J. Austin, *Comb. Chem. High Throughput Screen.* **2001**, *4*, 593.
221. Q.-L. He; H. Jiang; F. Zhang; H.-B. Chen; G.-L. Tang, *Bioorg. Med. Chem. Lett.* **2008**, *18*, 3995.
222. Y. Takakusagi; K. Ohta; K. Kuramochi; K. Morohashi; S. Kobayashi; K. Sakaguchi; F. Sugawara, *Bioorg. Med. Chem. Lett.* **2005**, *15*, 4846.
223. Y. Takakusagi; S. Kobayashi; F. Sugawara, *Bioorg. Med. Chem. Lett.* **2005**, *15*, 4850.
224. K. Morohashi; A. Yoshino; A. Yoshimori; S. Saito; S. Tanuma; K. Sakaguchi; F. Sugawara, *Biochem. Pharmacol.* **2005**, *70*, 37.
225. K. Morohashi; T. Arai; S. Saito; M. Watanabe; K. Sakaguchi; F. Sugawara, *Comb. Chem. High Throughput Screen.* **2006**, *9*, 55.
226. Y. Takakusagi; K. Takakusagi; K. Kuramochi; S. Kobayashi; F. Sugawara; K. Sakaguchi, *Bioorg. Med. Chem.* **2007**, *15*, 7590.
227. Y. Takakusagi; K. Kuramochi; M. Takagi; T. Kusayanagi; D. Manita; H. Ozawa; K. Iwakiri; K. Takakusagi; Y. Miyano; A. Nakazaki; S. Kobayashi; F. Sugawara; K. Sakaguchi, *Bioorg. Med. Chem.* **2008**, *16*, 9837.
228. Y. Takakusagi; Y. Kuroiwa; F. Sugawara; K. Sakaguchi, *Bioorg. Med. Chem.* **2008**, *16*, 7410.

229. K. Nishiyama; Y. Takakusagi; T. Kusayanagi; Y. Matsumoto; S. Habu; K. Kuramochi; F. Sugawara; K. Sakaguchi; H. Takahashi; H. Natsugari; S. Kobayashi, *Bioorg. Med. Chem.* **2009**, *17*, 195.
230. S. P. Goff, Retroviridae: The Retroviruses and their Replication. In *Fields' Virology*, 5th ed.; P. M. Howley, D. M. Knipe, M. L. Nibert, L. A. Schiff, Eds.; Lippincott Williams & Wilkins: Philadelphia, 2007; Vol. 2, p 1999.
231. C. J. Buchholz; L. J. Duerner; S. Funke; I. C. Schneider, *Comb. Chem. High Throughput Screen.* **2008**, *11*, 99.
232. P. D. Khare; A. G. Rosales; K. R. Bailey; S. J. Russell; M. J. Federspiel, *Virology* **2003**, *315*, 313.
233. S. Fields; O.-K. Song, *Nature* **1989**, *340*, 245.
234. C. T. Chien; P. L. Bartel; R. Sternglanz; S. Fields, *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 9578.
235. E. J. Licitra; J. O. Liu, *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 12817.
236. K. S. de Felipe; B. T. Carter; E. A. Althoff; V. W. Cornish, *Biochemistry* **2004**, *43*, 10353.
237. F. Becker; K. Murthi; C. Smith; J. Come; N. Costa-Roldan; C. Kaufmann; U. Hanke; C. Degenhart; S. Baumann; W. Wallner; A. Huber; S. Dedier; S. Dill; D. Kinsman; M. Hediger; N. Bockovich; S. Meier-Ewert; A. F. Kluge; N. Kley, *Chem. Biol.* **2004**, *11*, 211.
238. M. P. Schreuder; S. Brekelmans; H. van den Ende; F. M. Klis, *Yeast* **1993**, *9*, 399.
239. A. Kondo; M. Ueda, *Appl. Microbiol. Biotechnol.* **2004**, *64*, 28.
240. L. R. Pepper; Y. K. Cho; E. T. Boder; E. V. Shusta, *Comb. Chem. High Throughput Screen.* **2008**, *11*, 127.
241. A. Wadle; A. Mischo; J. Imig; B. Wuellner; D. Hensel; K. Waetzig; F. Neumann; B. Kubuschok; W. Schmidt; L. J. Old; M. Pfreundschuh; C. Renner, *Int. J. Cancer* **2005**, *117*, 104.
242. S. Bidlingmaier; B. Liu, *Mol. Cell Proteomics* **2006**, *5*, 533.
243. D. R. Bowley; A. F. Labrijn; M. B. Zwick; D. R. Burton, *Protein Eng. Des. Sel.* **2007**, *20*, 81.
244. Z. Wang; A. Mathias; S. Stavrou; D. M. Neville, Jr., *Protein Eng. Des. Sel.* **2005**, *18*, 337.
245. R. D. Blake; P. W. Hinds; S. Earley; A. L. Hillyard; G. R. Day, In *Proceedings of the Fourth Conversation in the Discipline Biomolecular Stereodynamics*; R. H. Sarma, N. H. Sarma, Eds.; State University of New York at Albany, Adenine Press: Guilderland, NY, 4–8 June, 1985; Biomolecular Stereodynamics, Vol. 4, p 271.
246. A. Charbit; J. C. Boulain; A. Ryter; M. Hofnung, *EMBO J.* **1986**, *5*, 3029.
247. P. S. Daugherty, *Curr. Opin. Struct. Biol.* **2007**, *17*, 474.
248. H.-C. Jung; J.-M. Lebeault; J.-G. Pan, *Nat. Biotechnol.* **1998**, *16*, 576.
249. S. Y. Lee; J. H. Choi; Z. Xu, *Trends Biotechnol.* **2003**, *21*, 45.
250. J. J. Rice; A. Schohn; P. H. Bessette; K. T. Boulware; P. S. Daugherty, *Protein Sci.* **2006**, *15*, 825.
251. M. Lunder; T. Bratkovic; B. Doljak; S. Kreft; U. Urleb; B. Strukelj; N. Plazar, *Appl. Biochem. Biotechnol.* **2005**, *127*, 125.
252. C. Kurland; J. Gallant, *Curr. Opin. Biotechnol.* **1996**, *7*, 489.
253. E. Veiga; V. de Lorenzo; L. A. Fernández, *Mol. Microbiol.* **2004**, *52*, 1069.
254. N. Narayanan; C. P. Chou, *Biotechnol. Prog.* **2008**, *24*, 293.
255. N. Kronqvist; J. Loefblom; A. Jonsson; H. Wernerus; S. Staahl, *Protein Eng. Des. Sel.* **2008**, *21*, 247.
256. B. Seed; A. Aruffo, *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 3365.
257. M. Ho; S. Nagata; I. Pastan, *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 9637.
258. M. Ho; S. Nagata; I. Pastan, *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 14543.
259. C. Tuerk; L. Gold, *Science* **1990**, *249*, 505.
260. A. D. Ellington; J. W. Szostak, *Nature* **1992**, *355*, 850.
261. H.-J. Thiesen; C. Bach, *Nucleic Acids Res.* **1990**, *18*, 3203.
262. L. C. Mattheakis; R. R. Bhatt; W. J. Dower, *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 9022.
263. G. M. Gersuk; M. J. Corey; E. Corey; J. E. Stray; G. H. Kawasaki; R. L. Vessella, *Biochem. Biophys. Res. Commun.* **1997**, *232*, 578.
264. M. He; M. J. Taussig, *Nucleic Acids Res.* **1997**, *25*, 5132.
265. J. Hanes; A. Plückthun, *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 4937.
266. K. C. Keiler; P. R. H. Waller; R. T. Sauer, *Science* **1996**, *271*, 990.
267. J. Hanes; L. Jermutus; S. Weber-Bornhauser; H. R. Bosshard; A. Pluckthun, *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 14130.
268. J.-M. Zhou; S. Fujita; M. Warashina; T. Baba; K. Taira, *J. Am. Chem. Soc.* **2002**, *124*, 538.
269. S. Fujita; S. Y. Sawata; R. Yamamoto-Fujita; Y. Endo; H. Kise; M. Iwakura; K. Taira, *J. Med. Chem.* **2002**, *45*, 1598.
270. S. Y. Sawata; K. Taira, *Protein Eng.* **2003**, *16*, 1115.
271. M. He; M. J. Taussig, *Nat. Methods* **2007**, *4*, 281.
272. S. Y. Sawata; E. Suyama; K. Taira, *Protein Eng. Des. Sel.* **2004**, *17*, 501.
273. T. Matsuura; A. Pluckthun, *FEBS Lett.* **2003**, *539*, 24.
274. J. Hanes; L. Jermutus; A. Pluckthun, *Methods Enzymol.* **2000**, *328*, 404.
275. T. Matsuura; A. Plueckthun, *Orig. Life Evol. Biosph.* **2004**, *34*, 151.
276. P. Amstutz; J. N. Pelletier; A. Guggisberg; L. Jermutus; S. Cesaro-Tadic; C. Zahnd; A. Plückthun, *J. Am. Chem. Soc.* **2002**, *124*, 9396.
277. B. Schimmele; N. Graefe; A. Plueckthun, *Protein Eng. Des. Sel.* **2005**, *18*, 285.
278. T. Kanagawa, *J. Biosci. Bioeng.* **2003**, *96*, 317.
279. L. Gold, *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 4825.
280. T. T. Takahashi; R. J. Austin; R. W. Roberts, *Trends Biochem. Sci.* **2003**, *28*, 159.
281. D. W. Allen; P. C. Zamecnik, *Biochim. Biophys. Acta* **1962**, *55*, 865.
282. D. Nathans, *Antibiotics (USSR)* **1967**, *1*, 259.
283. R. W. Roberts; J. W. Szostak, *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 12297.
284. N. Nemoto; E. Miyamoto-Sato; Y. Husimi; H. Yanagawa, *FEBS Lett.* **1997**, *414*, 405.
285. M. Kurz; K. Gu; P. A. Lohse, *Nucleic Acids Res.* **2000**, *28*, E83.
286. H. S. Jois; B. L. Manjunath; S. V. Rao, *J. Indian Chem. Soc.* **1933**, *10*, 41.
287. F. Dall'Acqua; S. Marciani; G. Rodighiero, *FEBS Lett.* **1970**, *9*, 121.

288. E. Miyamoto-Sato; H. Takashima; S. Fuse; K. Sue; M. Ishizaka; S. Tateyama; K. Horisawa; T. Sawasaki; Y. Endo; H. Yanagawa, *Nucleic Acids Res.* **2003**, *31*, e78.
289. P. W. Hammond; J. Alpin; C. E. Rise; M. Wright; B. L. Kreider, *J. Biol. Chem.* **2001**, *276*, 20898.
290. K. Horisawa; S. Tateyama; M. Ishizaka; K. M. N.; Horisawa; H. Takashima; E. Miyamoto-Sato; N. Doi; H. Yanagawa, *Nucleic Acids Res.* **2004**, *32*, e169.
291. E. Miyamoto-Sato; M. Ishizaka; K. Horisawa; S. Tateyama; H. Takashima; S. Fuse; K. Sue; N. Hirai; K. Masuoka; H. Yanagawa, *Genome Res.* **2005**, *15*, 710.
292. X. Shen; C. A. Valencia; J. Szostak; B. Dong; R. Liu, *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 5969.
293. S. Tateyama; K. Horisawa; H. Takashima; E. Miyamoto-Sato; N. Doi; H. Yanagawa, *Nucleic Acids Res.* **2006**, *34*, e27.
294. M. McPherson; Y. Yang; P. W. Hammond; B. L. Kreider, *Chem. Biol.* **2002**, *9*, 691.
295. N. Doi; H. Takashima; A. Wada; Y. Oishi; T. Nagano; H. Yanagawa, *J. Biotechnol.* **2007**, *131*, 231.
296. J. O. Bishop; J. G. Morton; M. Rosbash; M. Richardson, *Nature* **1974**, *250*, 199.
297. K. Horisawa; N. Doi; H. Yanagawa, *PLoS One* **2008**, *3*, e1646.
298. V. Taly; B. T. Kelly; A. D. Griffiths, *Chembiochem* **2007**, *8*, 263.
299. D. S. Tawfik; A. D. Griffiths, *Nat. Biotechnol.* **1998**, *16*, 652.
300. N. Doi; H. Yanagawa, *FEBS Lett.* **1999**, *457*, 227.
301. M. Yonezawa; N. Doi; Y. Kawahashi; T. Higashinakagawa; H. Yanagawa, *Nucleic Acids Res.* **2003**, *31*, e118/1.
302. M. Yonezawa; N. Doi; T. Higashinakagawa; H. Yanagawa, *J. Biochem.* **2004**, *135*, 285.
303. A. Sepp; D. S. Tawfik; A. D. Griffiths, *FEBS Lett.* **2002**, *532*, 455.
304. J. Bertschinger; D. Neri, *Protein Eng. Des. Sel.* **2004**, *17*, 699.
305. J. Bertschinger; D. Grabulovski; D. Neri, *Protein Eng. Des. Sel.* **2007**, *20*, 57.
306. H. Reiersen; I. Løbersli; G.Å. Løset; E. Hvattum; B. Simonsen; J. E. Stacy; D. McGregor; K. FitzGerald, *Nucleic Acids Res.* **2005**, *33*, e10/1.
307. M. Hashimoto; B. Mayers; P. Garstecki; G. M. Whitesides, *Small* **2006**, *2*, 1292.
308. D. B. Weibel; G. M. Whitesides, *Curr. Opin. Chem. Biol.* **2006**, *10*, 584.
309. G. M. Whitesides, *Nature* **2006**, *442*, 368.
310. A. J. You; R. J. Jackman; G. M. Whitesides; S. L. Schreiber, *Chem. Biol.* **1997**, *4*, 969.
311. R. E. Speight; D. J. Hart; J. D. Sutherland; J. M. Blackburn, *Chem. Biol.* **2001**, *8*, 951.
312. M. G. Cull; J. F. Miller; P. J. Schatz, *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 1865.
313. Y. S. Choi; S. P. Pack; Y. J. Yoo, *Biotechnol. Lett.* **2005**, *27*, 1707.
314. H. J. Muller, *Science* **1927**, *66*, 84.
315. C. Cogoni; G. Macino, *Curr. Opin. Genet. Dev.* **2000**, *10*, 638.
316. T. Gura, *Nature* **2000**, *404*, 804.
317. S. M. Hammond; A. A. Caudy; G. J. Hannon, *Nat. Rev. Genet.* **2001**, *2*, 110.
318. E. Anderson; Q. Boese; A. Khvorova; J. Karpilow, *Methods Mol. Biol.* **2008**, *442*, 45.
319. V. S. Gomase; S. Tagore, *Curr. Drug Metab.* **2008**, *9*, 241.
320. A. Kourtidis; C. Eifert; D. S. Conklin, *Ernst Schering Res. Found. Workshop* **2007**, *61*, 1.
321. G. Giaever; D. D. Shoemaker; T. W. Jones; H. Liang; E. A. Winzeler; A. Astromoff; R. W. Davis, *Nat. Genet.* **1999**, *21*, 278.
322. D. Xu; B. Jiang; T. Ketela; S. Lemieux; K. Veillette; N. Martel; J. Davison; S. Sillaots; S. Trosok; C. Bachewich; H. Bussey; P. Youngman; T. Roemer, *PLoS Pathog.* **2007**, *3*, 835.
323. K. Baetz; L. McHardy; K. Gable; T. Tarling; D. Reberioux; J. Bryan; R. J. Andersen; T. Dunn; P. Hieter; M. Roberge, *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 4525.
324. K. Balasubramanyam; R. A. Varier; M. Altaf; V. Swaminathan; N. B. Siddappa; U. Ranga; T. K. Kundu, *J. Biol. Chem.* **2004**, *279*, 51163.
325. L. Castagnoli; A. Zucconi; M. Quondam; M. Rossi; P. Vaccaro; S. Panni; S. Paoluzi; E. Santonico; L. Dente; G. Cesareni, *Comb. Chem. High Throughput Screen.* **2001**, *4*, 121.

Biographical Sketch

Peter Karuso was born in Sydney, Australia, and obtained his B.Sc. degree in organic chemistry from the University of Sydney. He went on to complete a Ph.D. under the guidance of W. C. Taylor in the Department of Organic Chemistry at the University of Sydney on Natural Products Chemistry, specifically marine natural products and the total synthesis of an aporphine alkaloid. After 6 years of postdoctoral research in New Zealand, the United States, and Germany as an Alexander von Humboldt Fellow, he returned to Sydney to take up his current position in the School of Chemistry at Macquarie University. His current research interests include the biomimetic synthesis of natural products, marine natural product isolation and structure elucidation, reverse chemical proteomics, and the development of new fluorescence technologies. His work on a fluorescent fungal natural product led to the establishment of a spin-off company (Fluorotechnics) that markets a range of products for proteomics and cellomics, all based on natural products. He is on the editorial board of four journals and a Fellow of the Royal Australian Chemical Institute and published over 70 papers in the general area of natural products chemistry and biology.

9.15 Bioinformatics

Yi-Ping Phoebe Chen, Deakin University, Melbourne, VIC, Australia

Elena P. Ivanova and Feng Wang, Swinburne University of Technology, Melbourne, VIC, Australia

Paolo Carloni, International School for Advanced Studies, Trieste, Italy

9.15.1 Introduction

Bioinformatics can be broadly described as the application of information technology to the field of molecular biology. It aims to solve practical problems arising from the management and analysis of biological data, such as natural products with pharmacological or biological activities.

In pharmacology, natural products have provided the inspiration for most of the active ingredients in medicines and have been the most productive source of leads for new drugs. For example, around 80% of medicinal products up to 1996 were either directly derived from naturally occurring compounds or were inspired by a natural product,[1] and more recent analyses confirm the continuing importance of natural products for drug discovery.[2] In an extensive review of new drugs introduced between 1981 and 2002, 28% of the 868 new chemical entities were natural products or derived from natural products, with another 24% created around a pharmacophore from a natural product.[3] In addition to launched products, at least 70 natural product-related compounds were in clinical trials in 2004[4] and exploration of the bioactivity of natural products continues to provide novel chemical scaffolds for further drug inventions.[4,5]

New approaches to source novel compounds from untapped areas of biodiversity coupled with the technical advances in analytical techniques (such as microcoil NMR and linked LC–MS–NMR) have removed many of the difficulties when using natural products in screening campaigns. As the 'chemical space' occupied by natural products is both more varied and more drug-like than that of combinatorial chemical collections, synthetic and biosynthetic methods are being developed to produce screening libraries of natural product-like compounds. A renaissance of drug discovery inspired by natural products can be predicted.[2]

There is an increasing need for the development and analysis of novel biologically active natural products. This is due to the emergence of resistant strains of pathogens, the appearance of new diseases, and the inadequacy or toxicity of current drugs.[6] In the search for novel and bioactive molecules for drug discovery, the source is largely from three major groups: the plant kingdom, the microbial world, and the marine world. In the last 30 years, there has been an explosive growth of natural products from the marine world that have been characterized by the isolation and structure elucidation of very diverse structures with no precedent in terrestrial natural products. For example, natural products from marine prokaryotes that are available in marine environments provide a tremendous opportunity for the discovery of novel therapeutic agents. Furthermore, the primary chemical diversity available from marine prokaryotes is most likely capable of delivering an even greater abundance of natural products. These include natural products from marine prokaryotes; in fact, the primary chemical diversity available from them is most likely capable of delivering a great abundance of active compounds. Complex structure and stereochemistry of these compounds limit efficient synthetic production and leave microbial fermentation as the predominant economical route of these products.[6]

Marine resources that represent approximately half of the global biodiversity offer a colossal resource for novel bacterial species that produce useful natural products.[7–10] Such tremendous bacterial diversity may be explained by the fact that bacteria are nearly as old as our planet, since these organisms have inhabited the Earth for more than 3.5 Ga (10^9 years). This fact is also a key to explain this massive diversity and evolution of bacterial metabolic pathways developed as survival, competition, and/or defense strategies, which, in turn, are translated in the production and secretion of natural products.[11]

Extracting information from gene expression data can be defined as a process of quantitative analysis of messenger RNA (mRNA) expression levels of genes under the influence of drug or disease perturbations. The advancement in DNA microarray technology has enabled simultaneous measurement of the expression levels of thousands of genes. This simultaneous quantification of expression levels for a large number of genes is an extremely important step in biology as it connects the static gene structure mappings to dynamic gene functional analysis for an experimental biological/pathological process (i.e., how genes react to the induction of drug or disease). The experimental results are accumulated in databases for many organisms, and the databases are growing continuously with the rapid progress of biological research. As a result, we face an overwhelming quantity of complex data and data structures. This makes mining and analyzing of gene expression data very difficult.

Current database technology cannot deal with gene expression data adequately because relational database query systems work on exact matches. That is, the query system returns data items that match the query keywords exactly. However, in mining gene expression databases, we are more inclined to perform the following type of queries: given a new gene expression pattern, find all items in the database that have similar gene expression profiles. In this case, the match is not exact; we are interested in returns that have similar (or partially similar) gene expression profiles from different biological processes.

Data mining is a research approach that combines database fundamentals and artificial intelligence techniques to reduce the complexity of data and to discover meaningful, useful patterns and relationships in data. Therefore, data mining is a potentially useful technique that can overcome the challenges encountered while searching for information from large sets of gene expression data. Data mining involves pattern recognition and pattern discovery. Traditionally, in bioinformatics, pattern recognition is most often concerned with the automatic classification of character sequences representative of the nucleotide bases or molecular structures, and 3D protein structures. Clustering techniques are often used to form the basis of pattern recognition. The pattern matching and pattern discovery components of data mining are often performed by machine learning techniques. The major focus of this chapter is, therefore, placed on the clustering techniques and some machine learning techniques used in extracting information from gene expression data.

Natural products control an enormous number of cellular functions by binding to their target macromolecules. This key event triggers complex cellular pathways characterized by intermolecular interactions between macromolecules. Unfortunately, however, structural information on the various components is often lacking, rendering a quantitative molecular description of pathways that are impossible. The problem is even more stringent for membrane proteins, for which very little structural information is available, although they constitute 30% of the total proteins expressed by the human genome.

Structural biology (SB) in bioinformatics (SB) is then the key approach for structural predictions. It uses biological concepts such as Darwinian evolution-based algorithms, along with algorithms taken from the theory of the information, to analyze sequences of biopolymers for predicting 3D structural models. That is, the structures of target proteins can be modeled using proteins with sizeable sequence identity as template.

As a case study, we exemplify the use of SB for the characterization of the largest membrane receptors family, the G-protein-coupled receptors (GPCRs). GPCRs are involved in an enormous variety of intra- and extracellular signaling, including detection of light, sense of smell, neurotransmission, inflammation, and cardiac and smooth muscle contractility.[12–13] They are of the utmost pharmaceutical relevance, being the targets of almost 30% of all marketed drugs.[14–15]

More than half of the GPCRs (about 900) are odorant receptors (ORs).[16] This points the crucial role of the sense of smell during evolution. For the last few years, we have been involved in the characterization of the membrane proteins involved in the OR pathway. Such investigation is succinctly presented here.

9.15.2 Measuring Biodiversity

9.15.2.1 Brief Overview of Natural Products of Microbial Origin

Among natural products of microbial origin a number of unique classes have been described; to name a few: the macrolactins,[16] highly brominated pyrrole compounds,[17] and antitumor depsipeptides.[18] One of the prominent groups of natural producers belongs to *Alteromonas or Pseudoalteromonas* and related bacteria of the *Gammaproteobacteria* class. Intensive research has demonstrated that these organisms are capable of producing a wide range of biologically active secondary metabolites (BASMs), for example, enzymes, antibiotics, cytotoxins, antibactericidal, bacteriolytic, autotoxic, antifouling, and biocontrolled compounds that are active against invertebrate larvae, algal spores, fungi, and diatoms.[9,19] Examples of characterized BASM comprise some of the first antibiotics isolated and characterized: 2,3,4-tribromo-5 (1′-hydroxy-2′,4′-dibromophenyl) pyrrole as extracted from a marine bacterium of seagrass *Thalassia*;[17] tetrabromopyrrole, 2-(2′-hydroxy-3′,5′-dibromophenyl)-3,4,5-tribromopyrrole, and hexabromo-2,2′-bipyrrole from *Chromobacterium*;[20] brominated compounds and cell-bound polyanionic antibiotics from *Pseudoalteromonas luteoviolacea*;[21] korormicin and thiomarinols from *Pseudoalteromonas* sp. F-420 and *Alteromonas rava*;[22] and 3,3′,5,5′-tetrabromo-2,2′-biphenyldiol from *Pseudoalteromonas phenolica*.[23]

Owing to the fact that the primary source of antibacterial agents for biomedical applications are secondary metabolites and chemical derivates of actinomycetes, fungi, and certain soil bacteria such as myxobacteria, most of the recent advances in molecular engineering have been made in relation to polyketide and peptide antibiotics (e.g., erythromycin and vancomycin).[24] Recent literature on this topic suggests a greater mechanistic diversity in biosynthetic potential of bacterial natural products that was believed previously.

The formation of natural products by *Archaea* is less studied. Halocin production may be expected to be of considerable ecological advantage, as the ability to compete for nutrients and other resources may be enhanced by excreting archaeocins. However, it has never yet been proven that halocins are indeed excreted by natural communities of halophilic *Archaea* in concentrations sufficient to inhibit the development of competing strains, thus substantiating their ecological role.[25] Protein antibiotics excreted by halophilic *Archaea* were first reported in 1982, when it was discovered that colonies of *Haloferax mediterranei* on agar plates inhibit growth of *Halobacterium salinarum* and other members of the *Halobacteriaceae*.[26] This type of BASM was referred to as halocins, halophilic ‘bacteriocins’ or better as archaeocins. Since then it has become clear that halocin production is an almost universal feature among the halophilic *Archaea*.[27–31] Comparative studies with large numbers of isolates have shown that there are several types of halocins with different activity spectra.[31]

A general overview of the halocins has been given by Shand *et al.*[30] Among the well-studied halocins, the following should be mentioned: heat-sensitive and salt-dependent 28-kDa halocin H4 of *H. mediterranei*,[32] a 31-kDa halocin H1 of *H. mediterranei* X1a3,[33–34] heat-resistant halocin H6 of *Haloferax gibbonsi*,[35] thermostable 3.8-kDa halocin Hal R1 of *Halobacterium* sp. GN 101,[36] halocin A4 produced by a halophilic archaeon obtained from Tunisia,[29] and halocin S8 of a yet uncharacterized rod-shaped halobacterium isolated from Great Salt Lake.[28] Halocin H4 interacts with the membrane of the target cells, where it probably causes permeability changes that result in an ionic imbalance, leading to death and cell lysis.[26–27,32,36] Sensitive cells become swollen and assume a spherical shape.[27] Halocin H6 inhibits the Na^+/H^+ antiporter activity of sensitive cells, thus targeting the central device used by the halobacteria to adapt to highly saline environments.[37]

9.15.2.2 Understanding Genetic and Epigenetic Components of Natural Compounds

Bioinformatics approaches in the search for natural products are a combination of molecular and chemical techniques. Important criteria of molecular approaches include phylogenetic resolution and potential to a large-scale screening. Application of comparative genome sequence analysis is essential for a better understanding of the genetic and epigenetic components of different bacterial taxa. With the increased numbers of fully sequenced microbial genomes, including those of well-known bacterial producers of natural products, it has become clear that the genomic and metabolic capacity of these microorganisms is much higher than initially anticipated. This is due to the discovery of ‘silent’ or ‘cryptic’ secondary metabolite gene clusters that encode the production of additional, unidentified compounds.

Recent examination of massive sequencing (metagenomics) approaches to analyze the composition of bacterial communities of complex milieu including sea water,[38] provide an abundant source of molecular sequence data for analysis. These data are useful in comparative genome analyses to identify genes directly involved, for example, in nonribosomal peptide synthesis (peptide synthetase), modifying enzymes, or other genes coding the production of certain natural products. Often, the complete set of specific genes involved in the synthesis of a particular natural product is contained in a single operon. For example, as the presence of conserved sequence motifs and a modular organization of nonribosomal peptide synthetases often assembled into single bacterial operons, a specific sequence search algorithm can be developed to screen public database resources. This enables a detailed analysis of evolutionary, structural, and functional aspect of natural products production based on the comparison of molecular sequences, molecular modeling, and simulation. For example, the situation of genomic colinearity of modular synthetase components might also facilitate the identification of the molecular components of natural products production as well as the reconstruction of natural products synthesis pathways. This will permit to clarify the details of natural production systems and may allow the simulation of these pathways to explore possible strategies for the optimization or engineering of natural product production systems.

Despite the enormous flexibility of genomes, the corresponding metabolic synthesis networks follow specific inherent rules that are responsible for their rigidity.[39] Evolutionary designed strategies are ideally suited to utilize this genomic flexibility to adapt desired phenotypes to balance the metabolic network states required for optimal performance. The identification of genes involved in the metabolic synthesis of natural products by genome sequence analysis can be complemented by the analyses and modeling of natural products production. Bioinformatics tools for the construction of metabolic networks from genome sequence (e.g., Pathway Tools developed by Karp and coworkers at the bioinformatics research group at SRI International (http://www.sri.com) and information from the literature can be used to infer and describe natural products synthesis pathways and analyze the production machinery of bacterial producers. It is generally recognized, particularly in systems responsible for the synthesis of diverse antibiotics, that, for example, nonribosomal peptide synthesis occurs within a molecular complex composed of modules or subunits grouping peptide synthetase modules and associated enzymatic activities.

9.15.2.3 Metagenomics Approach

Metagenomics, which is also called environmental genomics,[40] allows simultaneously to inspect genomes of several microorganisms in one sample.[41] Because the information regarding the regulatory networks and gene expression are in the genome sequence of a bacterium, metagenomics opens new perspectives for screening of natural products genes and/or bacterial producers. Metagenomics sequence data, which can be derived from a few different approaches, are also useful in assessing potential ecological functions and phylogenetic affiliation of uncultivable microbes. The first approach is a small insert library approach that has been technically well established: It clones from relatively small fragments of genomic DNA ('small inserts') into plasmid or phage vectors.[42] Such clone libraries are useful when probing for genes of interest,[40] or sequencing the ends of the insert.[38] This approach has been used in recent metagenomic studies such as the Global Ocean Survey[43] and the investigation of microbial populations of the Sargasso Sea[38] where nearly two million sequence runs have been conducted.

In the second approach, large fragments of microbial DNA are cloned into vectors that are able to accept fosmids of about 40 kbp and bacterial artificial chromosomes (BACs) of about 100 kbp of foreign DNA. Thus constructed clone libraries can be useful when probing for genes of interest. This approach is useful in investigating structure–function relations in the microbial domain.[44] The third approach is based on the direct sequencing of genomic DNA isolated from uncultivated microbial communities. This approach has received significant attention due to recent development of new sequencing technologies, for example, pyrosequencing. However, the pool bacterial DNA is used in direct sequencing; this approach is applicable for samples with low diversity.[44] The target in the fourth approach is the genome of a single cell that is usually isolated by the flow cytometry. This approach is not 'metagenomic' *sensu stricto*, but it remains to be included in the same group of approaches applied for studying uncaltivable microbes. Despite recent advances with a wide range of potential applications, this approach is a partial sequence approach[45] and remains technically challenging.

9.15.2.4 Pyrosequencing Technique

Pyrosequencing is one of the DNA sequencing techniques and has numerous potential applications.[46–48] It is based on the chain of enzymatic reactions during DNA synthesis when the inorganic pyrophosphate (PPi) is released as a result of nucleotide incorporated by polymerase. The light is generated as the result of luciferin oxidation by luciferase. The energy for this reaction is supplied via conversion of pyrophosphate to adenosine triphosphate (ATP). According to Ronaghi,[49] the overall reaction from polymerization to light detection usually takes 3–4 s at room temperature; 1 pmol of DNA could yield 6×10^{11} ATP molecules, which, in turn, generate more than 6×10^{9} photons at a wavelength of 560 nm. The light is detected by a photodiode, photomultiplier tube, or a charge-coupled device (CCD) camera.[49]

The sequence of the DNA template can be determined, as the added nucleotide is known. Primed DNA templates are commonly used because DNA polymerases are known to have higher catalytic activity than RNA polymerases. The enzymes employed in pyrosequencing include the Klenow fragment of *Escherichia coli* DNA Pol I,[50] ATP sulfurylase from a yeast called *Saccharomyces cerevisiae*,[51] and the luciferase from the American firefly *Photinus pyralis*. This is the so-called solid-phase pyrosequencing[46] where DNA immobilization and a washing step to remove the excess substrate after each nucleotide addition are essential. An improved pyrosequencing approach, the so-called liquid-phase pyrosequencing to eliminate a solid support and intermediate washing was also proposed by Ronaghi *et al.*[47] The pyrosequencing reaction can be performed in a single tube due to the employment of the fourth, a nucleotide-degrading enzyme from potato, apyrase. The power of the pyrosequencing approach lies in its capacity to generate massive amount of sequence data for relatively low costs.[48] The short-length (about 200 bp) sequences may be particularly suitable for screening specific genes involved in the synthesis of natural products. Bioinformatics analysis of pyrosequencing data can provide information regarding the biogeographical distribution of bacterial producers and their relative abundance in specific econishes.[52] Pyrosequencing data can be useful in other ecologically related areas, for example, modeling of experimental systems on the diversity–productivity relationships.[53–54]

9.15.3 Selected Data Mining Techniques for Gene Expression

9.15.3.1 Unsupervised Methods in Bioinformatics

Clustering methodologies represent unsupervised analyses that are not appropriate for the incorporation of prior knowledge about the observations. Clustering is the process of grouping the data into classes or clusters so that objects within a cluster have high similarity in comparison to one another but are very dissimilar to objects in other clusters.[55] Clustering techniques used in analyzing gene expression data can be categorized according to the internal algorithms used[55] as follows:

- Hierarchical clustering
- Partitional clustering
- Density-based methods
- Model-based methods

These clustering techniques are reviewed below.

9.15.3.1.1 Hierarchical clustering

Hierarchical clustering[56] is the most commonly used method for analyzing gene expression pattern. It works by grouping data into a tree of clusters, with each level of the tree representing a degree of similarity. By using certain techniques such as a tree structure to represent the affinity variations of gene expression data at different levels, a biologist can visualize and interpret the results easily.

Depending on whether the hierarchical decomposition is formed in a bottom-up or top-down fashion, it can be distinguished between agglomerative and divisive hierarchical clustering. Agglomerative clustering (see **Figure 1**) starts by putting each object into its own cluster (**Figure 1(a)**) and chooses two or more clusters that are correlated to build a bigger cluster (**Figure 1(b)**). This process is repeated until all clusters have merged to form one final cluster (**Figure 1(c)**). Divisive clustering is the opposite of agglomerative clustering. It starts by considering all objects in one large cluster (**Figure 2(a)**). Then it is broken down into smaller subsets (**Figure 2(b)**). The smaller clusters repeatedly divide themselves into smaller clusters (**Figure 2(c)**) until the smallest cluster that consists of only one single entity (node) (**Figure 2(d)**). The divisive hierarchical clustering is presented in **Figure 3**.

The main step used in hierarchical clustering algorithm in analyzing gene expression data is to compare every entity with all the other entities by calculating a distance. The calculation of the distance depends on the linkage method being implemented and the method of calculation of the actual distances. There are three major linkage methods:

Single linkage: the distance is defined as the minimum distance between any two clusters, where s,t defines the two clusters and $x_{s,i}$ defines the entity in cluster s and $x_{t,j}$ in cluster t.

$$d(s,t) = \min\left(\text{dist}\left(x_{s,i}, x_{t,j}\right)\right)$$

Complete linkage: the distance is defined as the maximum distance between any two clusters, where s,t defines the two clusters and $x_{s,i}$ defines the entity in cluster s and $x_{t,j}$ in cluster t.

$$d(s,t) = \max\left(\text{dist}\left(x_{s,i}, x_{t,j}\right)\right)$$

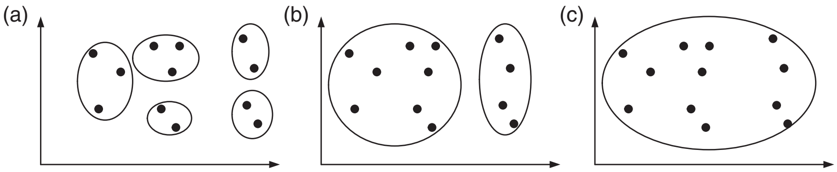


Figure 1 Agglomerative hierarchical clustering.

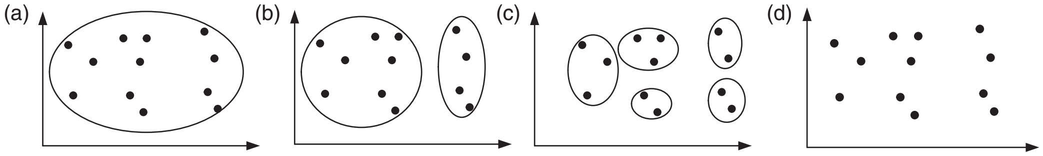


Figure 2 Divisive hierarchical clustering.

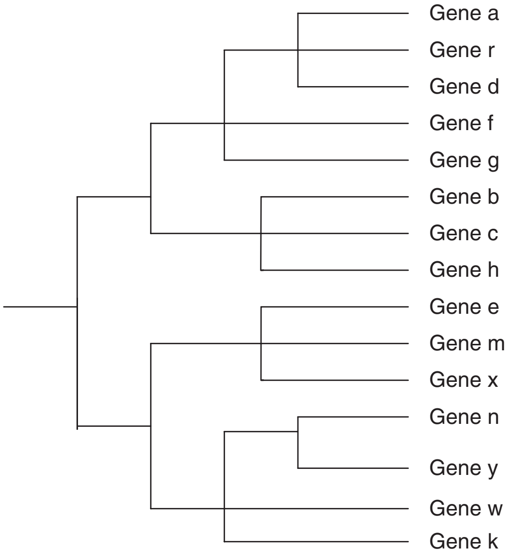


Figure 3 Divisive hierarchical clustering algorithm.

Average linkage: the distance is defined as the mean distance between all possible pairs of entities of the two clusters, where *s*,*t* defines the two clusters and $x_{s,i}$ defines the entity in cluster *s* and $x_{t,j}$ in cluster *t*,

$$d(s,t) = \frac{1}{n_s n_t} \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} \text{dist}\left(x_{s,i}, x_{t,j}\right)$$

Comparing with other two linkage methods, average linkage does not observe the chaining problem and outliers are not given more weight in the cluster decision, which are the drawbacks of the other two linkage methods; therefore, it makes the average linkage the most popular of the three linkage methods.

The hierarchical clustering techniques have been applied to many studies of gene expression patterns with some success.[57] However, the hierarchical tree cannot determine the optimal number of clusters in the dataset. The limitation of the hierarchical algorithms is that the number of classes is determined by cutting the tree structure at an *ad hoc* level selected by the user. Such an *ad hoc* level does not necessarily reflect the true nature of the underlying structure of the gene expression data.

The binary hierarchical clustering (BHC) algorithm[58] uses a hierarchical binary division-clustering framework for the clustering of gene expression data to avoid the above limitation of the traditional hierarchical algorithm. The BHC algorithm combines the fuzzy c-means algorithm and the average linkage hierarchical clustering algorithm with Fisher linear discriminator analysis to solve the two-class partitioning problem. By applying this BHC algorithm, the gene expression data is divided into many subgroups that are comparable. The visualization of the clustering result is easy due to the tree structure used by the BHC algorithm. The main advantage of the BHC algorithm is that it does not require to predetermine the number of clusters, or to make any assumption about the size of the cluster and the class distribution (which are the limitations of other hierarchical methods). The number of clusters can be estimated from the data directly. Hence, BHC avoids the limitation faced by the traditional hierarchical algorithms. An alternative way to improve the clustering quality of hierarchical clustering methods is to integrate hierarchical clustering with other nonhierarchical clustering methods. For example, the BIRCH algorithm[59] performs hierarchical clustering with a clustering feature (CF) tree before applying other techniques.

9.15.3.1.2 Partitional clustering

The most well-known and commonly used partitioning method is k-means clustering.[60–61] The k-means algorithm partitions genes into a defined set of discrete clusters by attempting to maximize the gene expression similarity of the genes in each cluster. The algorithm is initiated by randomly partitioning the genes into k

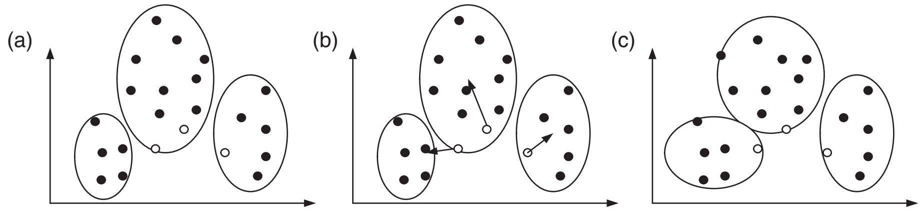


Figure 4 Figure of k-means clustering.

groups randomly (see **Figure 4(a)**). Each group is then represented by the mean expression pattern of genes in the group, and the genes are repartitioned to the cluster whose centroid is most similar to their expression pattern (see **Figure 4(b)**). The clustering repeats until all the clusters have no further change. Because the process could take forever (convergence is not guaranteed), the process can be stopped either after a given number of iterations or when changes are smaller than a given level. The end result of the algorithm is a set of k clusters of similarly expressed genes (see **Figure 4(c)**).

The main weakness of k-means algorithm is that each gene is assigned to one and only one cluster; therefore, it is difficult to understand the relationships between genes in different groups that are functionally related. This limitation is especially problematic when analyzing datasets of large gene expression that are collected over many experimental conditions, where many of the genes are likely to be similarly expressed with different groups in response to different subsets of the experiments.[62]

Fuzzy k-means clustering,[63] in contrast to standard k-means clustering, is less sensitive to the above problem since the genes can be assigned to more than one cluster rather than a single cluster. An application of the fuzzy k-means clustering method[62] has been successfully used to identify overlapping clusters of gene expression data obtained from datasets of well-known yeast genes. The result shows that the fuzzy k-means clustering is useful to gain an insight into the inner workings of the regulation of gene expression in yeast cells responding to environmental changes, where multiple overlapping clusters are often presented. The fuzzy clustering method can, therefore, be used to identify clusters of genes that are not identified by hierarchical or standard k-means clustering. Many of the genes that are clustered together by fuzzy k-means clustering are likely to be coregulated at the level of transcription in response to certain environmental conditions.

Despite the advantages of fuzzy k-means clustering, the method has the following limitations:

- The assignment of genes to the clusters requires a user-defined membership cut-off.
- It can fail to identify a small number of groups that are identified by hierarchical clustering.

The k-means clustering method is still a combinatorial optimization problem for assigning gene expression samples to different clusters. As the algorithm executes very slowly, it is not suitable to analyze large datasets. Fuzzy k-means clustering method is in the same class of k-means, and it is also very slow. However, fuzzy k-means can be combined with interactive visualization environment to improve the clustering process by allowing biologists to decide user-defined membership cutoff online.

Any of the ‘typical’ clustering methods such as hierarchical clustering and k-means clustering can be used when clusters are formed properly, the distance chosen has biological meaning, and all the conditions tested are equally relevant. However, this might not be possible when, for example, different tissues or different patients are analyzed. In that case, it could be interesting to find out genes that coexpress only in a subset of conditions. Biclustering algorithms, which were first used by Cheng and Church,[64] have been successfully used for these problems. Biclustering algorithms refer to a distinct class of clustering algorithms that perform simultaneous row–column clustering. Clustering can be applied to either the rows or the columns of the data matrix separately. On the other hand, biclustering algorithm performs clustering in these two dimensions simultaneously. This means that when clustering algorithms are used, each gene in a given cluster is defined using all the conditions. Similarly, each condition in a given cluster is characterized by the activity of all the genes that belong to it. However, each gene in a bicluster is selected using only a subset of the conditions and each condition in a bicluster is selected using only a subset of the genes. The goal of biclustering techniques is thus to identify subgroups of genes and subgroups of conditions, by performing simultaneous clustering of both rows and columns of the gene expression matrix, instead of clustering these two dimensions separately. Unlike clustering algorithms, biclustering algorithms identify groups of genes that show similar activity patterns under a specific subset of the experimental conditions. Therefore, biclustering approaches are the key technique to solve the problems of finding genes that coexpress only in a subset of conditions.

Graph-theoretic methods are also partitioning methods. They use a divisive approach to partition the space into subgraphs with respect to some geometric properties. Examples of this approach are the cluster affinity search technique (CAST),[65] enhanced cluster affinity search technique (E-CAST),[66] and the coupled two-way clustering (CTWC).[67] Ben-Dor and coworkers[65] undertook a graph-theoretic approach and makes no assumptions on the similarity functions or the number of clusters required. This algorithm relies on the average similarity (affinity) between unassigned vertices and the current cluster seed to make its next decision. The advantages of the CAST algorithm are:

- It removes spurious elements from cluster seeds and avoids repetition.
- It adds and removes elements from the current seed one at a time; therefore, it helps to strengthen the constructed seed.

The CAST algorithm handles more general inputs: it allows the user to specify both a real-valued similarity matrix and a threshold parameter that determines what is considered significantly similar. This parameter controls the number and the sizes of the produced clusters. Comparing with other five clustering algorithms on four gene expression datasets, Yeung *et al.*[68] demonstrated that CAST tends to have relatively high predictive power. The CAST algorithm is a successful technique in clustering gene expression data. However, it uses a fixed initial threshold value that is defined by the user to start the clustering; the size and number of the clusters produced by the CAST algorithm is directly affected by this factor.

In order to circumvent the abovementioned difficulty, the E-CAST algorithm[66] has been developed. E-CAST enhances the CAST algorithm by using a dynamic threshold whose value is computed at the beginning of each new cluster rather than using the fixed threshold value. Moreover, it does not need an expensive cleaning step, which is another drawback of using the CAST algorithm. Getz *et al.*[67] used CTWC to discover partitions and correlation in gene microarray data analysis. It allows attentions to be focused on minute details of small subsets of the huge expression patterns obtained from a large number of samples. A cellular process may only involve a subset of genes and samples and this would be masked by the full dataset. This means that only a limited number of genes in each experiment engage with useful information. CTWC provides a convenient method to discover partitions and correlations that are masked and hidden when the full dataset is used in the analysis by identifying relevant gene and sample subsets and focusing on them. It can be used in combination with any clustering algorithm to find stable clusters.

9.15.3.1.3 Density-based methods

A density-based method clusters gene expression data based on the notion of density. It grows clusters either according to the density of neighborhood clusters or according to some estimated density functions.[55]

Agrawal *et al.*[69] present a density-based clustering algorithm called CLIQUE to partition the data space into nonoverlapping rectangular cells by treating the attribute of each data record as one dimension. Each dimension is divided into a fixed number of bins of equal length. A bin is dense if the fraction of total data points contained in the bin is greater than a certain threshold. The algorithm also finds dense cells in lower-dimensional spaces and merges them to form clusters in higher-dimensional space. Most cluster analysis algorithms reviewed so far are based on distances of similarities such as Euclidean distance. However, genomic data often consists of features that are domains of the attributes; the features form a high-dimension space and they are crucial to study differentials with scaling and shifting factors in multidimensional space. The features are also used to discover pair-wise frequent patterns and cluster genomic data based on such frequent patterns. High dimensionality requires the user to specify the subspace (a subset of the dimensions) for clustering analysis. However, this requirement is difficult to be addressed very well by users as the user identification of subspaces is quite error prone.[69] The CLIQUE algorithm identifies dense clusters in subspaces of maximum dimensionality without requiring the user to guess subspaces that might have interesting clusters.

Comparing with other algorithms such as BIRCH,[59] which are designed for clustering in full dimensional space, CLIQUE has an advantage on finding clusters in the subspace of maximum dimensionality. Wang *et al.*[70]

also demonstrated a clustering analysis model – Pattern Cluster (pCluster) that uses pattern similarities (two objects are similar if they exhibit a coherent pattern on a subset of dimensions) to measure the distance between two objects. Furthermore, the clustering algorithm used in the pCluster model is to generate clusters in the highest dimensions, and then finds low-dimensional clusters not already covered by the high-dimensional clusters. It is more efficient than the CLIQUE algorithm because the combination of low-dimensional clusters to form high-dimensional ones are usually very expensive.[70]

9.15.3.1.4 Model-based clustering

These methods attempt to optimize the fit between the given data and some mathematical model.[55] In the traditional hierarchical methods, the clustering algorithms are largely heuristically motivated, which introduces the difficulties in determining the number of clusters. Ghosh and Chinnaiyan[71] use a mixture model-based approach for the analysis of microarray data to address the reliability of the clustering results. An attractive feature of the mixture modeling approach is that the strength of evidence measure for the number of true clusters in the data is computed. This assessment of reliability of clustering output is often an important question to biologists considering data from microarray studies. The advantages of the model-based approach to clustering are the following:

- It provides estimates of the number of classes as well as their parameters.
- It can directly compare the 'goodness' of clusters of different sizes.
- The cluster definitions can overlap, which allow degree of 'fuzziness' for samples that lay on the boundaries of different clusters.

However, the model-based clustering approach suffers the following weaknesses:

- It executes very slowly especially in large dataset, hence it cannot be combined with queries to a large search.
- It must specify underlying models for mixture.

Although the model-based clustering algorithm is evidently more powerful than the abovementioned clustering methods, it is very complex to implement and very slow to compute; therefore, it has very limited value.

Integrating several clustering methods as a classification tool has also been examined by some researchers. For example, Kim *et al.*[72] use the partitional k-partitioning matrix incision tree (MITree–K) algorithm based on the hierarchical matrix incision tree (MITree)[73] to form a theoretical framework and formalizations for the consistent development of clustering algorithms, which support both hierarchical and partitional structures. In contrast to k-means algorithm, MITree–K algorithm demonstrates higher cluster consistency and quality. MITree–K also provides the extra benefit of allowing quantitative high-dimensional visualization of the resulting clusters. Jain and Dubes[61] use a density-based hierarchical clustering (DHC) method to identify the clusters and the clustering results are of high quality and robustness. DHC organizes all objects in a dataset into an 'attraction tree' according to the density-based connectivity and then clusters to identify dense areas. In contrast to partition-based algorithms such as k-means algorithm, DHC automatically detects the number of clusters instead of using a number of clusters as parameter. Comparing to CAST, DHC can handle the embedded clusters and highly intersected clusters uniformly. Moreover, the mining results from DHC can be visualized and interpreted systematically.

9.15.3.1.5 Clustering combining visualization techniques

Clustering algorithms have been successfully used in analyzing the gene expression data. When users select different clustering algorithms, visualization and objective evaluation of clusters should be considered as well. Kim *et al.*[72] address this by applying a theoretical principle called 'matrix incision principle', to the development of cluster-optimization functions in combination with comprehensive data visualization strategies. Saffer *et al.*[74] apply integrated data visualization and mining framework (Omni Viz Pro) to integrate large amounts of data across different databases and information sources. This approach is fully interactive allowing ready access to the information and the relevant analytical tools. Torkkola *et al.*[75] use the self-organizing maps (SOM) to combine both aspects of gene analyses (clustering and visualization). Exploratory data analysis will become increasingly dependent on visualization methods and the integration of multiple databases can be approached by combining the clustering algorithms with the visualization strategies.

Kohonen[76] has developed a self-organizing network (also known as SOM). SOM uses a neighborhood of neuron to find the similarity within the data and therefore group similar data items together. The neurons are arranged in a matrix pattern called a map, and every input neuron is connected to every other neuron in this matrix. It consists of an input layer and a competitive output layer, which is normally organized into a two-dimensional grid of fully connected neurons. The input vectors are fed through an input layer and mapped into the competitive neurons in an output layer. The competition learning algorithm in the output layer ensures that similar input vectors are mapped into competitive neurons that are closer to each other in the grid than dissimilar ones. Hence, in SOM, input vectors in high-dimensional space are projected into two-dimensional output spaces according to their spatial similarities. Similar input patterns are clustered into one small region in the grid of the output layer. SOM works by initializing the topology of the map, which means initialization of the weight vectors to randomize the values. Then, it decides on the appropriate learning rate and neighborhood size values. For each input neuron, the learning algorithm finds the shortest distance to any output neuron. SOM, then modifies the winner neuron's weight according to the current state of the learning rate, and modifies the neighboring neuron's weights according to the current state of the neighborhood size. This process is repeated until all the input neurons have been reached, and then the learning rate and the neighborhood size are decreased for a sufficient number of iterations.

The SOM is widely used as a data mining and visualization method in bioinformatics. SOM[77] provides a more robust and accurate approach to the clustering of large amounts of noisy data than that of hierarchical clustering methods in analyzing the gene expression data. Tamayo *et al.*[78] ran the SOM on the yeast dataset sourced from Cho *et al.*,[79] and the clusters derived by SOM were similar to those that were determined by visual inspection by Cho *et al.*[79] This result shows that SOM is capable of finding meaningful biological patterns in gene expression data of thousands of genes. Torkkola *et al.*[75] use the Stanford yeast gene expression dataset to implement the analysis with SOM. They argue that the best performance of the gene expression analysis comes from the combination of clustering and visualization methods. SOM can be used at the same time both to reduce the amount of data by clustering, and to construct a nonlinear projection of the data onto a low-dimensional display. Therefore, SOM can be used to combine both aspects of gene analysis (clustering and visualization).

Nevertheless, SOM presents some problems.[80] For example, the SOM is a topology-preserving neural network and this means that the number of clusters is randomly chosen from the beginning. Therefore, the clustering obtained is not proportional. In addition, the lack of a tree structure makes it impossible to detect higher-order relationships between clusters of profiles. Herrero *et al.*[81] used the self-organizing tree algorithm (SOTA), described by Dopazo and Carazo,[82] which combines the advantages of both hierarchical clustering and SOM, and is free of the problems these methods present when applied to gene expression profiles, to the analysis of gene expression data coming from DNA array experiments. The advantages of SOTA are that the clustering obtained is proportional to the heterogeneity of the data and that the binary topology produces a nested structure in which nodes at each level are averages of the items below them. An alternative way to avoid the problem is to use fuzzy Kohonen neural networks[83] that combines a Kohonen network and a fuzzy c-means algorithm to benefit the advantages of both techniques and overcomes some of the shortcoming of each individual technique. The advantages of SOM can be attributed so that they map high-dimensional data to be a more comprehensible lower-dimensional space and it can execute very fast. It is potentially very useful for dealing with high dimensionality and large-scale databases in extracting information from gene expression data. However, it warrants further investigation as to whether it can be effectively combined with database queries. Finally, SOM has other limitations, namely, (1) there is no convergence guarantee and (2) the results may be nondeterministic and dependent on learning rates.

In conclusion, clustering techniques have been applied to identify clusters from gene expression data that share similar expression profiles and the results are moderately valuable. However, clustering methodologies represent a class of unsupervised analysis that is not capable of incorporating prior knowledge about the observations. They are more likely to be used as a preprocessing step in advanced data mining techniques.

9.15.3.2 Supervised Methods in Bioinformatics

In bioinformatics, machine learning techniques are used to support pattern matching and pattern discovery components of data mining. The spectrum of machine learning technologies applicable to mining genomic data includes neural networks and support vector machines (SVMs). Although clustering techniques have been applied to identify groups of genes sharing similar expression profiles and the results are valuable, the metrics normally used only allow capturing a subset of the relationships that potentially exist among various transcripts. Furthermore, clustering methodologies represent unsupervised analyses that are not appropriate for the incorporation of prior knowledge about the observations. Effective methods need to be developed for upgrading the information content of the larger amounts of data generated by microarray experiments. In this section, we provide an updated review of machine learning applications in bioinformatics with the most recent advances.

9.15.3.2.1 Support vector machines

SVMs are considered as a supervised learning method. The algorithm of SVMs appeared first in the book of Vapnik[84] and later in Cristianini and Shawe-Taylor.[85] Since SVMs are well known as training algorithm for learning classification from data, they are widely used for the applications of classification and pattern recognition in bioinformatics. Brown *et al.*[86] show application of the theory of SVMs to the classification of yeast microarray expression data. They compare the misclassification rates for SVMs and other machine learning approaches, and demonstrate that SVMs are the most superior methods. In addition to use for classifying microarray expression data, SVMs have been shown to perform well in multiple areas of biological analysis including detection of remote protein homologies,[87] and recognizing translation initiation sites. Recently, another effort has been proposed to use SVMs in analyzing expression data.[88,89] As gene expression data has, in general, high-dimensional data, it poses a serious problem to several machine learning methods. To circumvent this problem, dimensionality reduction can be used but it often leads to information loss and performance degradation. Fortunately, SVMs can overcome this problem as they can generalize well with high-dimensional data.[90]

As illustrated in Figure 5, SVMs work by constructing a hyper-plane in a higher-dimensional feature space of the input data,[91] and use the hyper-plane (represented as H_1 and H_2) to enforce a linear separation of input samples, which belong to different classes (represented as Class O and Class X). The samples that lie on boundaries of different classes are referred to as support vectors. The underlying principle behind SVM-based classification is to maximize the margin between the support vectors using kernel functions. In the case of

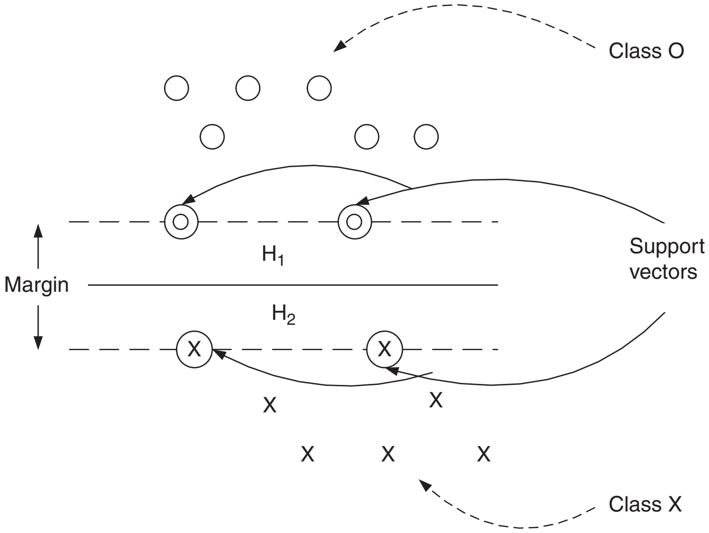


Figure 5 An illustration of support vector machine classification scheme.

extracting information from gene expression data, the expression data for each gene becomes an input vector that occupies a point in the input space. It is very hard to separate genes that have similar functions in this input space due to the complex interactions between genes. By using SVM-based classification approach, a kernel is carefully constructed to transform the input space of the gene expression data into some higher-dimensional feature space where the complex interactions between genes are taken into account, and hopefully, the gene of similar functions can be separated linearly in such a feature space. Once it is trained, an SVM classifier can be used as a predictive model for the given gene expression data; it can also be used to predict the class membership of any unknown gene.

9.15.3.2.2 Neural networks

Neural networks are parallel and distributed information processing systems that are inspired and derived from biological learning systems such as human brains. The architecture of neural networks consists of a network of nonlinear information processing elements that are normally arranged in layers and executed in parallel. This layered arrangement for the network is referred to as the topology of a neural network. These nonlinear information processing elements in the network are defined as neurons, and the interconnections between these neurons in the network are called synapse or weights. A learning algorithm must be used to train a neural network so that it can process information in a useful and meaningful way.

Most neural networks are trained with supervised training algorithms. This means that the desired output must be provided for each input used in the training. In other words, both the inputs and the outputs are known. In the supervised training, a network processes the inputs and compares its actual outputs against the expected outputs. Errors are then propagated back through the network, and the weights that control the network are adjusted with respect to the errors propagated back. This process is repeated until the errors are minimized; it means that the same set of data is processed many times as the weights between the layers of the network are refined during the training of the network. This supervised learning algorithm is often referred to as a back-propagation algorithm, which is useful for training multiple-layer preceptron neural networks (MLPs). Figure 6 demonstrates the architecture for a supervised neural network, which includes three layers, namely, input layer, output layer, and a hidden middle layer.

Neural networks are used in a wide variety of applications in pattern classification, language processing, complex systems modeling, control, optimization, and prediction.[92] Neural networks have also been actively used in many bioinformatics applications such as DNA sequence prediction, protein secondary structure prediction, gene expression profiles classification, and analysis of gene expression patterns.[93] Neural network has been applied widely in biology since the 1980s.[94] For example, Stormo *et al.*[95] reported prediction of the translation initiation sites in DNA sequences. Baldi and Brunak[96] used applications in biology to explain the theory of neural networks. The concepts of neural network used in pattern classification and signal processing

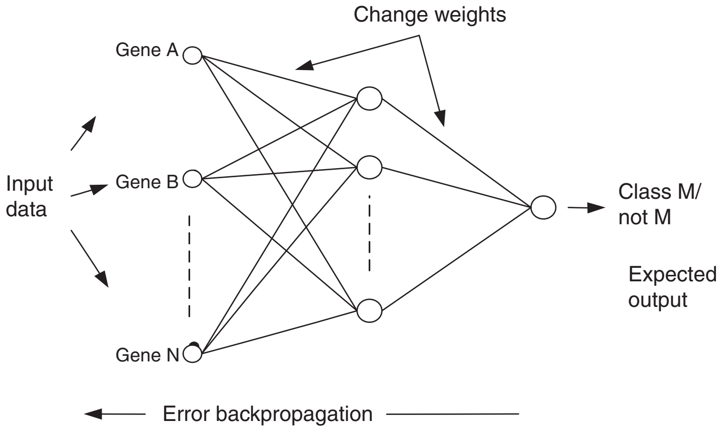


Figure 6 A sample structure of supervised neural network.

have been successfully applied in bioinformatics. Wu *et al.*[93,97–99] applied the neural networks to classify protein sequences. Wang *et al.*[100] applied neural networks to protein sequence classification by extracting features from the protein data and using them in combination with the Bayesian neural network (BNN). Qian and Sejnowski[101] predicted the protein secondary structure using neural networks. Neural networks have also been applied to the analysis of gene expression patterns as an alternative to hierarchical cluster methods.[75,100,102,103] Narayanan *et al.*[104] demonstrated the application of the single layer neural networks to analyze gene expression.

Besides SVMs and neural networks, there are also machine learning methods for gene selection such as 'discriminate analysis', which distinguishes a selected dataset from the rest of the data, and 'k-nearest neighbor (KNN) algorithm', which is based on a distance function for pairs of observations, such as the Euclidean distance. In this classification hypothesis, k nearest neighbors of a set of training data is computed. The similarities of one sample of testing data to the KNN are then aggregated according to the class of the neighbors, and the testing sample is assigned to the most similar class. One of the advantages of KNN is that it is well suited for multimodal classes as its classification decision is based on a small neighborhood of similar objects. So, even if the target class consists of objects whose independent variables have different characteristics for different subsets (multimodal), it can still lead to good accuracy. A major drawback of the similarity measure used in KNN is that it uses all features equally in computing similarities. This can lead to poor similarity measures and classification errors, when only a small subset of the features is useful for classification. Li *et al.*[105] successfully used an approach that combines a genetic algorithm (GA) and the KNN method to identify genes that can jointly discriminate between different classes of samples.

9.15.3.3 Data Mining Techniques for Bioinformatics Application: A Case Study

A novel hierarchical-partitioning framework[106,107] was proposed recently. It combines features of both categories of algorithms, which we called the binary hierarchical clustering (BHC).[108] In essence, this algorithm performs a successive binary subdivision of the data into smaller and smaller partitions in a hierarchical manner, until any further splitting of a partition into two smaller partitions is insignificant. The hierarchical structure is manifested in the binary tree structure[109] of the clustering result, where a parent node gives rise to two children nodes if the projection onto the optimal Fisher discrimination axis satisfies a certain threshold. The partitioning behavior of the algorithm is incorporated in the cluster splitting process, where the fuzzy c-means clustering algorithm is used to split a parent cluster into two children clusters.

In each stage of the binary partitioning module of the BHC algorithm, it uses the fuzzy c-means algorithm and the average linkage hierarchical clustering algorithm to split the data into two classes, then refine and verify the validity of the split by using the Fisher discrimination analysis. The binary hierarchical framework leads to a tree structure representation. The tree is constructed in such a way that adjacent clusters are more similar, in terms of the Mahalanobis distance, than nonadjacent clusters. By visualizing the clustering results using a tree structure, the relationship between each cluster, the adjacency between different clusters, as well as the variation within each cluster can be observed easily. The tree structure visualization allows visual interpretation of the clustering result using additional biological knowledge, in a manner similar to that in hierarchical clustering display. **Figure 7** shows clustering results using the cell cycle expression data of yeast from Spellman *et al.* (http://cellcycle-www.stanford.edu).[110] The dataset contains expression profiles for 6220 genes under different experimental conditions. Genes with similar expression profiles are seen to cluster successfully into the same group.

9.15.4 Software in Bioinformatics

9.15.4.1 Useful Databases and Repositories

Owing to the innate complexity of proteins, nucleic acids, and polysaccharides, a proper understanding of biological macromolecules requires a diverse array of experimental and theoretical techniques. As a result, a number of important databases containing the diverse array of information, either experimental or computational, are deposited into a number of databases, which are free and publicly available to the global community.

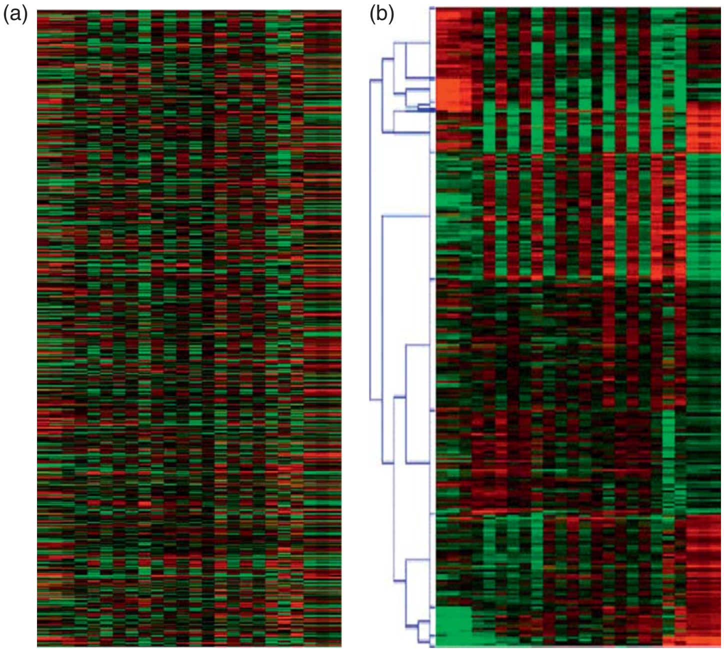


Figure 7 BHC clustering result for the cdc15 experiment dataset. (a) Original gene expression data. (b) Expression data after BHC clustering.

The Cambridge structural database (CSD, http://www.ccdc.cam.ac.uk/products/csd/) is a repository of small molecule crystal structures. The CSD records bibliographic, chemical, and crystallographic information for organic molecules and metal–organic compounds. The three-dimensional structures of small molecules in CSD have been determined using X-ray diffraction and neutron diffraction. Therefore, the CSD records results of single crystal studies and powder diffraction studies, which yield three-dimensional atomic coordinate data for all non-H atoms. The CSD does not contain compounds containing polypeptides and polysaccharides having more than 24 units. The latter can be found in the protein data bank (PDB).

The Worldwide Protein Data Bank (wwPDB, http://www.wwpdb.org/) is one of the most important databases. The wwPDB consists of organizations that act as deposition, data processing, and distribution centers for PDB data, which include the founding members of the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB, USA, http://www.rcsb.org/pdb/home/home.do), Macromolecular Structure Database Group (PDBe, Europe, http://www.ebi.ac.uk/msd/), and Protein Data Bank Japan (PDBj, Japan, http://www.pdbj.org/). The Biological Magnetic Resonance Data Bank (BMRB, USA, http://www.bmrb.wisc.edu/) joined the wwPDB in 2006. The number of protein structures deposited and released each year through RCSB PDB, PDBe, PDBj, and BMRB into wwPDB increases significantly, as shown in **Figure 8** (data based on wwPDB).

A number of public databases, such as DNA data bank of Japan (DDBJ, http://www.ddbj.nig.ac.jp/), the European Molecular Biology Laboratory (EMBL, http://www.embl.org/), and GenBank (http://www.ncbi.nlm.nih.gov/Genbank/) contain nearly all the known sequences. The latter can be retrieved using a few freely available softwares such as Blast, ACNUC, SRS, and Entrez. Sequence alignment and computation including phylogenetic trees reconstruction can be performed using tools provided at the Bioinformatics Organization, Inc. (USA, http://www.bioinformatics.org/). In addition, important and useful repositories, such as National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/), European Bioinformatics Institute (EBI, http://www.ebi.ac.uk/), Atomic Scale Design Network (ASDN, http://asdn.net/), and BioWareDB (http://biowaredb.org/site/index.php), collect exhaustive hyperlinked database links to the vast amounts of freely available and commercial bioinformatics and biocomputing software. For example, BioWareDB repository

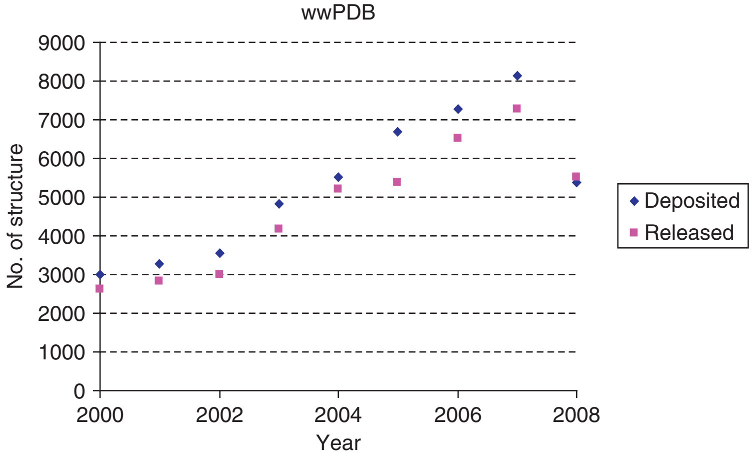


Figure 8 Number of structures deposited and released each year since 2000 (data as on 7 October 2008).

consists of records obtained from automated harvesting of online repositories of the software, from journals delving with bioinformatics software, and from frequently updated (a bimonthly basis) manual entries. Database such as the 3D Interacting Domains (3DID, http://3did.embl.de) is a collection of domain–domain interactions in proteins for which high-resolution three-dimensional structures are known.[111] It (3DID) exploits structural information to provide critical molecular details necessary for understanding how interactions occur. It also offers an overview on how similar structures interact between different members of the same protein family. The database also contains gene ontology (GO)-based functional annotations and interactions between yeast proteins from large-scale interaction discovery studies.

9.15.4.2 Limitations in Current Bioinformatics Analysis

Several issues in bioinformatics are relevant for analyzing sequencing data.[112] It appeared that the increasing amount of data submitted to public databases, such as DDBJ, the EMBL, and GenBank, makes it time consuming to retrieve sequences by keywords, or use Blast to retrieve sequences by similarity. As noted by Christen,[112] the current status of available databases for 'standard sequences', metagenomic data and new pyrosequencing data, almost any DNA sequence potentially coding for a protein open reading frame (ORF) is translated and deposited in the public database of proteins. The Universal Protein Resource (UniProt, http://www.uniprot.org/) that contains 4 949 164 entries, for example, is divided into two divisions of Swiss-Prot (http://www.ebi.ac.uk/swissprot/) with 276 256 entries that contain proteins duly identified and well annotated by experts, and Trembl (http://www.ebi.ac.uk/trembl/) that contains most of the remaining entries. Half of these proteins are from bacteria, and by the end of 2007, one-third of the bacterial proteins have been deposited during the past year. UniProt doubles only every year, as it does not include metagenome data. GenBank (updated 11 December 2007), an another example, contained 1 140 983 ribosomal sequences, 621 900 being annotated as 16S rRNA gene sequences. Note that these 16S rRNA gene sequences were mostly short to very short (50–500 nt, 248 972 entries), and only 201 076 entries had a length of 1200 nt or more, of which half were submitted in 2007. Only 32 880 entries of these long sequences belong to cultured strains comprising about 8000 different species.

A number of new developments are undertaken to facilitate retrieval of pyrosequencing data, for example, a new archive at NCBI, Short Read Archive, was created in 2007. To archive environmental data the Camera web site was designed. For 16S rRNA gene pyrosequencing data analysis general bioinformatics tools, for example, Blast and/or alignment programs, are not suitable due to the short fragment, hence new programs should be created.

9.15.5 Structural Bioinformatics Approaches for Signaling Processes: A Case Study

Natural products trigger complex cascades. In this section, we will briefly review recent advances on the olfactory receptors (ORs) cascades. The discovery of ORs in the cilia of olfactory sensory neurons (OSNs) awarded L. Buck and R. Axel the Nobel Prize in 2004. These are G-protein-coupled receptors (GPCRs) with high affinity for thousands of volatile molecules that are associated with odor, and are capable of discriminating ∼10 000 different odors.

As in most of the other GPCRs, ligand binding to the receptors activates a cascade of events producing an electrical signal as output (**Figure 9**). This activates its target G-protein (G), which, in turn, activates the adenylate cyclase (AC) enzyme. cAMP gates CNG (cyclic nucleotide-gated) channels by binding to them, causing an odorant-induced inward current carried by Na^+ and Ca^{2+} ions (**Figure 9**). The increased Ca^{2+} concentration causes the opening of Ca^{2+}-activated Cl^- channels and the subsequent Cl^- efflux. The resulting depolarization of the cell membrane is the action potential sent to the brain. This causes the sense of smell.

The increased intracellular Ca^{2+} concentration also has an inhibitory effect, which eventually shuts down the signal. The ion binds to calmodulin (CaM) lowering the ligand sensitivity of the cAMP-gated channels and activating the activity of a phosphodiesterase (PDE). Ca^{2+} is finally extruded by a Na^+/Ca^{2+} exchanger.

Most of the components of the cascade are membrane proteins, for which structural information is not available. As a result, structural bioinformatics (SB) is currently employed to predict structural determinants of most membrane proteins involved.

9.15.5.1 Odorant Receptors

As in all GPCRs, ORs' fold consists of seven transmembrane helices. The recent observation that structural features are well conserved across ORs, along the determination of a few GPCRs structures such as the β2-adrenergic receptor (B2AR)[113] has led to the conclusion that SB-based structural predictions may be reliable, provided that sufficient molecular biology data are available, and also included in the model.[114] The identification of nine crucial amino acids involved in ligand binding on helices TM3, TM5, and TM6 of ORs prompted to perform such structural prediction,[114] focusing on the ORs for which ligand binding data are available. The results were validated against experimental information. Based on the results, it is suggested that (1) residues in the positions indicated as I–IV in Figure 10, as well as IX, and V–VIII interact with ligands of Class 1 ORs114; (2) in Class 2 ORs114, some positions (I, II, and IX) are occupied mostly by polar groups, positions IV and VII by

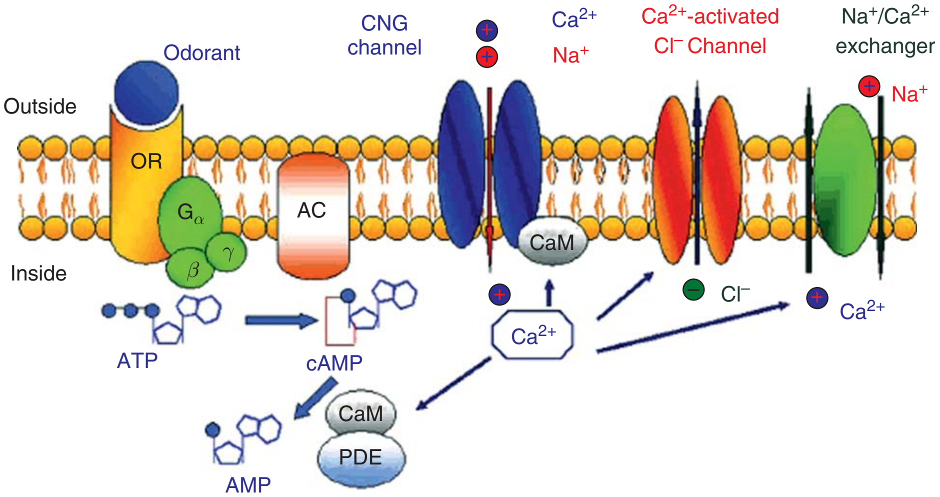


Figure 9 A scheme showing that ligand binding to the receptors activates a cascade of events, producing an electrical signal as output.

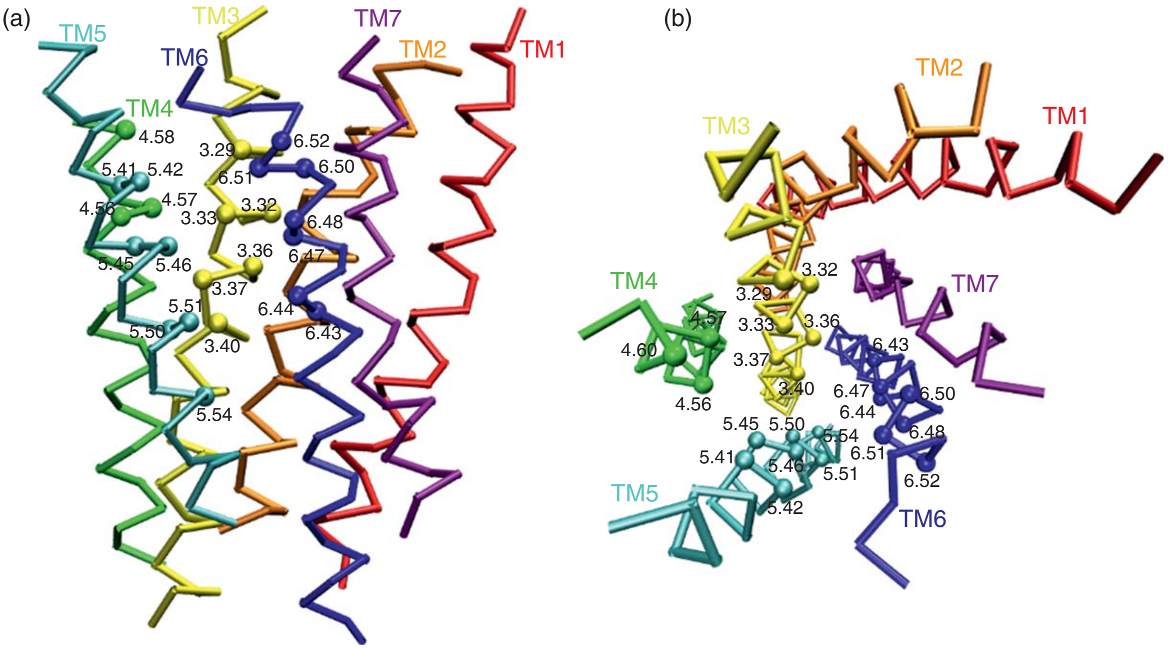


Figure 10 Side (a) and top (b) views of a representative structural model of ORs (Cα carbons in TM helices are only shown).

polar groups, and the nature of residues at positions III, V–VIII is similar to that Class 1 ORs; (3) several other positions (indicated as 1–14 in Figure 10), whose role did not emerge from the experimental studies, may be important for ligand binding. These conclusions are of course to be taken with great caution: in fact, inclusion of further mutagenesis experiments may dramatically improve the modeling.

9.15.5.1.1 Cyclic nucleotide-gated channels

Cyclic nucleotide-gated (CNG) ion channels are gated by cGMP and cAMP second messengers. They produce the electrical signal in response not only with odor stimulation but also with light in the vision process. They consist of two domains: (1) a transmembrane domain formed by six transmembrane helices (S1–S6) and a pore helix (P-helix) and (2) a cytoplasmic domain formed by the cyclic nucleotide binding domain (CNBD), which is linked to the transmembrane domain through the so-called C-linker region (**Figure 11**).

The alignment of the sequences across K^+ channels, along with experimental constraints, has been used to provide a structural basis of CNG channels.[115] As the focus is on the protein from bovine rod, the sequence identity (SI) between the latter and OR CNG channel is very high (58%). Therefore, the two channels are expected to have extremely similar structural determinants. The experimental constraints are obtained by cysteine scanning mutagenesis of residues present principally along the channel axis. Mutated channels are then studied by measuring the differences of current blockage upon the introduction of metals, such as Cd^{2+}, and agents capable of interacting with cysteines in the solution.

Gating has been proposed to occur via a rotational movement that begins in a portion of the cytoplasmatic C-linker region (red helices in **Figure 11**). This rotational movement is then transmitted upward, allowing for the rotation of the upper part of helices of the channel (called S6 in **Figure 11**). These are in contact with the pore lumen, at the outer mouth of the channel. As a result, the pore lumen opens and sodium and calcium ions enter the channel.

9.15.5.1.2 Chloride channels

Members of the bestrophin are Ca^{2+}-activated Cl^- channels that have been suggested to play a pivotal role for olfactory transduction.[116] According to topological models, the N- and C-terminal domains of bestrophins would be located at the intracellular side of the membrane and would be connected to four or five hydrophobic

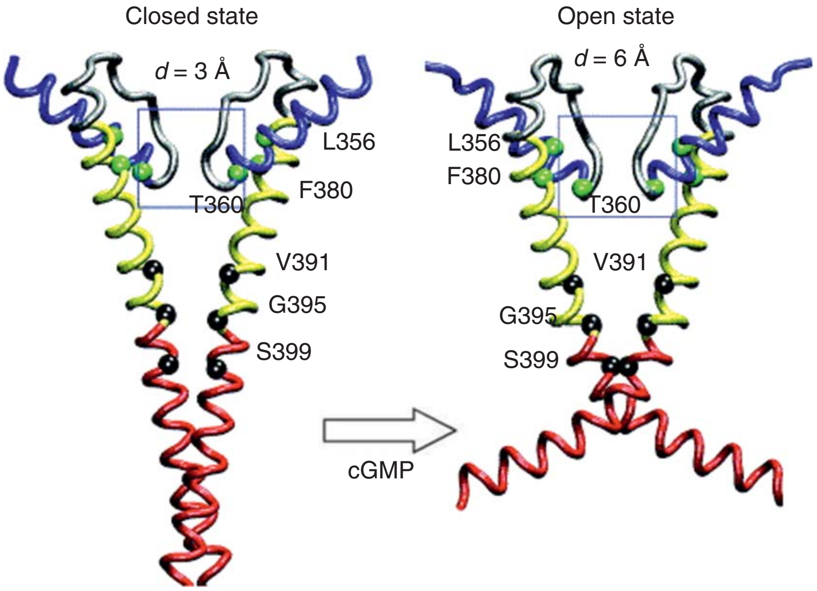


Figure 11 Structural models of CNG channels in the closed and open states. The helix S6 and of the pore lumen of the transmembrane domain along with the N-term position of their C-linker are shown. Only the C alpha atoms of two opposite subunits (out of four) are shown for the sake of clarity. Selected residues, for which experimental measurements were carried out, are shown. *d* is the shortest distance between opposite C alpha atoms in the pore. The blue box indentifies the narrowest region of the pore.

domains forming the channel. We have used structural bioinformatics and molecular simulation tools to identify Asp and Glu residues involved in Ca(II) ion binding. Our model has been validated by performing mutagenesis experiments, in which key acid residues identified by computational methods have been replaced by alanine.

9.15.6 Summary and Future Prospects

A comprehensive understanding of the structures of biologically active compounds has the potential to not only impact upon fundamental knowledge, but also to provide the beneficial effects nationally and internationally on issues that are on the whole of bigger economic scale, and the interplay with human activities. The search of novel producers of natural products is based on the efficient exploration of biological diversity using recent technological developments without damaging the environment.

Microarray technology is a prolific and highly effective method for analyzing gene expressions, which generates huge volumes of data. Data mining technology has been employed in searching for information from the large sets of gene expression data. The major data mining techniques used in analyzing gene expression data have been reviewed in this chapter. The advantage of hierarchical clustering is that it is relatively easy to implement and can be executed very fast. The disadvantages of hierarchical clustering are: (1) It is unclear about the optimal level of hierarchy to represent clusters. (2) As the distance function used in hierarchical clustering is one-dimensional, it can hide high-dimensional relationships.

A huge effort is currently underway in the field of biology to develop new techniques for genome analysis. Techniques such as mass spectrometry and biochips will revolutionize the way biological experiments to be conducted in the future. The introduction of these techniques will result in more increased volume of genomic data that need to be analyzed. As a consequence, new data mining techniques need to be developed to deal with an ever-increasing complexity and volume of genomic data efficiently. Data mining techniques that allow concurrent and distributed analyses over a distributed network have a promising future as researchers need to collaborate, analyze, and interpret data over a networked computer environment. In addition, model-based clustering techniques are expected to be developed in the next decade, through special distant measures that are biologically meaningful. SVMs have been shown to be a powerful tool to classify data. They are especially useful in situations where some of the input data are sparse. Again, SVM kernels that especially model biologically discriminators will be developed to allow classifications based on sound biological principles rather than just the input data.

Finally, we have turned our attention to structural biology of signaling associated with the binding of natural products to their target macromolecules. As a case study, we have presented aspects of the protein modeling involved in the OR cascade of events, triggered by odorant molecules. So far, the proteins in the cascade have been modeled as isolated entities. Challenges in the future include the development and applications of methods predicting the structural determinants of the protein complexes involved in the cascade. This might allow to reveal the aspects of regulation and allosteric effects that are crucial for the signaling. Advancements in experimental structural biology, multiscale modeling, and protein–protein docking algorithms make us confident that such challenges can be taken in a short time and the first insight on the molecular basis of the signaling in OR's signaling may be given.

Abbreviations

3DID	three-dimensional interacting domain
AC	adenylate cyclase
ASDN	Atomic Scale Design Network
ATP	adenosine triphosphate
B2AR	β2-adrenergic receptor
BAC	bacterial artificial chromosome
BASM	biologically active secondary metabolite
BMRB	Biological Magnetic Resonance Data Bank
CAST	cluster affinity search technique
CCD	charge-coupled device
CNG	cyclic nucleotide-gated
CSD	Cambridge structural database
CTWC	coupled two-way clustering
DDBJ	DNA data bank of Japan
EBI	European Bioinformatics Institute
E-CAST	enhanced cluster affinity search technique
EMBL	European Molecular Biology Laboratory
GA	genetic algorithm
GO	gene ontology
GPCR	G-protein-coupled receptor
KNN	k-nearest neighbor
LC–MS–NMR	liquid chromatography–mass spectrometry–nuclear magnetic resonance
MITree	matrix incision tree
MITree–K	k-partitioning matrix incision tree algorithm
MLP	multiple-layer perceptron neural network
mRNA	messenger RNA
NCBI	National Center for Biotechnology Information
NMR	nuclear magnetic resonance
OR	odorant receptor
ORF	open reading frame
OSN	olfactory sensory neuron
PDB	protein data bank
PDBe	Macromolecular Structure Database Group
PDE	phosphodiesterase
PPi	inorganic pyrophosphate

RCSB PDB	Research Collaboratory for Structural Bioinformatics Protein Data Bank
SOM	self-organizing map
SOTA	self-organizing tree algorithm
SVM	support vector machine
UniProt	Universal Protein Resource
wwPDB	Worldwide Protein Data Bank

References

1. W. Sneader, *Drug Prototypes and Their Exploitation*; Wiley: Chichester, 1996.
2. A. L. Harvey, *Natural Product Pharmaceuticals: A Diverse Approach to Drug Discovery*; PJB Publications: Richmond, Surrey, 2001.
3. D. J. Newman; G. M. Cragg; K. M. Snader, *J. Nat. Prod.* **2003**, *66*, 1022.
4. M. S. Butler, *Nat. Prod. Rep.* **2005**, *22*, 162.
5. Y.-W. Chin; M. J. Balaunas; H. B. Chai; A. D. Kinghorn, *AAPS* **2006**, *J 8*, E239.
6. A. L. Harvey, *Curr. Opin. Chem. Biol.* **2007**, *11*, 480–484.
7. W. Fenical; P. Jensen, *Nat. Chem. Biol.* **2006**, *2*, 666–673.
8. P. R. Jensen; T. J. Mincer; P. G. Williams; W. Fenical, *Antonie van Leeuwenhoek* **2005**, *87*, 43.
9. P. R. Jensen; W. Fenical, *Annu. Rev. Microbiol.* **1994**, *48*, 559.
10. D. J. Faulkner, *Nat. Prod. Rep.* **2002**, *17*, 7.
11. R. A. Aras; J. Kang; A. I. Tschumi; Y. Harasaki; M. J. Blaser, *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13579.
12. W. K. Kroeze; D. J. Sheffler; B. L. Roth, *J. Cell Sci.* **2003**, *116*, 4867.
13. T. P. Sakmar, *Curr. Opin. Cell Biol.* **2002**, *14*, 189.
14. Y. Landry; J. Gies, *Fundam. Clin. Pharmacol.* **2008**, *22*, 1.
15. S. Takeda; S. Kadowaki; T. Haga; H. Takaesu; S. Mitaku, *FEBS Lett.* **2002**, *520*, 97. Erratum in: *FEBS Lett. 523*, 257.
16. K. Gustafson; M. Roman; W. Fenical, *J. Am. Chem. Soc.* **1989**, *111*, 7519.
17. F. M. Lovell, *J. Am. Chem. Soc.* **1966**, *88*, 4510.
18. H. Ueda; H. Nakajima; Y. Hori; T. Fujita; M. Nishimura; T. Goto; M. Okuhara, *J. Antibiot.* **1994**, *47*, 301.
19. C. Holmström; S. Kjelleberg, *FEMS Microbiol. Ecol.* **1999**, *30*, 285–293.
20. R. J. Andersen; M. S. Wolfe; D. J. Faulkner, *Mar. Biol.* **1974**, *27*, 281.
21. M. J. Gauthier; G. N. Flatau, *Can. J. Microbiol.* **1976**, *22*, 1612–1619.
22. H. Shiozawa; A. Shimada; S. Takahashi, *J. Antibiot.* **1997**, *50*, 449.
23. A. Isnansetyo; Y. Kamei, *Antimicrob. Agents Chemother.* **2003**, *47*, 480.
24. R. Baltz, *Nat. Biotechnol.* **2006**, *24*, 1533.
25. T. Kis-Papo; A. Oren, *Extremophiles* **2000**, *4*, 35.
26. F. Rodriguez-Valera; G. Juez; D. J. Kushner, *Can. J. Microbiol.* **1982**, *28*, 151.
27. I. Meseguer; F. Rodriguez-Valera, *FEMS Microbiol. Lett.* **1985**, *28*, 177.
28. I. Meseguer; F. Rodríguez-Valera; A. Ventosa, *FEMS Microbiol. Lett.* **1986**, *36*, 177.
29. E. M. O’Connor; R. F. Shand, *J. Int. Microbiol. Biotechnol.* **2002**, *28*, 23.
30. R. F. Shand; L. B. Price; E. M. O’Connor, Halocins: Protein Antibiotics from Hypersaline Environments. In *Microbiology and Biogeochemistry of Hypersaline Environments*; A. Oren, Ed.; CRC Press: Boca Raton, 1999, pp 413–424.
31. M. Torreblanca; I. Meseguer; A. Ventosa, *Lett. Appl. Microbiol.* **1994**, *19*, 201.
32. I. Meseguer; F. Rodriguez-Valera, *J. Gen. Microbiol.* **1986**, *132*, 3061.
33. D. Cotton; L. Hauser; M. Keller, *Appl. Environ. Microbiol.* **2007**, *73*, 3205.
34. C. Haseltine; T. Hill; R. Montalvo-Rodriguez; S. K. Kemper; R. F. Shand; P. Blum, *J. Bacteriol.* **2001**, *183*, 287–291.
35. M. Torreblanca; I. Meseguer; F. Rodríguez-Valera, *J. Gen. Microbiol.* **1989**, *135*, 2655.
36. U. Rdest; M. Sturm, Bacteriocins from halobacteria. In *Protein Purification: Micro to Macro*; R. Burgess, Ed.; Alan R. Liss: New York, 1987.
37. I. Meseguer; M. Torreblanca; T. Konishi, *J. Biol. Chem.* **1995**, *270*, 6450.
38. J. C. Venter; K. Remington; J. F. Heidelberg; A. L. Halpern; D. Rusch; J. A. Eisen; D. Wu; I. Paulsen; K. E. Nelson; W. Nelson; D. E. Fouts; S. Levy; A. H. Knap; M. W. Lomas; K. Nealson; O. White; J. Peterson; J. Hoffman; R. Parsons; H. Baden-Tillson; C. Pfannkoch; Y. H. Rogers; H. O. Smith, *Science* **2004**, *304*, 66.
39. G. Stephanopoulos; J. Vallino, *Science* **1991**, *252*, 1675.
40. E. E. DeLong, *Nat. Rev. Microbiol.* **2005**, *3*, 459.
41. J. Handelsman, *Microbiol. Mol. Biol. Rev.* **2001**, *68*, 669.
42. T. M. Schmidt; E. F. DeLong; N. R. Pace, *J. Bacteriol.* **1991**, *173*, 4371.
43. D. B. Rusch; A. L. Halpern; G. Sutton; K. B. Heidelberg; S. Williamson; S. Yooseph; D. Wu; J. A. Eisen; J. M. Hoffman; K. Remington; K. Beeson; B. Tran; H. Smith; H. Baden-Tillson; C. Stewart; J. Thorpe; J. Freeman; C. Andrews-Pfannkoch; J. E. Venter; K. Li; S. Kravitz; J. F. Heidelberg; T. Utterback; Y.-H. Rogers; L. I. Falcn; V. Souza; G. Bonilla-Rosso; L. E. Eguiarte; M. Karl; S. Sathyendranath; T. Platt; E. Bermingham; V. Gallardo; G. Tamayo-Castillo; R. L. Ferrari; R. L. Strausberg; K. Nealson; R. Friedman; M. Fazier; J. C. Venter, *PLoS Biol.* **2007**, *5*, 398.
44. I. Brettar; M. G. Hofle, *Curr. Opin. Biotechnol.* **1999**, *19*, 274.

45. M. Podar; C. B. Abulencia; M. Walcher; D. Hutchison; K. Zengler; J. A. Garcia; T. Holland; D. Cotton; L. Hauser; M. Keller, *Appl. Environ. Microbiol.* **2007**, *73*, 3205.
46. M. Ronaghi; S. Karamohamed; B. Pettersson; M. Uhlen; P. Nyren, *Anal. Biochem.* **1996**, *242*, 84.
47. M. Ronaghi; B. Pettersson; M. Uhlen; P. Nyren, *BioTechniques* **1998**, *25*, 876.
48. R. A. Edwards; B. Rodriguerz-Brito; L. Wegley; M. Haynes; M. Breitbart; D. M. Peterson; M. O. Saar; S. Alexander; E. C. Alexander; R. Rohwer, *BMC Genomics* **2006**, *7*, 57.
49. M. Ronaghi, *Genome Res.* **2001**, *11*, 3.
50. S. J. Benkovic; C. E. Cameron, *Methods Enzymol.* **1995**, *262*, 257.
51. S. Karamohamed; J. Nilsson; K. Nourizad; M. Ronaghi; B. Pettersson; P. Nyren, *Protein Exp. Purif.* **1999**, *15*, 381.
52. Z. Liu; C. Lozupone; M. Harmady; F. D. Bushman; R. Knight, *Nucleic Acids Res.* **2007**, *35*, e120.
53. C. Pedros-Alio, *Trends Microbiol.* **2006**, *14*, 257.
54. X. Mou; S. Sun; R. A. Edwards; R. E. Hodson; M. A. Moran, *Nature* **2008**, *451*, 708.
55. J. Han; M. Kamber, *Data Mining: Concepts and Techniques*; Morgan Kaufmann Publishers: San Francisco, CA 9410, USA, 2001.
56. P. H. A. Sneath; R. R. Sokal, *Numerical Taxonomy*; San Francisco: Freeman, 1973.
57. M. B. Eisen; P. T. Spellman; P. O. Brown; D. Botstein, *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 14863.
58. L. K. Szeto; A. W.-C. Liew; H. Yan; S.-S. Tang, In *Gene Expression Data Clustering and Visualization Based on a Binary Heirarchical Clustering Framework*, Proceedings of the First Asia-Pacific Bioinformatics Conference (APBC 2003), Y. P. P. Chen, Ed.; Adelaide, Australia, 4–7 February 2003; pp 145–152.
59. T. Zhang; R. Ramakrishnan; M. Livny, *Proceedings of the ACM SIGMOD Conference on Management of Data*, Montreal, Canada, 1996; pp 45–53.
60. J. MacQueen, In *Some Methods for Classification and Analysis of Multivariate Observations*, Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, L. M. Le Cam, J. Neyman, Eds.; Statistics: Berkeley, USA, 1967; Vol. 1, pp 281–297.
61. A. K. Jain; R. C. Dubes, *Algorithm for Clustering Data*; Prentice Hall: Englewood Cliffs, NJ, 1988.
62. A. P. Gasch; M. B. Eisen, *Genome Biol.* **2002**, *3*, 1–22.
63. J. C. Bezdek, *Fuzzy Mathematics in Pattern Classification*; Cornell University: Ithaca, NY, 1973.
64. Y. Cheng; G. M. Church, In *Biclustering of Expression Data*, Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, P. E. Bourne, M. Gribskov, R. B. Altman, N. Jensen, D. Hope, T. Lengauer, J. C. Mitchell, E. Scheeff, C. Smith, H. Weissig, Eds.; La Jolla: San Diego, CA, USA, 2000; pp 93–103.
65. A. Ben-Dor; R. Shamir; Z. Yakhini, *J. Comput. Biol.* **1999**, *6*, 281.
66. A. Bellaachia; D. Portnoy; Y. Chen; A. G. Elkahloun, In *E-CAST: A Data Mining Algorithm for Gene Expression Data*, Proceedings of the 2nd ACM SIGKDD Workshop on Data Mining in Bioinformatics, M. J. Zaki, J. T.-L. Wang, H. Toivonen, Eds.; Edmonton: Alberta: Canada, 2002; pp 49–54.
67. G. Getz; E. Levine; E. Domany, *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 12079.
68. K. Y. Yeung; D. R. Haynor; W. L. Ruzzo, *Bioinformatics* **2001**, *17*, 309.
69. R. Agrawal; J. Gehrke; D. Gunopulos; P. Raghavan, *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Seattle, WA, USA, 1998; pp 94–105.
70. H. Wang; W. Wang; J. Yang; P. S. Yu, *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Madison, WI, USA, 3–6 June, 2002; pp 17–25.
71. D. Ghosh; A. M. Chinnaiyan, *Bioinformatics* **2002**, *18*, 275.
72. J. H. Kim; I. S. Kohane; L. Ohno-Machado, *J. Biomed. Informatics* **2002**, *35*, 25.
73. J. H. Kim; L. Ohno-Machado; I. S. Kohane, *Proceedings of the Pacific Symposium on Biocomputing*, Hawaii, USA, 2001; pp 79–86.
74. J. D. Saffer; C. L. Albright; A. J. Calapristi; G. Chen; V. L. Crow; S. D. Decker; K. M. Groch; S. L. Havre; J. M. Malard; T. J. Martin; N. E. Miller; P. J. Monroe; L. T. Nowell; D. A. Payne; J. F. R. Spinddla; R. E. Scarberry; H. J. Sofia; L. C. Stillwell; G. S. Thomas; S. J. Thurston; L. K. Williams; S. J. Zabriskie, In *Visualization and Integrated Data Mining of Disparate Information, Chemical Data Analysis in the Large: The Challenge of the Automation Age*, Proceedings of the Beilstein-Institut Workshop, M. G. Hicks, Ed.; Bozen, Italy, 22–26 May 2000; pp 107–113.
75. K. Torkkola; R. M. Gardner; T. Kaysser-Kranich; C. Ma, *Inf. Sci.* **2001**, *139*, 79.
76. T. Kohonen, *Biol. Cybern.* **1992**, *43*, 59.
77. T. Kohonen, *Proc. IEEE* **1990**, *78*, 1464.
78. P. Tamayo; D. Slonim; J. Mesirov; Q. Zhu; S. Kitareewan; E. Dmitrovsky; E. S. Lander; T. R. Golub, *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 2907.
79. R. J. Cho; M. J. Campbell; E. A. Winzeler; L. Steinmetz; L. A. Conway; L. Wodicka; T. G. Wolfsberg; A. E. Gabrielian; D. Landsman; D. J. Lockart; R. W. Davis, *Mol. Cell* **1998**, *2* (1), 65–73.
80. B. Fritzke, *Neural Netw.* **1994**, *7*, 1141.
81. J. Herrero; A. Valencia; J. Dopazo, *Bioinformatics* **2001**, *17*, 126.
82. J. Dopazo; J. M. Carazo, *J. Mol. Evol.* **1997**, *44*, 226.
83. M. Granzow; D. Berrar; W. Dubitzky; A. Schuster; F. J. Azuaje; R. Eils, *ACM SIGBIO Newsletter* **2001**, *21*, 16.
84. V. Vapnik, *The Nature of Statistical Learning Theory*; Springer: Germany, 1996.
85. N. Cristianini; J. Shawe-Taylor, *An Introduction to Support Vector Machines*; Cambridge University Press: Cambridge, 2000.
86. M. P. S. Brown; W. N. Grundy; D. Lin; N. Cristianini; C. W. Sugnet; T. S. Furey; L. Ares, Jr.; D. Haussler, *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 262.
87. T. Jaakkola; M. Diekhans; D. Haussler, In *Using the Fisher Kernel Method to Detect Remote Protein Homologies*, Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology, T. Lengauer, R. Schneider, P. Bork, D. L. Brutlag, J. I. Glasgow, H.-W. Mewes, R. Zimmer, Eds.; Heidelberg, Germany, 6–10 August 1999; pp 149–158.
88. T. S. Furey; N. Cristianini; N. Duffy; D. W. Bednarski; M. Schummer; D. Haussler, *Bioinformatics* **2000**, *16*, 906.
89. Y. Lee; C.-K. Lee, *Bioinformatics* **2003**, *19*, 1132.

90. G. Valentini, *Artif. Intell. Med.* **2002**, *6*, 281.
91. T. Joachims, In *Transductive Inference for Text Classification using Support Vector Machines,* International Conference on Machine Learning (ICML), 1999; pp 200–209.
92. R. P. Lippman, *IEEE Acoust. Speech Signal Process. Mag.* **1987**, *4*, 4 pp200–209.
93. C. H. Wu; J. W. McLarty, *Meth. Comp. Biol. Biochem.* **2001**, *1*, 1.
94. V. Brusic; J. Zeleznikow, *Knowl. Eng. Rev.* **1999**, *14*, 257.
95. G. D. Stormo; T. D. Schneider; L. Gold; A. Ehrenfeucht, *Nucleic Acids Res.* **1982**, *10*, 2997.
96. P. Baldi; S. Brunak, *Bioinformatics: The Machine Learning Approach*. MIT Press. Cambridge, MA 02142-1493, USA, 1998.
97. C. Wu; M. Berry; Y. S. Fung; J. McLarty, In *Neural Networks for Molecular Sequence Classification,* Proceedings of the 1st International Conference on Intelligent Systems for Molecular Biology, L. Hunter, D. Searls, J. Shavlik, Eds.; Bethesda, MD, USA, 1993; pp 429–437.
98. C. H. Wu; M. W. Berry; S. Shivakumar, *Mach. Learn.* **1995**, *21*, 177.
99. C. H. Wu, *Comput. Chem.* **1997**, *21*, 237.
100. J. T. L. Wang; Q. Ma; D. Shasha; C. H. Wu, *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Boston, MA, USA, 2000; pp 34–42.
101. N. Qian; T. J. Sejnowski, *J. Mol. Biol.* **1988**, *202*, 865.
102. S. Bicciato; M. Pandin; G. Didone; C. D. Bello, In *Analysis of an Associative Memory Neural Network for Pattern Identification in Gene Expression Data*, Proceedings of 1st Workshop on Data Mining in Bioinformatics, M. J. Zaki, H. T. T. Toivonen, J. T. L. Wang, Eds.; San Francisco, CA, USA, pp 22–30.
103. M. C. O'Neill; L. Song, *BMC Bioinformatics* **2003**, *4*, 211.
104. A. Narayanan; E. C. Keedwell; S. S. Tatineni; J. Gamalielsson, *Neurocomputing* **2004**, *61*, 217.
105. L. Li; C. R. Weinberg; T. A. Darden; L. G. Pedersen, *Bioinformatics* **2001**, *17*, 1131.
106. Y. P. P. Chen, *Bioinformatics Technologies*; Springer: Berlin/Heidelberg 2005; p 396.
107. A. W.–C. Liew; H. Yan; M. Yang; Y. P. P. Chen, *Microarray Data Analysis, Bioinformatics Technologies, Y. P. P. Chen, Ed.;* Springer: Berlin/Heidelberg 2005; Chapter 12, pp 353–388.
108. L. K. Szeto; A. W.–C. Liew; H. Yan; S.–S. Tang, *Proceedings of the 1st Asia-Pacific Bioinformatics Conference*, Adelaide, SA, Australia, 4 January, 2003; pp 145–152.
109. J. An; Y. P. P. Chen, *J. Biotechnol.* **2008**, *135*, 233.
110. P. T. Spellman; G. Sherlock; M. Q. Zhang; V. R. Iyer; K. Anders; M. B. Eisen; P. O. Brown; D. Botstein; B. Futcher, *Mol. Biol. Cell* **1998**, *9*, 3273.
111. A. Stein; A. Panjkovich; P. Aloy, *Nucleic Acids Res.* **2009**, *37*, D300.
112. R. Christen, *Curr. Opin. Biotechnol.* **2008**, *19*, 266.
113. V. Cherezov; D. M. Rosenbaum; M. A. Hanson; S. G. F. Rasmussen; F. S. Thian; T. S. Kobilka; H. J. Choi; P. Kuhn; W. I. Weis; B. K. Kobilka; R. C. Stevens, *Science* **2007**, *318*, 1258.
114. K. Khafizov; C. Anselmi; A. Menini; P. Carloni, *J. Mol. Model.* **2007**, *13*, 401.
115. A. Giorgetti; A. V. Nair; P. Codega; V. Torre; P. Carloni, *FEBS Lett.* **2005**, *579*, 1968.
116. S. Pifferi; G. Pascarella; A. Boccaccio; A. Mazzatenta; S. Gustincich; A. Menini; S. Zucchelli, *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 12929.

Biographical Sketches

Yi-Ping Phoebe Chen is an associate professor (reader) in Deakin University, Melbourne, Australia. Professor Chen is Bioinformatics Group leader in Deakin. She is the chief investigator and project leader in ARC Centre of Excellence in Bioinformatics. Professor Chen received her BInfTech degree with first class honors and Ph.D. in Computer Science (Bioinformatics) from the University of Queensland. Phoebe is currently working mostly on the emerging fields of Bioinformatics and Multimedia. Prior to that, she has done significant research in knowledge discovery, information retrieval, database query language and data visualization, as well as significant development work in data integration systems. Phoebe has published more than 100 refereed articles. She serves on the editorial boards of *IEEE Transactions on Multimedia*, *Current Bioinformatics*, and so on. She is a scientific advisor to Australia Bioinformatics. Phoebe has attracted over $5 million in competitive research and industry funding. Phoebe is a founder and steering committee chair of the Asia-Pacific Bioinformatics Conference (APBC).

Elena Ivanova is qualified in Microbiology (Sc.D. and Ph.D. in Microbiology, M.Sc. (Hons) in Biology/Biochemistry), and in Law (The Melbourne JD). Her professional interests are concentrated on the development and the coordination of collaborative research in fundamental and applied areas of Bio/Nanotechnology/Biochemistry and Microbiology. The research interests are focused in the area of interaction of microorganisms and biomolecules (proteins or DNA) with nanostructured surfaces for the fabrication of static and dynamic planar biodevices resulted in patent application. She has authored and coauthored for more than 155 published papers, 4 patents, a book, and 5 book chapters. Ivanova serves as a reviewer for 6 scientific journals, an external Ph.D. thesis examiner, an editor for *Microbes & Environments* (2009–2010), and a member of the editorial board of *The Open Biomaterials* journal. She is a recipient of Morrison Rogosa Award, American Society for Microbiology (1999); UNESCO Biotechnology Fellowship (1997); and AIST Fellowship, Japan, Foreign Researcher Invitation Program of the Agency of Industrial Science and Technology (1994–1997). She has also received a short-term JSPS Fellowship Japan (2008); Morrison Rogosa Award, American Society for Microbiology (1999); UNESCO Biotechnology Fellowship (1997); and AIST Fellowship, Japan, Foreign Researcher Invitation Program of the Agency of Industrial Science and Technology, (1994–1997).

Feng Wang received her Ph.D. from The University of Newcastle (NSW, Australia) in 1994 in theoretical chemistry and followed it with a two-year prestigious Canada International Postdoctoral Fellowship at the University of Waterloo. From 1996 to 2001, she worked as a research fellow in the School of Chemistry, The University of Melbourne. Feng joined Swinburne University of Technology in 2003 as a Senior Lecturer after nearly two years at Victoria Partnership for Advanced Computing (VPAC) as a computational scientist and senior computational scientist. Feng was promoted in 2004 to associate professor and again in 2008 to professor. Feng has published over 80 peer-refereed articles and book chapters, including an edited book. Feng has been Fellow of Royal Australia Chemical Institute since 2001 and received the Swinburne University of Technology Vice-Chancellor's Research Award in 2006.

Paolo Carloni obtained his Ph.D. in Computational Biophysics from the University of Florence, Italy in 1993 with a thesis on 'Theoretical Studies on Metalloproteins' supervised by L. Banci (University of Florence) and by Michele Parrinello (IBM Zurich Research Laboratory). Currently, he is a professor of Chemistry at the International School for Advanced Studies (SISSA), Trieste, Italy, head of the SISSA unit of the Italian Institute of Technology (www.iit.it), and head of the Statistical and Biological Physics Sector of SISSA. Paolo is using and developing molecular simulation and bioinformatics tools to gain insights in molecular medicine, structural genomics, and cell biology. He has published more than 130 papers and edited a book (h-index = 29). He has delivered more than 130 seminars and colloquia throughout the world at major universities, research and industrial laboratories, as well as at professional meetings. He has organized a dozen of international conferences. He is in the editorial board of the journal – *Proteins: Structure, Function and Bioinformatics*.

9.16 Natural Products Research and Metabolomics

V. Craige Trenerry and Simone J. Rochfort, DPI-Werribee Centre, Werribee, VIC, Australia

9.16.1 Introduction

Natural products research has its origins in ethnopharmacology and ethnoveterinary practice. Plants with medicinal properties have been used throughout history, and modern science has, in many cases, been able to isolate, identify, and elucidate the pharmacological mode of action (MOA) of the active agents. Modern natural products research encompasses a range of approaches ranging from the holistic to the reductionist. Metabolomics offers the natural product chemist new ways to discover and validate bioactives. Modern natural products research has relied on the purification of individual components and verification of activity in specific bioassays. Bioassays have changed significantly over time. For example, whole-animal studies such as mouse cancer models are now rarely carried out in discovery-phase work. In the pharmaceutical industry, there has been an emphasis on cellular or molecular systems to reduce time and cost of the screening process. More traditional approaches, such as herbal medicine, may employ whole extracts of one or more plant or animal species in a formulation.

In recent years, there has also been increased interest in functional foods – foods that provide health benefits beyond basic nutrition. Functional foods may include fortified foods where nutrients have been added, for example, calcium in orange juice, or foods where natural bioactive constituents have been enhanced through either addition or breeding strategies. Much like traditional medicines, these foods are complex mixtures of chemicals.

In terms of modern pharmaceutics, natural products have been most successful in the areas of cancer and antibiotics. It is therefore not surprising that much research in this area has focused on whole cells (human or microbial). This has led to many compounds or extracts being identified as bioactive but with little explanation of the MOA. Discovering the MOA is often extremely difficult and time consuming (see also Volume 3). For example, it took almost 20 years from the discovery and structure determination of camptothecin for its MOA to be determined (1966–85) and it was its unique MOA that rekindled interest in the compound and ultimately led to the development of the important commercial anticancer drugs topotecan and CPT-11.[1] A metabolomics approach may have significantly reduced this time. Molecular-based assays are a way to determine MOA but do not allow synergism between molecular activities. This can be important; for example, certain isoflavones such as genistein reportedly promote cancer cell death via a number of mechanisms including the activation of apoptosis via endoplasmic reticulum (ER) stress pathways (m-calpain, GADD153, GRP78, and caspase-12) and mitochondrial apoptotic pathways (Mcl-1 downregulation and Bad cleavage), as well as via their action on estrogen receptor β (ERβ) and nuclear factor-kappaB (NF-κB).[2]

Herbal medicines and functional foods are chemically complex and there may be synergistic effects of not only one compound across targets but also multiple compounds across multiple targets. This makes substantiation of pharmacological activity even more difficult. Considering that herbal medicines have been developed over centuries of observational evidence of effect on the patient, the problem of MOA can be even more complex since it may be different but specific organs that bioactives target to produce an overall effect. The same may be true for functional foods.

Metabolomics offers a new way of studying complex molecular problems and is particularly applicable for natural products research. Metabolomics is the study of global metabolite profiles in a system (e.g., cell, tissue, or organism) under a given set of conditions.[3–5] Metabolites are the result of the interaction of the system's genome with its environment and are not merely the end product of gene expression but also form part of the regulatory system in an integrated manner. The analysis of the metabolome – the set of metabolites synthesized by a biological system – is particularly challenging due to the diverse chemical nature of metabolites (see also Volumes 1 and 2). Generally, these include organic compounds, for example, amino acids, organic acids, fatty acids, vitamins, and lipids, as well as inorganic compounds. Metabolites constitute a diverse set of atomic arrangements and this provides wide variations in chemical (molecular weight, polarity, solubility) and physical (volatility) properties.

In the pharmaceutical industry, the techniques are being used to examine 'off-target effects' particularly for the early identification of toxicity. MOA can be studied through metabolomics and can also be used as a quality control tool for complex mixtures such as foods or herbal medicines. Similarly, the tools and expertise of natural products chemists are essential in metabolomics, particularly in biomarker discovery (see also Volume 9). Biomarker discovery via untargeted metabolomics can lead to metabolite signatures (nuclear magnetic resonance (NMR) spectroscopy, mass spectrometry (MS), etc.) that are not present in current metabolomics databases. This is particularly true for plant secondary metabolism studies and nonmammalian metabolites. Structure elucidation then becomes critical to understanding the metabolomics results and for biomarker development.

Metabolomics has its roots in early metabolite profiling studies but is now a rapidly expanding area of scientific research in its own right and has been labeled one of the new 'omics,' joining genomics, transcriptomics, and proteomics employed to the understanding of global systems biology.[4,5] Systems biology represents a major challenge in that it aims to integrate genomics, transcriptomics, proteomics, and metabolomics for a global understanding of biological systems with the aim to obtain a better understanding of how individual pathways or metabolic networks are related. Systems biology does not investigate individual genes, proteins, or metabolites one at a time, but rather investigates the behavior and relationships of all the elements in a particular biological system while it is functioning.

Metabolite analysis is complicated by the number of analytes, their diversity, and dynamic ranges. It is estimated that the metabolome extends over 7–9 orders of magnitude of concentration (pmol–mmol) and the number of metabolites in the plant system alone is estimated to exceed 200 000. Of more relevance and significance is the number and diversity of analytes likely to be encountered in a single system; for example, there are over 500 known flavonoids and these compounds commonly occur as sugar conjugates with up to 10 sugars being involved and more than 200 different aglycone–sugar conjugates have been identified in grapes of *Vitis vinifera*. Thus, a successful analytical approach in metabolomics must be capable of accurately measuring numerous known and unknown compounds that span a diverse chemical spectrum and a large concentration range. Sample preparation procedures and methods of quantification cannot yet meet these demands.[5]

Metabolomic approaches use analytical techniques such as high-field NMR spectroscopy and MS to measure populations of low-molecular-weight metabolites in biological samples. Advanced statistical and bioinformatics tools are then employed to maximize the recovery of information and interpret the large data sets generated.

Historical approaches to metabolite analysis include metabolite profiling, metabolite fingerprinting, and target analysis. Metabolite profiling is defined as the identification and quantification of a selected number of predefined metabolites, generally related to a specific metabolic pathway(s). Sample preparation and instrumentation are employed to isolate the compounds of interest from possible matrix effects prior to detection, normally with chromatographic separation prior to detection with MS. Metabolite fingerprinting is the high-throughput, rapid, global analysis of samples. Quantification and metabolite identification are generally not employed. This is widely used as a screening tool to discriminate between samples of different biological status or origin. Sample preparation is simple and, as chromatographic separation may be absent, analysis times can be rapid. Metabolite target analysis is the qualitative and quantitative analysis of one or more metabolites related to a specific metabolic reaction. Extensive sample preparation and separation from other related metabolites is required.[6] A summary of the different metabolomic strategies for sample preparation and sample analysis is depicted in Figure 1.

This chapter discusses the analytical tools of metabolomics – the instrumentation and data analysis techniques – and provides comments on how this emerging science can be employed in natural products research.

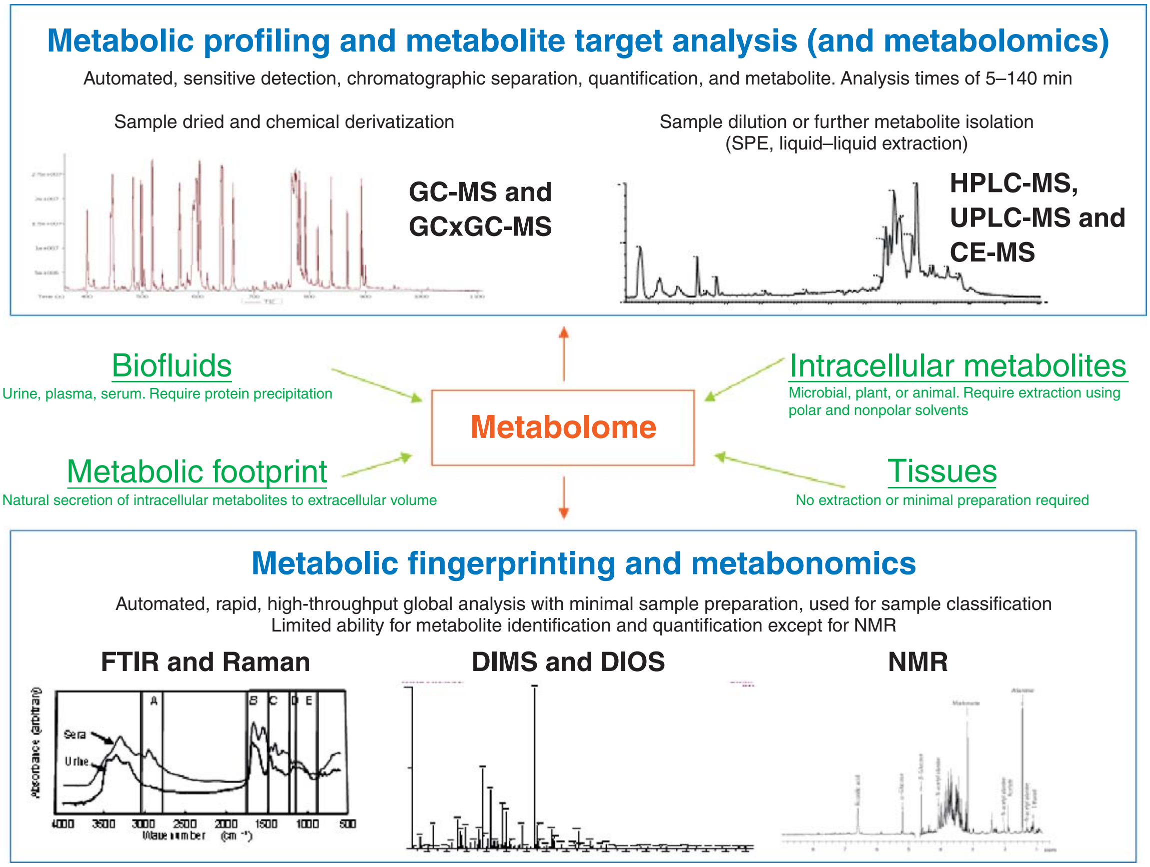

Metabolic profiling and metabolite target analysis (and metabolomics)
Automated, sensitive detection, chromatographic separation, quantification, and metabolite. Analysis times of 5–140 min
Sample dried and chemical derivatization
Sample dilution or further metabolite isolation (SPE, liquid–liquid extraction)
GC-MS and GCxGC-MS
HPLC-MS, UPLC-MS and CE-MS
Biofluids
Urine, plasma, serum. Require protein precipitation
Intracellular metabolites
Microbial, plant, or animal. Require extraction using polar and nonpolar solvents
Metabolome
Metabolic footprint
Natural secretion of intracellular metabolites to extracellular volume
Tissues
No extraction or minimal preparation required
Metabolic fingerprinting and metabonomics
Automated, rapid, high-throughput global analysis with minimal sample preparation, used for sample classification
Limited ability for metabolite identification and quantification except for NMR
FTIR and Raman
DIMS and DIOS
NMR
Absorbance (arbitrary)
Sera
Urine
A
B
C
D
E
4000
3500
3000
2500
2000
1500
1000
500
Wave number (cm⁻¹)

Figure 1 Summary of the different metabolomic strategies for sample preparation and sample analysis. Reprinted from W. B. Dunn; D. I. Ellis, *Trends Analyt. Chem.* **2005**, *24* (4), 285–294. Copyright (2005), with permission from Elsevier.

9.16.2 Analytical Techniques

It is generally accepted that a single analytical technique will not provide sufficient visualization of the metabolome and therefore multiple technologies are needed for a comprehensive view. Accordingly, many analytical technologies have been enlisted to profile the metabolome with the selection of the most suitable technology generally a compromise between speed, selectivity, and sensitivity.

MS (either stand alone or hyphenated with gas chromatography (GC), liquid chromatography (LC), or capillary electrophoresis (CE)) and high-field NMR spectroscopy are the main techniques used for metabolomic studies.[6–20] The main features of MS are high sensitivity, high resolution, wide dynamic range, coverage of a wide chemical diversity, robustness, and limited structure elucidation of unknown compounds. MS is inherently more sensitive than NMR spectroscopy. On the other hand, NMR is more rapid and selective. Hyphenated MS methods, for example, GC-MS and LC-MS, are commonly used for metabolomic studies and offer good sensitivity (GC-MS, 10^{-12} mol; LC-MS, 10^{-15} mol) and selectivity, but have relatively long analysis times. Fourier transform infrared (FTIR) also provides rapid, nondestructive assays and is a valuable tool for metabolite fingerprinting.[6]

9.16.2.1 Nuclear Magnetic Resonance Spectroscopy

NMR spectroscopy is one of the cornerstones of modern natural products research and is used routinely to monitor compound purity through the isolation procedure and to determine the structure of novel bioactives. NMR also has a long history of use in metabolism studies utilizing proton and phosphorus nuclei in both liquid and solid states and *in vivo* and *in vitro*. More recently, NMR has been applied in metabolomics (or metabonomics) for metabolite profiling, quantification, and structure elucidation.[4]

9.16.2.1.1 Instrumentation

Relatively low sensitivity is one of the major drawbacks of NMR analytical techniques compared to techniques such as MS. However, NMR does offer advantages over MS, in that it is nondestructive and less biased. In general, any soluble molecule with protons can be measured in a quantitative manner, unlike MS, where sensitivity relies on the ionization properties of the individual metabolites. However, signal overlap may cause problems in complex samples. Samples do not require derivatization as is often the case for GC or GC-MS applications, and NMR is more amenable to compounds such as sugars, amines, and nonreactive compounds.[21] Proton spectra are also relatively quick to acquire and though actual acquisition time varies depending on sample type, concentration, and instrumentation used, it is not unusual to have a 5-min acquisition time provide sufficient data for metabolomics. This is an obvious advantage compared to chromatography-based metabolomics technologies such as GC, CE, and LC and the associated hyphenated techniques (GC-MS, CE-MS, and LC-MS).

There have been several advances in recent years that significantly enhance the sensitivity of NMR techniques. The first is the increasing field strengths of commercially available NMR magnets. NMR instruments of 22.31 T or 950 MHz (the observation frequency for ^{1}H NMR spectra) are now commercially available and these instruments permit higher sensitivity and much greater spectral dispersion. The second major advance has been the advent of cryogenically cooled NMR probes. These probes are bathed in cryogenic liquids such that the electronics of the probes (the preamplifier and detector coil) are maintained at around 20 K while maintaining the sample in a liquid state at normal operating temperatures (typically around 298 K). This effectively reduces thermal noise and improves signal to noise by a factor of 3–5 and can significantly reduce the time for data acquisition.[22,23] Small-volume probes utilizing microcoils also offer significant increases in sensitivity (8- to 12-fold)[24] and can be applied to both liquid and solid samples.[25] These advances allow increasingly low level of metabolites or natural products to be detected and studied by NMR, moving from milligram to submicrogram amounts for structural elucidation.

Although the vast majority of NMR metabolomics experiments are carried out in the liquid state, the analysis of intact tissue is made possible by solid-state NMR. This technique offers the obvious advantage of simple sample preparation and removes bias associated with differential solubility of metabolites. Magic angle spinning (MAS) is used in solid-state NMR to remove effects such as chemical anisotropy. In solution-state NMR, this is achieved by the isotropic tumbling of molecules in solution, which averages out the effects of molecular orientation.[26] In solid-state NMR, this is achieved by spinning the sample at the 'magic angle' of approximately 54.7° to the magnetic field. The spinning rate can be varied and for tissue samples is generally between 4 and 8 kHz, which is sufficient to produce spectra of similar quality to solution-state spectra without rapidly damaging the tissue. Typical one-dimensional (1D) spectra can be obtained in 10 min at 600 MHz. Two-dimensional (2D) spectral acquisition may take several hours and tissue damage through heating can become a problem though this can be minimized by obtaining spectra at near freezing.[27]

High-throughput NMR analysis is of importance for both natural products research and metabolomics. Improvements in NMR field strength and probe design have meant that the time required to obtain NMR data has decreased. The introduction of automatic sample handling instrumentation has also offered significant benefits for high-throughput NMR. The instrumentation available ranges from sample changers that will automatically transport standard 5-mm tubes to and from the magnet, through liquid handling systems that utilize flow probes so that samples can be analyzed from 96-well plates or smaller, to systems that include solid phase extraction (SPE) sample cleanup before flow analysis. Utilization of the flow probes in NMR has allowed the development of hyphenated NMR techniques such as LC-NMR, SPE-LC-NMR, and LC-MS-NMR. The advantages these hyphenated techniques offer for natural products research have been the subject of a recent review.[28] Recently, the design of a multidimensional LC-SPE-NMR (LC^2-SPE-NMR) for complex mixture analysis, for example, drug impurities at the low microgram per milliliter level, was described by Alexander and Bernard.[29] While on-flow analysis is feasible for major constituents, the more usual approach is a stop-flow technique whereby the high-performance liquid chromatography (HPLC) is paused while data are acquired on part of the eluent. Of course, while rich in information, these hyphenated NMR techniques also increase the time taken for each experiment, reducing one of the benefits of NMR metabolomics compared to other techniques.

9.16.2.1.2 Experimental techniques

9.16.2.1.2(i) Solution NMR

Many metabolomics datasets are acquired in aqueous solutions and this determines the type of 1D proton NMR experiments that are acquired. Solvent suppression techniques such as presaturation, RECUR, 1D-NOE, or WATER GATE are therefore important. Spectra from blood or serum samples tend to show broad peaks from lipoprotein or protein signals and pulse sequences such as the Carr–Purcell–Meiboom–Gill (CPMG) can be employed to enhance or suppress the signals from macromolecules.[10] The pH is an important consideration, particularly for compound identification in aqueous samples since pH can significantly affect the chemical shifts of certain resonances. For metabolomics, pH is generally measured to within 0.05 units and can be done either with an external probe or with an internal chemical standard that is sensitive to pH changes (e.g., imidazole or difluorotrimethylsilanylphosphonic acid (DFTMP)). Quantitative NMR techniques for metabolomics are often acquired in water or D_2O and a chemical standard (e.g., dimethylsilapentanesulfonic acid (DSS) or 2, 2′, 3, 3′-deuterotrimethylsilylproprionic acid (TSP)) is added at a known level to enable quantification as well as to act as a chemical reference.[21] Despite its relative insensitivity, ^{13}C NMR has also been applied to metabolomics in the study of brain tissue metabolism using labeled substrates.[30] 2D NMR techniques can also be used for metabolomics; for example, urine profiles from cattle treated with anabolic steroids have been analyzed by heteronuclear multiple bond correlation (HMBC) techniques with linear discriminant analysis (DA).[31] However, the most widely used 2D NMR technique in metabolomics is 2D J-resolved spectroscopy. This technique separates chemical shift and coupling information onto different axes and produces less congested 1D projections that are advantageous for both peak assignment and statistical data analysis (**Figure 2**).[32,33]

9.16.2.1.2(ii) Magic angle spinning nuclear magnetic resonance spectroscopy

Preparation for MAS NMR is relatively simple. The tissue sample (typically 5–30 mg) is placed in the rotar and perfused with D_2O to provide a solvent lock. The D_2O may also include a chemical shift standard (e.g., DSS or TSP) or buffer (e.g., phosphate-buffered saline (PBS), saline). Tissue samples behave like a mixture of solids and liquids. Lipids and macromolecules tend to act more like solids while small molecules in the cytoplasm act more like solution samples. This means that many of the same types of experiments for solution-state NMR can be applied to tissue samples with good resolution. Water suppression techniques such as presaturation, CPMG, presaturation-nuclear overhauser effect spectroscopy (presat-NOESY), SEEN pulse sequence, and presaturation with a

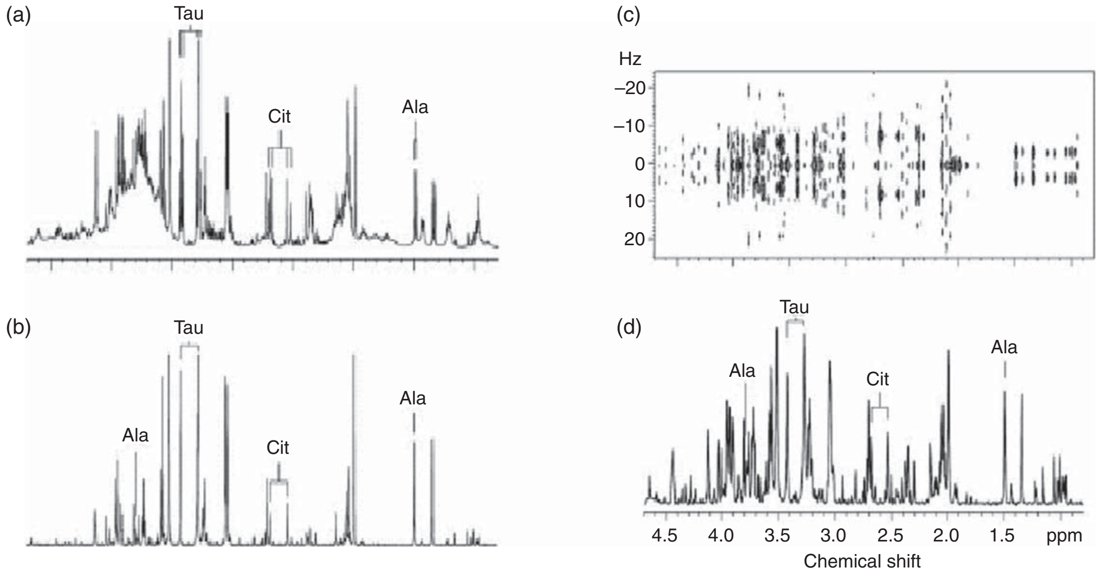


Figure 2 ^{1}H NMR data of stage 38 medaka embryo extracts, including (a) 1D spectrum, (b) the 1D skyline projection (p-JRES) of (c) a 2D JRES spectrum, and (d) the preprocessed p-JRES spectrum. Resonances due to citrate (Cit), taurine (Tau), and alanine (Ala) are indicated. Reprinted from M. R. Viant, *Biochem. Biophys. Res. Commun.* **2003**, *310* (3), 943–948. Copyright (2003), with permission from Elsevier.

pulse gradient stimulated echo sequence (presat-PGSTE) can all be used to acquire 1D MAS spectra.[27] 2D NMR experiments such as spin-echo correlated spectroscopy (COSY), totally correlated spectroscopy (TOCSY), and heteronuclear single-quantum correlation (HSQC) can equally well be obtained on tissue samples as solutions.[27] 1D techniques that have yielded some interesting insights into metabolomics include proton-decoupled ^{31}P experiments, which can be examined in conjunction with ^{1}H spectra.[34,35] For example, **Figure 3** illustrates the ^{1}H and decoupled ^{31}P spectra of liver tissue obtained during an investigation of galactosamine-induced hepatotoxicity.

9.16.2.1.3 Applications

NMR metabolomics and natural products techniques can be employed in a number of ways. In a study of the chemical diversity of two legume species (chickpeas and lentils), a methanolic extract was analyzed by NMR and principal components analysis (PCA) to determine chemical similarity (**Figure 4**). The set of legume cultivars displayed chemical diversity between cultivars but, not unexpectedly, greater diversity between legume species. Examination of the loadings of the model showed that the concentration of sugars varied and there were also significant differences between the secondary metabolites of the chickpeas compared to the legumes.

Natural product isolation techniques (C18 preparative HPLC) followed by 2D NMR allowed the isolation and identification of one of the discriminatory chickpea secondary metabolites, chromosaponin I (**Figure 5**).

The type of statistical approach used above to examine chemical diversity of legume cultivars can also be applied to chemotaxonomic approaches and is useful for bioactive natural products research when employed as part of a dereplication strategy. Pierens *et al.*[36] demonstrated this for plant and marine sponge extracts. High-throughput ^{1}H NMR (flow probe with samples introduced from a 96-well plate) was used to generate spectra, which were examined automatically using clustering techniques to identify extracts that contained the same bioactive compound or compounds. The authors found that multivariate analysis of the NMR data yielded statistically valid clustering of the extracts.

Metabolomics is often employed in early discovery projects and used for the identification of biomarkers. In mammalian systems, the metabolites so identified are often contained in NMR databases, which, depending on the database, are either commercially or freely available to the public (e.g., Human Metabolome Database).[37] There is considerably less nonmammalian NMR information available in these databases. While this is an obvious problem for plant biochemists, it can also impact biomarker discovery utilizing other organisms,

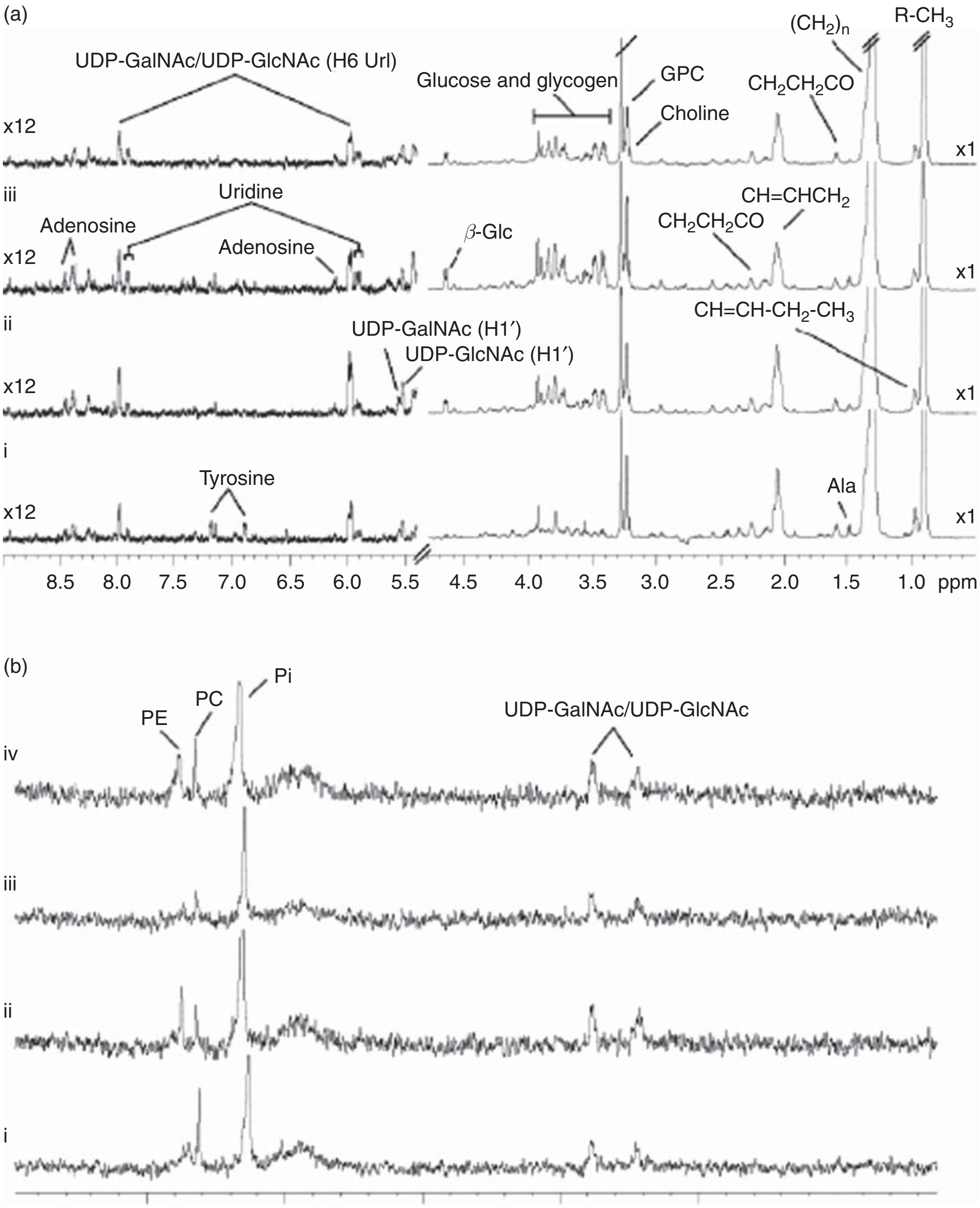


Figure 3 (a) Representative ^1H MAS NMR spectra of rat liver and (b) the corresponding ^{31}P-[^1H] MAS NMR spectra for (i) galN-treated, (ii) galN and 0.5 g kg^{-1} uridine, (iii) galN and 1.0 g kg^{-1} uridine, and (iv) galN and 2.0 g kg^{-1} uridine. Key: Ala, alanine; PC, phosphocholine; PE, phosphoethanolamine; P_i, inorganic phosphate; UDP-GlcNAc, uridine 5′-diphosphate-*N*-acetylglucosamine; UDP-GalNAc, uridine 5′-diphosphate-*N*-acetylgalactosamine. Reprinted with permission from M. Coen; Y. S. Hong; O. Cloarec; C. M. Rhode; M. D. Reily; D. G. Robertson; E. Holmes; J. C. Lindon; J. K. Nicholson, *Anal. Chem.* **2007**, *79* (23), 8956–8966. Copyright (2007) American Chemical Society.

particularly environmental toxicity studies that tend to focus on sentinel organisms such as earthworms for terrestrial environments or mussels for coastal environments. The unusual metabolite 2-hexyl-5-ethyl-3-furansulfonate has been identified as a major metabolite in the earthworm *Lumbricus rubellus* and a potential biomarker for metal toxicity.[38,39]

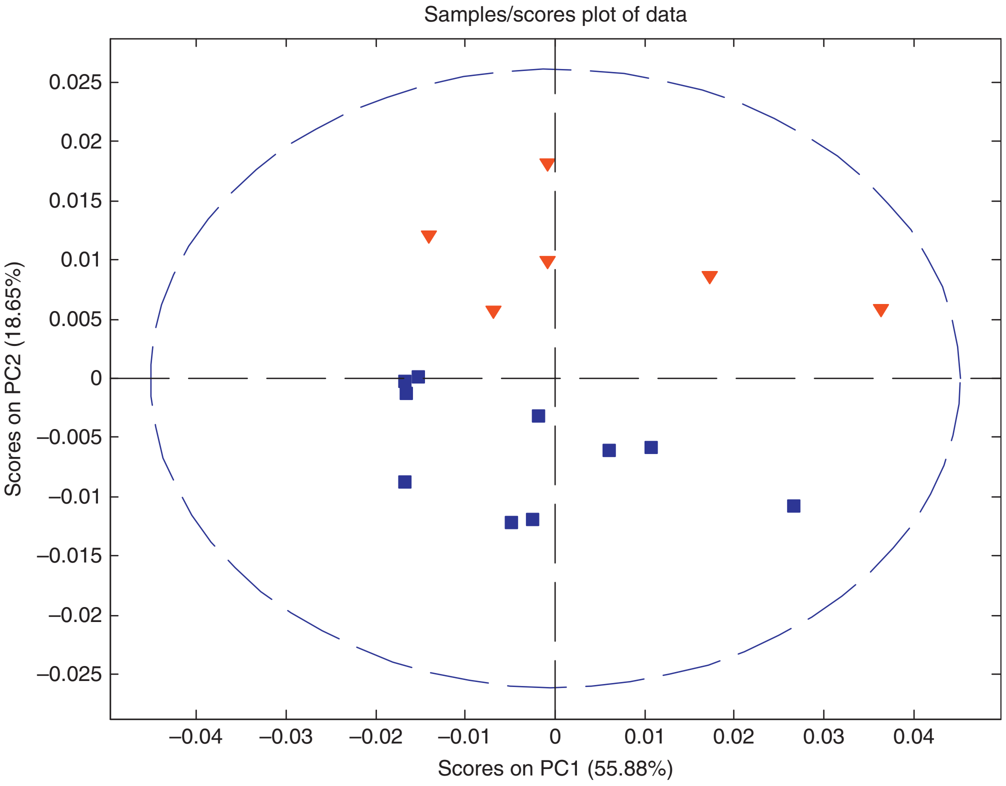


Figure 4 PCA analysis of the 1D ^{1}H NMR spectra of the chickpea (▲) and lentil (■) cultivars.

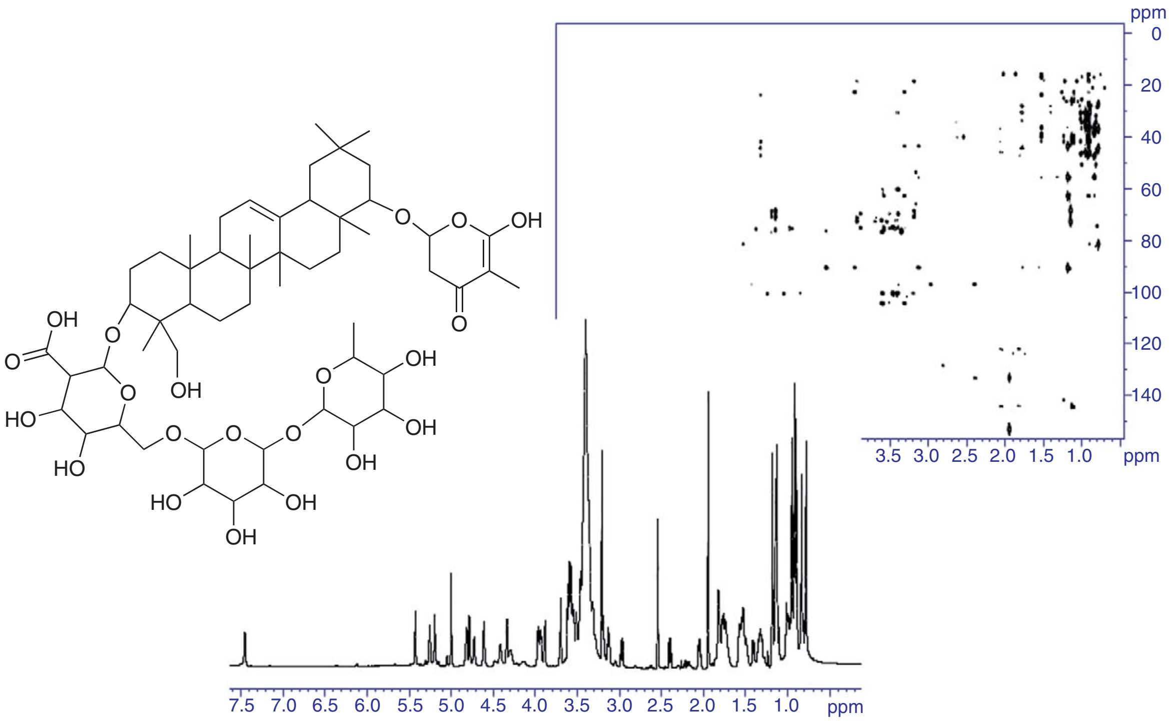


Figure 5 ^{1}H and gHMBC NMR spectra of chickpea metabolite chromosaponin I.

Metabolomics techniques can also be used as a quality control for complex natural products mixtures. Recently, NMR combined with PCA statistical clustering was applied to monitor the quality of claimed antimalarial herbal remedies. The authors demonstrate that the products incorporating *Artemisia afra* or *A. annua* could be distinguished and that the bioactive artemisinin was an important discriminant factor.[40]

NMR metabolomics has also been applied to investigate potential MOA. Plant researchers from BASF generated more than 400 ^{1}H NMR spectra from plants treated with different herbicides. Statistical analysis of the NMR spectra accurately clustered the extracts based on the metabolic action of the herbicides.[41] NMR metabolomics may therefore offer the natural product researcher new ways to investigate MOA for compounds where activity is demonstrated but MOA is undetermined.

NMR metabolomics increasingly makes use of natural product structure elucidation techniques such as 2D NMR, particularly for plant-based studies. In both natural products research and metabolomics, strategies such as quantitative NMR and low-volume NMR are employed, while the use of hyphenated NMR techniques is also expanding. These techniques have been applied in numerous studies; examples are given in **Table 1** (see also Volume 9).

As sensitivity increases and the availability of databases with chemical shift information improves, it seems likely that the use of NMR in metabolomics will increase.

9.16.2.2 Mass Spectrometry

There are many types of mass analyzers available for metabolomic studies, either as stand alone or hyphenated with GC, CE, or LC.[6–8,10,12,13,17,20,50] The common mass analyzers include single quadrupole (Q), triple quadrupole (QqQ), time-of-flight (TOF), ion trap, orbitrap, and Fourier transform ion cyclotron resonance (FT-ICR). In addition, there are a number of hybrid systems that combine two basic types of mass spectrometers, such as quadrupole time-of-flight (Q-TOF), quadrupole linear ion trap (Q-trap), and ion trap Fourier transform ion cyclotron resonance (IT-FT-ICR) mass spectrometers. Quadrupole instruments are robust, have a high linear dynamic range, and are capable of analyzing a m/z range of 50–4000. The single quadrupole mass analyzer is the simplest option, and it provides nominal mass resolution; however, its low duty cycle for full scan data acquisition reduces the sensitivity and limits its application for metabolic profiling. Single quadrupole systems are used as chromatographic mass detectors, whereas the QqQ mass spectrometer allows not only m/z information, but also the formation of fragment ions, allowing experiments such as neutral loss and product ion spectra to be performed. Using a process known as multiple reaction monitoring (MRM), the first and third quadrupoles monitor the parent and product ion, respectively, of a fragmentation transition that is specific for the target analyte. Along with providing highly selective detection, this arrangement avoids the duty cycle limitation. The MRM experiment, a targeted analysis experiment of highest sensitivity and accuracy, is the application of choice for clinical, forensic, and pharmaceutical chemistry and in pesticide/antibiotic residue testing.

Ion trap mass spectrometers are compact instruments that cover a m/z range of up to 6000 and are able to operate in full scan mode at high and low resolution. One of the benefits of ion trap instruments is the ability to perform successive fragmentation steps (MS^n). This provides more structural information than other mass analyzers. Quadrupole linear ion trap hybrid instruments (Q-trap or QqLIT) combine the MS^n capabilities of the ion trap instruments with the neutral loss and precursor ion scan capabilities of QqQ instruments. TOF instruments feature fast scanning capabilities, wide mass range, and high resolution (5000–20 000 full-width, half-maximum (FWHW)) and mass accuracy. These instruments are extremely useful for profiling complex metabolic mixtures. In order to perform MS–MS experiments with TOF instruments, another mass analyzer has to be combined. In Q-TOF instruments, the last quadrupole of the QqQ configuration is replaced by a TOF analyzer. These instruments combine the stability and robustness of the quadrupole analyzer with the fast scanning capabilities, accuracy (<5 ppm), and high sensitivity of TOF mass analyzers. Orbitrap and FT-ICR MS offer the highest resolution available (>100 000 FWHW) and high-accuracy fragment masses. For high-end hybrid instruments like Q-FT-ICR MS and Q-trap-FT-ICR MS instruments, ion selection and fragmentation can be performed outside the cyclotron of the FT-ICR MS.

Metabolic profiling and metabolite fingerprinting approaches require a sensitive full scan mode and exact masses. Therefore, Q-TOF or Q-trap-FT-ICR MS instruments are advantageous. In contrast, for targeted

Table 1 Selected NMR spectroscopy applications

Analyte(s)	*Matrix*	*Conditions/Experiment*	*Reference*
Drug impurities	Pharmaceutical preparations	Multidimensional LC-SPE-NMR	29
Alkaloids	*Corydalis solida*	LC-SPE-NMR	42
Reference materials	–	Automated microflow 5 µl	43
Amino acids, organic acids, carbohydrates, phenylpropanoids	Bacterially infected *Brassica rapa*	^{1}H NMR and 2D NMR	44
Terpenes	*Anisomorpha buprestoides* (walking stick insect) venom	Capillary NMR tube volume 10 µl, 1D and 2D NMR	45
Terpenes, indole alkaloids	*Catharanthus roseus* infected by phytoplasma	^{1}H NMR and PCA data analysis	46
Taxol	*Taxus brevifolia*	Quantitative NMR	47
Naphtodianthrones, phloroglucinols, flavonoids, phenolic acids	*Hypericum perforatum* (St. John's wort)	LC/DAD/SPE/NMR, ^{1}H NMR, COSY, TOCSY	48
Sterols, fatty acids, diacylglycerophospholipids, galactosyldiacylglycerols, sulfolipids, pheophytins, carotenoids, carbohydrates, polyols, organic acids, amino acids	*Lactuca sativa* (lettuce) (aqueous and lipophilic extract)	COSY, TOCSY, HSQC, HMBC identification and assignment	49

analysis of selected metabolites, QqQ instruments, in particular Q-trap instruments (with their capability for MRM), are frequently used.

Matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF MS) has the potential to contribute significantly to metabolomics. It offers the advantage of high throughput, the unique ability to generate singly charged ions of < 1000 Da, and is tolerant to moderate levels of salt. MALDI-TOF-MS/TOF tandem MS also permits the discrimination of isomeric molecular species that would not be possible using spectra of molecular ions alone.

9.16.2.2.1 Gas chromatography-mass spectrometry

GC-MS has been used extensively during the past few decades for the measurement of metabolite profiles of biological fluids in clinical trials and is one of the key technological platforms for metabolomics.[6,51,52] These instruments provide low initial start-up costs compared to other hybrid instruments (e.g., CE-MS, LC-MS, or LC-NMR), unsurpassed chromatographic reproducibility and resolution, highly repeatable mass spectral fragmentation, and few, if any, matrix effects. The main ionization techniques in GC-MS are electron ionization (EI) and chemical ionization (CI). EI, where the analyte vapor is subjected to bombardment by energetic electrons (typically 70 eV), is the oldest, most used, and probably best understood ionization technique. EI is performed in a high-vacuum source avoiding intermolecular collisions. Hence, spectra are highly reproducible, and as a consequence, several comprehensive databases are available; however, the interpretation of spectra from derivatized unknown compounds can be difficult. For CI, a reagent gas is applied to the ion source leading to CI, which results in spectra showing less fragmentation but a clear molecular ion, which is not always the case in EI. Both techniques can be combined with any mass analyzer, for example, Q or Q-TOF.

The advantage of GC-MS is that it can separate a large number of low-molecular-weight metabolites with high resolution. However, two main drawbacks of the GC-MS technique for the analysis of polar compounds in biological samples are the relatively long sample preparation time (extraction, derivatization) and long analytical run times. Many compounds can be derivatized so that they become more suitable for GC, but not all compounds will be amenable to this and thermolabile metabolites cannot be analyzed at all. In GC-MS, the majority of metabolites require chemical derivatization at room or elevated temperatures to provide volatility, and to cover the range of chemical functionality of metabolites, a two-stage derivatization can be employed. For example, carbonyl functional groups are converted to oximes with *O*-alkylhydroxylamine solutions, followed by formation of the trimethylsilyl (TMS) esters with silylating reagents (typically *N*-methyl-*N*-(trimethylsilyl)-trifluoracetamide), to replace exchangeable protons with TMS groups. Oxime formation is required to eliminate undesirable slow and reversible silylation reactions with carbonyl groups, whose products can be thermally labile. The presence of water can result in a breakdown of the TMS esters, although extensive drying and the presence of excess silylating reagents can limit the process. However, extensive sample drying can result in the loss of volatile metabolites. Ideally, an automated system employing a 'derivatization-when-required' approach is desirable to ensure maximum sample stability.[6]

Typically, GC-MS is performed with affordable single quadrupole mass analyzers; however, GC-TOF systems offer an attractive alternative to quadrupoles and provide greater m/z accuracies. These instruments have high scan speeds, which support ultrafast GC-MS and therefore the potential to profile increasingly complex mixtures. The separation of a wide range of metabolite classes, for example, aliphatic alcohols, acids, monosaccharides, fatty acids, sugar phosphates, sugar alcohols, disaccharides, trisaccharides, and sterols, in a methanol–chloroform–water plant extract is demonstrated in **Figure 6**.[9]

An unquestionable benefit of GC analysis is the possibility to correct retention time shifts easily by calculating retention time indices (Kovats index) or even by reanalyzing analytes on a second stationary phase resulting in a second confirmative index.

A significant development in the last decade is comprehensive 2D gas chromatography (GC × GC). Generally, the sample components are separated in the first column according to their volatilities, then small fractions of the effluent are trapped and focused using a modulator and sequentially released into the second column for further separation, this time based on polarity differences. GC × GC offers much greater peak capacities and is good for complex samples, particularly when hyphenated with fast scanning TOF instruments.

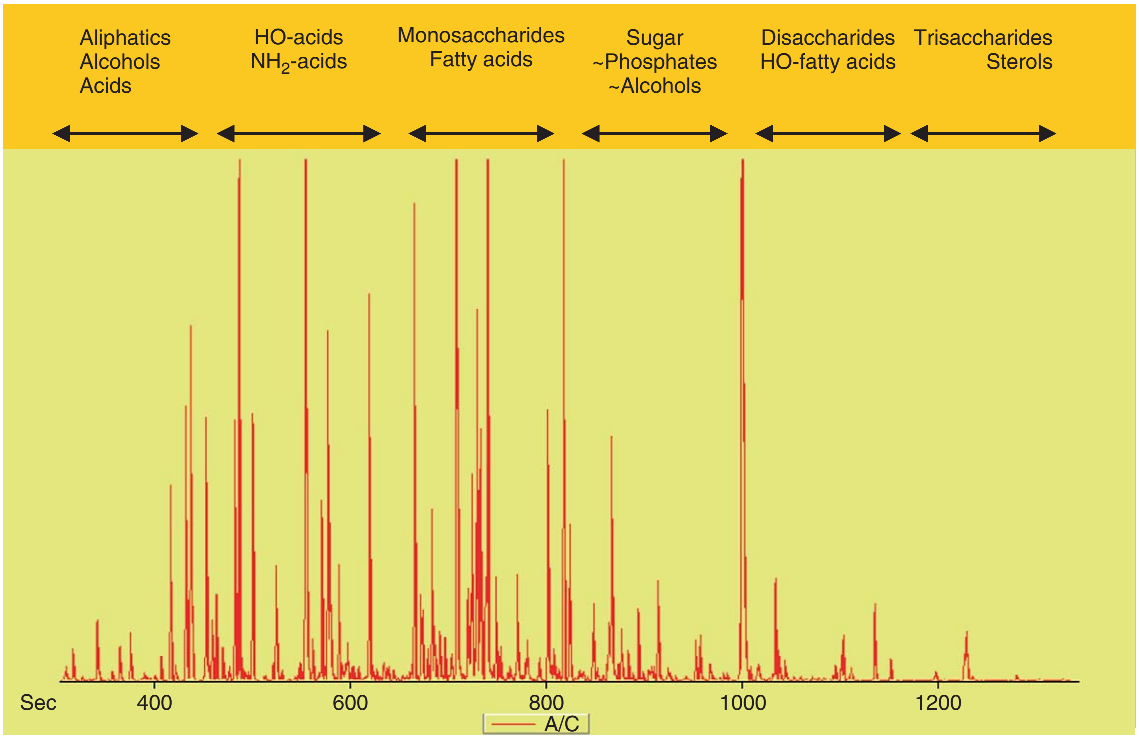


Figure 6 GC-TOF-MS analysis of a complete methanol–chloroform–water plant extract. The injection of a plant metabolite extract without separation of polar and hydrophobic phase is shown. Most of the metabolite compound classes are found in such a chromatogram. Reproduced with permission from M. Glinski; W. Weckwerth, *Mass Spectrom. Rev.* **2006**, *25* (2), 173–214. Copyright Wiley-VCH Verlag GmbH & Co. KGaA.

9.16.2.2.1(i) Data handling in GC-MS Metabolomics generates vast quantities of data and deconvoluting these data with conventional manual methods is very time-consuming and tedious, and requires skilled individuals. The increasing capability of chromatographic MS-based techniques, particularly those with increased signal-to-noise ratios (S/N) and higher peak capacity, means that the analyst expects to analyze hundreds of metabolites in a single assay. The extremely complex samples inevitably lead to differences in peak shapes, retention time drift, and variations in response for different compounds, which make deconvolution more complex and difficult. As a result, deconvolution is a major bottleneck of metabolomics. In recent years, tools have been developed to address the problems of coeluting interferences and to identify accurately as many peaks as possible. In a recent study by Lu *et al.*,[50] data sets from the analysis of a test mixture solution of 36 endogenous metabolites with a wide range of relative concentration ratios acquired using GC-TOF-MS were processed using three different commercial deconvolution software packages (LECO ChromaTOF, AMDIS, and SpectralWorks AnalyzerPro). Particular attention was paid to the extent of detection, identification, and agreement of qualitative results, and the flexibility and the productivity of these programs and their applications. None of the three software packages provided a comprehensive solution as there were differences in the number of components identified and the accuracy of deconvolution.

A representative set of examples of the application of GC-MS in metabolomics is given in Table 2.

9.16.2.2.2 Capillary electrophoresis-mass spectrometry

CE is an analytical separation technique capable of high-resolution separation of a diverse range of chemical compounds and is therefore well suited for metabolomics.[17,64] It is particularly suitable for the separation of polar and charged compounds and compounds with widely different structures, functional groups, physiochemical properties, and concentrations, for example, organic acids, amino acids, nucleic acids, steroids, carbohydrates, and flavonoids in various matrices. CE is complementary to GC and HPLC, and in many cases, samples that cannot be

Table 2 Selected GC-MS applications

Analyte(s)	*Matrix*	*Conditions*	*Reference*
Amino acids	*Escherichia coli* cell extracts	High-Resolution GC-MS	53
Amino acids, organic acids, sugars, sugar alcohols, aromatic amines	Potato tuber	GC-MS	54
Amino acids + other metabolites	*Sinorhizobium meliloti* 1021	GC-MS	55
Disaccharides, trisaccharides	Potato tuber	GC-TOF-MS	56
322 volatile metabolites (e.g., ketones, alcohols, aldehydes, esters, furans)	Tomato fruit	SPME-GC-MS	57
Amino acids, organic acids, sugars	Fruits and leaves of wild species of tomato	GC-MS	58
Amino acids, sugars, sugar alcohols, organic acids, amines (folate, *S*-adenosyl-L-methionine, *S*-adenosyl-L-homocysteine by LC-MS)	*Arabidopsis thaliana*	GC-MS + LC-MS	59
Flavonoids	Tomato	GC-MS, LC-MS	60
36 compounds comprising organic acids, amino acids, sugar alcohols, sugars, aromatic amines	Reference compounds	GC-TOF-MS	50
Sugars, amino acids, organic acids	Diploid yeast strain CEN.PK	GC × GC-TOF-MS	61
Sugars, sugar alcohols, organic acids, fatty acids, amino acids, biogenic amines	*E. coli*	GC-TOF-MS	62
Amino acids, sugars and polyols, organic acids	*Lotus japonicus*	GC/EI-TOF-MS	63

easily resolved by GC or HPLC can be separated by CE. Benefits of CE include the low cost of accessories/consumables, low organic consumption (or even none), the small amount of reagents needed for buffer preparation, and the use of cheap glass capillaries instead of more expensive GC and HPLC columns. An additional benefit of CE is that after analysis the CE column can easily be washed and is ready for a new run.

The major drawback of CE is the poor concentration sensitivity due to the limited amount of sample volume that can be introduced into the capillary (nanoliters) and the low absorption path length if UV detection is used. However, in the combination of electrospray ionization-mass spectrometry (ESI-MS), lower limits of detection can be obtained.

The main separation modes used in CE are capillary zone electrophoresis (CZE), micellar electrokinetic capillary chromatography (MEKC), capillary isotachophoresis, capillary gel electrophoresis, and capillary isoelectric focusing. CZE and MEKC are used most often. CE buffers are generally aqueous-based, though nonaqueous systems are exploited as well, particularly for analytes that are insoluble or sparingly soluble.

CE coupled to MS is suitable for a wide range of analytes; however, CE-MS is not as straightforward as GC-MS or HPLC-MS. A variety of examples of the application of CE-MS in metabolomics are given in **Table 3**.

The separation of 20 amino acids extracted from a root cell culture is shown in **Figure 7**[67] and a full description of the experimental conditions is given in Section 9.16.5.2.

For CE-MS to operate effectively, the instrument design requires a means of completing the electrical circuit for analyte separation, as well as simultaneously providing an electrical potential to the spray tip. This is generally accomplished using a sheath-flow or sheathless interface. The sheath-flow interface is more common because of its robustness and the ease at which it can be implemented. In this configuration, a coaxial sheath liquid, generally a hydro-organic solution, mixes with the effluent at the exit of the capillary. Its flow rate and composition can be varied to optimize the detection. The type and proportion of organic solvent and volatile acids (e.g., acetic acid and formic acid in the positive ion mode) and bases (e.g., ammonia and trimethylamine in the negative ion mode) affect the intensity of MS signals and can influence resolution, for example, a 50/50 2-propanol/water with 0.1% triethylamine solution was used for the analysis of hop acids and a mixture of 2-propanol and methanol offered better stability and sensitivity for liposaccharides. Pure organic solvents are compatible with both negative and positive ion modes, permitting the use of polarity switching during the CE run. To ensure a stable electrospray, the sheath liquid is usually introduced at a higher flow rate than the capillary effluent. However, sensitivity can be compromised due to dilution at the mixing point. This problem can be avoided by using a sheathless interface. Several designs are available, but technical difficulties have largely precluded their routine use.

The range of separation buffers for CE-MS is limited to volatile compounds, for example, acetic acid, formic acid, and ammonia, as these are more compatible with the electrospray ionization source of the mass spectrometer. These often do not provide the same separation potential as the normal phosphate/borate buffers that are used for UV detection. In addition, the concentrations must be kept reasonably low, so as not to impart ion production, and this generally results in lower efficiencies. Almost all types of mass analyzers have been coupled to CE. Initially, ion traps were used, as these instruments provided structural information through MS^n. However, higher performance instruments (e.g., TOF) and hybrids (e.g., Q-IT, Q-TOF) are increasingly used in CE-MS as they are capable of very fast scan speed, high mass range for expanded coverage, high mass accuracy and high resolution that is required to resolve closely migrating components with similar nominal masses, and/or multistage MS/MS capabilities for unequivocal metabolite identification.

9.16.2.2.2(i) Capillary electrophoresis on a chip Several designs for microchip electrophoresis have been fabricated to mix, react, concentrate, and separate analytes.[17] The small injection volumes, high electric fields and short separation lengths result in analysis times of seconds. CE microchips are fabricated mainly from glass, poly(methylmethacrylate), poly(dimethylsiloxane), polycarbonate, and poly(ethyleneterephthalate glycol). A disadvantage of these microchips is the adsorption of analytes to the wall. Dynamic or permanent coating of CE microchips with polymers can reduce analyte–wall interactions and provide a more rapid, efficient separation by adjusting the electroosmotic flow (EOF). For example, oligosaccharides adsorb strongly to the surface of a poly(methylmethacrylate) microchip, but this interaction can be suppressed by coating the microchip with a cellulose derivative, resulting in a highly efficient separation of 15 oligosaccharides within 45 s. The analysis times for CE microchips are very short but the separation efficiency is often still insufficient for the analysis of complex mixtures.

Table 3 Selected CE-MS applications

Analyte(s)	*Matrix*	*Conditions*	*Reference*
88 metabolites involved in glycolysis, tricarboxylic acid cycle, pentose phosphate pathway, photorespiration, amino acid biosynthesis	Rice leaves	CE-PDA, CE-MS	65
233 cationic, anionic, and mononucleotide metabolites related to energy metabolism	*Bacillus subtilise*	CE-MS	66
Amino acids	Plant cell cultures	CE-MS	67
Isoquinoline alkaloids	Central European *Corydalis* species	Nonaqueous CE-ESI-MS	46
375 charged hydrophilic intermediates in primary metabolisms – quantitative data on 198 metabolites: amino acids, glycolysis, tricarboxylic acid cycle, nucleotide biosynthesis pathways	*Escherichia coli*	CE-TOF MS	68
Cationic metabolites	*Desulfovibria vulgaris* Hildenborough	CE-FT-ICR MS	69
Amino acids	Variety of matrices	CE	70

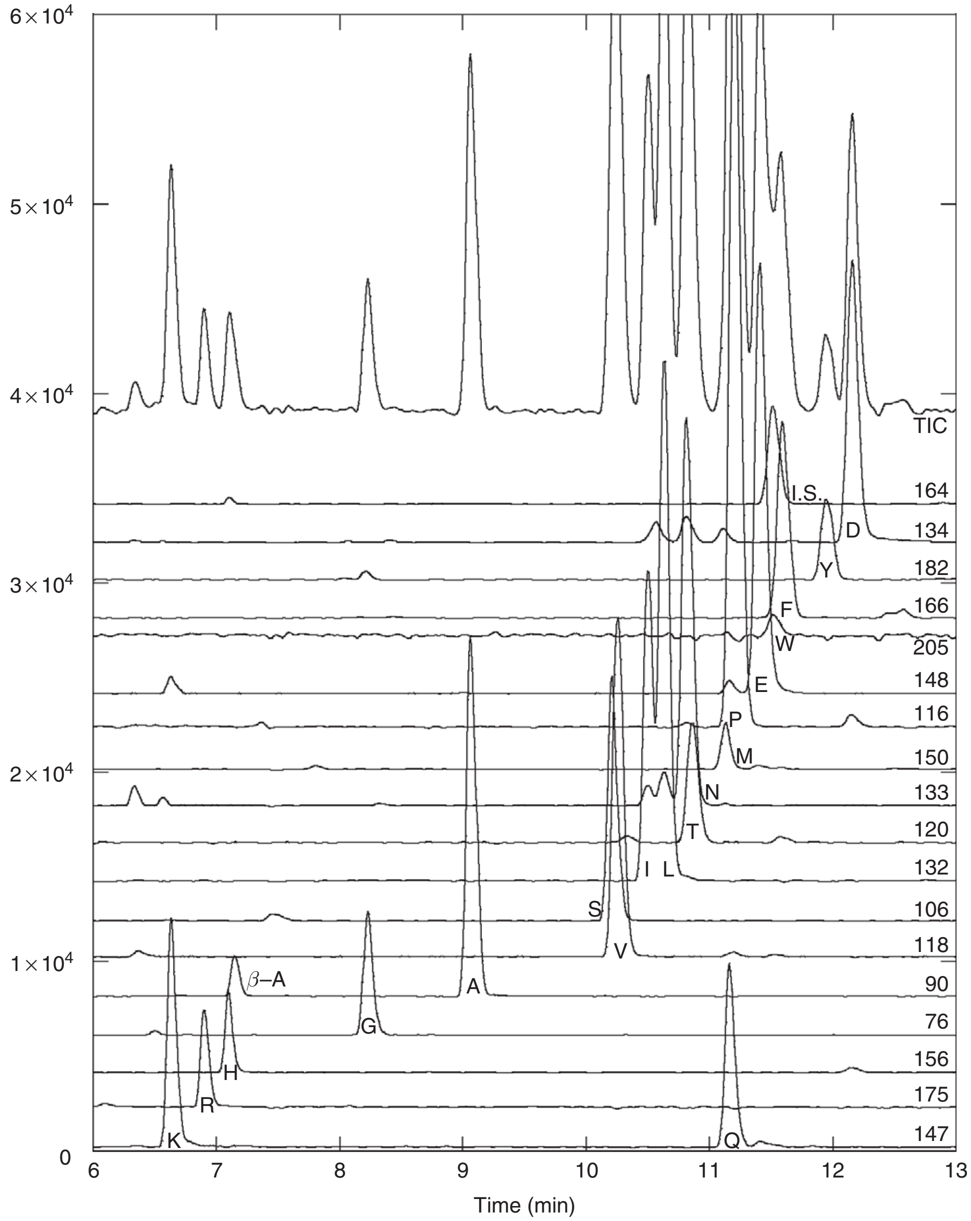


Figure 7 CE-MS electropherogram of control *Medicago truncatula* root cell culture extract at time = 0.0 h. The extracted ion chromatograms for each *m/z* are indicated on the right. Reproduced with permission from B. D. Williams; C. J. Cameron; R. Workman; C. D. Broeckling; L. W. Sumner; J. T. Smith, *Electrophoresis* **2007**, *28* (9), 1371–1379. Copyright Wiley-VCH Verlag GmbH & Co. KGaA.

9.16.2.2.2(ii) Capillary electrophoresis-mass spectrometry for comprehensive analysis Highly complex samples cannot be completely characterized by CE-MS because of instrumental limitations related to dynamic range and other analytical parameters that favor the separation and detection of some compounds over others. For comprehensive analyses, an attractive approach is to use two or more sets of conditions that have been optimized for different compounds and then concatenate the results. In this manner, wider sample coverage can be obtained as demonstrated by the works of Soga.[64] Carboxylic acids, phosphorylated carboxylic acids, phosphorylated saccharides, nucleotides, and nicotinamide and flavin adenine coenzymes were analyzed as anions using an alkaline background solution (BGS) and a cationic polymer-coated SMILE(+) capillary to reverse the EOF and prevent deleterious current drops. This was further enhanced by replacing the SMILE(+) capillary with a neutral capillary to prevent the anionic species from adsorbing onto the wall and applying air pressure to the capillary inlet during electrophoresis to provide a constant flow of liquid toward the anode. Under these conditions, citrate isomers, nucleotide, dinucleotides, and CoA compounds could be separated and detected well. By using a TOF instrument, the number of runs per sample was reduced to three (one for cations, one for anions, and one for nucleotides). A total of 1692 metabolites from exponentially growing *Bacillus subtilise* cells were cataloged this way.

9.16.2.2.3 Liquid chromatography-mass spectrometry

HPLC coupled to MS is a powerful alternative to the traditional coupling of HPLC with diode array detection (DAD) as it offers higher selectivity and sensitivity.[12,15,20] ESI is the most common ionization technique in LC-MS. To obtain a comprehensive profile, ionization must be performed in both positive and negative ion modes. ESI uses a high electric field to produce charged droplets from a liquid solution, ultimately leading to the formation of gas-phase ions. The main advantages of the ESI ion source are soft ionization, no need for derivatization, ability to ionize compounds of large mass range, suitability for nonvolatile and polar compounds, excellent quantitative analysis, and high sensitivity. A drawback of the ESI process is its ability for ion suppression due to competition effects in the ionization process. Three alternate solution-based ionization strategies to ESI are also being used for HPLC-MS-based metabolomics, namely nanoESI, atmospheric pressure chemical ionization (APCI), and atmospheric pressure photoionization (APPI). NanoESI LC, performed at low flow rates (approximately 200 nl min^{-1}) enhances sensitivity and dynamic range. Using nanoLC/nanoESI-MS methods, the chemical noise entering the mass spectrometer is decreased and the chromatographic separations improved. The application of capillary HPLC/nanoESI-MS methods provides better separations and hence a smaller chance of coelution with competitors for ionization. Chip-based nanoelectrospray systems are another tool to reduce matrix effects and improve sensitivity. APCI and APPI typically induce little or no fragmentation, and are robust and relatively tolerant of high buffer concentrations. These approaches can be valuable for the analysis of nonpolar and thermally stable compounds such as lipids. There is also a trend toward a single ionization source containing combinations of ESI and APCI or ESI and APPI. Nordstrom *et al.*[71] demonstrated the effectiveness of multiple ionization MS with both ESI and APCI in the positive and negative ionization modes as well as offline MALDI and desorption ionization on silicon (DIOS) for metabolomics. Complementing ESI with APCI resulted in an approximately 20% increase in unique ions, and DIOS accounted for 50% of the unique ions detected.

The difficulty in exchanging a spectra library from LC-MS analysis is a major drawback of this technology, particularly for metabolite profiling, where identification of a large number of compounds is desirable. HPLC can be simultaneously coupled to both DAD and MS to provide multiple levels of information for structure elucidation as well as quantification of low-level metabolites.[72] HPLC-MS also has some limitations, especially when applied simultaneously to a wide range of compounds. In general, metabolomics investigations by HPLC-MS have been performed using solvent gradients, on reverse-phase packing materials, 2.0, 3.0, or 4.6 mm inner diameter (i.d.) columns, of lengths between 5 and 25 cm containing 3–5 μm packing materials. The physical properties of small endogenous metabolites vary widely such that, in reverse-phase LC using conventional C8 and C18 columns, the highly polar metabolites are (nearly) unretained and as a result not identified by the MS. Other types of columns have been used to circumvent these problems; for example, monolithic silica-based columns and hydrophilic interaction liquid chromatography (HILIC) have been used for highly polar metabolites of plant origin. Ion-exchange chromatography can also be used for the separation of ionic or polar compounds, but the combination of this with MS is not favorable due to the high concentrations of nonvolatile salts in the mobile phase. Generally, matrix effects can lead to difficulties in the detection of certain compounds by MS, so to minimize these interferences, sample preparation is often essential for sensitive and reliable LC-MS analyses.

9.16.2.2.3(i) Capillary high performance liquid chromatography-mass spectrometry

Conventional HPLC-MS is characterized by reasonable resolution and moderate throughput. Capillary HPLC-MS provides higher chromatographic resolution, higher peak capacity, and increased signal to noise than conventional HPLC-MS due to more concentrated peaks as well as reduced ion suppression. In addition, the reduction in the amount of sample required for capillary HPLC easily enables the analysis of very small samples (a few microliters). Capillary HPLC-MS using monolithic C18-bonded silica columns (0.2 mm i.d. and 30–90 cm in length) was used by Lu *et al.*[11] to study *Arabidopsis thaliana* extracts.

9.16.2.2.3(ii) Ultra-performance liquid chromatography-mass spectrometry Ultra-performance liquid chromatography (UPLC) is a combination of a 1.7 μm reverse-phase packing material and a chromatographic system that can operate at pressures in the 6000–15 000 psi range (conventional HPLC uses 3–5 μm packing material and operates between 2000 and 4000 psi). There is a greater (S/N) due to the reduction in band broadening, and thus an increase in sensitivity. This has enabled better chromatographic peak resolution and increased speed and sensitivity to be obtained for complex mixture separation. The typical peak widths generated by UPLC are in the order of 1–2 s for a 10 min separation. Because of the much improved chromatographic resolution of UPLC, the problem of ion suppression from coeluting peaks is greatly reduced. UPLC coupled to a Q-TOF mass spectrometer is a powerful tool for analyzing complex mixtures as seen in **Figure 8**.

The increased separating potential of UPLC (coupled to a Q-TOF mass spectrometer) was demonstrated by Xie *et al.*[73] for the analysis of five *Panax* herb varieties, in which 25 saponins were identified in individual samples of *P. ginseng* (Chinese ginseng), *P. notoginseng* (Sanchi), *P. japonicus* (Rhizoma), *P. quinqeufolium* L. (American ginseng), and *P. ginseng* (Korean ginseng) in less than 20 min. This contrasted with 11 saponins being identified by traditional HPLC-MS with a run time of 80 min. The resulting UPLC-QTOF MS spectra and data showed excellent accurate mass information and permitted the distinct differentiation of the five *Panax* herbs.

A representative set of examples of LC-MS in metabolomics are given in **Table 4**.

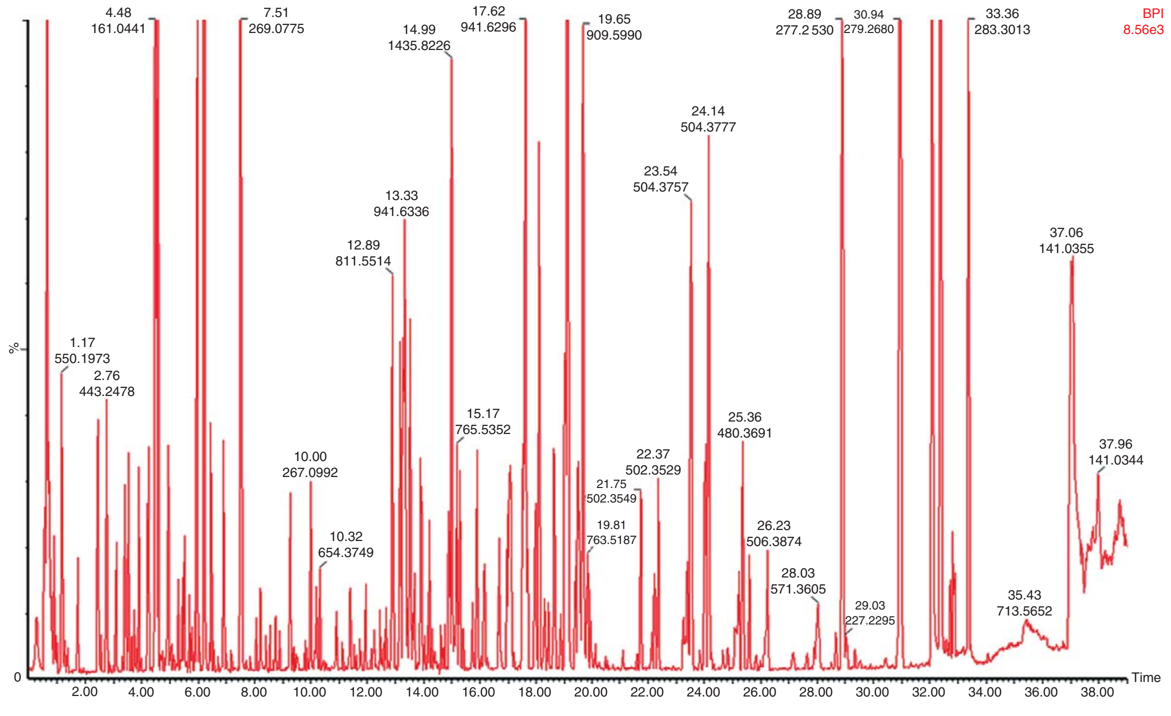


Figure 8 UPLC-QTOF-MS base peak ion chromatogram obtained for the combined methanol extracts from soybean and *Medicago truncatula* (CV Jemalong A17). Separations were achieved using a Waters Acquity UPLC 2.1 × 100 mm, BEH C18 column with 1.7 μm particles, a flow of 600 μl min^{-1}, and a linear gradient of 0.1% acetic acid:acetonitrile (5:95 to 30:70 over 30 min). Mass spectra were collected on a Waters QTOFMS Premier. (Data generated by David Huhman.) Reprinted from M. Bedair; L. W. Sumner, *Trends Analyt. Chem.* **2008**, *27* (3), 238–250, Copyright (2008), with permission from Elsevier.

Table 4 Selected LC-MS applications

Analyte(s)	*Matrix*	*Conditions*	*Reference*
Carotenoids, tocopherols, chlorophylls	Fruits and vegetables	HPLC-PDA	74
Flavonoids, isoprenoids (carotenoids, tocopherols, lycopene), ascorbic acid, phenylalanine, fruit volatiles – e.g., aldehydes, alcohols (GC-MS)	Tomato	LC-PDA, LC-QTOF MS + GC-MS	75
Carotenoids, flavonoids	Tomato	LC-PDA	76
Lipids (GC-MS), isoprenoids (LC-PDA)	Tomato flesh and seeds	LC-PDA + GC-MS	77
Anthocyanins, flavones, amino acids, organic acids	*Perilla frutescens*	LC-PDA-MS, LC-fluorescence (amino acids), CE (organic acids)	78
Anthocyanins	Berries	LC-PDA-ESI-MS	79
Dihydrocaffeoyl polyamines	Potato tubers	LC-PDA, LC-MS, LC-MS/MS	80
Psoralen, isopsoralen, corylidin, psoralidin, bavachin, bavachalcone, neobavachalcone, bavachromene, corylifol B, corylifol A, corylin, bakuchiol	*Psoralea corylifolia*	LC-MS, MALDI-TOF MS	81
Flavonoids	*Populus*	LC-PDA	82
Flavonoid conjugates, aromatic carboxylic acids, alkaloids, phenylpropanoids	Tomato peel and flesh	LC-PDA-TOF MS	83
Flavonoid conjugates	*Arabidopsis thaliana*	LC-UV-MS/MS	84
Flavonoid conjugates	Pak choi	LC-PDA-ESI-ion trap MS	72
Glucosinolates	Brassica species	LC-PDA-ESI-ion trap MS	85
Adenosine, AMP, cytosine, flavin adenine dinucleotide, guanosine, leucine/isoleucine, NAD, phenylalanine, riboflavin, tryptophan, tyrosine, valine	*Cyanothece* sp. ATCC 51142	capLC-Orbitrap MS	86
Phenolic compounds	*Vaccinium angustifolium* (lowbush blueberry)	HPLC-DAD-APCI-MS	87
Phospholipids, phosphatidic acid, phosphatidylserine	Biological samples	LC-MS	88
Saponins – chromosaponin L	Chickpeas	LC-PDA-MS, NMR	89
Saponins – ginsenosides	*Panax notoginseng*	UPLC-TOF MS	90
Saponins – ginsenisides	*P. ginseng* (Chinese and Korean), *P. notoginseng*, *P. japonicus*, *P. quinqeufolium* L.	UPLC-TOF MS	73

9.16.2.2.3(iii) Online multidimensional high performance liquid chromatography Despite their high resolving power, sensitivity, precision, and practicability, HPLC analyses are restricted by the pretreatment and processing of highly complex matrices. In many cases, metabolites are present in trace amounts and biological samples are too complex or incompatible with conventional HPLC phase systems to permit an analysis by direct injection into an analytical column. Thus, simplification of such multicomponent mixtures as well as metabolite enrichments is needed prior to analysis. In general, this is obtained by prefractionation or class separation and preconcentration steps. To solve these problems, an HPLC column-switching technique has been developed. These techniques always use two or more columns that are connected in parallel or in series and thus allow the selective prefractionation and subsequent analysis of the target compounds.[11]

9.16.2.2.4 Direct injection mass spectrometry

Direct injection mass spectrometry (DIMS) is a high-throughput screening tool often assaying hundreds of samples per day (1 min analysis time). Crude samples are injected or infused into the electrospray mass spectrometer resulting in one mass spectrum per sample, which is representative of the sample. This is limited to a semiquantitative screening process as metabolome coverage depends on the ability of the metabolite to be ionized. The mass spectrum or mass list is used for sample classification. Metabolite identification can be tentatively performed using high-resolution instruments and accurate mass measurements (e.g., TOF, orbitrap, FT-ICR). FT-ICR is particularly useful as it allows for the separation of compounds with the same nominal mass but different monoisotopic mass (e.g., glutamine (MW 146.0 689 Da) and lysine (MW 146.1 052 Da)). All possible metabolite peaks can be resolved and their mass accuracy calculated.[6] Some applications of DIMS in metabolomics are given in **Table 5**.

9.16.2.3 Fourier Transform, Infrared, and Raman Spectroscopy

FTIR spectroscopy is an established, yet constantly developing analytical technique that enables a rapid, nondestructive, reagentless, and high-throughput analysis of a diverse range of sample types.[6] In relation to biological applications, strong absorption bands can be directly related to specific compounds, for example, spectral vibrations from CH_3 and CH_2 groups associated with fatty acids can be found within the spectral region 3050–2800 cm^{-1}. FTIR is a valuable metabolic fingerprinting tool owing to its ability to analyze carbohydrates, amino acids, lipids as well as proteins and polysaccharides simultaneously. FTIR is also a highly versatile technique that requires minimal sample preparation. Samples can simply be spotted (typically 0.5–20 μl) onto a variety of plates for high-throughput analyses (spectra are collected in seconds). FTIR does have a drawback; the IR absorption of water is very intense, which however can be overcome with dehydration of the samples (a prerequisite for samples prepared on the plates mentioned above) or the water signal can be subtracted simultaneously or by other techniques (e.g., attenuated total reflectance). Portable FTIR instruments are also available allowing the analyst much scope in spectral collection in a variety of environments. Sensitivity and selectivity are not as high as the other methods. To date, the majority of metabolomic studies undertaken using vibrational spectroscopy have been with FTIR. Recently, FTIR has been introduced as a metabolic fingerprinting technique within the plant sciences, for example, fingerprinting of salt-stressed tomatoes where potentially important functional groups were identified in response to the impact of salinity to tomato.[99] Near-infrared (NIR) and Raman spectroscopy have also shown significant potential for monitoring metabolites.

9.16.3 Modeling and Data Analysis

Data analysis and sensibly applied statistical tools are of crucial importance for metabolomics. Good experimental design is of course a fundamental first requirement. There have been a number of books and research papers written recently discussing statistics use and models for data analysis of metabolomics.[100–104] Statistical and experimental robustness have been the focus of metabolomics and demonstrated in a study of NMR protocols and multivariate statistical batch processing, which were examined for consistency over six different centers. The data were shown to be sufficiently robust to generate comparable results across each center.[105]

Table 5 Selected stand-alone DIMS applications

Analyte(s)	*Matrix*	*Conditions*	*Reference*
Organic acids, amino acids, amines, hexose sugars, sucrose and isomers, ascorbic acid	Green tomato	ESI-TOF MS	91
Amino acids, organic acids, 4-aminobutyric acid, putrescine, D-glucose	*Saccharomyces cerevisiae*	Laser desorption mass spectrometry	92
Metabolites from biosynthetic pathways of fatty acids, branched-chain amino acids, aromatic amino acids, carotenoids	Plant tissue	FT-ICR MS	93
Carotenoids	Plant cells	MALDI-TOF MS	94
Sucrose, glucose, fructose, starch, citric acid species	Strawberry, banana, and grape tissues	MALDI-TOF MS	95
Organic acids, flavonoids, glucose, fructose, sucrose, fatty acids	Apples and strawberries	Graphite-assisted laser desorption/ionization mass spectrometry (GALDI-IMS)	96
20 essential amino acids and sulfur metabolites	*S. cerevisiae* S288c	LTQ-Orbitrap MS	97
Polyphenols, vitamin C	Berries	LCQ-Deca ion trap MS	98

PCA is one of the most common statistical methods to analyze metabolomics data sets. This is an unbiased approach and hence is particularly suitable for biomarker or diversity studies. The related partial least squares (PLS) analysis can be used similarly or can incorporate 'class' information in a DA. PLS-DA is a guided or supervised statistical method and as such there is the risk of overfitting data. Models built using such techniques should always include test data which were withheld from the original model. Data can be preprocessed either before they are introduced into a statistical model (e.g., log functions applied to NMR data,[33] peak alignment and filtering in LC-MS and GC-MS applications) or within the model. Orthogonal signal correction (OSC) is a filter used with PLS-DA modeling to reduce the influence of signals that do not correlate with class distinction. Modeling techniques such as clustering tools continue to develop. In a recent study, a new supervised modeling technique, uncorrelated linear discriminant analysis (ULDA), was compared to PLS-DA and PCA. The new model outperformed PCA and PLS-DA in the examples given and the explanation was that the ULDA model removes interference from linear correlations in data (such as multiple resonances for one compound in NMR data).[106] These types of clustering models tend to be best represented in the literature; however, there is increasing use of evolutionary computational-based methods such as genetic algorithms and genetic programming. These models appear to work well with high-dimensional metabolomics data and generate useful relationships, rules, and predictions.[101]

Interpretation of metabolomics data is not always straightforward. Even if a resonance in an NMR spectrum or a peak in an LC-MS chromatogram is identified as being statistically important, it may not be straightforward to determine the chemical structure of the metabolite of interest. The preceding sections of this chapter have addressed how natural product structure elucidation techniques using MS and NMR can be used in metabolomics (see also Volume 9). Compound libraries with searchable databases are also important, though these have been lacking for metabolomics. The NIST compound libraries for GC-MS are perhaps the best known, but more recently others have been developed. Some of these are commercial products (e.g., AURELIA/AMIX and Chemonx NMR Suite for NMR data), but there are also public, freely available databases emerging. The human metabolome database contains both NMR and MS data for many mammalian metabolites[21] and is being expanded to incorporate drug and food metabolite information. The Madison Metabolomics Consortium Database (MMCD) is another online resource for metabolomics researchers with both NMR and MS data searchable.[107] A public LIMS system allowing documentation of experimental metadata as well as annotation of GC-MS data (via a second relational database system, BinBase) has recently been developed by the Fiehn group at UCDavis and incorporates many plant metabolites not found in the NIST databases.[108]

Metabolomics aims to provide information about a system, and often to put the chemical data into context there is a requirement to understand the relationships between the metabolites and the flux dynamics. Systems biology approaches also aim to integrate metabolomics data with other 'omic' data sets (e.g., proteomics, genomics, transcriptomics). Computational models and software that can meet these requirements are under development. A number of tools have been developed for plant scientists. Lange *et al.*[109] recently reviewed these tools, which included a number of online databases, and introduced their own tool, BioPathAt, for the model plant *A. thaliana*.[110] More recently, an online resource has been developed for crop plants. MetaCrop describes pathway diagrams, reactions, locations, transport processes, reaction kinetics, taxonomy, and literature for six major crop plants. The same research group has also developed VANTED (visualization and analysis of networks containing experimental data).[111] VANTED is a tool for visual exploration and statistical analysis of complex biochemical data sets and allows the integration of data from different areas of genome, proteome, and metabolome research. There are several online tools that also describe pathway and metabolic information, for example, KEGG, MetaCyc, EcoCyc, ArMet, and HumanCyc. In general, these databases present qualitative information but cover many species.

Advances are continually being made in the technology, experimental design, data analysis, and data integration that will advance metabolomics. The science of metabolomics can be applied to many research areas including natural products research where it has already shown potential for enhancing the understanding of MOA, chemotaxonomy, and for the optimization of microbial processes for the production of enhanced levels of antibiotics.

9.16.4 Sampling and Sample Preparation

In the metabolomics experiment, sampling provides a picture or snapshot of the metabolome at one point in time.[6] The requirement of sampling and sample preparation that are not biased toward groups of metabolites presents challenges, which currently have not been fully resolved. The time and method of sampling can greatly influence the reproducibility of the analytical sample, as can the section of a plant sampled. Finally, the storage of the sample is important, as the continual freeze/thawing of samples can be detrimental to stability and composition. All these influence precision, accuracy, and reproducibility of results. However, generally, it is observed that biological variability is greater than analytical variability even when controlled sampling and sample preparation are employed. For example, data suggest that the average biological variance for *Medicago truncatula* is approximately 50%. These large biological variations represent the major limitations of the resolution of the metabolomics approach. One way to reduce biological variance is to pool samples either by analyzing different tissues of the plant within a single sample or by pooling multiple replicate plants. This helps minimize random variations through statistical averaging; however, many variations in metabolite levels often have biological significance and result from functional differentiation of tissues. Pooling tissues can, therefore, result in undesirable dilution of site- or tissue-specific up/downregulated metabolites. An alternative approach is to start with a homogeneous tissue such as well cultures, but this has obvious restrictions since the syntheses of some plant metabolites, particularly natural products, may be linked to cellular differentiation. Plant growth stage, environmental parameters, and sampling are critical. Therefore, strategies need to be incorporated to minimize variation.

The extraction of intracellular metabolites provides a snapshot of the metabolome, can be time consuming, and is subject to certain difficulties when compared to other sampling strategies. The very process of metabolite extraction disturbs the *status quo* of an organism, and chemical changes brought about by exposure to oxygen, solvents, and change in pH are particularly common. Thus, the metabolites isolated from natural sources are not necessarily the metabolites that are present in the living tissue.

Metabolic processes are rapid (reaction time <1 s), so rapid inhibition of enzymatic processes is required, generally by freeze clamping or freezing in liquid nitrogen after harvesting, and subsequent storage at −80 °C. Freezing gives rise to specific issues, such as the loss of metabolites, the release of touch- or wound-induced metabolites, or the nonreversible loss of compounds by absorption to cell walls. Polar/nonpolar extractions are the most frequently applied methods and are performed by physical/chemical disruption of the cells, removal of the cell pellet by centrifugation, and distribution of the metabolites to polar (methanol/water) and nonpolar (chloroform) solvents. Hot alcoholic extractions have also been performed. Analysis of volatile compounds is normally achieved by headspace GC-MS analysis; for example, Tikunov *et al.*[57] used an automated sequential headspace solid-phase microextraction (SPME)-GC-MS for large-scale profiling of tomato fruit volatiles. The volatile compounds were liberated from the sample matrix by incubating in a sealed container for a fixed period of time. Calcium chloride was then added to stop enzyme activity and force the volatiles into the headspace, which were then assayed by GC-MS.

9.16.5 Examples

9.16.5.1 Profiling and Quantification of Free Amino Acids in Plant Cell Cultures by CE-MS and GC-MS

The development of metabolomics has yielded increased interest in the analysis of many nonvolatile metabolites, including amino acids, and higher throughput assays are now required.[67] Most practical applications involving amino acid analysis are performed using HPLC and more recently GC-MS. HPLC methods are robust; however, the amino acids require derivatization either before or after chromatographic separation and the analytical run times can be quite long. Another drawback of HPLC is the relatively large quantities of protein or peptide required for analysis, since significant efforts are often required to isolate these compounds in such amounts. Amino acids also require derivatization for GC analysis and the comprehensive analysis of all amino acids can prove difficult and a vast number of methods for the preparation of 'volatile' amino acids have been reported. CE-MS provides a reasonably rapid and robust assay for the complete separation of the 20 or more closely related amino acids. The key advantage is that CE-MS allows for the direct analysis of underivatized amino acids. Amino acids can be analyzed under acidic conditions to form $[M+H]^+$ ions or under basic conditions to yield $[M–H]^-$ ions.

Williams *et al.*[67] examined the utility of CE-MS in the analysis of more than 500 root cell culture extracts from *M. truncatula* as part of an integrated functional genomics study of this model legume and its response to stress at all cellular levels (transcriptome, proteome, and metabolome). A CE-MS method was developed and optimized to yield maximum analytical sensitivity combined with high sample throughput. The method used a relatively inexpensive benchtop Q-ESI mass spectrometer. Sample preparation was minimal as CE-MS required no derivatization or extensive sample cleanup to remove unwanted proteins, peptides, or other biogenic interferences. In a parallel analysis of the same cell culture tissue, GC-MS was used to profile polar metabolites including a large portion of the amino acids. The interlaboratory comparison shows excellent correlation between the amino acid levels determined by CE-MS and GC-MS.

Briefly, triplicate biological samples were collected from both control and elicited cell cultures at each time point, with each sample collected from separate culture flasks. Cells were harvested by vacuum filtration, immediately frozen by liquid nitrogen, and stored at −80 °C. The frozen tissue was then lyophilized for 48–72 h and stored at −80 °C until further distribution for the various analyses. In total, the yeast elicitation time course contained 126 culture flasks. Each biological replicate was analyzed in triplicate to establish a complete statistical view of both biological and analytical variation.

9.16.5.1.1 Tissue analysis by capillary electrophoresis-mass spectrometry

Dried tissue (50 mg) was added to a 2 ml centrifuge tube followed by 1.5 ml of HPLC grade water and 50 μl of ethionine (internal standard). The aqueous mixture was placed on ice and homogenized for 30 s using an Ultra-Turrax model T8 homogenizer equipped with a microrotor. The homogenized solution was subsequently mixed vigorously for 5 min using a vortex mixer and then centrifuged for 10 min at 8000 × g. A 1 μl aliquot of the supernatant was withdrawn and placed into a screw cap centrifuge tube and then dried in a vacuum centrifuge at ambient temperature. Dried samples were stored at −20 °C until analysis. Immediately before analysis, the dried samples were resuspended in 100 μl of 0.01 mol l^{-1} HCl and sonicated for 10 min. Subsequently, 20 μl aliquots were placed in 100 μl conical CE sample vials prior to analysis. No further preparation was required.

The amino acids were analyzed using an Agilent G1600 CE connected to an Agilent 1946A single quadrupole mass spectrometer with an ESI source using the Agilent CE spray needle adapter. A 70 cm × 50 mm i.d. fused silica capillary from Polymicro Technologies was used for the assay. A sheath liquid composed of a 1:1 v/v mix of 2-propanol and HPLC grade water with 5 mmol l^{-1} formic acid was supplied to the coaxial sheath–fluid interface at a flow rate of 4 $\mu\text{l min}^{-1}$ using an Agilent 1100 binary pump fitted with a 1:100 splitter. The separation was performed using a +25 kV potential and samples were injected hydrostatically for 250 mbar s (6.1 nl). Mass spectra were collected in the SIM mode using the $[M+H]^+$ m/z for each amino acid. The running electrolyte was 1.0 mol l^{-1} formic acid prepared in HPLC grade water with no pH adjustment. Prior to each sample injection, the capillary was preconditioned with a 4.0-bar rinse of the running electrolyte for 1 min followed by an additional 4.0-bar rinse with a separation voltage applied for 2 min. A post column rinse sequence, comprising 1.0 mol l^{-1} NaOH for 2.0 min, HPLC grade water for 1 min, and 0.1 mol l^{-1} HCl in HPLC grade water for 2 min with a pressure of 4.0 bar for each rinsing step, followed each separation.

9.16.5.1.2 Tissue analysis by gas chromatography-mass spectrometry

For the GC-MS analysis, dried tissue samples were homogenized with a glass rod, and 6–6.05 mg of dried tissue was weighed into a 4.0 ml glass vial. Chloroform (1.5 ml) containing 10.0 mg l^{-1} docosanol (internal standard) was added to the dried tissue. The samples were vortexed and incubated for 45 min at 50 °C. After equilibrating to room temperature, 1.5 ml of HPLC grade water containing 25.0 mg l^{-1} ribitol was added to the chloroform. The sample was then vortexed and incubated for a second 45 min period. The biphasic solvent system was then centrifuged at 2900 × g for 30 min at 4 °C to separate the layers. One milliliter of each layer was collected and transferred to individual 2 ml autosampler vials. The chloroform layer was dried under nitrogen and the aqueous layer was dried in a vacuum centrifuge at ambient temperature. Dried polar extracts were methoximated in pyridine with 120 μl of 15.0 mg l^{-1} methoxyamine-HCl, briefly sonicated, and incubated at 50 °C until the residue was resuspended. Metabolites were then derivatized with a commercial derivatization solution containing 120 μl of *N*-Methyl-*N*-trifluoroacetamide (MSTFA) + 1% Trimethylchlorosilane (TMCS) for 1 h at 50 °C. The sample was subsequently transferred to a 300 μl glass conical insert and analyzed using an Agilent 6890 GC coupled to a 5973 MSD instrument scanning from *m*/*z* 50–650. Samples were injected at a 15:1 split ratio and the inlet and transfer line were held at 280 °C. Separation was achieved with a temperature program of 80 °C (2 min) to 315 °C (12 min) at 5 °C min^{-1} on a 60-m DB-5MS column (J&W Scientific; 0.25 μm i.d., 0.25 mm film thickness) held at a constant flow of 1.0 ml min^{-1}.

Using the conditions described in Section 9.16.5.1.1, automated CE-MS was established and used for the unattended analysis of large sample sets. In a 3-month period, over 500 separate tissue extracts were analyzed with triplicate analytical replicates. The entire study was performed using a single CE capillary and the only maintenance that was performed was the wiping down of the ESI source on a weekly basis. A typical electropherogram for *M. truncatula* liquid suspension cell culture extracts is shown in **Figure 7** (see Section 9.16.2.2.2). The first amino acid (lysine) elutes in the 6.5–6.8 min window and the last amino acid (aspartic acid) elutes in the 11.8–12.8 min window. These variations in migration times were observed with extensive rinsing with both strong base and acid at the end of each analysis. Peak identification and quantification were performed using extracted ion chromatograms. The extracted ion *m*/*z* (extracted ion electropherograms (XIEs)) values for each amino acid assayed by CE-MS and GC-MS are listed in **Table 6**.

As shown in **Figure 7**, all of the amino acids are clearly resolved. Wide retention windows can be used in the peak i.d. table to help overcome the variations in migration times. The only exceptions were the isomers isoleucine and leucine at 132 *m*/*z*. Isoleucine always elutes first in this pair, as determined using authentic compounds. The peak area ratio reproducibility for the analytical replicates of the same biological sample varied from a coefficient of variation (CV) of less than 1% for the strongest signals to almost 5% for the amino acids at lower concentrations. In a previous GC-MS report on the same metabolic sample set, the biological variation for a large group of metabolites including amino acids ranged from 27.4 to 33.3% CV.[112] The cell culture extraction procedure described for CE-MS was also optimized for profiling of free carbohydrates using CE-LIF following derivatization of the extracts with 2-aminoacridone. HPLC grade water was used to resuspend the dried extracts for this work.

Table 6 List of amino acids profiled by CE-MS and GC-MS

Amino acid	*One-letter symbol*	*m/z* for $[M+H]^+$	*CE-MS*	*GC-MS*
Glycine	G	76	+	+
Alanine	A	90	+	+
β-Alanine	β-A	90	+	+
Valine	V	118	+	+
Leucine	L	132	+	+
Isoleucine	I	132	+	+
Methionine	M	150	+	+
Proline	P	116	+	+
Phenylalanine	F	166	+	+
Tryptophan	W	205	+	−
Asparagine	N	133	+	+
Glutamine	Q	147	+	−
Serine	S	106	+	+
Threonine	T	120	+	+
Tyrosine	Y	182	+	−
Aspartic acid	D	134	+	+
Glutamic acid	E	148	+	+
Lysine	K	147	+	+
Arginine	R	175	+	+
Histidine	H	156	+	−
Ethionine	I.S.*	164	+	

* internal standard used in CE-MS only.
+ represents amino acids detected by each methodology; − represents amino acids not detected by each methodology.

9.16.5.2 Profiling, Identification, and Quantification of Glucosinolates in Brassica Plant and Seed Material

Rochfort *et al.*[85] used LC-ESI-ion trap MS to profile, identify, and quantify a range of glucosinolates present in plant and seed material. Glucosinolates are naturally occurring anionic secondary plant metabolites incorporating a thioglucosidic link to the carbon of a sulfonated oxime. There are a large number of naturally occurring glucosinolates and they are found in relatively large quantities in many plant species within the family Cruciferae. These metabolites are of interest for both their anticancer and flavor properties and in the study of nitrogen and sulfur metabolism in model plants such as *Arabidopsis*. For high-glucosinolate breeding studies, the ability to rapidly screen prospective parental lines for glucosinolate content is of value. On the other hand, with oil/forage crops such as canola, the aim of breeding is to reduce the total content of glucosinolates.

LC-ion trap MS experiments on a range of standard compounds and seed/plant extracts showed that all of the glucosinolates present in these samples fragmented in the mass spectrometer to a common ion of *m*/*z* 259, through either MS^2 or MS^3 fragmentation. This observation led to the development of a rapid screening procedure for the presence of glucosinolates in seed/plant material through a 'parent ion mapping' experiment. 'Parent ion mapping' is a novel analytical MS approach and offers a sensitive, qualitative technique to rapidly and simultaneously detect any parent ion that gives rise to the daughter ion at *m*/*z* 259 – the ion that is consistently generated in the ion trap fragmentation of glucosinolates. **Figure 9(a)** shows the MS^2 fragmentation of glucoraphanin to produce ions at *m*/*z* 372 and 259 and **Figure 9(b)** shows the MS^3 fragmentation to give *m*/*z* 259 as the dominant ion.

For the 'parent ion mapping' experiment, the aqueous plant/seed extract is infused directly into the mass spectrometer and each experiment takes between 2 and 3 min. Infusion of the broccoli seed extract generates an intensity map of the parent ions and a spectrum view of the parent ions detected (**Figure 10**). The major glucosinolates detected by this experiment are sinigrin (*m*/*z* 358), gluconapin (*m*/*z* 372), progoitrin (*m*/*z* 388), glucoerucin (*m*/*z* 420), glucoiberin (*m*/*z* 422), glucoraphanin (*m*/*z* 436), and 4-methoxyglucobrassicin/neoglucobrassicin (*m*/*z* 477). This analysis results in a very clean spectrum, devoid of confounding ions caused by other metabolites.

Similarly, the foremost parent ion (*m*/*z* 358) detected in the spectrum generated from the infusion of mustard seed extract was due to sinigrin. The method is also extremely sensitive. Serial dilution analysis of a sinigrin standard allowed detection of the parent ion at $3\,\mu g\,l^{-1}$ (ppt) of extract.

LC-ion trap MS was also used for the targeted analysis of the glucosinolates present in Brassica sprouts via extraction of the *m*/*z* 259 ion, and this is displayed in **Figure 11**.

LC-MS^n provided positive identification and quantification of individual glucosinolates in the extracts. The levels of sinigrin in mustard seeds and glucoraphanin in broccoli florets were determined using the specific target ions at *m*/*z* 358 and 259 for sinigrin and at *m*/*z* 436, 372, and 259 for glucoraphanin. The method was both robust and reproducible, with good area and retention time CVs and linearity over five ranges. For glucoraphanin, triplicate analyses of various concentrations resulted in an average CV of 5% or less for the molecular ion and the fragment ions at *m*/*z* 372 and 259 (2.5–3.4, 1.5–2.6, and 3.3–9.1%, respectively). For sinigrin, a CV of 6% was recorded for the molecular ion and the *m*/*z* 259 ion reflecting greater variation due to its early retention time. The method allowed sufficient data points for detection and quantification using both the parent ions and the *m*/*z* 259 ion. Seven replicate injections of both the standards and the plant extracts indicate that good CVs are typical across the chromatogram. The levels of sinigrin and glucoraphanin for mustard seeds and broccoli florets, respectively, compared favorably with the levels determined concurrently using the in-line PDA detector monitoring the λ_{max} of the glucosinolates: for example, sinigrin in mustard seeds *m*/*z* 358, $69\,g\,kg^{-1}$; *m*/*z* 259, $71\,g\,kg^{-1}$; $\lambda_{max\ 224\,nm}$ $71\,g\,kg^{-1}$ and glucoraphanin in broccoli florets *m*/*z* 436, $1.7\,g\,kg^{-1}$; *m*/*z* 372, $1.8\,g\,kg^{-1}$; *m*/*z* 259, $1.2\,g\,kg^{-1}$; $\lambda_{max\ 230\,nm}$ $1.7\,g\,kg^{-1}$.

9.16.5.2.1 Tissue analysis by liquid chromatography-ion trap mass spectrometry

9.16.5.2.1(i) Samples and reference compounds Canola-quality *Brassica juncea* (condiment mustard) cv AC Vulcan, canola *Brassica napus* cv AV, and Sapphire broccoli seeds (*Brassica oleracea* cv italica) were used in these experiments. Sapphire broccoli sprouts were grown in the laboratory and commercial 'brassica' sprouts were obtained from a local supplier. Pure glucosinolate reference compounds were purified by ion-exchange

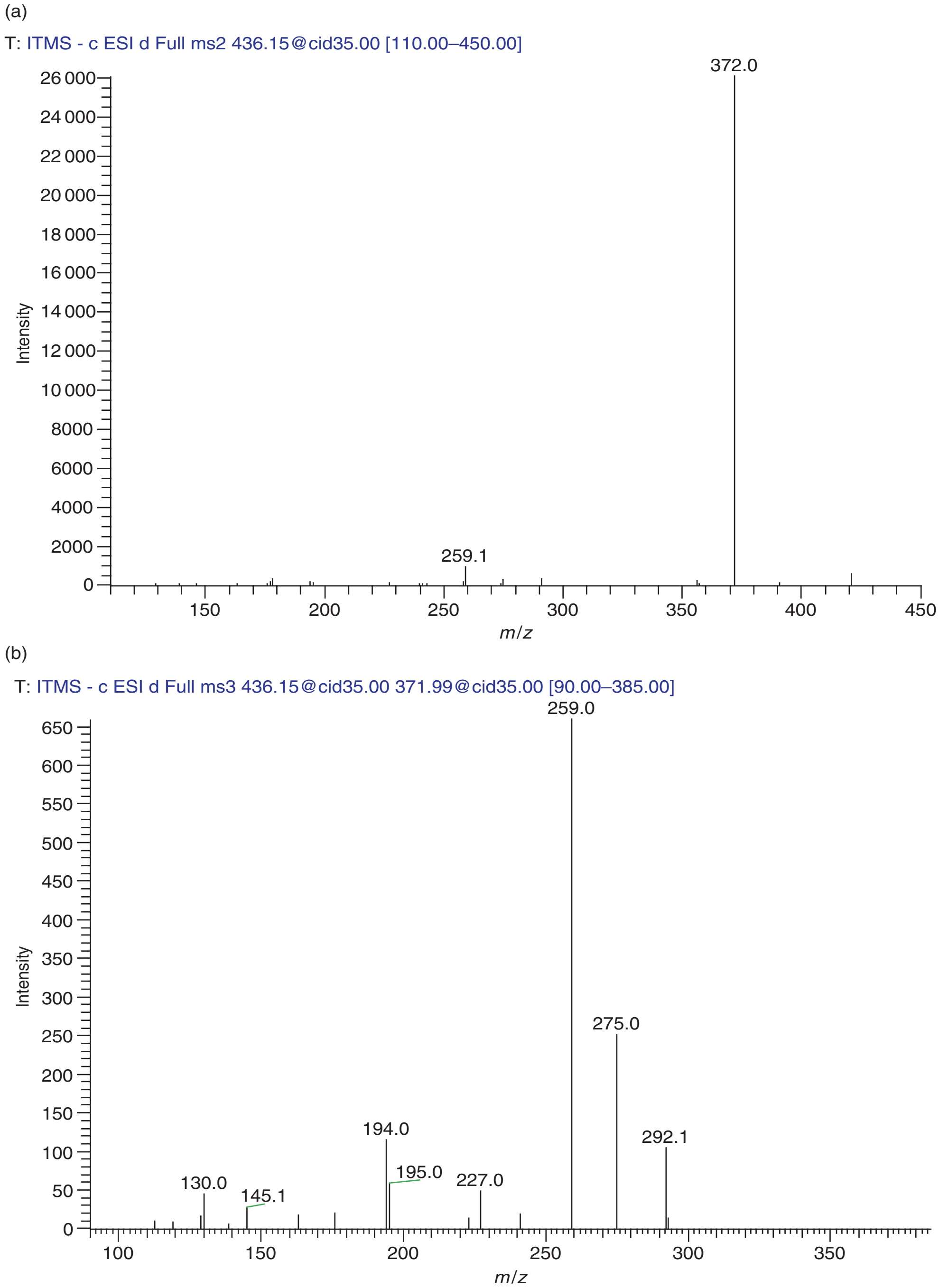


Figure 9 Fragmentation of glucoraphanin: (a) MS^2 fragmentation results in a dominant ion at *m/z* 372 and (b) fragmentation of the *m/z* 372 results in an MS^3 spectrum with *m/z* 259 as the dominant ion. Reprinted from S. J. Rochfort; V. C. Trenerry; M. Imsic; J. Panozzo; R. Jones, *Phytochemistry* **2008**, *69* (8), 1671–1679. Copyright (2008), with permission from Elsevier.

chromatography and combined to form a standard mix for LC-ESI-MS. Glucosinolates were isolated from the following sources: glucoerucin from Rocket seeds (*Eruca sativa*); glucosinalbin from white mustard seeds (*Sinapsis alba*); glucotropaeolin from garden cress (*Lepidium sativum*); glucoiberin, glucoraphanin, and neoglucobrassicin from broccoli seeds (*B. oleracea* cv italica).

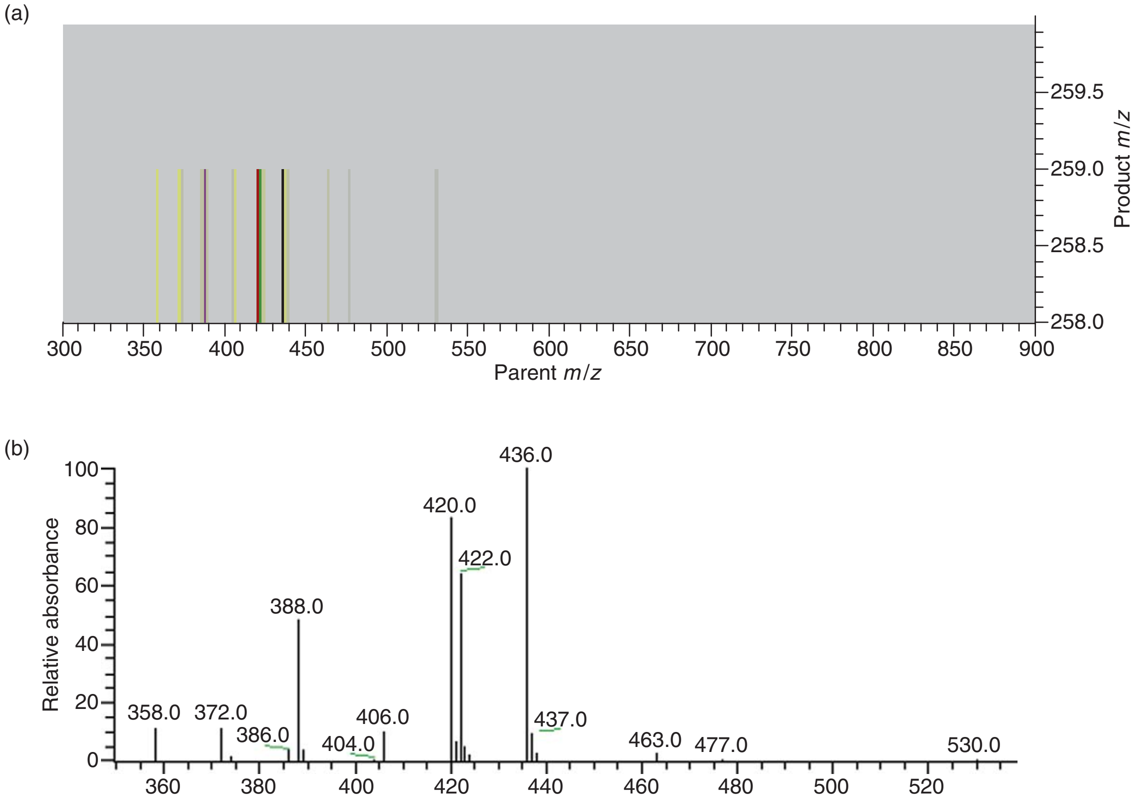


Figure 10 Ion mapping result from a broccoli seed extract: (a) the map view generated from the experiment and (b) the spectrum view of the same data. Reprinted from S. J. Rochfort; V. C. Trenerry; M. Imsic; J. Panozzo; R. Jones, *Phytochemistry* **2008**, *69* (8), 1671–1679. Copyright (2008), with permission from Elsevier.

9.16.5.2.1(ii) Extraction of plant material

9.16.5.2.1(ii)(a) Mustard, canola, and broccoli seeds A 15 ml volume of hot water (90 °C) was added to 1 g of seeds and the solution boiled for 5 min to destroy the enzyme myrosinase. The mixture was transferred to a mortar and the seeds/water ground to a paste. The paste was transferred to a 100 ml volumetric flask with water (90 ml) and the mixture sonicated for 5 min. For mustard and broccoli seeds, the solution was made to volume (100 ml), mixed thoroughly, and allowed to separate (20 min). A portion of the upper layer was filtered through a 0.45 μm cellulose acetate syringe filter disc for analysis. For canola seeds, the final volume of the solution was 50 ml.

9.16.5.2.1(ii)(b) Brassica/broccoli sprouts and florets A 70 ml volume of hot water (90 °C) was added to 10 g of sprouts/florets and the solution boiled for 5 min. The mixture was then blended with a Bamix blender for 5 min and transferred to a 100 ml volumetric flask with water (90 ml) and the mixture sonicated for 5 min. The solution was made to volume (100 ml), mixed thoroughly, and allowed to separate (20 min). A portion of the upper layer was filtered through a 0.45 μm cellulose acetate syringe filter disc for analysis.

9.16.5.3 Parent Ion Mapping Experiment

Prior to data acquisition, the system was tuned using a 250 mg l^{-1} standard of sinigrin infused into the Thermo Fisher LTQ linear ion trap mass spectrometer at a flow rate of 10 μl min^{-1}. The ion map data were acquired with a parent mass range of 300–900, with a 1 mass unit step and an isolation width of 1 mass unit. Normalized collision energy was set at 35 with an activation time of 30 ms. The product mass scanned for was *m*/*z* 259 with a mass width of 1.5. Sample for analysis was introduced into the spectrometer with an infusion flow rate of 10 μl min^{-1}.

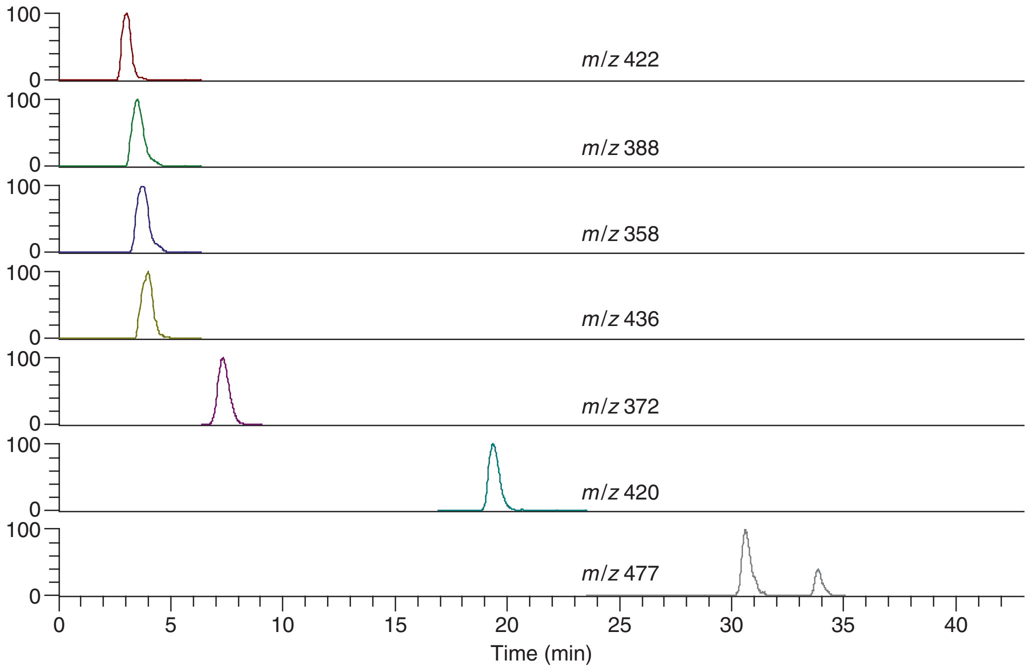


Figure 11 Targeted analysis of Brassica sprouts depicting parent ion and MS^n identification of glucosinolates via extraction of the *m/z* 259 ion: *m/z* 422, glucoiberin; *m/z* 388, progoitrin; *m/z* 358, sinigrin; *m/z* 436, glucoraphanin; *m/z* 372, gluconapin; *m/z* 420, glucoerucin; *m/z* 477, 4-methoxyglucobrassicin and neoglucobrassicin. Reprinted from S. J. Rochfort; V. C. Trenerry; M. Imsic; J. Panozzo; R. Jones, *Phytochemistry* **2008**, *69* (8), 1671–1679. Copyright (2008), with permission from Elsevier.

9.16.5.3.1 High-performance liquid chromatography analysis

The mixtures were analyzed using a 150 mm × 2.1 mm BDS Hypersil 3 µm C18 HPLC column fitted to an Agilent series 1100 high-performance liquid chromatograph (quaternary gradient pump, cooled autosampler maintained at 4 °C, column heater maintained at 30 °C, and diode array detector). For the unbiased analysis, the compounds were eluted from the column using a gradient mobile phase consisting of a mixture of three solvents, A (50 mmol l^{-1} ammonium acetate in water), B (water), and C (methanol) – see Table 7 for gradient details. For the targeted analysis, the compounds were eluted from the column using a gradient mobile phase consisting of a mixture of two solvents, A (0.1% ammonium acetate in water) and B (0.1% ammonium acetate in methanol) – see Table 8 for gradient details. For quantifying sinigrin in mustard seed and glucoraphanin in broccoli florets, the mobile phase consisted of 0.1% formic acid in water (isocratic). Flow rates were maintained at 0.2 ml min^{-1}. The compounds were detected with a Thermo Fisher LTQ ESI-ion trap mass spectrometer operating in the negative ion mode. MS grade solvent, water with 0.1% ammonium acetate, and methanol with 0.1% ammonium acetate were used for targeted studies.

Table 7 Gradient information for HPLC analysis

Time (min)	*Flow rate (ml min^{-1})*	*Mobile phase A*	*Mobile phase B*	*Mobile phase C*
0	0.2	80	20	0
4	0.2	80	20	0
14	0.2	80	10	10
20	0.2	80	10	10
25	0.2	75	0	25
35	0.2	75	0	25
36	0.2	80	20	0
43	0.2	80	20	0

Table 8 Gradient information for targeted analysis

Time (min)	*Flow rate (ml min^{-1})*	*Mobile phase A*	*Mobile phase B*
0	0.2	100	0
4	0.2	100	0
14	0.2	90	10
20	0.2	90	10
25	0.2	75	25
35	0.2	75	25
36	0.2	100	0
43	0.2	100	0

9.16.6 Conclusion

Metabolomics is a very demanding field. Currently, the 'ideal' analytical platform does not exist and is unlikely to occur in the near future. Different methodologies have distinct advantages that can be exploited for investigating different metabolite classes. The resulting information is then put together to obtain a better characterization of the metabolome. The low proportion of analyte identification, due to the lack of structural information obtainable from hyphenated chromatographic/electrophoretic-MS, presents the greatest challenge in MS-based metabolite profiling. In addition to using arrays of reference compounds to generate comprehensive spectral databases, strategies have to be developed and implemented to allow for the rapid characterization of 'general unknowns.' NMR spectroscopy has the highest potential in identifying small organic molecules. The introduction of miniaturized hyphenated NMR instruments such as capillary NMR or LC-SPE-NMR has increased the range of metabolite profiling technologies. The combination of NMR- and MS-based technologies with state-of-the-art chromatographic systems presents the best chance to develop metabolic profiling toward a sturdy, robust analytical platform for the comprehensive analysis of complex natural products present in biological systems.

Abbreviations

1D	one-dimensional
2D	two-dimensional
APCI	atmospheric pressure chemical ionization
APPI	atmospheric pressure photoionization
CE	capillary electrophoresis
CI	chemical ionization
CPMG	Carr–Purcell–Meiboom–Gill
CZE	capillary zone electrophoresis
DA	discriminant analysis
DAD	diode array detection
DIMS	direct injection mass spectrometry
DIOS	desorption ionization on silicon
EI	electron ionization
EOF	electroosmotic flow
ER	endoplasmic reticulum
ERβ	estrogen receptor β
ESI	electrospray ionization
FT-ICR	Fourier transform ion cyclotron resonance
FTIR	Fourier transform infrared
FWHW	full-width, half-maximum
GALDI-IMS	graphite-assisted laser desorption/ionization mass spectrometry

GC	gas chromatography
HILIC	hydrophilic interaction liquid chromatography
HPLC	high-performance liquid chromatography
IT-FT-ICR	ion trap Fourier transform ion cyclotron resonance
LC	liquid chromatography
MALDI-TOF	matrix-assisted laser desorption ionization time-of-flight
MAS	magic angle spinning
MEKC	micellar electrokinetic capillary chromatography
MMCD	Madison Metabolomics Consortium Database
MOA	mode of action
MRM	multiple reaction monitoring
MS	mass spectrometry
NF-κB	nuclear factor-kappaB
NIR	near infrared
NMR	nuclear magnetic resonance
PCA	principal components analysis
PLS	partial least squares
presat-NOESY	presaturation-nuclear overhauser effect spectroscopy
presat-PGSTE	presaturation with a pulse gradient stimulated echo sequence
Q	single quadrupole
QqQ	triple quadrupole
SPME	solid-phase microextraction
TMS	trimethylsilyl
TOF	time-of-flight
ULDA	uncorrelated linear discriminant analysis
UPLC	ultra-performance liquid chromatography
VANTED	visualization and analysis of networks containing experimental data
XIE	extracted ion electropherogram

References

1. N. H. Oberlies; D. J. Kroll, *J. Nat. Prod.* **2004**, *67* (2), 129–135.
2. S. Rochfort; J. Panozzo, *J. Agric. Food Chem.* **2007**, *55* (20), 7981–7994.
3. O. Fiehn, *Plant Mol. Biol.* **2002**, *48* (1), 155–171.
4. S. Rochfort, *J. Nat. Prod.* **2005**, *68* (12), 1813–1820.
5. D. Ryan; K. Robards, *Anal. Chem.* **2006**, *78* (23), 7954–7958.
6. W. B. Dunn; D. I. Ellis, *Trends Analyt. Chem.* **2005**, *24* (4), 285–294.
7. M. Bedair; L. W. Sumner, *Trends Analyt. Chem.* **2008**, *27* (3), 238–250.
8. O. Fiehn, *Trends Analyt. Chem* **2008**, *27* (3), 261–269.
9. M. Glinski; W. Weckwerth, *Mass Spectrom. Rev.* **2006**, *25* (2), 173–214.
10. J. C. Lindon; J. K. Nicholson, *Trends Analyt. Chem.* **2008**, *27* (3), 194–204.
11. X. Lu; X. Zhao; C. Bai; C. Zhao; G. Lu; G. Xu, *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* **2008**, *866* (1–2), 64–76.
12. A. Marston, *Phytochemistry* **2007**, *68* (22–24), 2786–2798.
13. T. O. Metz; J. S. Page; E. S. Baker; K. Tang; J. Ding; Y. Shen; R. D. Smith, *Trends Analyt. Chem.* **2008**, *27* (3), 205–214.
14. M. R. Monton; T. Soga, *J. Chromatogr. A* **2007**, *1168* (1–2), 237–246.
15. W. M. Niessen, *J. Chromatogr. A* **2003**, *1000* (1–2), 413–436.
16. H. Pham-Tuan; L. Kaskavelis; C. A. Daykin; H. G. Janssen, *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* **2003**, *789* (2), 283–301.
17. R. Ramautar; A. Demirci; G. J. D. Jong, *Trends Analyt. Chem.* **2006**, *25* (5), 455–466.
18. E. J. Song; S. M. Babar; E. Oh; M. N. Hasan; H. M. Hong; Y. S. Yoo, *Electrophoresis* **2008**, *29* (1), 129–142.
19. L. W. Sumner; P. Mendes; R. A. Dixon, *Phytochemistry* **2003**, *62* (6), 817–836.
20. G. Theodoridis; H. G. Gika; I. D. Wilson, *Trends Analyt. Chem.* **2008**, *27* (3), 251–260.
21. D. S. Wishart, *Trends Analyt. Chem.* **2008**, *27* (3), 228–237.
22. D. J. Russell; C. E. Hadden; G. E. Martin; A. A. Gibson; A. P. Zens; J. L. Carolan, *J. Nat. Prod.* **2000**, *63* (8), 1047–1049.

23. M. W. Voehler; G. Collier; J. K. Young; M. P. Stone; M. W. Germann, *J. Magn. Reson.* **2006**, *183* (1), 102–109.
24. R. E. Hopson; W. Peti, *Methods Mol. Biol.* **2008**, *426*, 447–458.
25. A. P. Kentgens; J. Bart; P. J. van Bentum; A. Brinkmann; E. R. van Eck; J. G. Gardeniers; J. W. Janssen; P. Knijn; S. Vasa; M. H. Verkuijlen, *J. Chem. Phys.* **2008**, *128* (5), 052202.
26. M. J. Duer, *Introduction to Solid-State NMR Spectroscopy*; Blackwell: Oxford, Malden, MA, 2004.
27. J.-H. Chen; S. Singer, High-Resolution Magic Angle Spinning NMR Spectroscopy. In *The Handbook of Metabonomics and Metabolomics*; J. C. Lindon, J. K. Nicholson, E. Holmes, Eds.; Elsevier: Amsterdam, 2007; pp 113–148.
28. V. Exarchou; M. Krucker; T. A. van Beek; J. Vervoort; I. P. Gerothanassis; K. Albert, *Magn. Reson. Chem.* **2005**, *43* (9), 681–687.
29. A. J. Alexander; F. X. C. Bernard, *Magn. Reson. Chem.* **2006**, *44* (1), 1–6.
30. C. Rae; C. El-Hajj Moussa; J. L. Griffin; W. A. Bubb; T. Wallis; V. J. Balcar, *J. Neurochem.* **2005**, *92* (2), 405–416.
31. M. E. Dumas; C. Canlet; F. Andre; J. Vercauteren; A. Paris, *Anal. Chem.* **2002**, *74* (10), 2261–2273.
32. S. Tiziani; A. Lodi; C. Ludwig; H. M. Parsons; M. R. Viant, *Anal. Chim. Acta* **2008**, *610* (1), 80–88.
33. M. R. Viant, *Biochem. Biophys. Res. Commun.* **2003**, *310* (3), 943–948.
34. M. Coen; Y. S. Hong; O. Cloarec; C. M. Rhode; M. D. Reily; D. G. Robertson; E. Holmes; J. C. Lindon; J. K. Nicholson, *Anal. Chem.* **2007**, *79* (23), 8956–8966.
35. Y. Wang; O. Cloarec; H. Tang; J. C. Lindon; E. Holmes; S. Kochhar; J. K. Nicholson, *Anal. Chem.* **2008**, *80* (4), 1058–1066.
36. G. K. Pierens; M. E. Palframan; C. J. Tranter; A. R. Carroll; R. J. Quinn, *Magn. Reson. Chem.* **2005**, *43* (5), 359–365.
37. D. S. Wishart; D. Tzur; C. Knox; R. Eisner; A. C. Guo; N. Young; D. Cheng; K. Jewell; D. Arndt; S. Sawhney; C. Fung; L. Nikolai; M. Lewis; M. A. Coutouly; I. Forsythe; P. Tang; S. Shrivastava; K. Jeroncic; P. Stothard; G. Amegbey; D. Block; D. D. Hau; J. Wagner; J. Miniaci; M. Clements; M. Gebremedhin; N. Guo; Y. Zhang; G. E. Duggan; G. D. Macinnis; A. M. Weljie; R. Dowlatabadi; F. Bamforth; D. Clive; R. Greiner; L. Li; T. Marrie; B. D. Sykes; H. J. Vogel; L. Querengesser, *Nucleic Acids Res.* **2007**, *35* (Database issue), D521–D526.
38. J. G. Bundy; E. M. Lenz; N. J. Bailey; C. L. Gavaghan; C. Svendsen; D. Spurgeon; P. K. Hankard; D. Osborn; J. M. Weeks; S. A. Trauger; P. Speir; I. Sanders; J. C. Lindon; J. K. Nicholson; H. Tang, *Environ. Toxicol. Chem.* **2002**, *21* (9), 1966–1972.
39. J. G. Bundy; J. K. Sidhu; F. Rana; D. J. Spurgeon; C. Svendsen; J. F. Wren; S. R. Sturzenbaum; A. J. Morgan; P. Kille, *BMC Biol.* **2008**, *6*, *25*.
40. F. Van der Kooy; R. Verpoorte; J. J. Marion Meyer, *S. Afr. J. Bot.* **2008**, *74* (2), 186–189.
41. K. H. Ott; N. Aranbar; B. Singh; G. W. Stockton, *Phytochemistry* **2003**, *62* (6), 971–985.
42. S. Sturm; C. Seger; H. Stuppner, *J. Chromatogr. A* **2007**, *1159* (1–2), 42–50.
43. A. Jansma; T. Chuan; R. W. Albrecht; D. L. Olson; T. L. Peck; B. H. Geierstanger, *Anal. Chem.* **2005**, *77* (19), 6509–6515.
44. M. Jahangir; H. K. Kim; Y. H. Choi; R. Verpoorte, *Food Chem.* **2008**, *107* (1), 362–368.
45. F. Zhang; A. T. Dossey; C. Zachariah; A. S. Edison; R. Bruschweiler, *Anal. Chem.* **2007**, *79* (20), 7748–7752.
46. Y. H. Choi; E. C. Tapias; H. K. Kim; A. W. M. Lefeber; C. Erkelens; J. T. J. Verhoeven; J. Brzin; J. Zel; R. Verpoorte, *Plant Physiol.* **2004**, *135* (4), 2398–2410.
47. G. F. Pauli; B. U. Jaki; D. C. Lankin, *J. Nat. Prod.* **2007**, *70* (4), 589–595.
48. E. C. Tatsis; S. Boeren; V. Exarchou; A. N. Troganis; J. Vervoort; I. P. Gerothanassis, *Phytochemistry* **2007**, *68* (3), 383–393.
49. A. P. Sobolev; E. Brosio; R. Gianferri; A. L. Segre, *Magn. Reson. Chem.* **2005**, *43* (8), 625–638.
50. H. Lu; Y. Liang; W. B. Dunn; H. Shen; D. B. Kell, *Trends Anal. Chem.* **2008**, *27* (3), 215–227.
51. O. Fiehn; G. Wohlgemuth; M. Scholz; T. Kind; Y. Lee do; Y. Lu; S. Moon; B. Nikolau, *Plant J.* **2008**, *53* (4), 691–704.
52. S. Seger; S. Sturm, *J. Proteome Res.* **2007**, *6* (2), 480–497.
53. M. Dauner; U. Sauer, *Biotechnol. Prog.* **2000**, *16* (4), 642–649.
54. U. Roessner; C. Wagner; J. Kopka; R. N. Trethewey; L. Willmitzer, *Plant J.* **2000**, *23* (1), 131–142.
55. A. Barsch; T. Patschkowski; K. Niehaus, *Funct. Integr. Genomics* **2004**, *4* (4), 219–230.
56. G. S. Catchpole; M. Beckmann; D. P. Enot; M. Mondhe; B. Zywicki; J. Taylor; N. Hardy; A. Smith; R. D. King; D. B. Kell; O. Fiehn; J. Draper, *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102* (40), 14458–14462.
57. Y. Tikunov; A. Lommen; C. H. de Vos; H. A. Verhoeven; R. J. Bino; R. D. Hall; A. G. Bovy, *Plant Physiol.* **2005**, *139* (3), 1125–1137.
58. N. Schauer; D. Zamir; A. R. Fernie, *J. Exp. Bot.* **2005**, *56* (410), 297–307.
59. J. K. Kim; T. Bamba; K. Harada; E. Fukusaki; A. Kobayashi, *J. Exp. Bot.* **2007**, *58* (3), 415–424.
60. A. Bovy; E. Schijlen; R. Hall, *Metabolomics* **2007**, *3* (3), 399–412.
61. R. E. Mohler; B. P. Tu; K. M. Dombek; J. C. Hoggard; E. T. Young; R. E. Synovec, *J. Chromatogr. A* **2008**, *1186* (1–2), 401–411.
62. C. L. Winder; W. B. Dunn; S. Schuler; D. Broadhurst; R. Jarvis; G. M. Stephens; R. Goodacre, *Anal. Chem.* **2008**, *80* (8), 2939–2948.
63. D. H. Sanchez; F. Lippold; H. Redestig; M. A. Hannah; A. Erban; U. Kramer; J. Kopka; M. K. Udvardi, *Plant J.* **2008**, *53* (6), 973–987.
64. T. Soga, *Methods Mol. Biol.* **2007**, *358*, 129–137.
65. S. Sato; T. Soga; T. Nishioka; M. Tomita, *Plant J.* **2004**, *40* (1), 151–163.
66. N. Ishii; T. Soga; T. Nishioka; M. Tomita, *Metabolomics* **2005**, *1* (1), 29–37.
67. B. D. Williams; C. J. Cameron; R. Workman; C. D. Broeckling; L. W. Sumner; J. T. Smith, *Electrophoresis* **2007**, *28* (9), 1371–1379.
68. Y. Ohashi; A. Hirayama; T. Ishikawa; S. Nakamura; K. Shimizu; Y. Ueno; M. Tomita; T. Soga, *Mol. Biosyst.* **2008**, *4* (2), 135–147.
69. E. E. Baidoo; P. I. Benke; C. Neususs; M. Pelzing; G. Kruppa; J. A. Leary; J. D. Keasling, *Anal. Chem.* **2008**, *80* (9), 3112–3122.
70. P. Iadarola; F. Ferrari; M. Fumagalli; S. Viglio, *Electrophoresis* **2008**, *29* (1), 224–236.
71. A. Nordstrom; E. Want; T. Northen; J. Lehtio; G. Siuzdak, *Anal. Chem.* **2008**, *80* (2), 421–429.
72. S. J. Rochfort; M. Imsic; R. Jones; V. C. Trenerry; B. Tomkins, *J. Agric. Food Chem.* **2006**, *54* (13), 4855–4860.
73. G. Xie; R. Plumb; M. Su; Z. Xu; A. Zhao; M. Qiu; X. Long; Z. Liu; W. Jia, *J. Sep. Sci.* **2008**, *31* (6–7), 1015–1026.
74. J. Burns; P. D. Fraser; P. M. Bramley, *Phytochemistry* **2003**, *62* (6), 939–947.
75. R. J. Bino; C. H. Ric de Vos; M. Lieberman; R. D. Hall; A. Bovy; H. H. Jonker; Y. Tikunov; A. Lommen; S. Moco; I. Levin, *New Phytol.* **2005**, *166* (2), 427–438.

76. M. Long; D. J. Millar; Y. Kimura; G. Donovan; J. Rees; P. D. Fraser; P. M. Bramley; G. P. Bolwell, *Phytochemistry* **2006**, *67* (16), 1750–1757.
77. F. Mounet; M. Lemaire-Chamley; M. Maucourt; C. Cabasson; J.-L. Giraudel; C. Deborde; R. Lessire; P. Gallusci; A. Bertrand; M. Gaudillère; C. Rothan; D. Rolin; A. Moing, *Metabolomics* **2007**, *3* (3), 273–288.
78. M. Yamazaki; J.-i. Nakajima; M. Yamanashi; M. Sugiyama; Y. Makita; K. Springob; M. Awazuhara; K. Saito, *Phytochemistry* **2003**, *62* (6), 987–995.
79. J. I. Nakajima; I. Tanaka; S. Seo; M. Yamazaki; K. Saito, *J. Biomed. Biotechnol.* **2004**, *2004* (5), 241–247.
80. A. J. Parr; F. A. Mellon; I. J. Colquhoun; H. V. Davies, *J. Agric. Food Chem.* **2005**, *53* (13), 5461–5466.
81. X. Chen; L. Kong; X. Su; C. Pan; M. Ye; H. Zou, *J. Chromatogr. A* **2005**, *1089* (1–2), 87–100.
82. K. Morreel; G. Goeminne; V. Storme; L. Sterck; J. Ralph; W. Coppieters; P. Breyne; M. Steenackers; M. Georges; E. Messens; W. Boerjan, *Plant J.* **2006**, *47* (2), 224–237.
83. S. Moco; R. J. Bino; O. Vorst; H. A. Verhoeven; J. de Groot; T. A. van Beek; J. Vervoort; C. H. de Vos, *Plant Physiol.* **2006**, *141* (4), 1205–1218.
84. M. Stobiecki; A. Skirycz; L. Kerhoas; P. Kachlicki; D. Muth; J. Einhorn; B. Mueller-Roeber, *Metabolomics* **2006**, *2* (4), 197–219.
85. S. J. Rochfort; V. C. Trenerry; M. Imsic; J. Panozzo; R. Jones, *Phytochemistry* **2008**, *69* (8), 1671–1679.
86. J. Ding; C. M. Sorensen; Q. Zhang; H. Jiang; N. Jaitly; E. A. Livesay; Y. Shen; R. D. Smith; T. O. Metz, *Anal. Chem.* **2007**, *79* (16), 6081–6093.
87. C. S. Harris; A. J. Burt; A. Saleem; P. M. Le; L. C. Martineau; P. S. Haddad; S. A. L. Bennett; J. T. Arnason, *Phytochem. Anal.* **2007**, *18* (2), 161–169.
88. H. Ogiso; T. Suzuki; R. Taguchi, *Anal. Biochem.* **2008**, *375* (1), 124–131.
89. S. J. Rochfort; V. C. Trenerry, *Chem. Aust.* **2006**, *73* (6), 11–14.
90. E. C. Y. Chan; S.-L. Yap; A.-J. Lau; P.-C. Leow; D.-F. Toh; H.-L. Koh, *Rapid Commun. Mass Spectrom.* **2007**, *21* (4), 519–528.
91. W. B. Dunn; S. Overy; W. P. Quick, *Metabolomics* **2005**, *1* (2), 137–148.
92. S. Vaidyanathan; D. Jones; D. Broadhurst; J. Ellis; T. Jenkins; W. Dunn; A. Hayes; N. Burton; S. Oliver; D. Kell; R. Goodacre, *Metabolomics* **2005**, *1* (3), 243–250.
93. D. Ohta; D. Shibata; S. Kanaya, *Anal. Bioanal. Chem.* **2007**, *389* (5), 1469–1475.
94. P. D. Fraser; E. M. A. Enfissi; M. Goodfellow; T. Eguchi; P. M. Bramley, *Plant J.* **2007**, *49* (3), 552–564.
95. Y. Li; B. Shrestha; A. Vertes, *Anal. Chem.* **2007**, *79* (2), 523–532.
96. H. Zhang; S. Cha; E. S. Yeung, *Anal. Chem.* **2007**, *79* (17), 6575–6584.
97. G. Madalinski; E. Godat; S. Alves; D. Lesage; E. Genin; P. Levi; J. Labarre; J. C. Tabet; E. Ezan; C. Junot, *Anal. Chem.* **2008**, *80* (9), 3291–3303.
98. D. Stewart; G. J. McDougall; J. Sungurtas; S. Verrall; J. Graham; I. Martinussen, *Mol. Nutr. Food Res.* **2007**, *51* (6), 645–651.
99. H. E. Johnson; D. Broadhurst; R. Goodacre; A. R. Smith, *Phytochemistry* **2003**, *62* (6), 919–928.
100. H. Antti; T. M. D. Ebbels; H. C. Keun; M. E. Bollard; O. Beckonert; J. C. Lindon; J. K. Nicholson; E. Holmes, *Chemom. Intell. Lab. Syst.* **2004**, *73* (1), 139–149.
101. R. Goodacre, *J. Exp. Bot.* **2005**, *56* (410), 245–254.
102. R. Goodacre; D. Broadhurst; A. Smilde; B. Kristal; J. Baker; R. Beger; C. Bessant; S. Connor; G. Capuani; A. Craig; T. Ebbels; D. Kell; C. Manetti; J. Newton; G. Paternostro; R. Somorjai; M. Sjöström; J. Trygg; F. Wulfert, *Metabolomics* **2007**, *3* (3), 231–241.
103. P. J. Mulquiney; P. W. Kuchel, Models of Human Erythrocyte Metabolism. In *Modelling Metabolism with Mathematica*; CRC Press: Boca Raton, FL, 2003; pp 175–196.
104. D. M. Rocke, *Semin. Cell Dev. Biol.* **2004**, *15* (6), 703–713.
105. H. Antti; M. E. Bollard; T. Ebbels; H. Keun; J. C. Lindon; J. K. Nicholson; E. Holmes, *J. Chemom.* **2002**, *16* (8–10), 461–468.
106. D. Yuan; Y. Liang; L. Yi; Q. Xu; O. M. Kvalheim, *Chemom. Intell. Lab. Syst.* **2008**, *93* (1), 70–79.
107. Q. Cui; I. A. Lewis; A. D. Hegeman; M. E. Anderson; J. Li; C. F. Schulte; W. M. Westler; H. R. Eghbalnia; M. R. Sussman; J. L. Markley, *Nat. Biotechnol.* **2008**, *26* (2), 162–164.
108. M. Scholz; O. Fiehn, *Pac. Symp. Biocomput.* **2007**, *12*, 169–180.
109. E. Grafahrend-Belau; S. Weise; D. Koschutzki; U. Scholz; B. H. Junker; F. Schreiber, *Nucleic Acids Res.* **2008**, *36* (Database issue), D954–D958.
110. B. M. Lange; M. Ghassemian, *Phytochemistry* **2005**, *66* (4), 412–451.
111. B. H. Junker; C. Klukas; F. Schreiber, *BMC Bioinformatics* **2006**, *7*, 109.
112. C. D. Broeckling; D. V. Huhman; M. A. Farag; J. T. Smith; G. D. May; P. Mendes; R. A. Dixon; L. W. Sumner, *J. Exp. Bot.* **2005**, *56* (410), 323–336.

Biographical Sketches

V. Craige Trenerry completed his Ph.D. in synthetic organic chemistry at Adelaide University in 1979. He gained postdoctoral experience with Professor John Bowie at Adelaide University and Professor Jim Coxon at the University of Canterbury before joining the Australian Government Analytical Laboratories in 1982 as a forensic scientist. In 1985, he was transferred to the food group where he continued to develop skills in HPLC, GC, and CE in food and illicit drug analysis. He joined the Victorian Department of Primary Industries in 2001 and is currently a Principal Research Scientist in the Future Farming Research Division. His current interests are in the application of GC/HPLC/CE-MS to the study of nutritionally important bioactives in food.

Simone J. Rochfort completed her Ph.D. in marine natural products chemistry at the University of Melbourne in 1996. After a postdoctoral fellowship with Dr. Jeffrey Wright, National Research Council of Canada, she returned to Australia to take up a research position with AstraZeneca R&D Griffith University where she worked on the discovery of natural products for human pharmaceuticals. Dr. Rochfort's research in the pharmaceutical and biotech industries continued until 2004 when she joined the Victorian Department of Primary Industries. She is currently employed as a Principal Research Scientist and applies her natural products research interests to metabolomics and the substantiation of functional foods.

9.17 Small Molecules as Versatile Tools for Activity-Based Protein Profiling Experiments

Stephan A. Sieber, Thomas Böttcher, Isabell Staub, and Ronald Orth, Ludwig-Maximilians-Universität München, Munich, Germany

9.17.1 Introduction

Genome sequencing projects have provided a wealth of information on gene identities in prokaryotic and eukaryotic organisms. Currently, a total of 720 genome sequences are completed and a large number is still in progress (http://www.ncbi.nlm.nih.gov). The daunting challenge for proteomic research now is to assign the molecular, cellular, and (patho)physiological functions for the full complement of proteins encoded by the genome.[1] Since significant fractions of sequenced genomes encode uncharacterized enzymes, the analysis of genome sequences alone will not be sufficient to achieve this task especially if one considers that a single gene can in principle code for several proteins as a result of posttranscriptional and posttranslational processing. New technologies have been developed to characterize gene transcription on the mRNA level and on the protein expression level of cells including RNA microarrays and two-dimensional sodium dodecyl sulfate (SDS) gel electrophoresis, respectively. However, these established technologies focus on the abundance of RNA or proteins not taking into account that the cell uses in many cases additional posttranslational processing steps to generate the active protein that participates in its dedicated physiological or pathological cellular function (**Figure 1**).[2] The necessity of additional tools for the determination of activity and function can be illustrated by the example of the protease enzyme family: Many proteases are key players of crucial cellular processes requiring a tight regulation of their activity in order to keep the balance between needed proteolytic power and uncontrolled proteolytic degradation.[3] Under physiological conditions, the cell has developed a sophisticated system that regulates protease activity by the expression of catalytically inactive enzymes that need to be posttranslationally processed in order to generate an active species (**Figure 2**). For instance, cleavage of propeptides leads to the active enzyme that can subsequently engage in its dedicated physiological function. However, to keep this proteolytically active population under control, in some cases the activity can be tuned by the action of natural protein inhibitors (e.g., TIMPs (tissue inhibitors of metalloproteinases)).[4] This complex regulatory mechanism allows the cell to control the proteolytic power and minimizes the chance of unwanted degradative processes that could be harmful for the cellular viability. In contrast, it is believed that in many pathological processes such as in cancer, uncontrolled proteolytic activity leads to the degradation of tissue and cell membranes causing invasive processes such as metastasis.[5] In order to determine the proteases that are responsible for these pathological processes, established genomic or proteomic methods, which focus only on the abundance of certain proteins, are not useful to distinguish silent inactive enzymes from the aggressive active ones.[6] To measure protein activity directly, a complementary technology, termed ‘activity-based protein profiling (ABPP)’ has been systematically explored by Cravatt and coworkers about a decade ago. This technology has evolved rapidly to a standard tool for the identification and functional characterization of individual proteins in complex proteomes and has been the subject of recent reviews.[7–15] One striking feature of this approach is the combination of several disciplines including synthetic, biological, and analytical chemistry, which together build a powerful platform for the labeling, visualization, and identification of enzymes in their active state. The principal concept of this technology is the covalent active site labeling of distinct enzyme classes by functionalized small molecules, which are referred to as chemical probes. These probes contain a reporter molecule (e.g., biotin, fluorescent dyes, or radioactive isotopes) that allows visualization of the labeled enzymes via analytical methods. Common chemical reactive group scaffolds of these probes include mechanism-based inhibitors, general electrophiles, and natural products, which usually target only a small subset of active proteins (10–100) in complex proteomic mixtures with ten thousands of proteins present. In contrast to solely proteomic methods, the reduced information content enables the focused analysis of single enzyme families to elucidate their function and regulation in more detail. One of the first examples of an activity-based probe dates back into the 1970s where Strominger and coworkers used radioactive penicillins as probes for the detection of penicillin-binding proteins (PBPs).[16,17] Since then, fundamental advancements in analytical sciences have allowed to substitute radioactivity by fluorescent dyes or affinity tags for visualization and

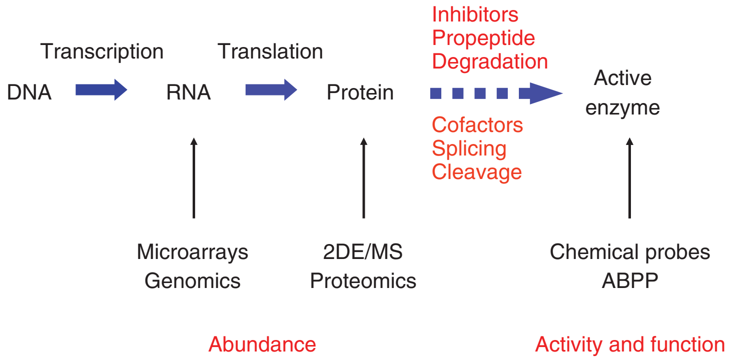


Figure 1 Overview of genomic and proteomic methods. Classical methods focus on the abundance of RNA and proteins. In addition, ABPP explores the activity and function of proteins by chemical–proteomic methods.

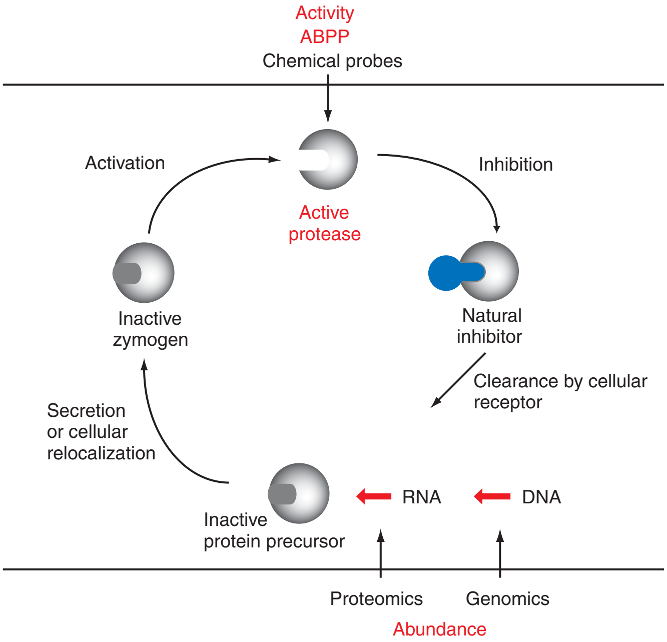


Figure 2 Regulation of proteolytic activity by the cleavage of propeptides and inhibition with natural inhibitors. While classical genomic and proteomic methods fail to detect these vital changes in activity during physiological or pathological processes, ABPP fills the methodological gap.

identification, respectively. In addition, the number and scope of addressable enzyme classes has been expanded to a dozen families including proteases, kinases, phosphatases, glycosidases, and oxidoreductases. The investigation of these enzymes has largely contributed to a better understanding of their role in physiological and pathological processes on a proteome-wide scale. In this chapter we try to capture the main aspects of this novel technology with a special emphasis on probe design and analytical target identification technologies.

9.17.2 Principles of Activity-Based Protein Profiling

In this section we introduce the basic principles of ABPP and provide an overview of its broad applicability and utility for many challenging biological questions. At the beginning we introduce concepts of probe and tag design, which were the basis for rapid developments in inhibitor discovery, imaging, and *in vivo* studies

(Section 9.17.4). Moreover, several analytical platforms have been developed in the past years, which will be discussed as well and put in the context of their individual value in proteomic and medicinal research.

9.17.2.1 Probe Design and Proteomic Labeling

The design of specific probes is a critical part of all ABPP experiments. The choice of the inhibitor, covalent or noncovalent, defines the success of all later experiments in the proteome and heavily depends on the demand of the biological task. For example, if an inhibitor does not display the desired selectivity, for example, it is too reactive, many enzymes will be randomly labeled presumably also at residues that are not within their active site. On the other hand, if the inhibitor is too selective, essential enzyme activities might be missed during the analysis. The choice of the 'right' inhibitor always depends on the biological question that is addressed. The early onset of ABPP experiments pioneered by Cravatt and coworkers for the profiling of the serine hydrolase enzyme family by fluorophosphonate (FP) probes represents an elegant approach to a selective and frequently applied probe class, which will be discussed here as a successful example to introduce the principles of probe design.[18–20]

The general layout of a probe consists of three general elements (**Figure 3**). The first is the active site-directed group (e.g., FP), which defines the affinity for the dedicated enzyme class or classes of the probe. The FP group represents an outstanding example of a selective inhibitor that targets preferentially only the serine hydrolase enzyme family that makes up to 1% of all proteins in a eukaryotic cell.[21] Attached to this inhibitor group is a spacer molecule that is usually a short alkyl or polyethyleneglycol tether that separates the inhibitor from the tag. The spacer can vary in its hydrophobicity in order to account for the different enzyme environments. While some enzymes are better accessed by hydrophobic compounds, others prefer more hydrophilic spacers. The third element of a probe is a proper tag for enrichment, identification, or visualization of protein targets. Initially, radioactive labels have been used for target visualization. These, however, display

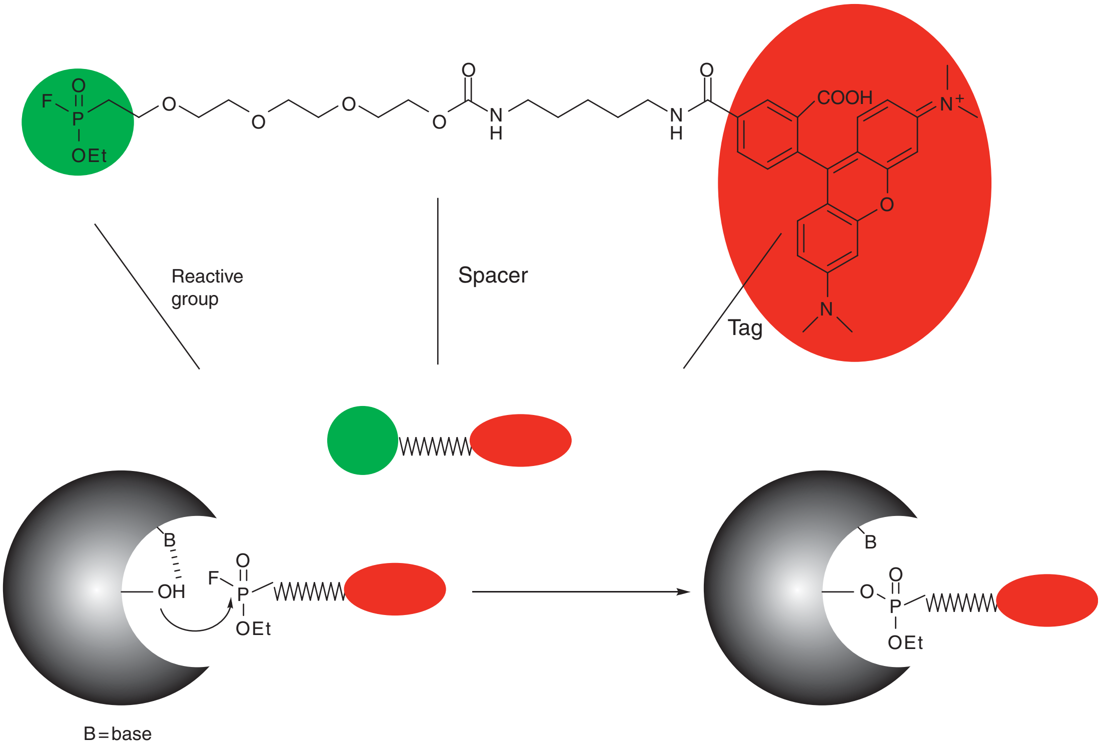


Figure 3 Design of an ABPP probe. The design is driven by the choice of the active site-directed group (e.g., fluorophosphonate (FP), shown in green) and a fluorescent tag (rhodamine, shown in red) separated by a spacer group. FPs react covalently with the active site serine residue of the serine hydrolase enzyme family (see also **Figure 9**). B, base. Reproduced by permission of The Royal Society of Chemistry.

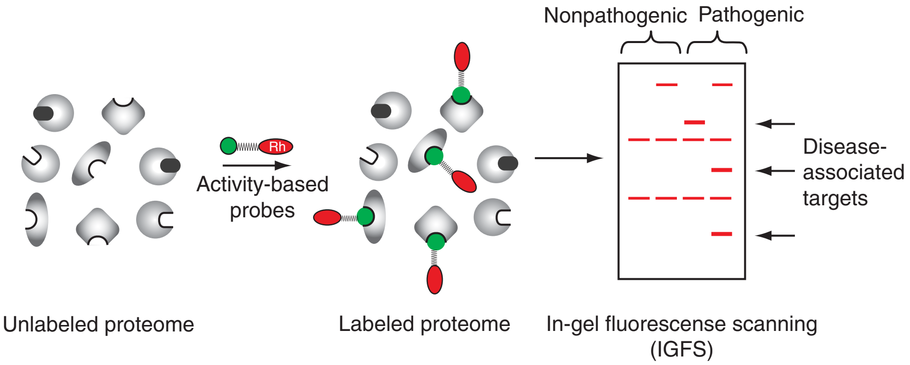


Figure 4 Comparative profiling of enzymatic activity states in the proteomes of nonpathogenic and pathogenic origin. Disease-associated targets can be determined based on the activity pattern revealed by analytical methods such as SDS gel electrophoresis coupled with in-gel fluorescence scanning. Reproduced by the permission of The Royal Society of Chemistry.

limited applicability due to their time-consuming and hazardous experimental procedures and thus have been largely replaced by biotin and fluorescent tags such as tetramethylrhodamine (TAMRA). In ABPP experiments, these tags are of paramount importance due to their great sensitivity, which allows visualization of targets of very low abundance. Moreover, biotin tags facilitate the visualization of labeling events by avidin blotting and enable the mass spectrometry (MS)-based characterization by enrichment with avidin-conjugated beads (see Section 9.17.2.3.2). As we will see in the next section, new technologies have been established to introduce the bulky tag after target labeling, which can display several advantages for enzyme recognition and cell permeability of probes.

In a typical proteomic experiment, the probes are incubated with proteomes, for instance, of two different pathogenic states. During the incubation, the fluorescent probe binds directly to the active site of its dedicated target enzyme(s). Subsequent analytical procedures, such as SDS gel electrophoresis, and in-gel fluorescence scanning (IGFS), allow a rapid comparative analysis of different cell or tissue states that can reveal disease-associated enzyme activities (**Figure 4**). These disease-associated enzymes represent promising candidates for subsequent follow-up studies and could serve as therapeutic targets or diagnostic markers. Later, we will discuss examples of probe–protein interactions that have indeed revealed important enzyme activities with relevance in cancer, infectious diseases, and metabolic disorder (Section 9.17.4).

9.17.2.2 The Introduction of Bio-Orthogonal Chemistry for *In Vivo* Studies

Covalently tethered bulky tags display several shortcomings including low cell permeability, which limited the scope and applicability of initial ABPP studies that were predominantly carried out in cell lysates. One general problem of proteomic experiments in cell lysates is the disruption of organelles and fine compartments, which lead to the release of activators or inhibitors that artificially affect enzymatic activity.[22] Improved methodologies for bio-orthogonal chemical reactions, such as the Cu(I)-catalyzed Huisgen's [3+2] azide–alkyne cycloaddition (click chemistry, CC)[23] or the Staudinger ligation between azides and methylester-modified triphenylphosphines,[24] were separately introduced several years ago and have now become standard procedures for ABPP (**Figure 5**). The major advantage of these new methods for ABPP is the decoupling of the bulky tag from the probe, which allows the smaller probe to label proteins within living cells before the tag is attached in the corresponding cell lysates.[25,26] Usually, the active site-directed group contains an alkyne moiety for CC or an azide moiety for the Staudinger ligation. The corresponding tags are attached to an azide moiety or methylester-modified triphenylphosphines, respectively. While this is the general and preferred layout, it should be noted that in some cases the order of alkyne and azide can be reverse for CC reactions. However, in these cases, increased background labeling was observed.[26] These chemical reactions are compatible with aqueous environments and exhibit low reactivity toward other biomolecules such as DNA or proteins. In a typical CC experiment, the 'tag-free' alkyne probes were

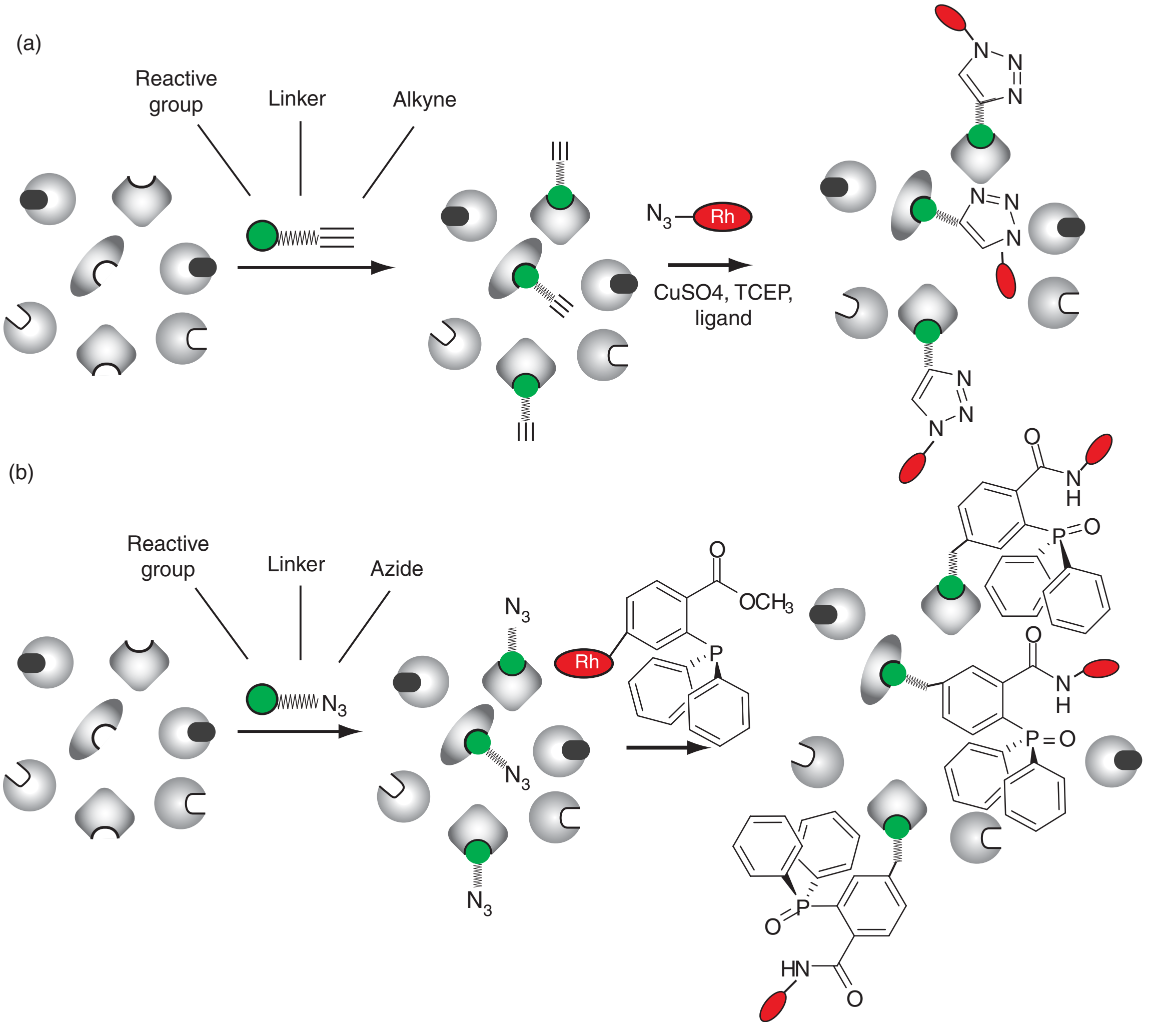


Figure 5 Principles of (a) click chemistry and (b) Staudinger ligation in ABPP experiments.

applied to living cells and after a certain incubation time, the cells were harvested, and the lyzed protein-bound probes chemically reacted with the complementary 'clickable' reporter tag. The product of this reaction is a 1,4-disubstituted triazole that connects the probe-labeled enzyme to the reporter tag. As we will see in later sections, this technology has gained widespread applications in the past years. Recently, a copper-free [3+2] cycloaddition was introduced by Bertozzi and coworkers, which opens up the way to new chemoselective ligations within living cells.[27]

The methodological impact of CC for ABPP was tremendous. Several comparisons of *in vitro* and *in vivo* experiments have revealed that in many cases the activity profiles differ significantly between enzymes in living cells and in homogenates. For instance, the general utility of this approach was demonstrated by profiling cell cultures and living mice with an azide-derivatized ('tag-free') phenylsulfonate (PS) ester probe (Section 9.17.3.2.1) and subsequent reaction of the treated proteomes upon cell lysis with an alkyne–rhodamine reporter group under CC conditions.[25] An optimized version of this approach resulted in the identification of several enzyme activities in living breast cancer cells that were labeled only *in vivo* but not *in vitro* emphasizing the great utility of 'tag-free' ABPP for target identification in cancer research (Section 9.17.4).[26] These results suggest that *in vivo* ABPP assess a dimension of the functional proteome whose integrity is dependent on native cellular environments.

9.17.2.3 Analytical Platforms for Activity-Based Protein Profiling

The specific properties of ABPP experiments had a significant impact on the design of postlabeling detection and identification methods, which will be summarized below. All methods represent complementary approaches that display individual strengths and weaknesses in important proteomic disciplines such as throughput, sensitivity, and sample conservation.

9.17.2.3.1 In-gel fluorescence scanning

IGFS represents the most mature standard method for ABPP, which displays the advantages of being simple and suitable for high-throughput analysis (hundreds of proteomes can be analyzed per day). In this method, typically 30–50 μg of proteome are treated with a fluorescent activity-based probe and applied to an SDS polyacrylamide gel. Subsequent separation by one-dimensional gel electrophoresis and visualization of labeled proteins by fluorescence scanning result in the identification of characteristic proteome-specific activity profiles. Typical sensitivity limits for fluorescently labeled proteins on gels are around 10 pmol mg^{-1} corresponding to approximately 0.05% of the total protein, which is, in some cases, below the detection limit of physiological relevant enzymes. Another limitation is the low resolution of one-dimensional gel electrophoresis in which proteins are separated based on only their molecular weight. Homologous enzymes tend to have a similar molecular mass, so members of a protein family reacting with ABPP probes will be likely to comigrate and obscure or confuse the signal. Therefore, fluorescently labeled proteins of same or similar size can appear as only one band. Moreover, the detection of proteins on a gel does not automatically reveal their identity. This requires additional steps such as cutting and purifying gel bands prior to MS analysis. Furthermore, to reduce high background levels of nontarget proteins in these gel bands and simultaneously provide sufficient target protein for liquid chromatography–mass spectrometry (LC–MS) analysis, a biotin–avidin affinity enrichment step is in many cases required. Nevertheless, IGFS represents a routine application of ABPP for many biological samples in parallel, permitting, for example, the rapid classification of human cancer cell lines into phenotypically relevant groups based on their shared enzyme activity profiles. However, technical limitations in sensitivity and resolution associated with IGFS have led to the development of complementary ABPP analysis platforms that will be discussed below.

9.17.2.3.2 Liquid chromatography–mass spectrometry

Recently, several LC–MS-based methods have emerged as novel, gel-free analysis tools for ABPP.[28–30] The first report was published by Adam *et al.*[28] describing a consolidated single-step identification of both protein targets and the specific amino acid residues labeled by a nondirected PS ester probe (Section 9.17.3.2.1). In this approach, specific methodological steps were required in order to be compatible with the requirements of LC–MS analysis. These included proteome denaturation, reduction, and alkylation of thiols with dithiothreitol and iodoacetamide, respectively. Subsequent tryptic digestion revealed peptide mixtures that were incubated with antirhodamine antibodies to affinity capture PS–rhodamine-labeled peptides (**Figure 6(a)**). Liquid chromatography–tandem mass spectrometry (LC–MS/MS) and a computer search algorithm were used to concurrently identify recombinantly expressed protein targets of PS–rhodamine and the specific residues labeled by this probe. For each enzyme examined, probe labeling was found to occur on a conserved active site residue, including catalytic nucleophiles as well as residues of yet unknown function. A comparison of this gel-free methodology with gel-based ABPP reveals several advantages: The consolidated single step identification of protein targets and active site residues is in particular useful for screening new probes with unknown protein specificity and reactivity profiles. Gel-free ABPP also offers a complementary method for resolving probe-labeled enzymes that comigrate by SDS gels. This additional resolution results from differences in tryptic peptide fragments derived from diverse active sites, which should be sufficiently separated by LC. However, since this gel-free method discards the rest of the proteomic digest, it is difficult to differentiate proteins of high similarity and prohibits molecular analysis of entire protein sequences including posttranslational modifications.

A modified version of gel-free ABPP, introduced by Speers and Cravatt,[29] overcomes these limitations by the application of a tandem orthogonal proteolysis (TOP) strategy for the parallel and full characterization of probe-labeled proteins and sites of modification. In this approach, proteomes were first labeled with a

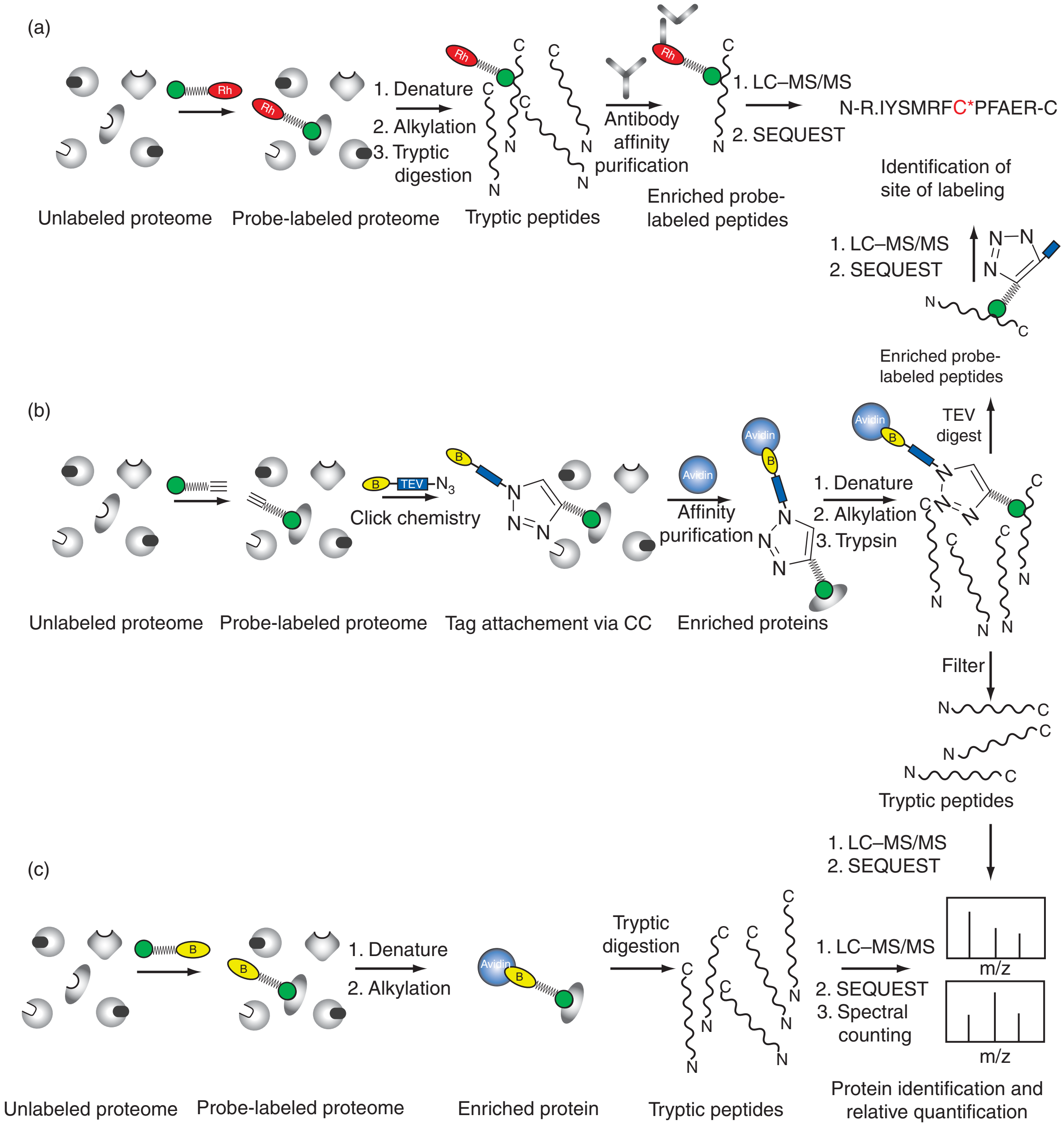


Figure 6 Methods for gel-free ABPP target identification via LC–MS. (a) Active site peptide profiling. Rhodamine probe-labeled enzymes are digested and active site peptides enriched with antirhodamine antibodies. Subsequent LC–MS/MS analysis reveals the protein identity and site of labeling. (b) Tandem orthogonal proteolysis (TOP). The proteome is treated with an alkyne probe that becomes attached to a biotin azide tag containing a TEV protease cleavage site. The biotin-tagged proteins are enriched with avidin beads and subsequently digested by trypsin. Avidin-bound active site peptides are released by the TEV protease to reveal the site of labeling by LC–MS/MS analysis. In addition, the soluble peptides are also analyzed by LC–MS/MS to identify the corresponding proteins. (c) Multidimensional peptide identification (MudPIT). Biotin-labeled proteins are enriched with avidin beads and after subsequent tryptic digest analyzed by LC–MS/MS.

PS–alkyne probe, followed by a CC reaction to introduce a biotin tag with a tobacco etch virus (TEV) protease cleavage site. Tagged proteins were then enriched with streptavidin beads and subsequently digested by trypsin. The supernatant contained peptide fragments of the enriched proteins and was, in contrast to the previous procedure, saved for subsequent analysis by multidimensional protein identification technology (MudPIT), a novel two-dimensional LC–MS/MS analysis method that will be discussed in more detail below. In the last step, the probe-labeled active site peptides were eluted from the beads by cleavage with TEV protease and subsequently also applied to MudPIT (**Figure 6(b)**). The two separate LC–MS/MS runs of the tryptic digest and the TEV elution have the advantage of displaying two independent data sets, which must both yield the same results to validate protein targets. For example, protein hits found only in the TEV elution but not in the tryptic digest and vice versa can be regarded as false assignment and background, respectively. This novel double control strategy ensures a high fidelity of data derived from LC–MS-based high content chemical proteomics. Indeed, a large-scale TOP–ABPP analysis with the PS–alkyne probe of mouse heart proteome resulted in the identification of 32 specifically labeled proteins. These included a family of acyl-CoA dehydrogenases, modified on their Glu catalytic bases, thiolase and ALDH6, modified on Cys nucleophiles, and a set of proteins in which the modified residues had no purely catalytic functions.[29] In addition, TOP–ABPP has proven useful for the assignment of nucleophilic residues in hypothetical proteins with yet unknown function. Downstream studies of these proteins could potentially reveal their catalytic function within the proteome. Finally, it is also important to note that the tryptic digest alone already contains all the data that provides a means to distinguish protein isoforms and to comparatively quantify proteins (e.g., by spectral counting).

This premise of comparing relative levels of enzyme activities by multidimensional LC–MS/MS was accomplished recently by Jessani *et al.*[30] A prerequisite for the success of this methodology was that the abundance of probe-enriched enzymes could be directly correlated to MS spectral counting as was shown recently.[31,32] The ABPP–MudPIT approach involves treatment of proteomes with a biotinylated activity-based probe, enrichment of probe-labeled proteins with avidin-conjugated beads, on bead trypsin digestion, and finally multidimensional LC–MS/MS analysis of the resulting tryptic peptide mixture (**Figure 6(c)**). The two-dimensional LC column, consisting of an ion exchange material followed by reversed-phase material, provides an optimal resolution of peptides based on their charge and hydrophobicity, which improves spectral quality of the tandem MS analysis.

All ABPP LC–MS-based methods discussed in this chapter share some common features that differ from gel-based approaches. For instance, the multidimensional LC–MS/MS methodology exhibits remarkable sensitivity (1 pmol mg^{-1}) and resolution especially suitable for the identification of low-abundance proteins in complex samples. In contrast to gel-based approaches, ABPP LC–MS facilitates protein separation and identification in one step and additionally delivers comprehensive information about site of labeling and relative levels of activity. This technology, however, is time consuming (many hours per sample), difficult to perform in parallel, and requires large quantities of proteome (~1 mg or more). Therefore, a combination between a rapid and low sample-consuming separation step for clustering complex proteomic information into phenotypically relevant groups and a subsequent deeper analysis of these subgroups by multidimensional LC–MS/MS seems to be a promising compromise.

9.17.2.3.3 Capillary electrophoresis

Capillary electrophoresis (CE) represents a rapid, robust, sensitive, and high-resolution separation strategy for small peptide fragments and was recently introduced as a new analysis platform for ABPP by Okerberg *et al.*[33] In order to be compatible with standard ABPP labeling procedures, several methodological adjustments prior to CE had to be established. A postlabeling sample preparation procedure was developed that included disulfide reduction, thiol alkylation, and gel filtration steps to prevent thiol oxidation and to remove excess probe, respectively. Subsequently, proteins were digested and fluorescently labeled peptide fragments applied to a CE with a laser-induced fluorescence detection system (CE–LIF) (**Figure 7(a)**). The versatility of this approach was first shown by a reproducible and uniform detection of eight fluorescently labeled serine hydrolases added to background proteomes. Obtained sensitivity limits of 0.1 pmol mg^{-1} combined with the low sample consumption of only 20 nl allowed the analysis of samples that were previously intractable through other known methods. Another important feature of this new methodology is the improved resolution compared to gel electrophoresis. Profiling of several mouse proteomes revealed distinct serine hydrolase and cathepsin activities of similar molecular weight, which all appeared in one indistinguishable band on the SDS gel. One disadvantage of CE–ABPP is that, similar to IGFS, the identities of probe-labeled proteins is not immediately apparent. However, CE–ABPP experiments can be run in parallel with the LC–MS/MS technology in order to assign enzyme identities for each obtained peak. Since LC–MS/MS still requires high amounts of material and also limits the sample throughput, an ultimate ABPP analysis platform is desired, in which detection and identification occur in one simple, sensitive, high-resolution step. One possible solution could be antibody microarrays (discussed below).

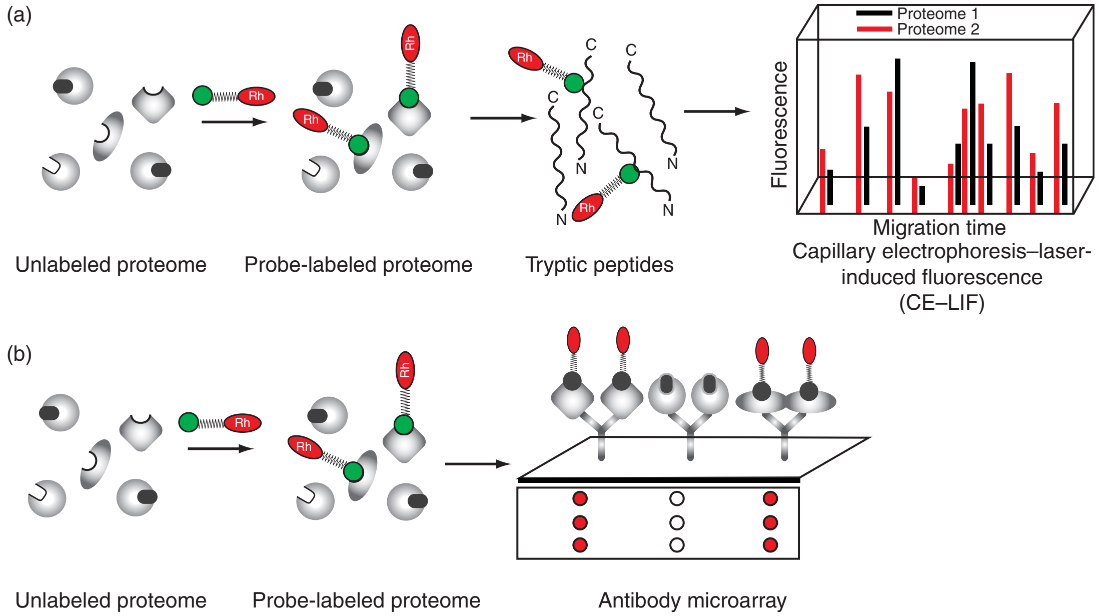


Figure 7 (a) Capillary electrophoresis (CE) for ABPP. The probe-labeled proteome is trypsinized and peptide fragments are analyzed via CE. (b) General strategy for antibody-based ABPP microarrays. Proteomes are labeled in solution with fluorescent activity-based probes and captured on glass slides arrayed with enzyme-specific antibodies. Reproduced by permission of The Royal Society of Chemistry.

9.17.2.3.4 Antibody microarrays

A microarray platform specifically customized for the needs of ABPP with orthogonal strategies for the labeling and capture of enzyme activities was introduced by Sieber *et al.*[34] In this method, proteomes were first treated with fluorescent activity-based probes, and then captured and visualized by fluorescence scanning on glass slides displaying antienzyme antibodies (**Figure 7(b)**). With this methodology, several enzyme activities could be profiled in parallel on individual slides by arraying a complementary set of antibodies. Since antibodies are spatially separated on the glass slide and specific for only one enzyme species, the defined locus of fluorescence on the slide reveals the enzyme identity avoiding the limitations of gel electrophoresis (resolution) and MS (sample consumption) procedures. The general value of this methodology was shown on the example of four cancer-associated enzymes: the three serine hydrolases – tissue plasminogen activator (tPA), urokinase (uPA), and prostate-specific antigen (PSA) – as well as the matrix metalloprotease 9 (MMP9). Detection of these enzymes in the background of breast cancer proteomes revealed very distinct fluorescent signals at the locations of the corresponding antibodies. Notably, the PSA detection limit for ABPP microarrays was only 2–8 ng ml^{-1}, which is in the range of endogenous serum level for this protease. In contrast, gel-based methods display a 50–100-fold reduced sensitivity limit for PSA.

Hermetter and coworkers used a modified ABPP micorarray platform in which target lipases and esterases were directly spotted onto glass slides and subsequently labeled with fluorescent phosphonate ester probes.[35] In addition, biotinylated probes were immobilized on streptavidin slides to capture enzymes from the solution.

Collectively, ABPP microarrays enable the parallel analysis of many enzyme activities in whole proteomes with a sensitivity (1 pmol mg^{-1}) and resolving power that greatly exceeds gel-based methods. Additionally, ABPP microarrays consume much less proteomic material than gel- or LC–MS-based methods (~1 vs. 15–1000 μg per experiment, respectively). Since ABPP microarrays consolidate the detection and identification of probe-labeled enzymes into a single assay, they greatly increase the throughput of proteomic experiments. This technology also eliminates the need for random protein labeling and/or secondary antibodies in conventional proteomic antibody microarrays. However, due to the limitation of available high-quality antibodies, this methodology can so far only be applied for a small subset of the total proteome and stresses the need for additional ABPP microarray compatible antibody production. Moreover, as conventional proteomic antibody microarrays, this method cannot be utilized for the identification of uncharacterized enzyme species.

9.17.2.4 Cleavable Linker

One strength of ABPP is its ability to enrich single proteins within complex proteomes thereby simplifying the subsequent analysis procedures, which are no longer limited by the abundance of proteins. Biotin is commonly used as a high-affinity tag binding to immobilized avidin or strepavidin thus allowing the enrichment of any targeted protein. However, one disadvantage of this technology is the requirement of harsh denaturing conditions to disrupt the strong biotin–avidin interaction. Moreover, contamination of avidin by unspecific protein binding and by endogenously biotinylated proteins can limit the specific detection of low-abundance proteins leading to time-consuming hit validation procedures. To address some of these limitations, cleavable linkers based on acid,[36] proteolytic,[29] or oxidative cleavage[37] have been introduced. One novel approach, which was specifically designed for the needs of ABPP, has been recently established by Bogyo and coworkers.[38] This method is based on a diazobenzene linker that can be cleaved to the corresponding anilines using sodium dithionite, a mild reducing agent (**Figure 8**). The utility of this approach was demonstrated for the MS analysis of cysteine and serine proteases, which revealed significant improvements in the quality of data with the chemoselective cleavage system compared to the previous heat denaturation procedures.[39]

9.17.2.5 Mechanism- and Affinity-Based Probes

One important aspect of all ABPP probes is the chemical design of the inhibitor moiety, which defines the specificity and sensitivity for the targeted enzyme class or classes. Two different types of binding groups for ABPP have been established. The first type uses electrophilic entities that exploit the enzyme mechanism resulting in a covalent modification of key catalytic residues in the active site. These residues are usually nucleophilic and form a covalent intermediate with their endogenous substrates in the corresponding enzymatic reaction. Prominent nucleophiles, which are targeted by the majority of mechanism-based probes, are, for example, thiols (Cys) and alcohols (Ser, Thr) as part of a charge relay system (usually catalytic dyad or triad) that activates these residues by partial abstraction of their hydrogen atom. In the case of covalent mechanism-based probes, the enzyme is directly labeled and this linkage remains stable during all subsequent denaturing procedures such as SDS gel analysis (see Section 9.17.3). However, many other enzymes of biological relevance do not use protein-bound nucleophiles for catalysis, making the design of ABPP reagents for these enzyme classes particularly challenging. These noncovalent substrate-binding enzymes are targeted by the second type of ABPP probes that are affinity based. The incorporation of broad-spectrum, high-affinity binding groups can direct probes to enzymes that exhibit the matching active site structural features for a tight binding. Since these noncovalent complexes would fall apart during harsh and denaturing analytical procedures, a covalent link between the high-affinity probe and the enzyme has to be established. In ABPP, this has been achieved by the use of diverse photocross-linkers (see Section 9.17.3), such as benzophenone and diaziridin.[40] Benzophenone-based photoaffinity scanning has emerged as a preferred method for mapping the bimolecular interface of ligand–receptor complexes.[41] On exposure to long-wavelength ultraviolet (UV) light, the benzophenone carbonyl group forms a triplet biradical that can abstract hydrogen from the protein to which the molecule is bound, ultimately causing the two molecules to be covalently linked. The exited triplet state is reversible, which leads to several advantageous properties of the benzophenone moiety compared to arylazide and other photoreactive groups.[42] These properties include low reactivity with water and favorable photokinetics with high cross-linking specificity and efficiency.

9.17.3 Chemical Probes for ABPP

One challenge in the design of activity-based probes is the selection of the appropriate chemical molecules that provide the necessary interaction with the dedicated protein target(s). Ideally, the probe should specifically bind to the active site of a desired protein and be inert toward other reactive species within a cell.[14] In this chapter, we summarize the chemical design of established probes, which is the fundamental basis to understand their function and application in biology. In many cases, the inhibitor moiety of probes is inspired by natural products and these examples will be discussed here too.

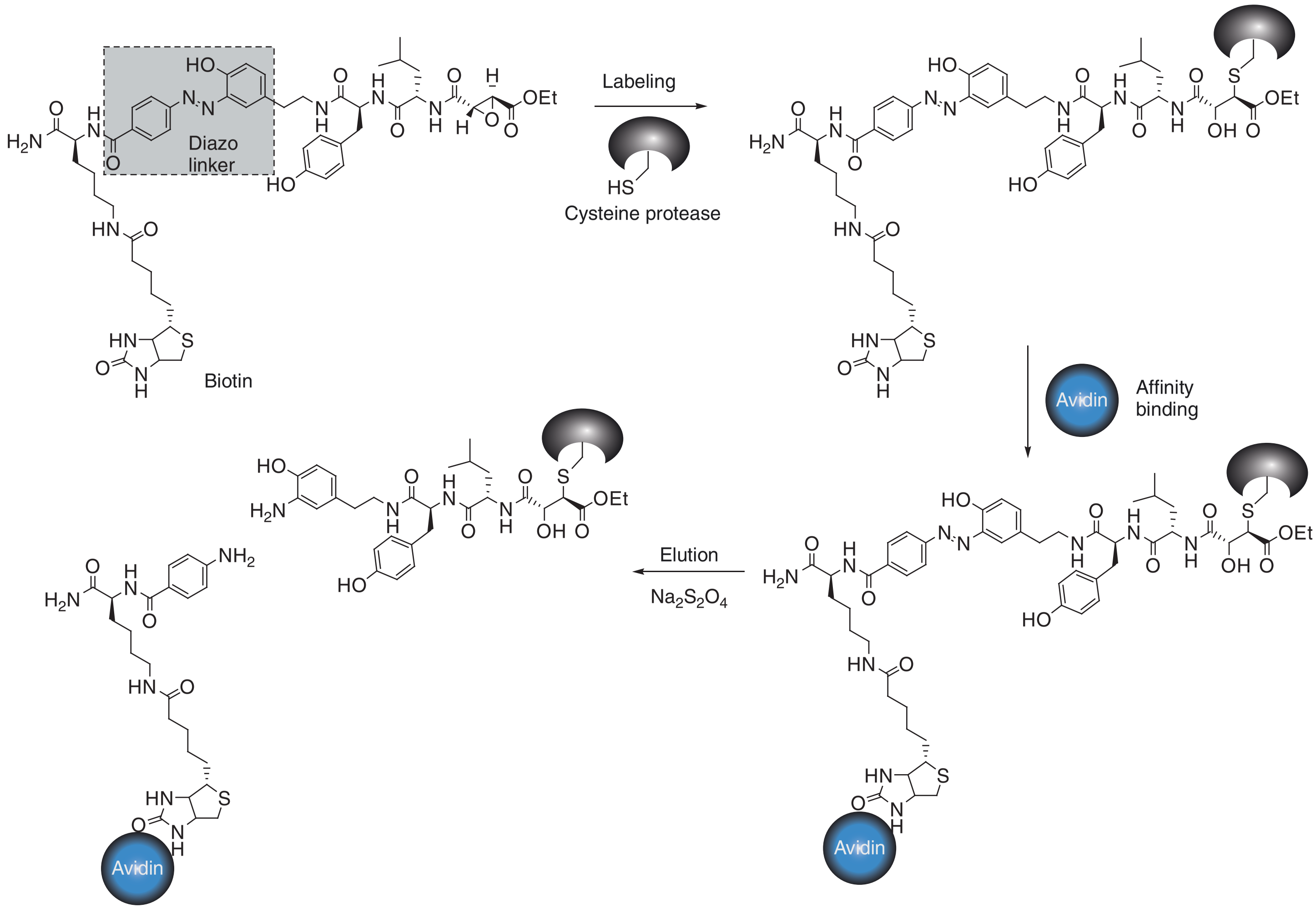


Figure 8 Cleavable tags for ABPP via diazobenzene linkers. Proteomes are labeled, enriched, and cleaved with $Na_2S_2O_4$.

9.17.3.1 Enzyme Classes Addressable by Directed Electrophilic Probes

9.17.3.1.1 Serine hydrolases

The serine hydrolase enzyme family is one of the largest and most diverse enzyme classes including proteases, peptidases, lipases, esterases, and amidases, which account for approximately 1% of the predicted protein products encoded by many eukaryotic genomes.[21] These enzymes play important roles in numerous physiological and pathological processes such as blood clotting, cancer, inflammation, and diabetes.[21,43,44] The activity of these enzymes is tightly regulated by the cell in order to control their proteolytic power under physiological conditions. However, under many pathological conditions, this regulation gets out of balance and it is assumed that the corresponding unregulated proteolytic activity contributes in disease progression. Serine hydrolases degrade their substrates by the hydrolytic cleavage of ester/thioester/amide bonds. This is achieved by the activation of a conserved serine nucleophile that attacks the corresponding electrophilic substrate bond to form a covalent acyl–enzyme intermediate (**Figure 9(a)**).[45] The covalent adduct is subsequently liberated by water-catalyzed hydrolysis regenerating the enzyme for a new round of catalysis. The greatly enhanced nucleophilicity of the catalytic serine residue distinguishes it from all other noncatalytic serine residues and makes it a desired residue for modification by electrophiles, including FPs, aryl phosphonates, sulfonyl fluorides, and carbamates.[21] FPs have been well-known inhibitors for serine hydrolases for 80 years and have been first applied in chemical warfare as potent acetylcholine esterase inhibitors (sarin, soman). These compounds do not resemble a peptide substrate and hence are nonselective toward a particular serine hydrolase that is desired for enzyme activity profiling of the whole enzyme class. In addition, FPs exhibit minimal cross-reactivity with other classes of mechanistically distinct hydrolases such as cysteine-, metallo-, and aspartylhydrolases.[46] FPs react only with the active species, not with the inactive zymogen, which allows the desired discrimination for comparative proteomic studies. All these beneficial features of FPs as selective serine hydrolase inhibitors have stimulated the design of corresponding activity-based probes. Cravatt and coworkers prepared a panel of FP probes with various linkers and tags such as fluorophores or biotin. These probes enable the visualization, identification, and functional characterization of catalytically active serine hydrolases in complex proteomes consisting of hundred thousands of different proteins (**Figures 3 and 9(b)**).[18–20] In various biological experiments (Section 9.17.4), these reporter-tagged FPs showed indeed remarkably broad reactivity with enzymes of the serine hydrolase family.[30,47,48] Since the beginning of FP probe–proteome screening, more than 80 distinct serine hydrolases have been identified with a large number of uncharacterized enzymes regarding their endogenous substrates and function. In certain instances, the assignment of metabolic and cellular functions to previously uncharacterized members of the serine hydrolase family by ABPP was possible (see Section 9.17.4).[49] In general, FP probes are suitable for a class-wide profiling of the serine hydrolase family. However, there are few examples of serine proteases that display restricted substrate selectivities that reduced the success of FP probe labeling. In these cases, peptidic arylphosphonate probes have been used to label individually the subclass of trypsin-like serine proteases and granzymes A and B with high specificity (**Figure 9(b)**).[50,51] Together, this set of available phosphonate tools should allow to label the majority of active serine hydrolases.

9.17.3.1.2 Threonine proteases

Similar to serine hydrolases, threonine proteases utilize a catalytic charge relay system to activate their secondary hydroxyl nucleophile for catalysis.[21] These residues are important active sites in multiple catalytic subunits of the proteasome. Nazif and Bogyo[52] have designed ABPP probes with a vinyl sulfone moiety as a reactive group for covalent modification of the active site threonine. In order to generate specificity for the proteasome subunits, they used a peptide library and gained insight into the substrate recognition properties of specific proteasomal subunits. This methodology was refined by Ovaa *et al.*[53] who used cell-permeable vinyl sulfone probes with an azide tag that could be modified after cell lysis with the corresponding phosphine reporter tag via the Staudinger ligation (**Figure 10**). In a follow-up study, Berkers *et al.*[54] investigated the specificity and mode of action of the proteasome inhibitor bortezomib *in vitro* and *in vivo.* They synthesized a bortezomib analogue of the previously established vinyl sulfone probe with a dansyl-sulfonamidohexanoyl

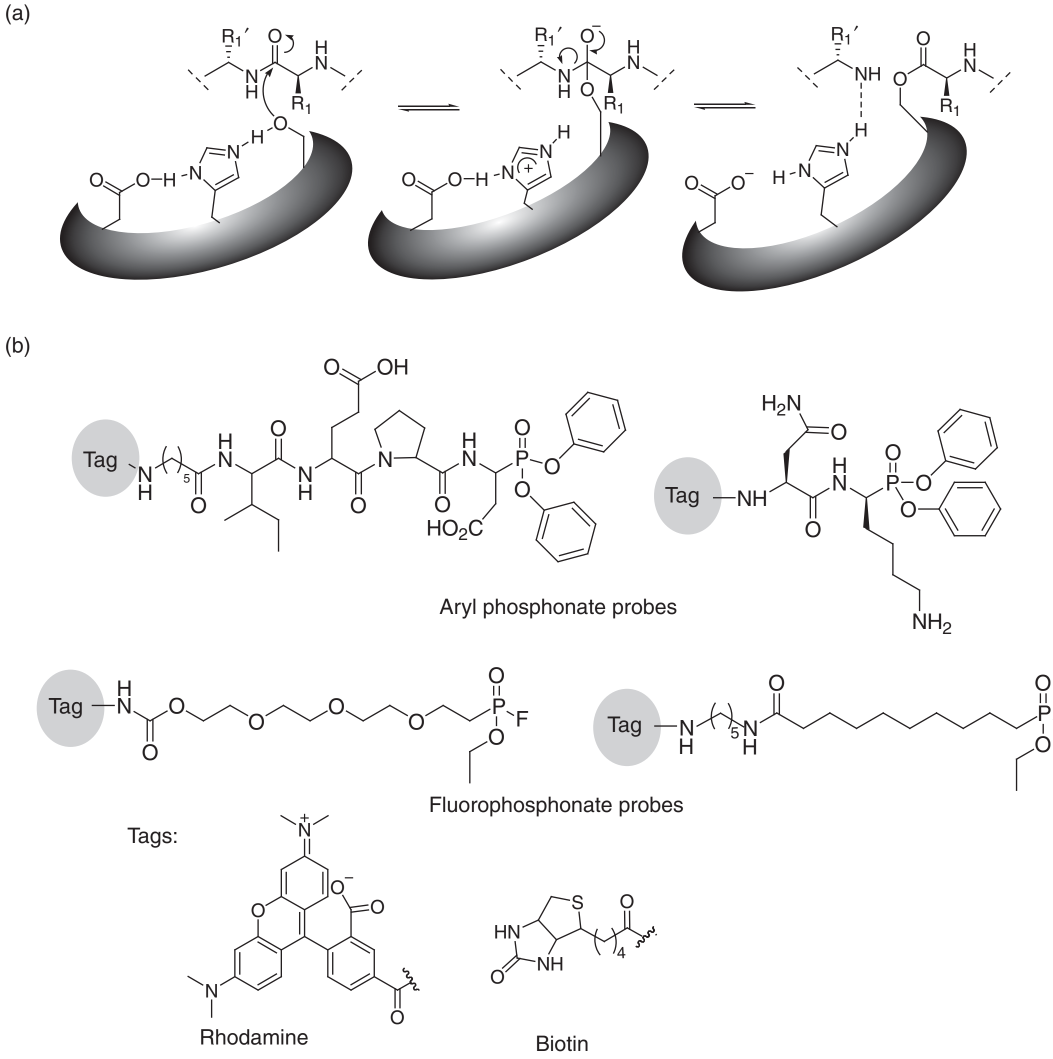


Figure 9 Serine hydrolases as targets of ABPP. (a) Catalytic mechanism of peptide substrate cleavage via the action of a catalytic triad. (b) Aryl phosphonate and fluorophosphonate probe scaffolds.

fluorescent tag as a fast and easy detection strategy that allowed to monitor and visualize the proteasome activity in living cells by fluorescence microscopy. This probe exhibited distinct labeling of individual proteasome subunits and showed substantial differences between activity pattern in living cells and cultured cell extracts. This shows that the use of fixed *in vitro* assay conditions does not necessarily reflect the conditions found *in vivo*, emphasizing the value of *in vivo* ABPP studies.

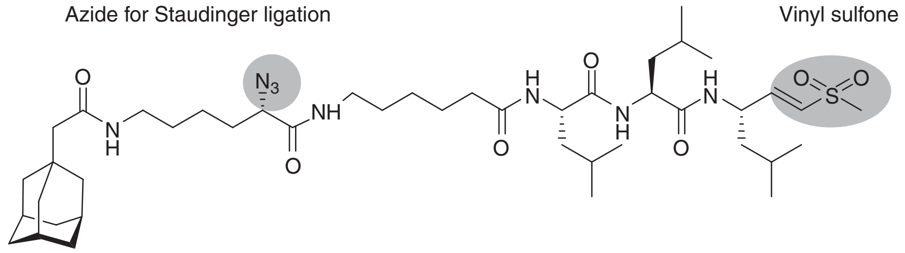


Figure 10 Cell-permeable peptidomimetic vinyl sulfone probe for the labeling of proteasomal subunits.

9.17.3.1.3 Cysteine proteases

Similar to serine hydrolases, cysteine proteases represent a large class of enzymes, which are involved in numerous physiological and pathological cellular events including bone remodeling,[55] cancer invasion,[56,57] and malaria.[58] Their posttranslational activation results in expression profiles that do not directly correlate with activity. This designates them as dedicated targets for ABPP analysis to explore their regulation and function. Cysteine proteases employ a nucleophilic cysteine residue (activated by an adjacent histidine as a general base for proton transfer) in their active site, which makes them susceptible to inactivation by different electrophilic chemotypes.[21] In order to develop class-specific ABPP probes for cysteine proteases, diverse reactive groups have been incorporated including epoxides,[59,60] vinyl sulfones,[61] diazomethyl ketones,[62] α-halo ketones,[63] and acyloxymethyl ketones (AOMKs),[64] which display a high reactivity toward the nucleophilic cysteine residue.[12] The cysteine protease class can be subdivided into six different clans based on the structure of their active sites.[21] Extensive studies have been carried out with the papain family, which includes cathepsins that belong to the CA clan of cysteine proteases. Cathepsins are efficiently inactivated by the natural product E-64 that contains an electrophilic epoxide group that reacts covalently with the active site cysteine nucleophile.[60] The first activity-based probe for cathepsins was introduced by Bogyo and coworkers by the attachment of diverse reporter tags, such as radioactive ^{125}I, biotin, and fluorescent BODIPY on the E-64 scaffold (**Figures 11(a) and 11(b)**).[65,66] These probes exhibited a broad applicability for the functional analysis of cathepsins in many biological samples. E-64 does not inhibit serine proteases, aspartic proteases, or metalloproteases. However, not all cysteine proteases such as legumain and caspases, which belong to the CD clan, are inhibited by E-64.[21]

In order to expand the profiling repertoire of ABPP for these cysteine proteases, additional reactive groups have been explored. Caspases, which play crucial roles in apoptotic-mediated cell death, have been successfully targeted by a number of ABPP probes such as peptidic α-halo[67] and AOMKs.[68,69] Selectivity for caspases was achieved by the incorporation of a negatively charged substituent in the P1 position. The AOMK-derived reactive groups have been shown to display exceptional selectivity for cysteine proteases including members from the CD as well as CA clan.[21] The broad applicability of these probes for versatile cysteine proteases was recently demonstrated by Bogyo and coworkers, who generated structurally diverse libraries of AOMK probes (**Figure 11(c)**).[68] While in some cases broad target coverage of cysteine proteases is desired, other probes were designed to exhibit a focused target affinity. For example, Bogyo and coworkers introduced dipeptide vinyl sulfone probes in order to achieve selective labeling of cathepsin C in compex proteomic mixtures.[70]

Another subclass of cysteine proteases, termed ubiquitin (Ub)-specific proteases or deubiquitinating enzymes, is involved in the cleavage of Ub and ubiquitin-like (Ubl) modifications acting as an additional level of control over the Ub–proteasome system.[71] Borodovsky *et al.*[72,73] have made use of the ABPP technology to design deubiquitinating enzyme-specific probes to identify new members of this enzyme class and explore their complex regulation and function. These protein probes contained Ub and Ubl sequences in order to gain specificity for the dedicated enzyme class. Since an irreversible bond is needed for analytical procedures, such as MS and gel analysis, the researchers introduced various electrophilic Michael acceptors and alkyl halides to the Ub/Ubl recognition sequence for covalent modification of the active site cysteine (**Figure 12**). Visualization of probe binding was achieved by the incorporation of a hemagglutinin tag that could be detected and isolated with specific antibodies. These probes indeed targeted several distinct sets of deubiquitinating enzymes including novel, redundant classes.

9.17.3.1.4 Metallohydrolases

Another family of enzymes that have been functionally linked to the degradation of connective tissue and dissolution of epithelial and endothelial basement membrane are the MMPs.[74,75] These enzymes are multidomain, zinc-containing, neutral endopeptidases including collagenases, stromelysin, gelatinases, and membrane-type metalloproteases. Each MMP exhibits a preferred substrate specificity toward individual matrix proteins, but there is overlapping specificity within the whole family.[4] A striking feature of these

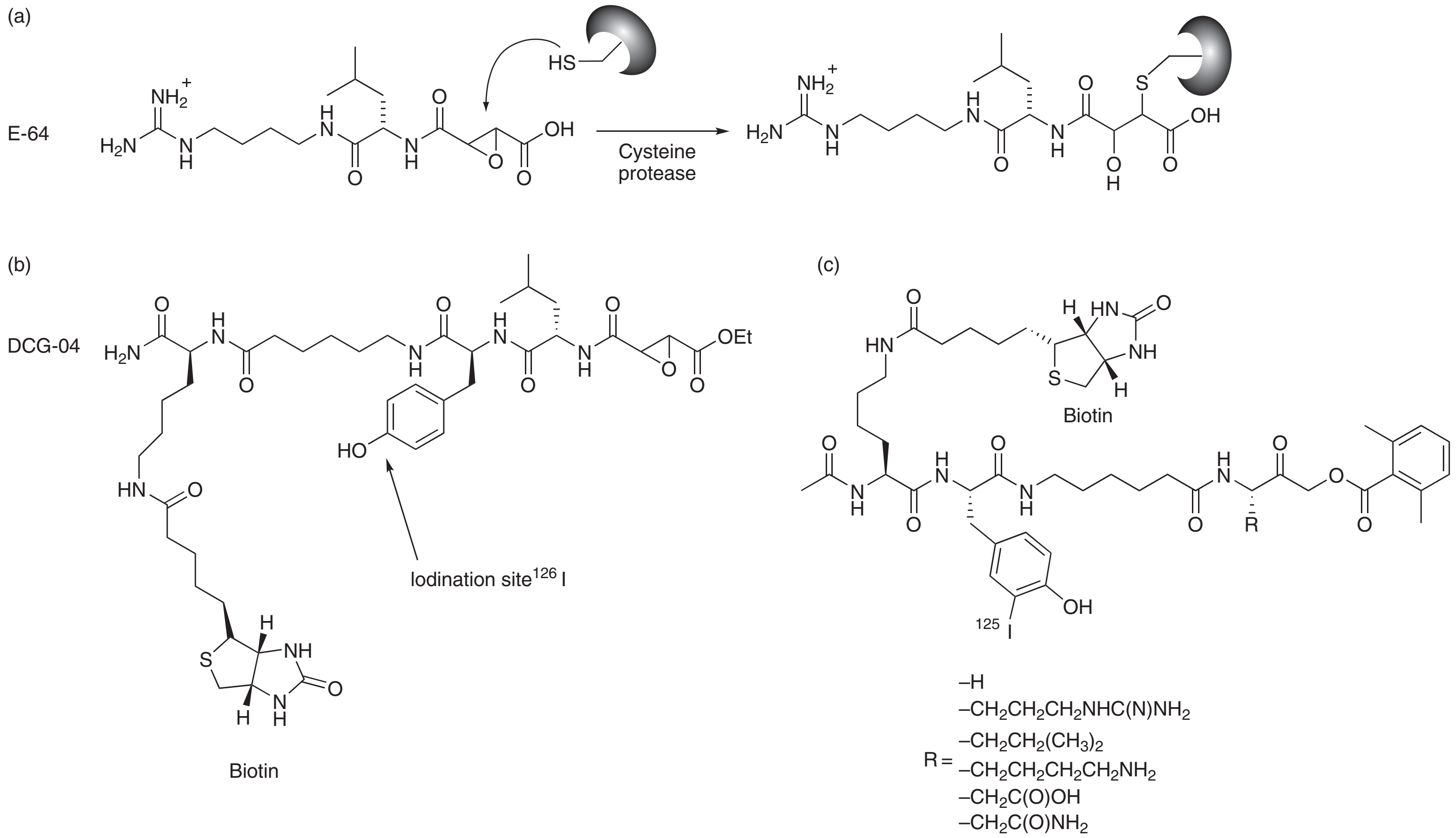


Figure 11 Mechanism-based probes for cysteine protease active site labeling. (a) E-64-mediated inhibition of cysteine proteases via nucleophilic attack of the catalytic cysteine at the epoxy group. (b) E-64-based probe with biotin reporter tag and tyrosine as ^{125}I iodination site for radioactive labeling. (c) Acyloxymethylketone probe for the complementary labeling of cysteine proteases.

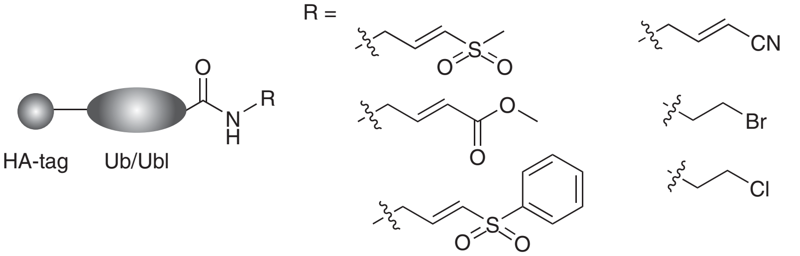


Figure 12 Design of a probe library for the labeling of deubiquitinating enzymes. A variety of reactive groups (R) are appended to the ubiquitin recognition site (Ub/Ubl). A hemaglutinin tag (HA-tag) is used for antibody-based detection of target enzymes.

MMP enzymes is their tight regulation of catalytic activity. MMPs are expressed as inactive zymogens and require an activation process to convert them into active enzymes.[76] These active enzyme species are again tightly regulated by a family of proteins, the TIMPs that are widespread in tissues and block MMP activity effectively.[77] In contrast to serine and cysteine hydrolases, metalloproteases do not use a covalent protein-bound intermediate for catalysis making the design of ABPP reagents for this enzyme class particularly challenging. These enzymes incorporate a zinc ion in their active site, which coordinates and activates a water molecule to carry out a nucleophilic attack on the corresponding substrate. Since nearly all metalloproteases are susceptible to reversible inhibition by hydroxamate-based compounds, which complex the active site zinc atom in a bidentate manner, this functional group was incorporated into the structure of the first generation of a metalloprotease-directed probe by Saghatelian *et al.*[76] (**Figure 13(a)**). To promote covalent modification of metalloprotease active sites, the probe also contained a photoreactive (benzophenone) group that forms a covalent bond with the enzyme active site upon UV irradiation. The rhodamine-tagged hydroxamate benzophenone probe (HxBP-Rh) was found to selectively label active, but not zymogen- or inhibitor-bound forms of metalloproteases. A similar approach was described by Yao and coworkers, who prepared libraries of photoreactive peptidyl hydroxamate probes and showed labeling of a panel of yeast metalloproteases[78] and MMPs.[79] More recently, optimized probes for the labeling of diverse metalloproteases enzyme families were introduced.[80] In this approach, an alkyne group was attached to the probe scaffold, which facilitated the versatile coupling of reporter groups (rhodamine or biotin) via CC either before or after proteome labeling (**Figure 13(b)**). Comprehensive profiling of several tissue and cancer cell proteomes revealed many metalloprotease activities including members from all of the major subfamilies of this enzyme class. These results highlighted the necessity of using a diverse set of side chains to saturate the structurally different active site pockets of this diverse enzyme family.

Acetylation and deacetylation of lysine residues in histones represent a central mechanism of transcriptional activation and repression.[81] The removal of acetyl residues from lysine is catalyzed by histone deacetylases (HDACs) that fall into three different classes of enzymes. Class I and II are zinc-dependent metallohydrolases, whereas class III are NAD^+-dependent deacetylases. Inhibitors of class I and II have been described in the literature with potent effects against cancer.[82,83] One of these inhibitors is suberoylanilide hydroxamic acid (SAHA) that was employed by Salisbury and Cravatt[84] as the core scaffold for specific HDAC ABPP probes (**Figure 13(c)**). The final SAHA–BPyne probe was elegantly incorporated into the existing SAHA structure by an extension of the SAHA benzyl group to an alkyne-functionalized benzophenone. This moderate variation of the inhibitor scaffold minimized unfavored target interactions with the large benzophenone group. The HDAC probe was found to target indeed multiple class I and II enzymes in proteomes. Remarkably, these enzymes were also labeled *in vivo*, which represents the first example of an affinity probe that is activated by light inside a living cell. Interestingly, in the course of studies several HDAC-associated enzymes were identified, indicating that these proteins were located in close proximity to the HDAC active sites and within reach of the photocross-linking reaction. It was therefore speculated that they could be involved in substrate recognition and regulation of activity. More recently, Salisbury and Cravatt[85] tested several structurally different HDAC

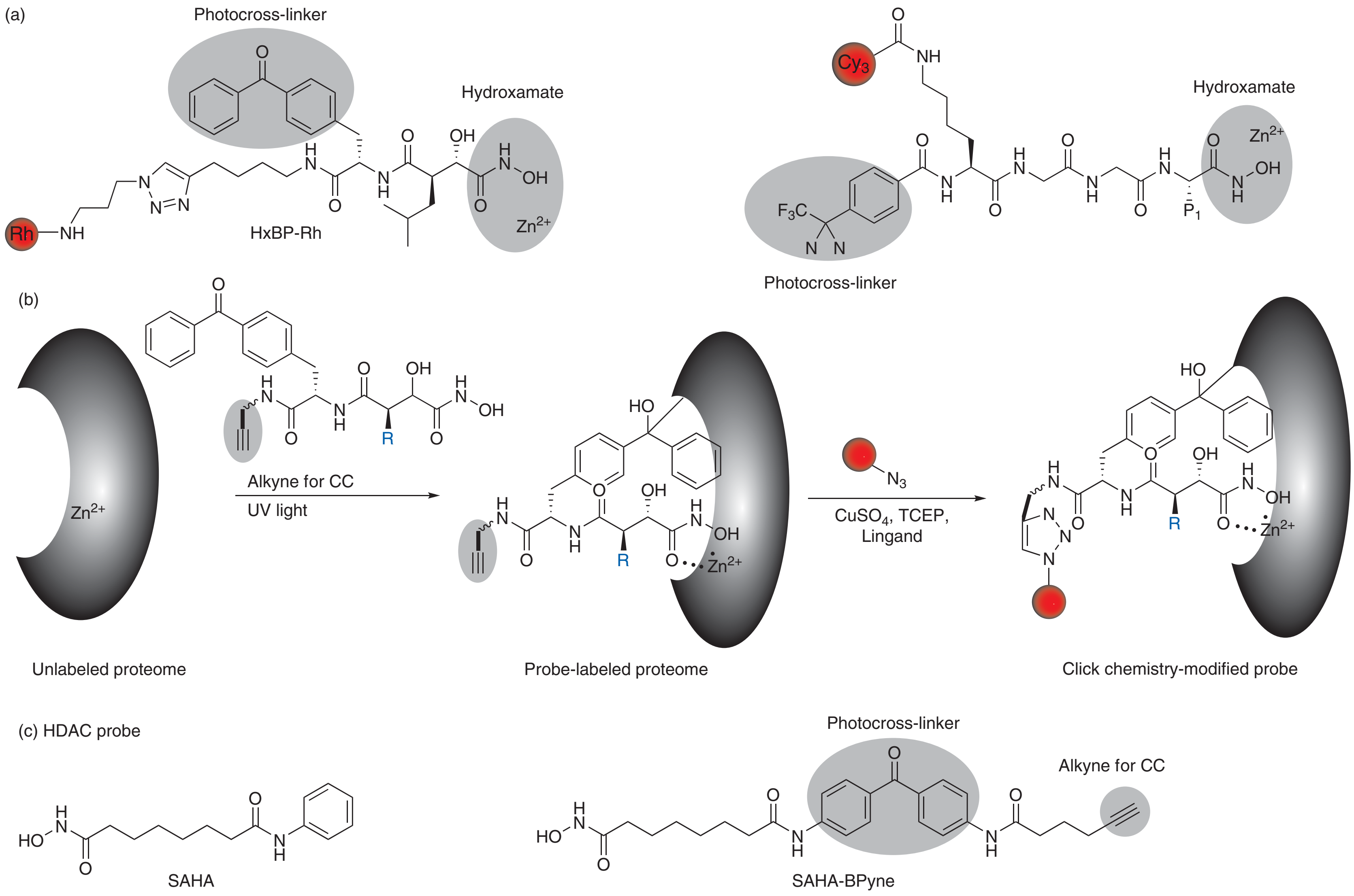


Figure 13 ABPP probes for metalloproteases. (a) First generation of hydroxamate-based probes with rhodamine (Rh) and Cy3 as fluorescent tags and photocross-linkers for covalent modification. (b) Active site-directed labeling mechanism for a clickable second-generation probe. (c) The potent SAHA inhibitor for HDAC classes I and II and its derived ABPP probe.

probes that were derived from various potent inhibitors. While some of these probes turned out to be good inhibitors of HDACs *in vitro*, the initial SAHA–BPyne probe remained the best for *in vivo* activity studies.

9.17.3.1.5 Aspartyl proteases

Compared to serine, cysteine, and metallohydrolases, aspartyl proteases represent a much smaller subset of proteolytic enzymes. However, some of these enzymes play crucial roles in diseases such as the integral membrane proteins β- and γ-secretase, which cleave the amyloid precursor protein to generate the Aβ amyloidogenic peptides, the major constituents of the amyloid plaques of Alzheimer's disease.[86] Similar to metallohydrolases, aspartyl proteases utilize an activated water molecule for catalysis (**Figure 14(a)**).[87] To label active aspartyl proteases involved in Alzheimer's disease, Li *et al.*[88] used a hydroxyethylene dipeptide isostere as a transition state analogue, which they linked to a benzophenone photocross-linker and a biotin tag (**Figure 14(b)**). This probe was successfully applied to label and identify presenilin 1 and 2 as γ-secretases, while the zymogen form of presenilin 1 was not labeled confirming the high fidelity of ABPP probes for the correct readout of the corresponding enzyme activity state. Interestingly, although presenilin 1 and 2 contain no recognizable aspartyl protease motifs, the ABPP methodology was successful in their class assignment.

9.17.3.1.6 Protein kinases

Protein kinases catalyze the transfer of the γ-phosphate moiety of ATP to the corresponding protein substrate. More than 30% of all eukaryotic proteins are phosphorylated, with the majority of the modifications occurring on serine or threonine residues.[89] Kinases constitute the largest enzyme class in eukaryotic proteomes with more than 500 members encoded in the human genome[90] and play a central role in most signal transduction pathways

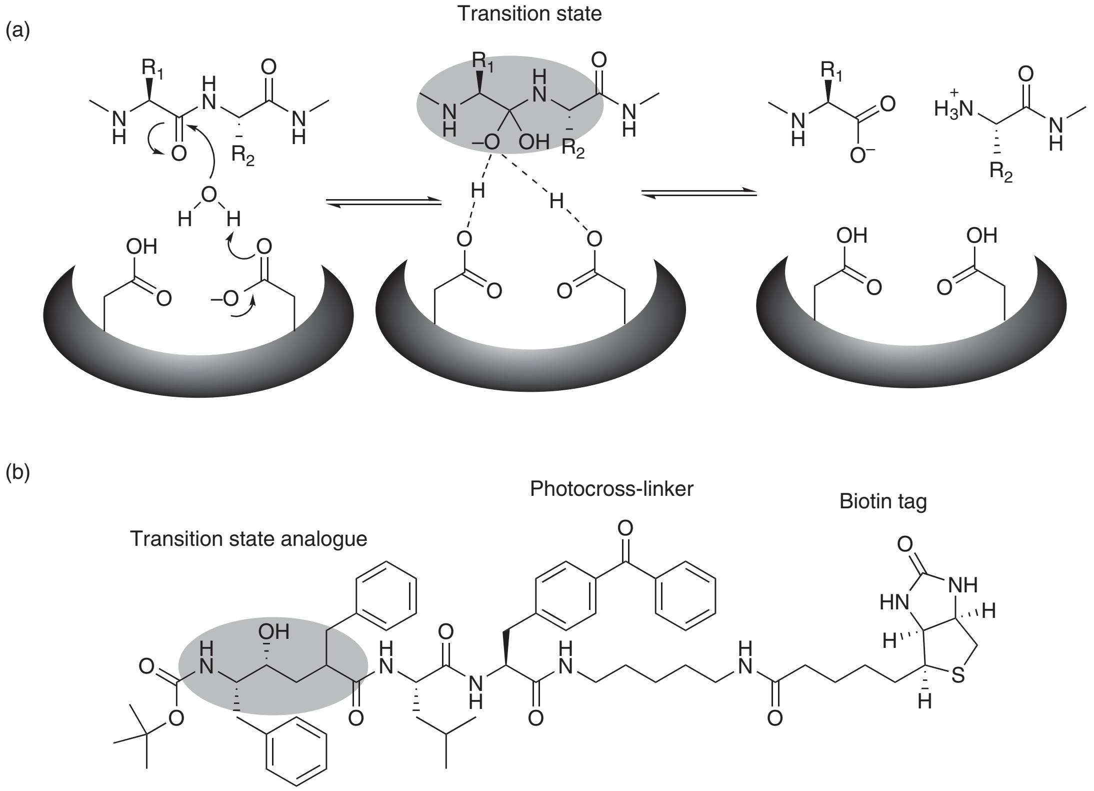


Figure 14 Profiling of aspartyl protease activity. (a) Catalytic mechanism of aspartyl proteases that cleave peptide bonds via activation of a water molecule. (b) A hydroxyethylene dipeptide isostere-based probe binds tightly to the active site and gets covalently attached via UV irradiation.

with major implications for the regulation of cell physiology and pathology. Aberrantly regulated kinases have been shown to play causative roles in several diseases such as cancer and diabetes and serve as promising drug targets for therapeutic intervention.[91–93] In order to develop kinase-specific drugs, a detailed understanding of protein phosphorylation (and dephosphorylation, Section 9.17.3.1.7) is a central task of modern medicinal chemistry efforts. However, functional analysis of kinases by conventional genetic or pharmacological approaches is difficult.[89] Gene knockouts are often susceptible to compensatory changes and the development of specific inhibitors has remained, with a few exceptions, a problematic task.[94,95] This is attributed to the high sequence homology of the active site region of kinases that all bind a common substrate, ATP. To monitor activity and function of kinases, several approaches for specific ABPP probes have been applied. One challenge in probe design is the selection of a specific inhibitory unit for the kinase active site. This active site lacks a common highly nucleophilic residue, essential for catalysis as observed in serine hydrolases and cysteine proteases. However, conserved active site lysines, which are observed in the majority of kinases, represent a potential anchor for ABPP probes.[90] These lysines are in close proximity to the γ- and β-phosphates of the bound ATP and counterbalance their negative charge. In a very elegant approach, Patricelli *et al.*[96] developed kinase-specific probes that were based on the ATP or ADP scaffolds for broad enzyme coverage modified with an acyl phosphate group at the position of the ATP γ- and β-phosphate to covalently bind the conserved lysine residue in close proximity (**Figure 15(a)**). The acyl phosphate reactive group was chosen due to several advantages including its appropriate reactivity with amines and at the same time sufficient stability in aqueous solutions. It reacts with primary amines by the addition of the amine nitrogen to the carbonyl group and elimination of the corresponding adenosine phosphate (**Figure 15(b)**). The third essential element of this kinase probe is the biotin tag for enrichment of bound targets, which comprises a 6-C spacer to minimize negative steric effects. The acyl phosphate probes displayed indeed favorable selectivity and reacted covalently at the ATP sites of about 75% of human kinases in cell lysates. Subsequently, the labeled kinases were tryptically digested and biotinylated peptides enriched with streptavidin agarose beads to determine their protein sequence and the site of labeling by LC–MS (Section 9.17.2). This platform was potent enough to investigate the functional state of many protein kinases in complex proteomes and revealed target specificity and selectivity of known kinase inhibitors by competitive ABPP (Section 9.17.4).

In contrast to this polyspecific ABPP probe suitable for profiling a large fraction of the protein kinase family, several other approaches have been introduced, which focus on specific subfamilies. Wortmannin, for example, is a potent natural product specific for the inhibition of phosphoinositide 3- and polo-like kinases *in vitro* and *in vivo*. Its role in ABPP probe design will be discussed in the natural product-derived probe section (Section 9.17.3.3).[97,98]

Another potent approach for the targeting of a single kinase was introduced by Shokat and coworkers who developed a chemical proteomic strategy that utilized both inhibitor design and protein kinase engineering to generate a specific probe–protein interaction.[99] Kinases of interest were mutated at their gatekeeper position to incorporate a smaller glycine residue, which generates a larger binding pocket that can in turn accommodate inhibitors with bulky aryl groups. These bulky groups act as a selectivity element that discriminates between genetically engineered kinases and native kinases for which they are sterically occluded. As discussed before, selective labeling of enzymes by electrophilic probes requires a nucleophilic residue in the active site. Therefore, a second mutation was introduced into the kinase active site that provides a nucleophilic cysteine residue as an irreversible anchor point for the electrophilic Michael acceptor-based kinase inhibitor (**Figure 16**). Using this technology, a linear correspondence between inhibition of epidermal growth factor receptor (EGFR) kinase activity in intact cells and of its downstream effectors could be demonstrated.

In a reverse approach, Taunton and coworkers used structural bioinformatics to identify kinases that have a cysteine in their ATP-binding pocket and can accommodate a bulky fluoromethylketone inhibitor. Native ribosomal S6 kinase (RSK)1 and RSK2 kinases were identified as suitable targets in cell lysates of mammalian cells since they exhibit cysteine residues and large binding pockets that could accommodate the corresponding bulky inhibitor.[100] An alkyne-modified ‘clickable’ probe variant of this compound indeed revealed selective and saturable labeling of endogenous RSK1 and RSK2 kinases in intact cells and cell lysates.[101] The probe was used to monitor RSK modification and its effects on phosphorylation and downstream signaling.

Another approach to study kinase activity was introduced by Hagenstein *et al.*[102] who designed an affinity-based probe equipped with a benzophenone photocross-linker for covalent modification of plant kinases.

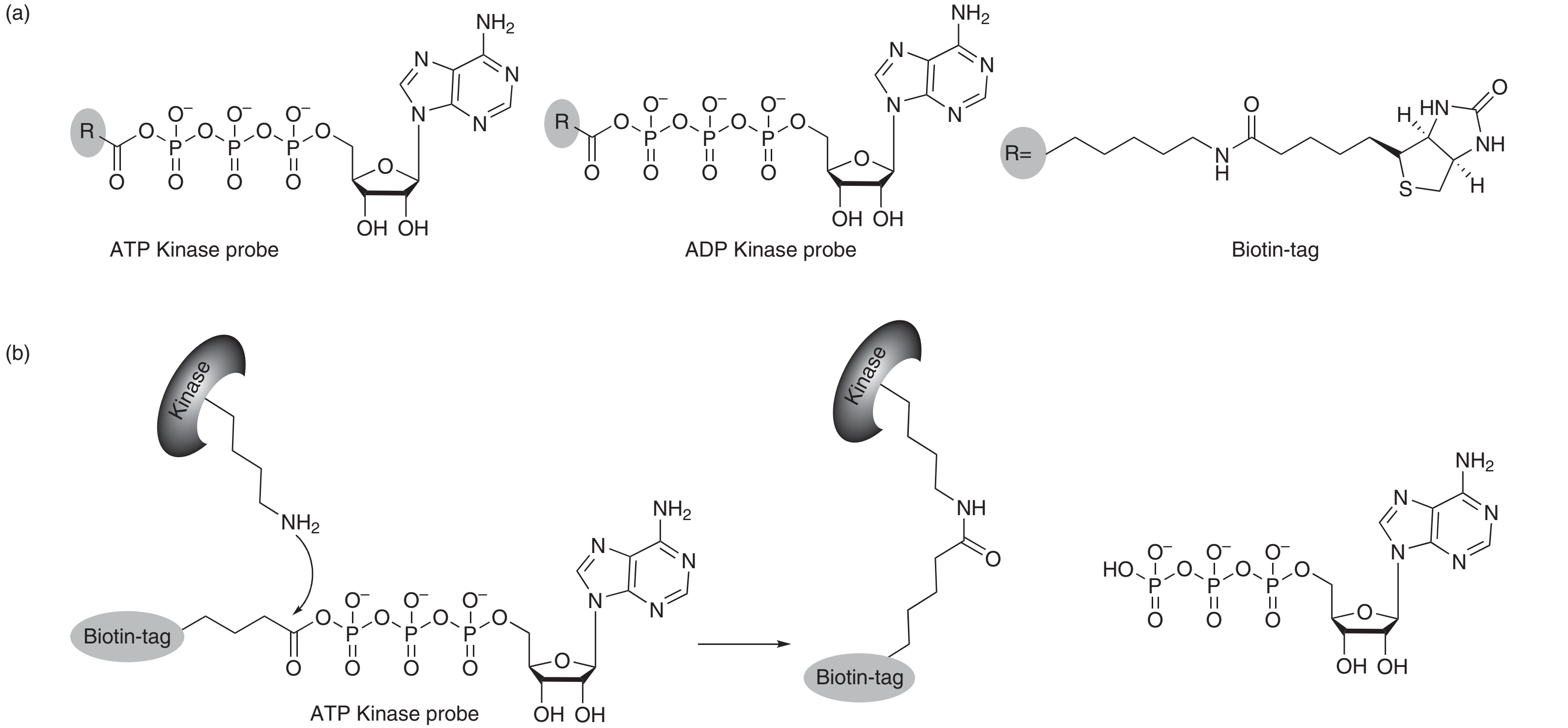


Figure 15 Acyl adenosine phosphate probes for the selective labeling of kinases at their ATP-binding site. (a) Probes were designed on the basis of ATP or ADP. (b) Mechanism of kinase labeling. A conserved active site lysine attacks the acyl phosphate resulting in acylation of the enzyme by a biotin tag.

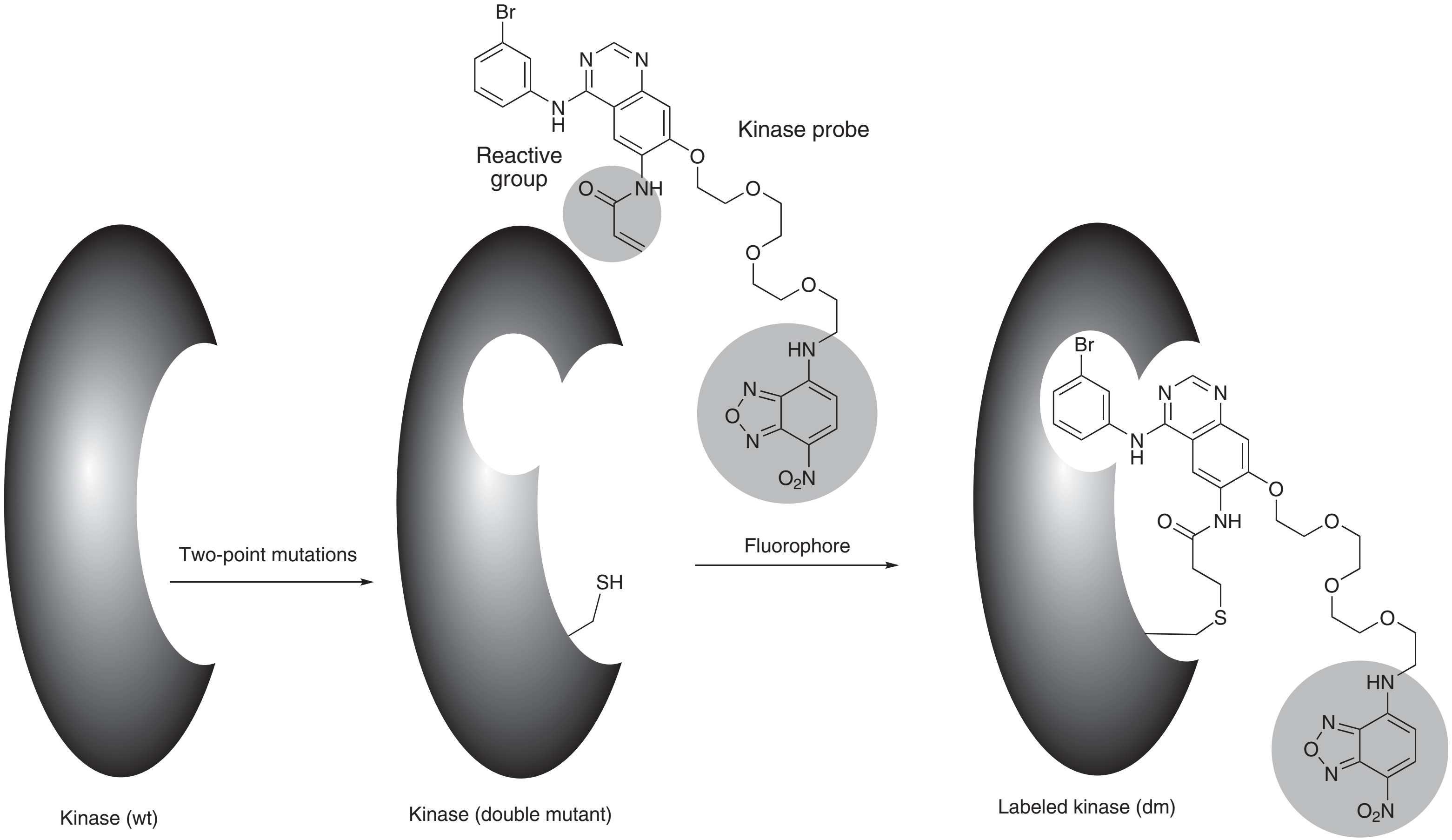


Figure 16 Genetic engineering of kinases in combination with customized probe design allowed the selective labeling of kinases via a Michael acceptor as reactive group.

Isoquinolinesulfonamides have been selected for broad kinase inhibition by occupying the ATP-binding sites. The probes exhibited indeed specific labeling of several purified kinases including hexokinase and creatine kinase.

9.17.3.1.7 Phosphatases

Protein phosphorylation is a reversible posttranslational process. Two major classes of phosphatases are responsible for the removal of phosphate groups either from serine/threonine or tyrosine residues. ABPP technologies have been developed to monitor their regulation and activities. Protein tyrosine phosphatases (PTPs) are critical modulators of signaling events and contribute to several human diseases including cancer, diabetes, and obesity.[103] Initial ABPP probes for PTPs were introduced by Lo *et al.*,[104] which were based on the 4-fluoromethylaryl phosphate moiety. PTP-mediated hydrolysis of the 4-fluoromethylaryl phosphate group generates a highly reactive quinone methide intermediate that alkylates nucleophiles at the active site of phosphatases. One major drawback of this probe is the high reactivity of the quinone methide that can escape the active site pocket by diffusion and label other proteins in the proteome.

Zhang and coworkers introduced a more specific probe for PTPs consisting of an α-bromobenzylphosphonate moiety as a PTP-specific trapping device that was connected via a linker to a biotin or rhodamine tag (**Figure 17(a)**).[105,106] This probe reacts irreversibly with an active site cysteine residue that is highly conserved in PTPs (**Figure 17(b)**). However, although these probes were specific for PTPs and unreactive toward other proteins in the proteome, the low stability in water and limited cell permeability rendered its application in ABPP studies. More recently, the same research group reported a new generation of their PTP probes inspired by aryl vinyl sulfonate and sulfone mechanism-based inhibitors.[107] These electrophilic Michael acceptor-based probes covalently reacted with the conserved cysteine residues of a broad selection of PTPs (**Figure 17(c)**).

In addition to PTPs, serine/threonine phosphatases, which are metal-dependent enzymes and share no sequence similarities to PTPs, have also been subject of ABPP studies by the functionalization of the natural product microcystin that will be discussed in the natural product section (Section 9.17.3.3.5).[97,108]

9.17.3.1.8 Glycosidases

Glycoside hydrolases represent a large enzyme family that together with glycosyltransferases are the major machinery for the formation and hydrolysis of glycosidic bonds. Two mechanistic classes of glycosidases, inverting and retaining, exist based on the stereochemical configuration at the anomeric center upon hydrolysis of the glycosidic bond.[109] Due to their essential role in metabolism, antibacterial defense, and pathogenesis, glycosidase probe design for ABPP represents an important task in chemical proteomic research and was addressed by several research groups. For many glycosidases, the design of active site-directed probes is complicated due to the pocket-shaped active site in which extensive interactions between substrate and enzyme leave little space for large probes and tags.[109] First studies with glycosidase probes were initiated by Tsai *et al.*[110] who reported the synthesis of a biotin-functionalized *p*-hydroxybenzylic fluoride moiety linked through a β-glycosidic linkage to glucose. This chemistry was adapted from a previously introduced PTP probe (Section 9.17.3.1.7).[104] Upon glycosidic bond cleavage by β-glycosidases, the probe undergoes 1,6-elimination to remove a fluoride and generates a reactive quinone methide. Similar to the previous phosphatase probe, it worked with purified enzymes but produced cross-labeling with other proteins in more complex mixtures, limiting its application in proteomic studies. This approach was subsequently extended to the surface glycoprotein neuramidase by the attachment of its natural substrate, sialic acid, to the reactive quinone methide-generating moiety (**Figures 18(a)** and **18(b)**).[111] Also, this approach worked well for purified neuramidase from *Arthobacter ureafaciens.*

A different approach to the specific labeling of β-glycosidases was introduced by Vocadlo *et al.*[112,113] who designed small azide-functionalized 2-fluorosugar probes that fit well into the pocket-shaped active site. Most retaining β-glycosidases utilize a double displacement catalytic mechanism in which two key catalytic active site carboxylic acid residues form a covalent glycosyl–enzyme intermediate that is subsequently broken down to release the hydrolyzed sugar. In this process, fluorosugars act as mechanism-based inactivators of retaining β-glycosidases by trapping the covalent glycosyl–enzyme intermediate (**Figure 19(a)**). Upon enzyme binding, the azide group was modified with the corresponding tag via Staudinger ligation (Section 9.17.2.2). The value of this approach was demonstrated by the labeling of several glycosidases of different families. An improved version of this probe led to the labeling and detailed characterization of NagZ, which is involved in murein recycling.[114]

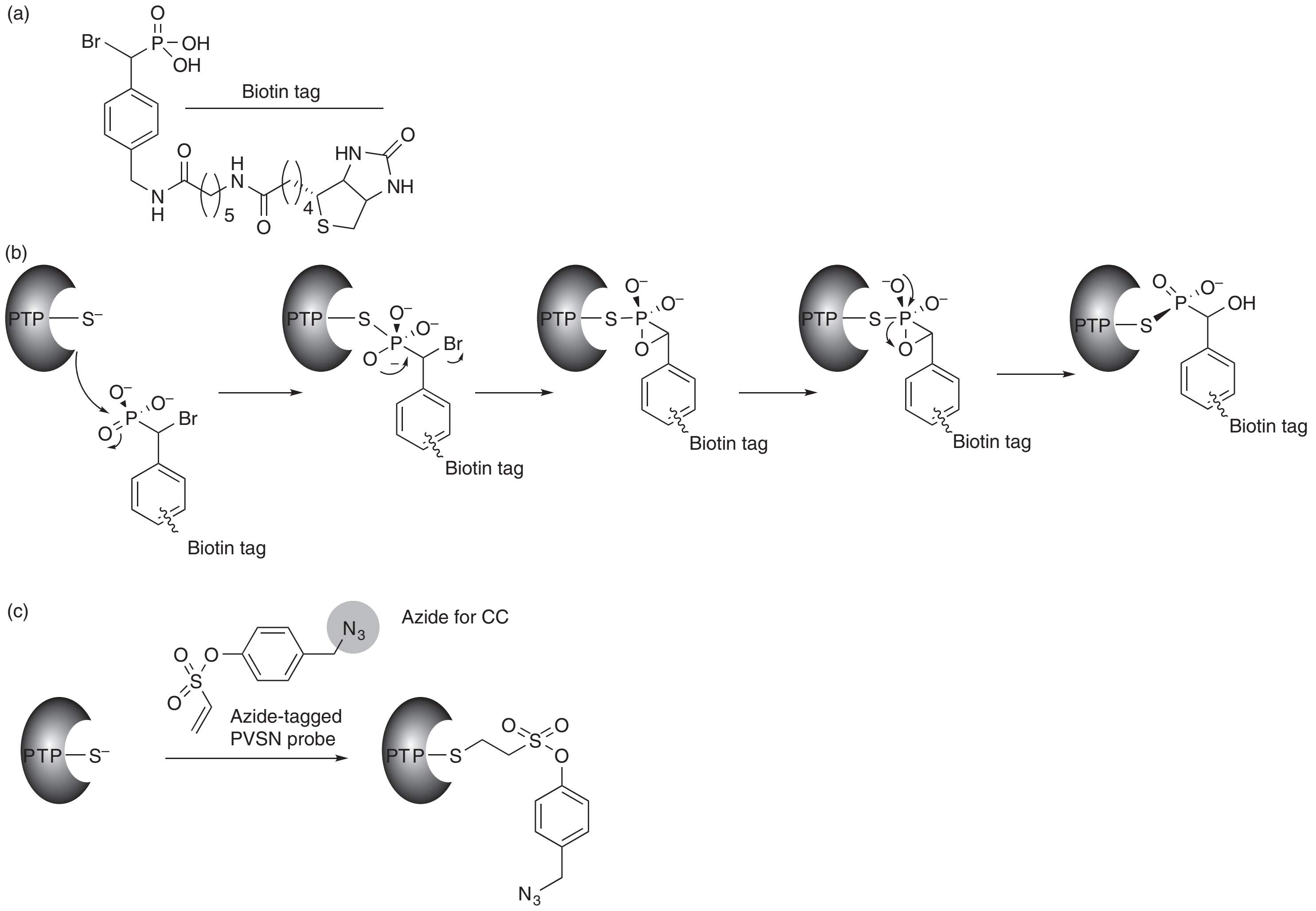


Figure 17 Probes for protein tyrosine phosphatase (PTP) targeting. (a) Design of a specific PTP probe based on an α-bromobenzylphosphonate group. (b) Proposed mechanism of inhibition. (c) Aryl vinyl sulfonates as probes for PTP labeling. The azide moiety can be modified via click chemistry with the appropriate reporter tag after target labeling.

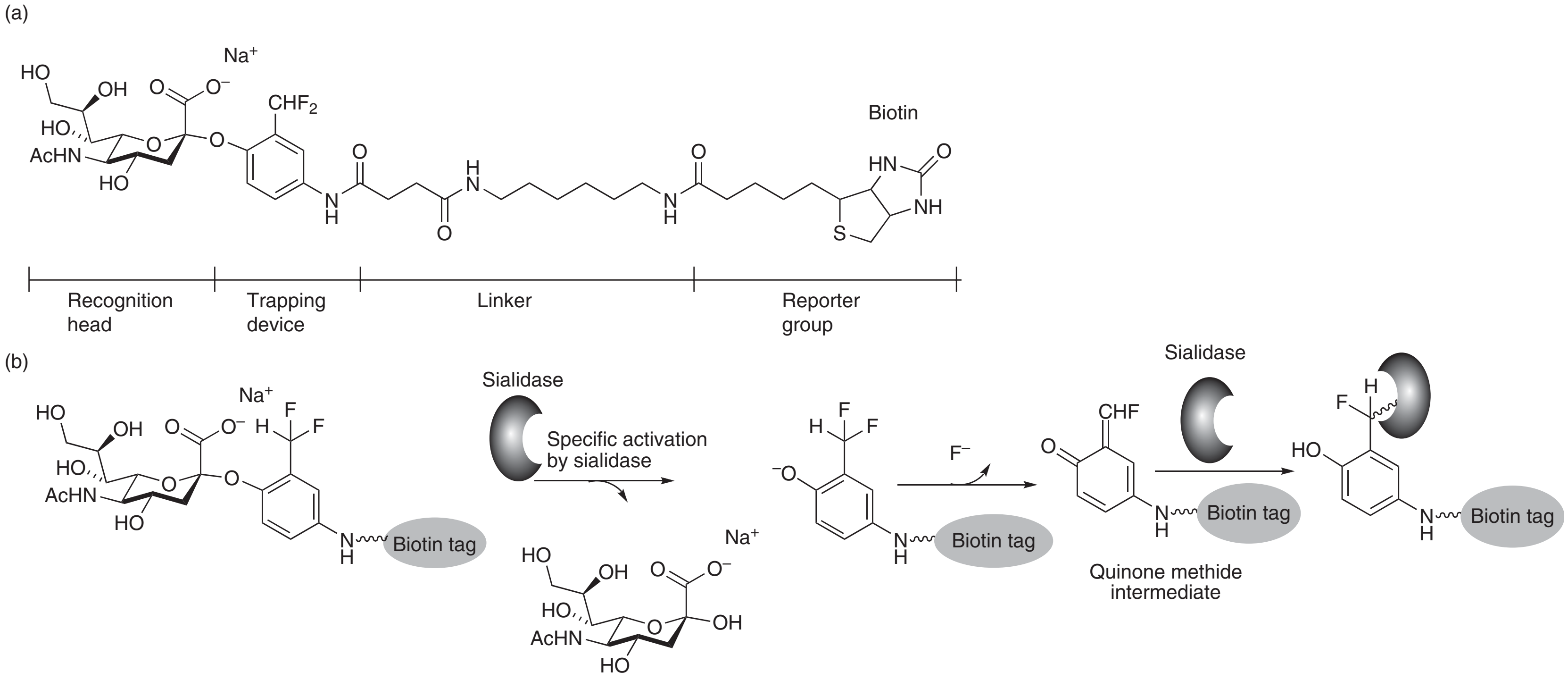


Figure 18 ABPP probe for glycosidase labeling. (a) Specific probe design using sialic acid as recognition head and a quinone methide as reactive group precursor. (b) Upon activation by glycosidase-mediated cleavage of sialic acid, a reactive quinone methide intermediate is generated that rapidly reacts with residues in the active site.

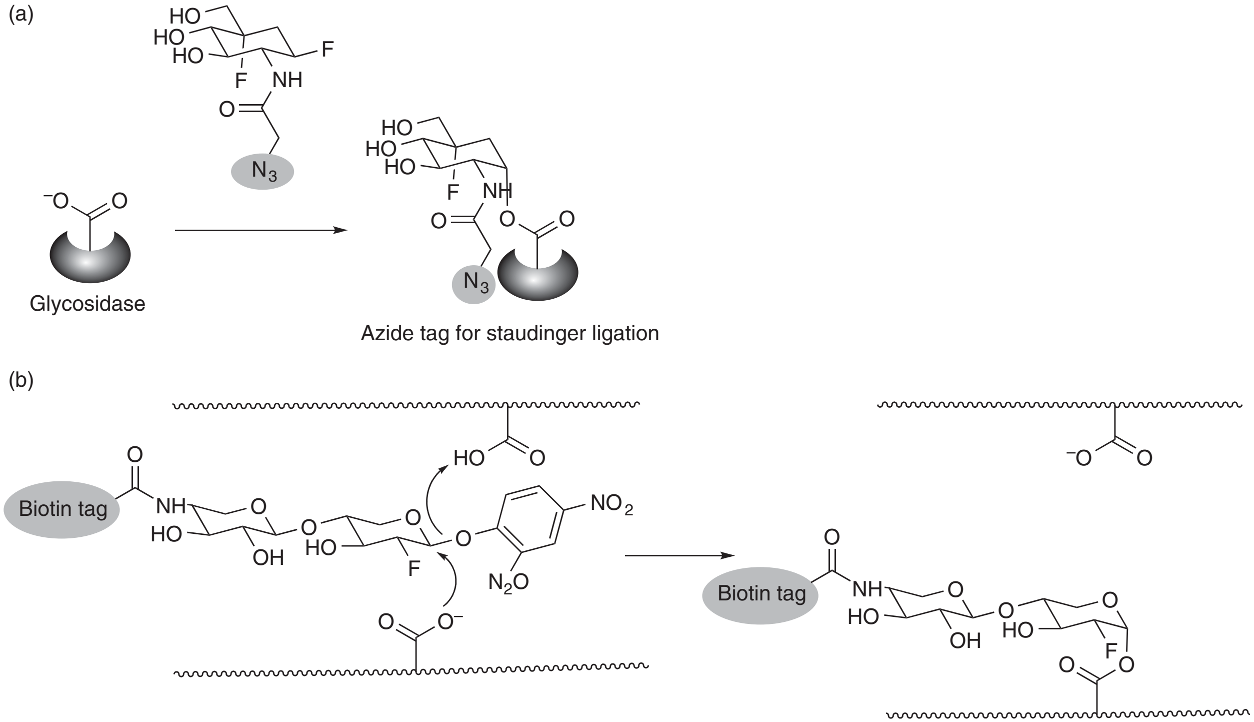


Figure 19 Mechanism of β-glycosidase active site labeling with probes for (a) retaining and (b) inverting β-glycosidases.

Recently, Hekmat *et al.*[115] reported the synthesis of novel sugar probes that are based on a 1-(2,4-dinitrophenyl)-2-fluorosugar inhibitor that traps the catalytic nucleophile of inverting β-glycosidases (**Figure 19(b)**). This probe was successfully applied to profile several glycosidases in the extracellular proteome of the soil bacterium *Cellulomonas fimi* by MS methods.

9.17.3.1.9 Cytochrome P-450

Cytochrome P-450 enzymes represent a large and diverse protein family with important functions in the metabolism of endogenous signaling molecules, xenobiotics, and drugs.[116] These enzymes are diverse and share only 16% sequence identity. Only five human enzymes contribute to 90% of metabolism of current clinical drugs and some enzymes are involved in biosynthesis of estrogen, making them important targets for the treatment of hormone-sensitive cancers.[117,118] Moreover, the activity of these enzymes is tightly regulated by multiple factors *in vivo*. These and other aspects make cytochrome P-450 enzymes prime targets for ABPP studies and functional characterization. Cravatt and coworkers recently introduced an elegant strategy to label and identify P-450 enzymes and monitored their activity even *in vivo*.[119] The chemical design of the probes was guided by a known broad-spectrum, mechanism-based P-450 inhibitor, 2-ethynylnaphthalene (2EN), that was appended to an alkyne tag for visualization via CCy (**Figure 20**). The naphthalene moiety of 2EN directs the probe into the hydrophobic pocket of the P-450 enzymes, wherein enzyme-catalyzed oxidation of the conjugated 2-acetylene generates a highly reactive ketene that acylates nucleophilic residues within the P-450 active site. The 2EN-derived probe was successfully applied to label several P-450 enzymes in rodent liver in an NADPH-dependent manner and allowed to monitor drug induction as well as inhibition of these enzymes *in vivo*.

9.17.3.1.10 Arginine deiminase

Protein arginine deiminase 4 (PAD4) plays a crucial role in the pathophysiology of rheumatoid arthritis in which the activity of this enzyme is upregulated.[120] In order to understand the role and functional regulation of PAD4, Thompson and coworkers used haloacetamidine-based lead compounds, including F-amidine and Cl-amidine, as mechanism-based inhibitors for the design of specific PAD4 probes.[121] These inhibitors covalently modified a catalytic cysteine residue in PAD4 only in the active, calcium-bound form (**Figure 21**). The value of this approach was demonstrated by the specific labeling of recombinant PAD4 in complex proteomes.

9.17.3.2 Undirected Electrophilic Probes

As seen in the previous sections, directed probe scaffolds already cover a large set of important protein classes with crucial functions in many diseases. However, for a number of protein families, selective active site-directed probes are difficult to design in part due to a lack of knowledge concerning their precise catalytic mechanisms and a lack of corresponding mechanism-based inhibitors or high-affinity labels. In order to close this gap, several undirected probes have been developed and will be discussed in this section.

9.17.3.2.1 Sulfonate esters

To expand the number of enzyme families susceptible to analysis by activity-based profiling methods, Cravatt and coworkers developed an undirected, combinatorial strategy in which libraries of candidate probes were screened against complex proteomes.[122] Sulfonate esters represent moderately reactive electrophiles that were unexplored as protein labeling reagents and easily modified with a variety of chemical structures. The probe library consisted of several aromatic and aliphatic side chains that were directly attached to the sulfonate ester moiety and linked via a spacer to biotin for visualization and enrichment. The initial library showed indeed distinct reactivity profiles of individual probes with complex proteomes and resulted in the identification of an aldehyde dehydrogenase, an enzyme class that had not been subject to ABPP probes before. Subsequent in-depth studies with an extended sulfonate ester library revealed the additional labeling of six mechanistically distinct enzyme classes (including enzymes like thiolase, oxidoreductase, hydratase, and gluthathione *S*-transferase) in several proteomes and provided evidence that the labeling occurred in the active site of the targeted proteins (**Figure 22(a)**).[123] These enzymes contain nucleophilic residues in their active site, such as cysteine and

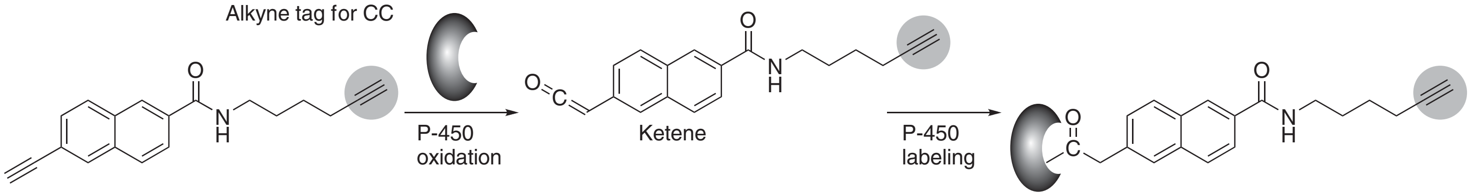


Figure 20 Mechanism-based probe for cytochrome P-450 targeting. Oxidative activation of the probe by the enzyme generates a reactive ketene that covalently binds within the active site.

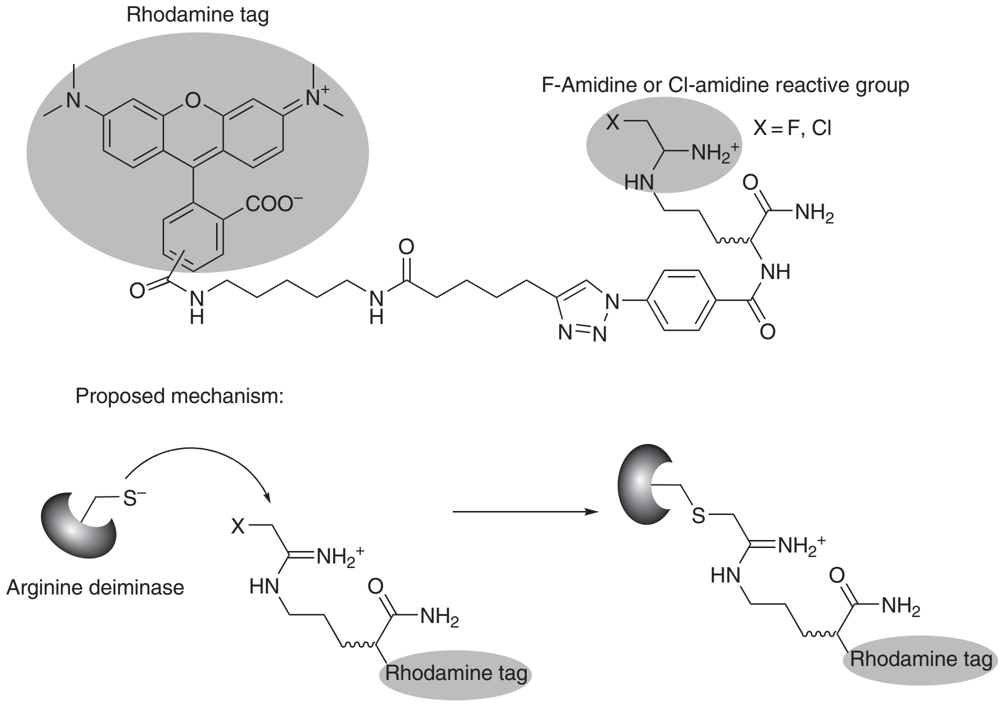


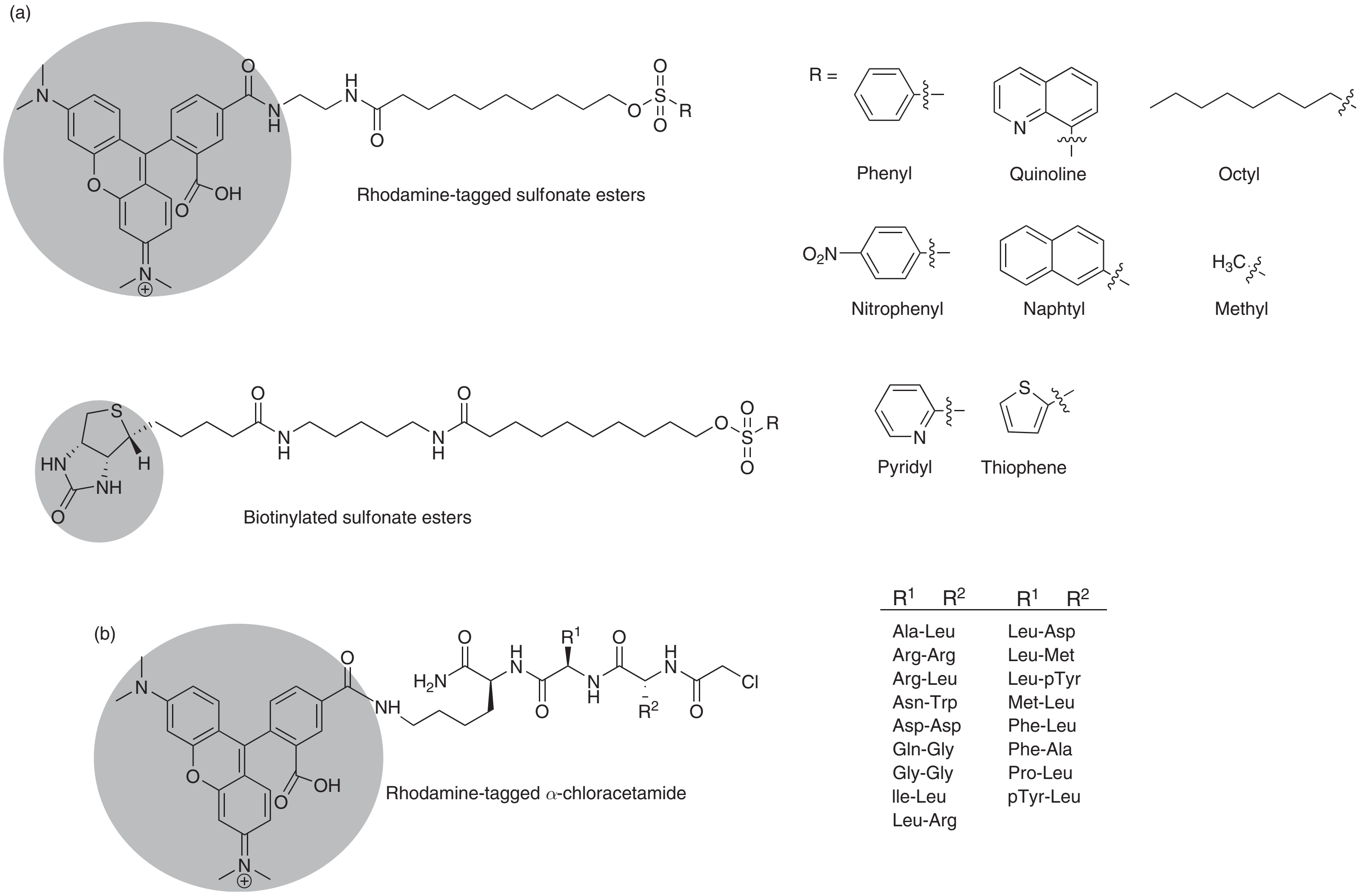
Figure 21 Haloacetamide-derived probes as potent tools for the labeling of arginine deaminase PAD4. The probe design is based on arginine as a natural substrate that is converted into a F/Cl reactive group. Active site labeling occurs via a nucleophilic substitution of the halogene by the catalytically active cysteine of PAD4.

aspartate, which are likely to react with the electrophilic compounds of the probe library. A striking feature of this approach is that none of the sulfonate-labeled enzyme targets were previously studied by directed proteomic probes, which emphasized for the first time the value of undirected ABPP studies (see also Sections 9.17.2.2 and 9.17.2.3.2).

9.17.3.2.2 α-Chloracetamide

In addition to sulfonate esters, Cravatt and coworkers explored the potential of α-chloracetamide (α-CA) as a reactive group for undirected ABPP libraries.[124] Several properties of the α-CA group including its small size, which does not interfere with individual affinity elements of the library and the unbiased reactivity with a broad range of nucleophilic amino acid residues, are beneficial for its dedicated application in complex proteomes. The α-CA-based probes were composed of a dipeptide binding element that was varied with small, bulky, hydrophobic, and charged amino acid groups. Two reporter tags, biotin and rhodamine, were directly appended to the library for target visualization and identification (**Figure 22(b)**). The library was screened against several mouse proteomes and more than 10 different classes of enzymes were identified as specific targets of the α-CA probe library including fatty acid synthase, hydroxypyruvate reductase, and malic enzyme, most of which were not labeled by previously described ABPP probes. A bioinformatic approach allowed to classify members of the library into subgroups based on shared proteome reactivity profiles that generated an optimal probe set for focused biological studies and helped to limit the consumption of precious materials.

Out of the library, one Leu-Asp α-CA dipeptide probe attracted special attention due to its selective labeling of a 44 kDa mouse liver enzyme. In a follow-up study, Barglow and Cravatt[125] further characterized the probe-labeling profile of this protein that was subsequently identified as a member of the nitrilase superfamily. Interestingly, other members of this family were also labeled and revealed evidence that dipeptide α-CA probes are suitable for the effective labeling of multiple members of the nitrilase family. The nitrilase family is a phylogenetically ancient class of enzymes that perform important functions in a wide range of organisms including humans. The specific labeling of individual members of this enzyme class by dipeptide α-CA probes



R^1 R^2	R^1 R^2
Ala-Leu	Leu-Asp
Arg-Arg	Leu-Met
Arg-Leu	Leu-pTyr
Asn-Trp	Met-Leu
Asp-Asp	Phe-Leu
Gln-Gly	Phe-Ala
Gly-Gly	Pro-Leu
Ile-Leu	pTyr-Leu
Leu-Arg	

Figure 22 Undirected ABPP probes exploiting sulfonate esters and α-chloracetamide as reactive groups. (a) A probe library with diverse sulfonate esters appended via a linker either to rhodamine or biotin. (b) α-Cloracetamide probes with a variable peptide backbone linked to a rhodamine tag.

demonstrates the possibility to achieve customized highly specific probe design by screening undirected probe libraries, a technological approach that could, in principle, achieve further success in future studies for other less-explored enzyme classes.

9.17.3.2.3 Michael acceptors

In a recent study, Weerapana *et al.*[126] investigated the proteome reactivity profiles of several carbon electrophiles including the above-discussed α-CA and sulfonate ester groups as well as a Michael acceptor system. All reactive groups were attached via a short carbon spacer to an alkyne moiety for postlabeling tag attachment via CC (**Figure 23**). Out of the set of undirected probes under investigation the Michael acceptor probe (UK) demonstrated the highest reactivity and revealed substantial protein labeling. To gain insights into the target preferences and site of labeling of these reactive groups, the authors made use of the TOP–ABPP MS platform (Section 9.17.2.3.2) and screened four different mouse tissue proteomes. In general, a very distinct reactivity profile was observed for the α-CA and UK probes, which selectively labeled cysteine residues. In contrast, the sulfonate ester group revealed unique labeling events with several active site amino acid side chains such as aspartate, glutamate, cysteine, tyrosine, and histidine. The labeling of amino acid side chains in proteins was compared to the labeling of free amino acids in solution, which revealed unanticipated trends of reactivity in the case of the UK probe that reacted with free cysteine, lysine, and histidine in solution and only with cysteine residues in proteins. The proteome coverage of UK and α-CA in this study included a variety of enzymes with cysteine nucleophiles such as fatty acid synthase, UDP-glucose-6-dehydrogenase, and multiple nitrilases. This study therefore represents a comprehensive comparison between several undirected probe scaffolds and their reactivity in proteomes, which could help to design customized ABPP probes for a wide range of proteins.

9.17.3.3 Natural Product-Based Probes

Natural products constitute a large class of pharmacological relevant molecules, which contribute to about 60% of all drugs released to the market including prominent examples such as aspirin and penicillin.[127–129] A plethora of yet uncharacterized bioactive small molecules with privileged structures awaits elucidation of their cellular targets and mechanism of action.[130] In this regard, the ABPP technology has made progress toward a systematic investigation of such natural product-derived molecular scaffolds, which will be the focus of this section. Some of the probes summarized here exhibit specificities for single enzyme classes and therefore represent tools for directed enzyme profiling (Section 9.17.3.1). As we will see, other natural product-derived probes label enzymes across different classes, which, as seen for undirected α-CA probes, in some cases lead to class selectivity by the appropriate decoration of the core scaffold with high-affinity ligands.

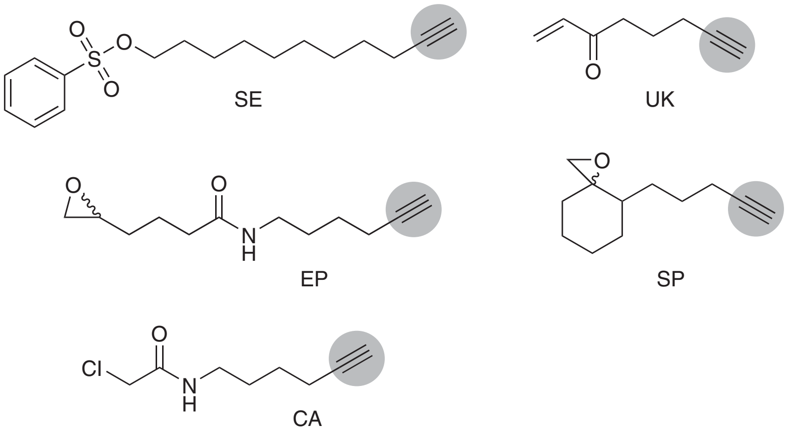


Figure 23 A set of undirected probes with diverse electrophilic reactive groups such as sulfonate esters (SE), epoxides (EP), α-chloracetamide (CA), Michael acceptor (UK), and spiroepoxide (SP).

9.17.3.3.1 Spiroepoxides

Some natural products exhibit their biological effects through covalent modification of proteins by moderately reactive electrophilic groups. Chemical reactivity is usually considered in terms of the electrophilic nature of the acyl center and the leaving group ability of the group displaced. In addition to these electronic effects, the reactivity may be modified and adjusted by steric and strain effects of the accompanying ligands.[131] Evans *et al.*[132] designed and synthesized a library of protein-reactive small molecules with a central spiroepoxide (1-oxa-spiro[2.5]octane) reactive group that was inspired by natural products such as fumagillin, luminacin D, and FR901464, linked to diverse ligands and to an alkyne moiety as a benign tag for the attachment of biotin or rhodamine reporter groups via CC (**Figure 24**). The incorporation of the moderately reactive spiroepoxide electrophilic group into the appropriate molecular scaffolds allowed a comprehensive cell-based screening approach for antiproliferation activity against the invasive human breast cancer cell line MDA-MB-231. Subsequent *in situ* proteome profiling led to the discovery of one potent compound (MJE3) that had a significant antiproliferative effect by targeting the glycolytic enzyme phosphoglycerate mutase 1. Interestingly, this enzyme could only be labeled by *in vivo* experiments with living cells but not in the corresponding lysates, suggesting that factors inside the cell may influence the interaction between the compound and the protein. This result emphasizes the importance of CC-based ABPP within living cells and shows the value of the application of probe libraries in cell-based screens for lead discovery.

9.17.3.3.2 β-Lactones

In addition to spiroepoxides, other small heterocyclic ring systems that exhibit moderate electrophilicity have been adapted for ABPP studies. One large class of such compounds are the β-lactones that represent promising biologically active privileged structures poised to react covalently with certain enzyme active sites.[133] Although β-lactones have been proven to exhibit antibiotic activities, for example, obafluorin,[134] and hymeglusin,[135] their biological enzyme targets remained largely unexplored. In order to identify the dedicated enzyme targets β-lactones in bacterial proteomes Böttcher and Sieber synthesized a small library of *trans*-β-lactone probes with diversity introduced at the C-3 position and an alkyne tag at the C-4 position for modification with a reporter tag via CC after labeling of dedicated targets.[136,137] Inspired by naturally occurring β-lactones, the biomimetic library comprised compounds with aliphatic or aromatic substitutions varying in length and branching (**Figure 25(a)**). Target identification revealed distinct labeling of about 20 different enzymes belonging to four major families comprising ligases, oxidoreductases, hydrolases, and transferases with major differences depending on the decoration of the probe. All of these families require a nucleophilic residue in their active site for catalysis (Cys or Ser), which likely attacked the electrophilic β-lactone ring. Several of the identified enzymes were of medicinal interest in the quest for new antibiotic targets including the virulence-associated enzyme caseinolytic protein protease (ClpP) that will be discussed in the biological application section (Section 9.17.4.1.3).

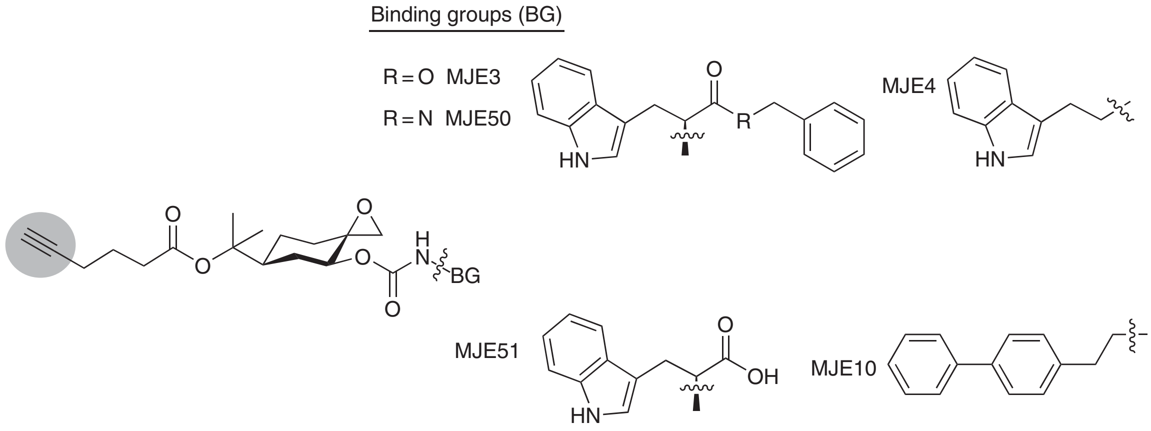


Figure 24 Design of spiroepoxide probes with diverse ligand decorations.

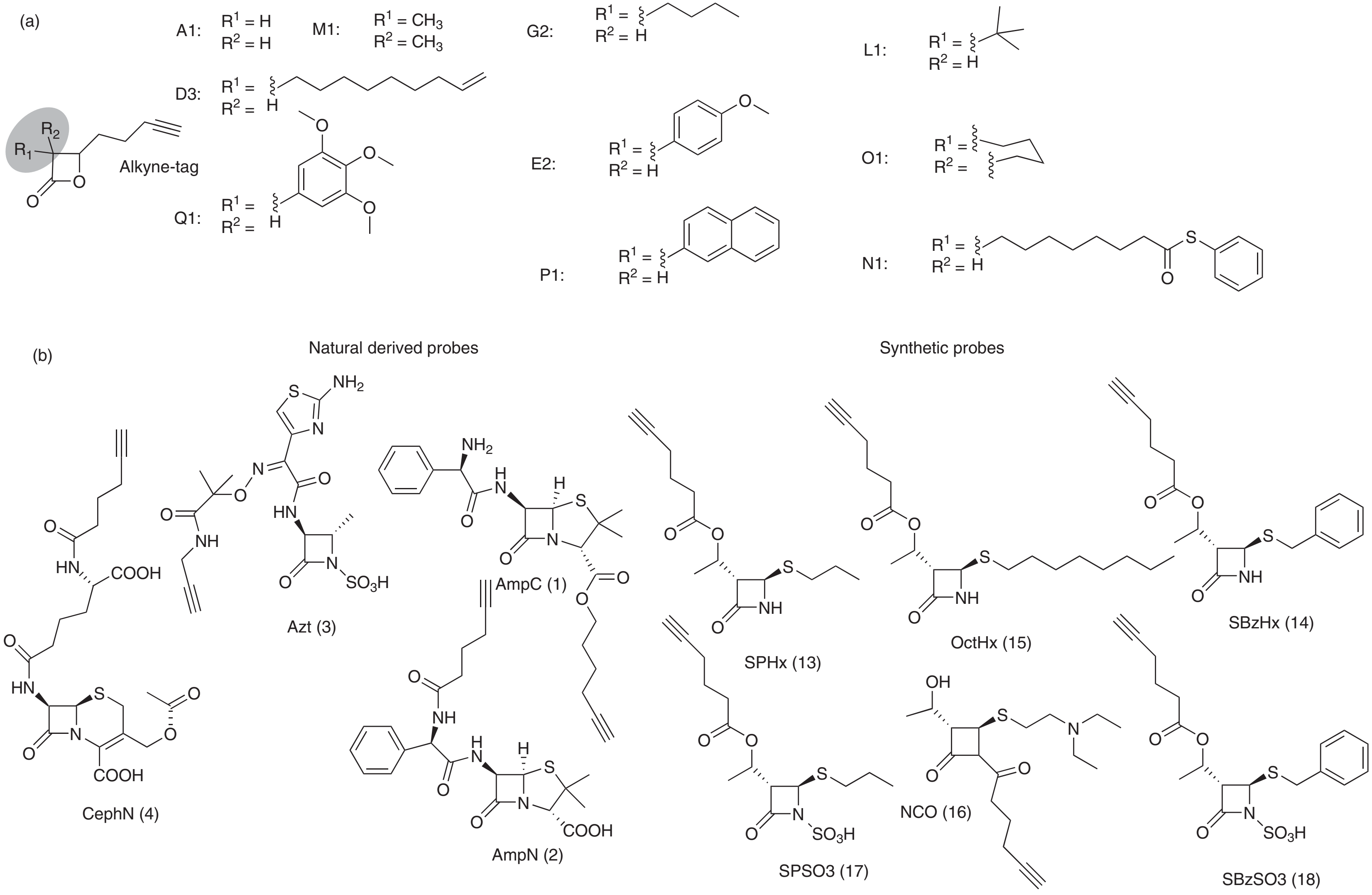


Figure 25 β-Lactones and β-lactams as natural product-derived probes for ABPP. (a) A biomimetic β-lactone probe library with an alkyne tag and several diversity elements. (b) β-Lactam probes derived from known antibiotics and synthetic approaches.

To monitor protease activities in plant extracts, Wang *et al.*[138] generated biotinylated peptides that contained a β-lactone reactive group. The probes labeled several enzymes in leaf proteomes of *Arabidopsis thaliana.* Interestingly, these studies led to the identification of a papain-like protease called RD21 that has the unexpected ability to ligate donor molecules such as peptides or lactones, probably through a thioester intermediate, to unmodified N termini of acceptor molecules.

9.17.3.3.3 β-Lactams

β-Lactams represent one of the most important groups of antibiotics prescribed for antibacterial treatment today. They stop bacterial growth by inhibiting PBPs that are indispensable for the cross-linking process during cell wall biosynthesis.[139,140] Previously, β-lactams with radioactive or fluorescent tags were used to label and visualize active PBPs in membrane preparations.[141,142] However, these methods were either limited by their time-consuming and hazardous procedures (radioactivity) or by the attachment of large fluorescent dyes onto the core scaffold, which lower target affinity and reduce cell permeability important for *in vivo* studies. To explore the role of PBPs and other yet unknown targets of β-lactams in their native environment Staub and Sieber[143] designed small β-lactam probes appended with a short alkyne handle as a benign tag for visualization and enrichment of labeled proteins. Although related in structure and size, the β-lactam probes exhibited in part different proteome reactivities and target preferences compared to the β-lactones. In their two-tiered strategy, the authors first utilized a selection of conventional antibiotics for labeling of diverse PBPs and secondly introduced a new synthetic generation of β-lactam probes that labeled and inhibited a selection of additional PBP unrelated bacterial targets (**Figure 25(b)**). Among these a resistance-associated β-lactamase was labeled and inhibited by selected probes indicating that the specificity of β-lactams can be adjusted to versatile enzyme families with important cellular functions.

9.17.3.3.4 Wortmannin

Wortmannin is a fungal metabolite that was identified as a potent and selective inhibitor for phosphoinositide 3-kinases (PI3Ks) and PI3K-related enzymes.[144,145] Kinases of this family covalently react via a conserved active site lysine nucleophile with an electrophilic extended Michael acceptor system in wortmannin. The remarkable selectivity for this important kinase subclass has stimulated the design of wortmannin-based ABPP probes (**Figure 26(a)**). Addressing this challenge, two studies have been reported that chemically modified the wortmannin core structure in order to be compatible with ABPP studies. Yee *et al.*[98] prepared several wortmannin analogues that were appended with biotin or fluorescent dyes (BODIPY and rhodamine) via a deacylated alcohol side chain at C-11 leading to the formation of an ester bond. These wortmannin probes exhibited remarkable reactivity for PI3K and PI3K-related enzymes in lysates as well as in living cells.

Similarly, Liu *et al.*[97] modified wortmannin via the C-11 alcohol with a rhodamine dye and used this probe for proteome profiling. Interestingly, in addition to targeting the expected PI3Ks, the probe was found to label and inhibit the polo-like kinase 1, a serine/threonine protein kinase with natural substrate preferences for proteins rather than small molecules. This result emphasizes again the value of ABPP studies in which all off-targets of individual bioactive compounds can be evaluated. In case of polo-like kinase 1, the ABPP results suggested that wortmannin, which was assumed to inhibit PIKs alone, may exhibit its pharmacological activity in part through polo-like kinase 1 inhibition.

9.17.3.3.5 Microcystin

Similar to wortmannin, which functions as a complementary probe for targeting a subset of the large kinase family, other natural products can be used as privileged inhibitor scaffolds for directed ABPP probe design. Microcystin, a cyclic peptide isolated from cyanobacteria, is a known covalent inhibitor of many serine and threonine phosphatases that binds into the phosphatase active site via covalent modification of a conserved cysteine residue by an electrophilic Michael acceptor embedded in the molecule.[146,147] In order to utilize this compound for ABPP, Shreder *et al.*[108] identified an arginine residue in microcystin, which was not essential for its potency against phosphatases and was therefore selected for the modification with a fluorescent rhodamine-1,3-diketone tag (**Figure 26(b)**). The probe was tested in Jurkat cell lysates and worked well for the specific labeling and identification of several phosphatases. Moreover, these probes

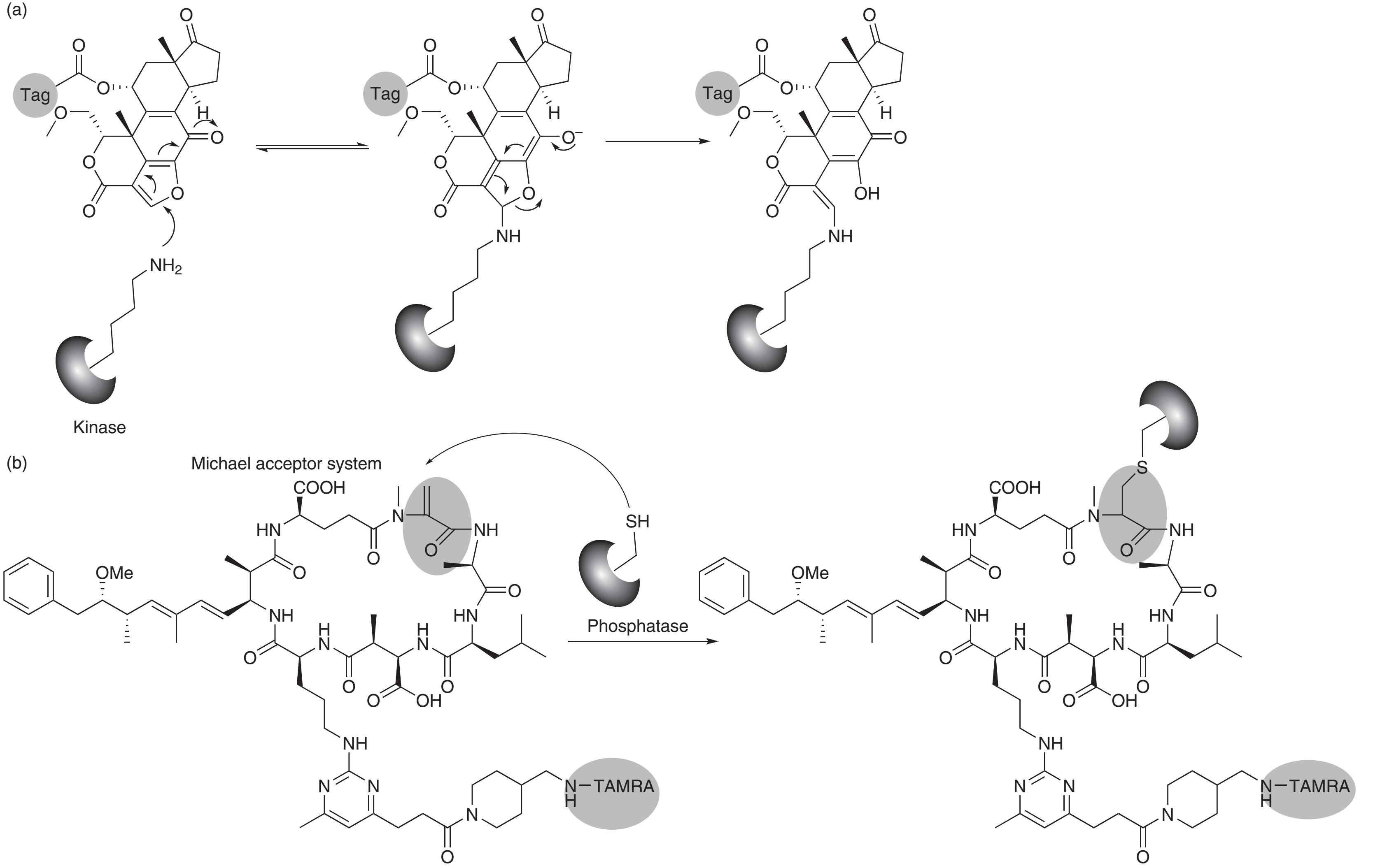


Figure 26 Wortmannin- and microcystin-based probes and their mechanistical action via embedded Michael acceptors. (a) The Michael acceptor system of wortmannin and its derived probes irreversibly inhibits its kinase target by covalent modification of the conserved active site lysine. (b) The cyclic peptide microcystine and its derived probes react covalently with the active site cysteine of phosphatases.

were successfully used to monitor changes in phosphatase activities upon stimulation of Jurkat cells with calyculin A, another potent protein phosphatase inhibitor.

9.17.3.3.6 FR182877

FR182877 is a natural product that has been reported to display several interesting biological activities including antitumor effects in mouse models.[148] The molecule contains two electrophilic groups, a strained bridgehead olefin and a lactone group. One or both of these groups represent possible sites for protein modification. In order to discover the targets of this natural product, Adam *et al.*[149] attached an azide moiety onto a secondary alcohol of FR182877 (**Figure 27(a)**). The probe was subsequently tested in mouse proteomes and carboxyesterase 1 was identified as the specific and predominant target of this natural product. This finding raises the possibility that FR182877 exhibits its antitumor activity via carboxyesterase 1 inhibition, as carboxyesterases are expressed in a variety of cancers. Further analysis will be required to investigate this notion in detail.

9.17.3.3.7 HUN-7293

While previous examples utilized reactive electrophilic natural products for ABPP probe development, the following example illustrates that in principle noncovalent binding affinity-based natural products can also be applied for ABPP design. HUN-7293 is a fungal cyclodepsipeptide that inhibits vascular cell adhesion molecule (VCAM) expression by binding to an unidentified target. In order to identify the natural targets of HUN-7293, Taunton and coworkers designed a photoaffinity probe in which photoleucine, an alkyl diazirine leucine mimic, replaces the native leucine at position 4 in the natural cyclodepsipeptide. A propargyl substituent was attached to position 1 to enable CC after photocross-linking (**Figure 27(b)**.[150] Following incubation with endoplasmatic reticulum microsome fractions revealed Sec61α as a selective target. Sec61α is thought to form a channel through which all proteins transit as they enter the secretory pathway. HUN-7293-mediated inhibition of Sec61α is therefore the probable mechanism by which the compound exhibits its observed potency that emphasizes again the utility of ABPP for target discovery.

9.17.4 Biological Applications

The success of the ABPP technology has led to significant progress in biological applications. The fundamental tools for monitoring the system-wide activity changes in the proteome have been developed to capture information on the function of proteins leading to mechanistic insights and helping to treat human diseases. As we will see in this chapter, the molecular properties of several diseases such as cancer, diabetes, and bacterial

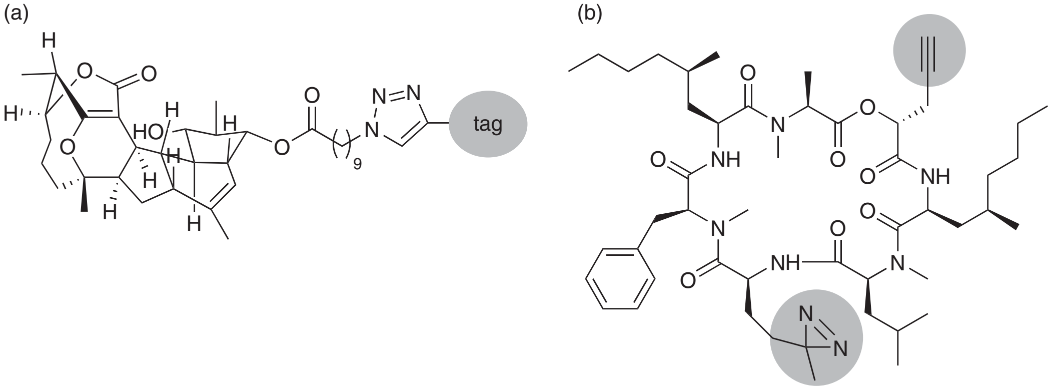


Figure 27 (a) Probe derived from the natural product FR182877. (b) Probe derived from the natural product HUN-7293 with a photoleucine cross-linker.

infections have been investigated. Moreover, specialized versions of ABPP for inhibitor discovery and biological imaging have been developed and will be discussed as well.

9.17.4.1 Enzyme Activities in Human Diseases

9.17.4.1.1 Cancer

Proteases are considered to play crucial roles in cancer. Their proteolytic activities can contribute to the degradation of connective tissue and dissolution of epithelial and endothelial basement membranes. Their physiological role involves remodeling processes that occur during tumor invasion and metastasis.[5,151] Utilizing the chemical tools for hydrolase profiling, several ABPP studies have been carried out in order to understand their activity, function, and regulation as well as to identify crucial targets for diagnostic studies and therapeutic intervention.

Jessani *et al.*[47] applied their previously established FP probes (Section 9.17.3.1.1) to systematically profile variations in serine hydrolase activities in human breast and melanoma cancer cell lines. The global analysis of these enzyme activities resulted in the identification of a cluster of proteases, lipase, and esterases that distinguished cancer lines based on their tissue of origin or state of invasiveness. Identification of these enzymes with a biotinylated FP probe revealed known markers of cancer malignancy including urokinase and uncharacterized enzymes, such as the integral membrane hydrolase KIAA1363 for which no previous link to cancer could be established. The activity of the latter enzyme was also found to be highly elevated in aggressive estrogen receptor-negative primary human breast tumors in ABPP–MudPIT studies (Section 9.17.2.3.2), emphasizing its important role in cancer.[30] In subsequent proteomic and metabolomic studies, Chiang *et al.*[49] utilized a potent and selective carbamate inhibitor for KIAA1363, which was identified via competitive ABPP (Section 9.17.4.3) to explore its biological role by a chemical knockout. Aggressive cell lines that were treated with this compound revealed a decrease in monoalkylglycerol ether levels. Biochemical analysis confirmed that KIAA1363 regulated the formation of these lipids that are subsequently processed into alkyl-lysophospholipids that are known to play crucial roles in cancer.[152] The disruption of KIAA1636 activity reduced cancer cell migration and tumor growth *in vivo*, which makes this enzyme a promising target for therapeutic intervention.

In addition to profiling serine hydrolase activities of cancer cell lines, Cravatt and coworkers investigated to what extent these enzyme activities reflect the situation in living organisms, where specific host factors may influence tumor biology.[48] The human breast cancer cell line MDA-MB-231 was grown in culture and as orthotopic xenograft tumors in the mammary fat pad of immunodeficient mice. Comparison of serine hydrolase activities in these two different models monitored by FP probes enabled the identification of both tumor- and host-derived enzyme activities, which indicated that the *in vivo* environment of the mouse mammary fat pad stimulated the growth of human breast cancer cells with elevated tumorigenic properties.

Further support for the crucial roles of proteases in cancer was provided by Joyce *et al.*[153] who first analyzed data from gene expression profiling of pancreatic islet tumors in a transgenic RIP1-Tag2 mouse model that revealed an upregulation of cathepsin cysteine proteases and then utilized cathepsin-directed E-64-based ABPP probes (Section 9.17.3.1.3) to assess their activity by SDS gel analysis and *in vivo* imaging. These studies showed elevated activities (more than three orders of magnitude) of several cathepsins in tumors compared to normal islets, including cathepsin B, C, L, and Z. Cell-permeable, fluorescent (BODIPY) probes were applied to localize cathepsin activity within tumor cells revealing elevated protease activity, for example, at invasive fronts of carcinomas. Additional studies with mice treated with broad-spectrum cathepsin inhibitors revealed impaired angiogenisis and tumor growth, which further supported the crucial role of cathepsins in tumorigenesis and emphasizes the potential of these enzymes as valuable drug targets.

9.17.4.1.2 Metabolic disorders

Obesity and diabetes are crucial metabolic disorders with increased incidences in modern societies. Defects in liver function are suspected to be a major cause of these diseases and the identification of new diagnostic markers and targets in liver proteomes represents a promising starting point for therapeutic intervention. Following this notion, Barglow and Cravatt[124] utilized an optimal probe set of nondirected α-CA dipeptide probes (Section 9.17.3.2.2) to profile liver tissue from lean (wild-type) and obese mice. With this optimal probe set, several enzyme activities with altered expression in lean and obese mice livers, including hydroxypyruvate reductase that is involved in the unusual biosynthesis of glucose from serine, were identified. These results suggested that nonclassical pathways for glucose synthesis may be effective in obesity-related diseases such as type II diabetes.

9.17.4.1.3 Infectious diseases

Malaria is a devastating disease caused by the parasite *Plasmodium falciparum*. Proteases are promising drug targets for antimalaria therapy due to their large occurrence in *P. falciparum* and their essential role during the parasite's life cycle. Bogyo and coworkers realized the potential value of cysteine proteases as putative drug targets and systematically profiled their activity with cysteine protease-specific E-64-based ABPP probes (Section 9.17.3.1.3) throughout various stages of the parasite's life cycle.[154] Dramatic changes in the activities of the cysteine proteases falcipain 1, 2, and 3 were observed with falcipain 1 being the most active during the invasive merozoite phase of the life cycle. Specific inhibitors of this enzyme, which have been previously identified by screening of chemical libraries, blocked parasite invasion into host erythrocytes. These experiments demonstrated the essential role of falcipain 1 for host invasion and established a potential new target for therapeutical intervention. In addition to their efforts to understand protease-mediated host invasion, Bogyo and coworkers recently introduced a new model on how proteases of *Plasmodium* regulate erythrocyte rupture.[155] A focused library of covalent irreversible serine and cysteine protease inhibitors, based on vinyl sulfones (Section 9.17.3.1.2), acyloxymethylketones (Section 9.17.3.1.3), isocoumarin, epoxides (Section 9.17.3.1.3), and phosphonates (Section 9.17.3.1.1), were applied to identify compounds that specifically block the release of the parasite from host red blood cells (**Figure 28**). Hit compounds from this library were appended with a biotin tag for target enrichment. Two proteases, the subtilisin family serine protease PfSUB1 and the cysteine protease dipeptidyl peptidase 3, were identified as primary regulators of this process, suggesting that these two proteases regulate the processing of downstream substrates that are required for efficient release of parasites from host red blood cells.

More recently, ABPP was applied to bacterial infections, which, with the evolution of multidrug-resistant bacterial pathogens, pose a serious threat to public health. Especially the opportunistic pathogen *Staphylococcus aureus* has gained importance through the dramatically increasing appearance of methicillin-resistant (MRSA) strains. The major reasons for this daunting problem are the excessive use of conventional antibiotics, the limited number of essential cellular targets addressed by these compounds, and their paramount selective pressure exerted on bacterial viability leading to resistance development. Böttcher and Sieber utilized the ABPP technology to develop specific β-lactone-based inhibitors for the ClpP, a phylogenetically highly conserved serine protease that was found to be crucial for virulence of many bacterial pathogens.[137,156,157,158] The most potent lactone inhibitor was able to completely abolish hemolytic and proteolytic activities and showed a dramatic decrease in the expression of lipase and DNAse exoprotein activities in *S. aureus*. Moreover,

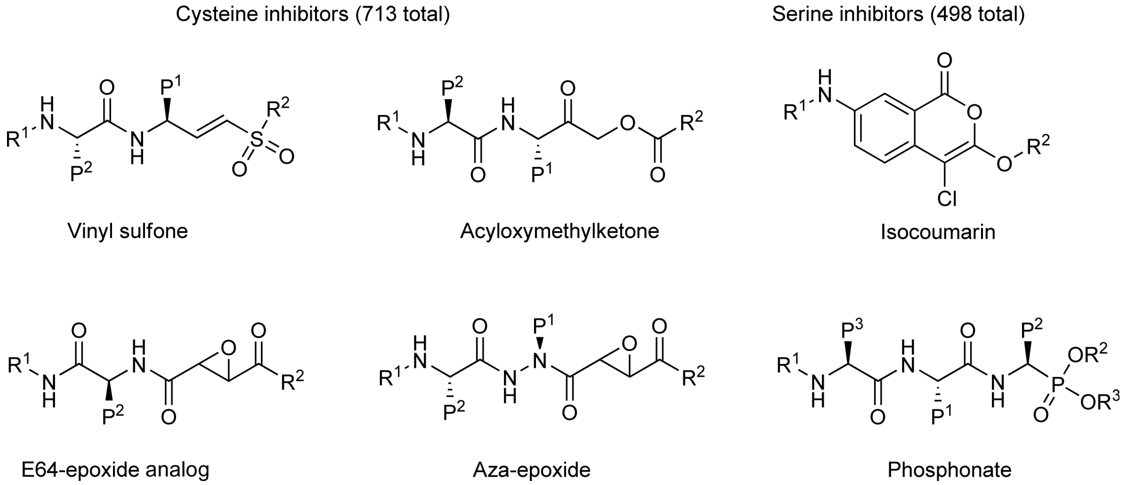


Figure 28 Cysteine and serine protease inhibitors for antimalaria screening. Lead compounds were converted into the corresponding probes for target discovery.

virulence properties were significantly reduced even in MRSA strains. Targeting a central virulence regulator may therefore represent an attractive strategy for neutralizing the harmful effects of bacterial pathogens and help the host immune response to eliminate the disarmed bacteria.

Another approach was introduced by Ploegh and coworkers who utilized an azide-modified E-64 cysteine protease probe to label and visualize active cathepsin B of *Salmonella typhimurium*, a facultative intracellular bacterial pathogen, during infection of primary macrophages. The studies demonstrated that active cathepsin B was excluded from *Salmonella*-internalized host vacuoles, which suggested a potential role in virulence.[159]

9.17.4.2 Class Assignment of Uncharacterized Enzymes

Many gene products represent uncharacterized enzymes that lack a specific class assignment. ABPP can be used to identify sequence unrelated members of enzyme families on the basis of their reactivity with mechanism-based inhibitors. So far, several uncharacterized enzyme activities have been identified by ABPP including several serine hydrolases such as KIAA1363[49] (discussed in Section 9.17.4.1.1) and the aspartyl protease presenilin-1[88] (discussed in Section 9.17.3.1.5). In another study, Jessani *et al.*[160] report the in-depth characterization of a specific FP probe target, sialic acid 9-*O*-acetylesterase (SAE), which was selectively expressed in melanoma cell lines but shared no homology with serine hydrolases or any other enzyme class. In order to assign this enzyme to a family, the site of FP probe labeling was investigated by LC–MS/MS analysis (Section 9.17.2.3.2), which turned out to be a conserved serine residue. Mutation of this serine to alanine resulted in a lack of catalytic activity and no FP labeling occurred supporting the role of this residue as an active site nucleophile. These results indicate that SAE constitute a novel branch of the serine hydrolase enzyme family and may enable the design of specific inhibitors.

9.17.4.3 Competitive Profiling for Inhibitor Discovery

The development of broad class-specific ABPP probes that target many enzymes of the same family in parallel affected the traditional way of drug discovery. Classical screening of small molecule libraries was performed against single predefined and purified targets by using *in vitro* assays that usually revealed only little information about target selectivity and potential off-targets in complex proteomic mixtures. In contrast, ABPP methods offer the advantage of testing enzyme–inhibitor interactions in their native environment without the need for recombinant expression, purification, and the development of specific substrate assays for each individual target. This technology was first applied for the identification of irreversible inhibitors by Bogyo and coworkers.[161] Their competitive profiling procedure involved the synthesis of an epoxide library of which individual members were preincubated with crude tissue extracts followed by the addition of an E-64 epoxide-based fluorescent probe that is directed against the active sites of the cysteine protease enzyme family (Section 9.17.3.1.3). Binding of an inhibitor to one or more dedicated enzymes was measured as a reduction in the fluorescence intensity of probe labeling relative to an untreated control on SDS gels (**Figure 29**). This rapid screening approach resulted in the identification of a new cathepsin B selective inhibitor.

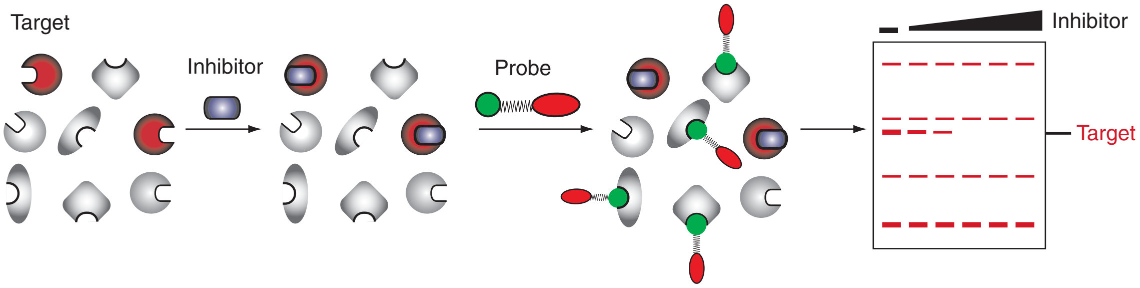


Figure 29 Competitive ABPP. Inhibitors are screened in whole proteomes and potent compounds are identified by their ability to reduce probe labeling of the target enzyme.

Recently, Li *et al.*[162] used ABPP competitive profiling to screen libraries of carbamate compounds to identify specific inhibitors for uncharacterized serine hydrolases. The carbamate reactive group reacted irreversibly with the active site serine, which, in turn, blocked subsequent labeling with a fluorescent FP probe. The carbamates were decorated with various side chains in order to find the optimal composition for dedicated target binding. Utilizing this method, a potent and selective inhibitor for α/β-hydrolase domain 6, which may play important functions in nervous system metabolism or signaling, was discovered.

Competitive ABPP does not only work for the discovery of irreversible inhibitors and has also been successfully applied for the identification of reversible high-affinity compounds. In principle, the methodology for reversible compound competition is similar to the abovementioned irreversible protocols, with the exception that the kinetics of probe–proteome reactions must be taken into account. Reversible inhibitors affect probe labeling of an enzyme only for a restricted period of time, depending on the affinity of the inhibitor and the rate of probe reactivity. Cravatt and coworkers established a methodology for the identification of specific and potent inhibitors in mouse tissue proteomes by competitive analysis with serine hydrolase-directed FP probes.[163] FP probe labeling of proteomes was carried out at conditions under which the extent of probe labeling for the majority of enzymes was incomplete and competitive inhibition could be monitored at a single, kinetically relevant time point. The success of this technology was demonstrated by the identification of several high-affinity inhibitors for the endocannabinoid-degrading enzyme fatty acid amide hydrolase (FAAH), triacylglycerol hydrolase (TGH), and an uncharacterized membrane-associated hydrolase.

9.17.4.4 Imaging of Enzyme Activities

Protein activity inside cells is regulated by many factors including spatial and temporal expression or activation mediated by posttranslational modifications, small molecule, or cofactor binding. So far, molecular imaging tools have provided a number of methods to monitor the dynamics of spatial and temporal distribution of proteins and bioorganic molecules.[164,165] The green fluorescent protein (GFP) has become a key element of many protein-based probes designed to monitor molecular and cellular events.[164] This approach has been particularly useful for imaging proteins and receptors, but not for analyzing enzyme function and activity in living cells or organisms. Addressing this challenge, Bogyo and coworkers monitored cysteine protease activity in an *in vivo* mouse model for pancreatic cancer by injecting a fluorescent E-64-based probe into living mice and visualized active enzymes by fluorescence microscopy.[153] One limitation of this method is the permanent fluorescent signal that is obtained regardless whether the probe is bound to its molecular target or free in solution. To avoid this limitation, Bogyo and coworkers introduced a second generation of imaging probes for the visualization of cathepsin activities in living cells.[166] An acyloxymethylketone reactive group was appended with a cell-permeable BODIPY fluorophore and a quencher moiety that is a part of the leaving group that gets liberated upon protease binding (**Figure 30**). In the quenched state, the compound does not exhibit any fluorescence until the quencher is displaced by the protease resulting in a strong fluorescent signal. These probes were successfully applied for real-time imaging of cysteine protease activity in living cells. In a recent approach, Blum *et al.*[167] extended this quenched ABPP concept for monitoring cathepsin activity in mice bearing grafted tumors by whole-body imaging.

9.17.5 Conclusions and Outlook

During the past decade, ABPP has become a mature and established standard technology for the rapid, sensitive, and selective identification of dedicated enzyme activities and inhibitor properties in complex proteomes. Its interdisciplinary setup, involving techniques ranging from organic chemistry over cell biology to analytical sciences, is the key to its successful application for biological tasks that could not be addressed by a single discipline alone. Today, a wealth of chemical probes for several important enzyme classes are available that have been proven successful for the functional characterization of many diseases such as cancer, infections, and metabolic disorders. In combination with applications for competitive inhibitor screening and molecular imaging, ABPP represents a key technology not only important for the functional characterization of enzyme activity but also indispensable for drug discovery programs. One fundamental breakthrough is the application

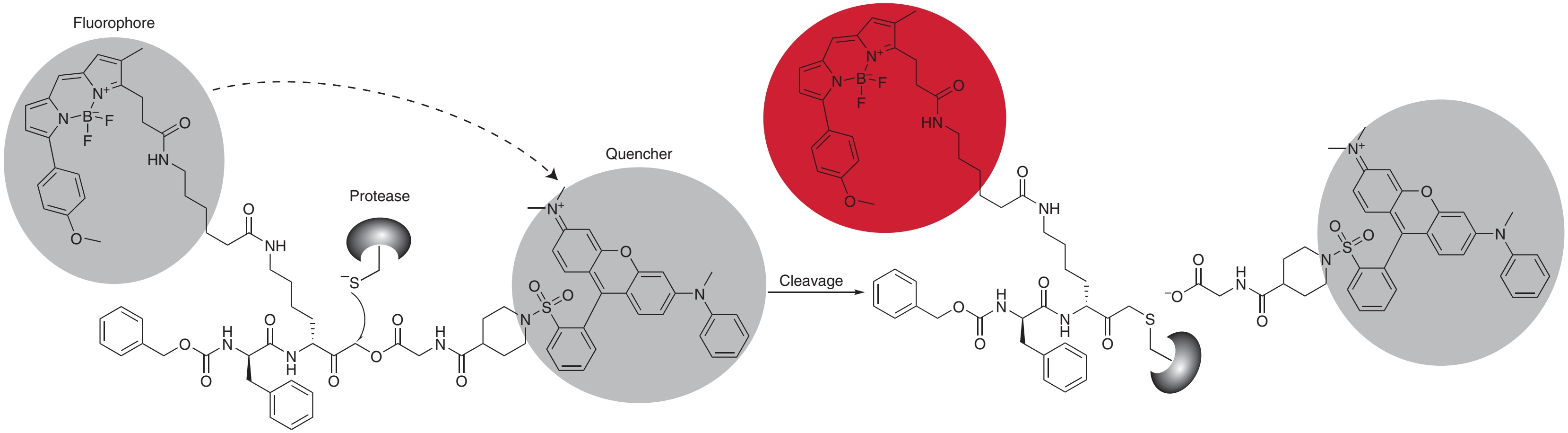


Figure 30 Design and mechanism of a FRET ABPP probe for life cell imaging of enzyme activities. Upon target binding, the quenching group gets displaced and fluorescence of the BODIPY fluorophore can be monitored.

of bio-orthogonal labeling strategies that facilitate the detection of enzyme activities within living cells and organisms and reveal functional differences compared to activities in lysates. Moreover, approaches to continuously monitor the regulation of enzymes by life cell imaging, provided unprecedented insights into native processes within living cells and whole organisms in real time.

A major task for future ABPP development will be the expansion of available probes against an almost complete set of physiological and pharmacological relevant enzyme and protein classes, including those that have not been addressed at all so far such as receptors, ion channels, and structural proteins. In addition, the chemical design of noncovalent, reversible probes appended with small and almost 'invisible' reactive groups that form a covalent bond with the enzyme active site (e.g., photocross-linkers) and do not lower the binding affinity, are desired. Looking back to the rapid and successful advances of the past, these challenges seem to be addressable by modern chemistry methods and will lead to continuous success of this technology for future applications in systems biology, medicine, and chemistry.

References

1. A. Saghatelian; B. F. Cravatt, *Nat. Chem. Biol.* **2005**, *1* (3), 130–142.
2. C. T. Walsh; S. Garneau-Tsodikova; G. J. Gatto, Jr., *Angew. Chem. Int. Ed. Engl.* **2005**, *44* (45), 7342–7372.
3. C. Lopez-Otin; J. S. Bond, *J. Biol. Chem.* **2008**, *283*, 30433–30437.
4. M. Whittaker; C. D. Floyd; P. Brown; A. J. H. Gearing, *Chem. Rev.* **1999**, *99*, 2735–2776.
5. D. Hanahan; R. A. Weinberg, *Cell* **2000**, *100* (1), 57–70.
6. M. G. Paulick; M. Bogyo, *Curr. Opin. Genet. Dev.* **2008**, *18* (1), 97–106.
7. A. E. Speers; B. F. Cravatt, *ChemBioChem* **2004**, *5*, 41–47.
8. N. Jessani; B. F. Cravatt, *Curr. Opin. Chem. Biol.* **2004**, *8*, 54–59.
9. S. A. Sieber; B. F. Cravatt, *Chem. Commun. (Camb.)* **2006**, (22), 2311–2319.
10. M. J. Evans; B. F. Cravatt, *Chem. Rev.* **2006**, *106* (8), 3279–3301.
11. B. F. Cravatt; A. T. Wright; J. W. Kozarich, *Annu. Rev. Biochem.* **2008**, *77*, 383–414.
12. M. Fonovic; M. Bogyo, *Curr. Pharm. Des.* **2007**, *13* (3), 253–261.
13. K. T. Barglow; B. F. Cravatt, *Nat. Methods* **2007**, *4* (10), 822–827.
14. D. A. Jeffery; M. Bogyo, *Curr. Opin. Biotechnol.* **2003**, *14* (1), 87–95.
15. H. Schmidinger; A. Hermetter; R. Birner-Gruenberger, *Amino Acids* **2006**, *30* (4), 333–350.
16. P. M. Blumberg; J. L. Strominger, *Methods Enzymol.* **1974**, *34*, 401–405.
17. J. W. Kozarich; J. L. Strominger, *J. Biol. Chem.* **1978**, *253* (4), 1272–1278.
18. M. P. Patricelli; D. K. Giang; L. M. Stamp; J. J. Burbaum, *Proteomics* **2001**, *1*, 1067–1071.
19. Y. Liu; M. P. Patricelli; B. F. Cravatt, *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 14694–14699.
20. D. Kidd; Y. Liu; B. F. Cravatt, *Biochemistry* **2001**, *40*, 4005–4015.
21. J. C. Powers; J. L. Asgian; O. D. Ekici; K. E. James, *Chem. Rev.* **2002**, *102* (12), 4639–4750.
22. A. Baruch; D. A. Jeffery; M. Bogyo, *Trends Cell Biol.* **2004**, *14* (1), 29–35.
23. V. V. Rostovtsev; J. G. Green; V. V. Fokin; K. B. Sharpless, *Angew. Chem. Int. Ed. Engl.* **2002**, *41*, 2596–2599.
24. E. Saxon; C. R. Bertozzi, *Science* **2000**, *287*, 2007–2010.
25. A. E. Speers; G. C. Adam; B. F. Cravatt, *J. Am. Chem. Soc.* **2003**, *125*, 4686–4687.
26. A. E. Speers; B. F. Cravatt, *Chem. Biol.* **2004**, *11*, 535–546.
27. J. M. Baskin; J. A. Prescher; S. T. Laughlin; N. J. Agard; P. V. Chang; I. A. Miller; A. Lo; J. A. Codelli; C. R. Bertozzi, *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104* (43), 16793–16797.
28. G. C. Adam; J. J. Burbaum; J. W. Kozarich; M. P. Patricelli; B. F. Cravatt, *J. Am. Chem. Soc.* **2004**, *126*, 1363–1368.
29. A. E. Speers; B. F. Cravatt, *J. Am. Chem. Soc.* **2005**, *127* (28), 10018–10019.
30. N. Jessani; S. Niessen; B. Q. Wei; M. Nicolau; M. Humphrey; Y. Ji; W. Han; D. Y. Noh; J. R. Yates, III; S. S. Jeffrey; B. F. Cravatt, *Nat. Methods* **2005**, *2* (9), 691–697.
31. H. Liu; R. G. Sadygov; J. R. Yates, III, *Anal. Chem.* **2004**, *76* (14), 4193–4201.
32. W. M. Old; K. Meyer-Arendt; L. Aveline-Wolf; K. G. Pierce; A. Mendoza; J. R. Sevinsky; K. A. Resing; N. G. Ahn, *Mol. Cell Proteomics* **2005**, *4* (10), 1487–1502.
33. E. S. Okerberg; J. Wu; B. Zhang; B. Samii; K. Blackford; D. T. Winn; K. R. Shreder; J. J. Burbaum; M. P. Patricelli, *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102* (14), 4996–5001.
34. S. A. Sieber; T. S. Mondala; S. R. Head; B. F. Cravatt, *J. Am. Chem. Soc.* **2004**, *126* (48), 15640–15641.
35. H. Schmidinger; H. Susani-Etzerodt; R. Birner-Gruenberger; A. Hermetter, *ChemBioChem* **2006**, *7* (3), 527–534.
36. P. van der Veken; E. H. Dirksen; E. Ruijter; R. C. Elgersma; A. J. Heck; D. T. Rijkers; M. Slijper; R. M. Liskamp, *ChemBioChem* **2005**, *6* (12), 2271–2280.
37. M. Shimkus; J. Levy; T. Herman, *Proc. Natl. Acad. Sci. U.S.A.* **1985**, *82* (9), 2593–2597.
38. S. H. Verhelst; M. Fonovic; M. Bogyo, *Angew. Chem. Int. Ed. Engl.* **2007**, *46* (8), 1284–1286.
39. M. Fonovic; S. H. Verhelst; M. T. Sorum; M. Bogyo, *Mol. Cell Proteomics* **2007**, *6* (10), 1761–1770.
40. C. Lopez-Otin; C. M. Overall, *Nat. Rev. Mol. Cell Biol.* **2002**, *3* (7), 509–519.
41. G. Bitan; L. Scheibler; H. Teng; M. Rosenblatt; M. Chorev, *J. Pept. Res.* **2000**, *55* (3), 181–194.
42. P. J. Weber; A. G. Beck-Sickinger, *J. Pept. Res.* **1997**, *49* (5), 375–383.
43. Y. A. DeClerck; S. Imren; A. M. P. Montgomery; B. M. Mueller; R. A. Reisfeld; W. E. Laug, *Adv. Exp. Med. Biol.* **1997**, *425*, 89–97.

44. J. D. Clark; A. R. Schievella; E. A. Nalefski; L. L. Lin, *J. Lipid Mediat. Cell Signal.* **1995**, *12*, 83–117.
45. C. T. Walsh, *Enzymatic Reaction Mechanisms*; W.H. Freeman and Company: New York, 1979; pp 53–107.
46. B. N. Bouma; L. A. Miles; G. Beretta; J. H. Griffin, *Biochemistry* **1980**, *19* (6), 1151–1160.
47. N. Jessani; Y. Liu; M. Humphrey; B. F. Cravatt, *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 10335–10340.
48. N. Jessani; M. Humphrey; W. H. McDonald; S. Niessen; K. Masuda; B. Gangadharan; J. R. Yates, III; B. M. Mueller; B. F. Cravatt, *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101* (38), 13756–13761.
49. K. P. Chiang; S. Niessen; A. Saghatelian; B. F. Cravatt, *Chem. Biol.* **2006**, *13* (10), 1041–1050.
50. Z. Pan; D. A. Jeffery; K. Chehade; J. Beltman; J. M. Clark; P. Grothaus; M. Bogyo; A. Baruch, *Bioorg. Med. Chem. Lett.* **2006**, *16* (11), 2882–2885.
51. S. Mahrus; C. S. Craik, *Chem. Biol.* **2005**, *12* (5), 567–577.
52. T. Nazif; M. Bogyo, *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98* (6), 2967–2972.
53. H. Ovaa; P. F. Van Swieten; B. M. Kessler; M. A. Leeuwenburgh; E. Fiebiger; A. M. Van Den Nieuwendijk; P. J. Galardy; G. A. Van Der Marel; H. L. Ploegh; H. S. Overkleeft, *Angew. Chem. Int. Ed. Engl.* **2003**, *42*, 3626–3629.
54. C. R. Berkers; M. Verdoes; E. Lichtman; E. Fiebiger; B. M. Kessler; K. C. Anderson; H. L. Ploegh; H. Ovaa; P. J. Galardy, *Nat. Methods* **2005**, *2* (5), 357–362.
55. B. D. Gelb; G. P. Shi; H. A. Chapman; R. J. Desnick, *Science* **1996**, *273* (5279), 1236–1238.
56. B. F. Sloane; S. Yan; I. Podgorski; B. E. Linebaugh; M. L. Cher; J. Mai; D. Cavallo-Medved; M. Sameni; J. Dosescu; K. Moin, *Semin. Cancer Biol.* **2005**, *15* (2), 149–157.
57. S. Yan; M. Sameni; B. F. Sloane, *Biol. Chem.* **1998**, *379* (2), 113–123.
58. B. R. Shenai; P. S. Sijwali; A. Singh; P. J. Rosenthal, *J. Biol. Chem.* **2000**, *275* (37), 29000–29010.
59. A. J. Barrett; A. A. Kembhavi; M. A. Brown; H. Kirschke; C. G. Knight; M. Tamai; K. Hanada, *Biochem. J.* **1982**, *201* (1), 189–198.
60. A. J. Barrett; A. A. Kembhavi; K. Hanada, *Acta Biol. Med. Ger.* **1981**, *40* (10–11), 1513–1517.
61. J. T. Palmer; D. Rasnick; J. L. Klaus; D. Bromme, *J. Med. Chem.* **1995**, *38* (17), 3193–3196.
62. E. Shaw, *Methods Enzymol.* **1994**, *244*, 649–656.
63. E. Shaw; H. Angliker; P. Rauber; B. Walker; P. Wikstrom, *Biomed. Biochim. Acta* **1986**, *45* (11–12), 1397–1403.
64. D. H. Pliura; B. J. Bonaventura; R. A. Smith; P. J. Coles; A. Krantz, *Biochem. J.* **1992**, *288* (Pt. 3), 759–762.
65. D. Greenbaum; K. F. Medzihradszky; A. Burlingame; M. Bogyo, *Chem. Biol.* **2000**, *7*, 569–581.
66. D. Greenbaum; A. Baruch; L. Hayrapetian; Z. Darula; A. Burlingame; K. F. Medzihradszky; M. Bogyo, *Mol. Cell. Proteomics* **2002**, *1*, 60–68.
67. L. Faleiro; R. Kobayashi; H. Fearnhead; Y. Lazebnik, *EMBO J.* **1997**, *16*, 2271–2281.
68. D. Kato; K. M. Boatright; A. B. Berger; T. Nazif; G. Blum; C. Ryan; K. A. Chehade; G. S. Salvesen; M. Bogyo, *Nat. Chem. Biol.* **2005**, *1* (1), 33–38.
69. N. A. Thornberry; E. P. Peterson; J. J. Zhao; A. D. Howard; P. R. Griffin; K. T. Chapman, *Biochemistry* **1994**, *33* (13), 3934–3940.
70. F. Yuan; S. H. Verhelst; G. Blum; L. M. Coussens; M. Bogyo, *J. Am. Chem. Soc.* **2006**, *128* (17), 5616–5617.
71. K. R. Love; A. Catic; C. Schlieker; H. L. Ploegh, *Nat. Chem. Biol.* **2007**, *3* (11), 697–705.
72. A. Borodovsky; H. Ovaa; N. Kolli; T. Gan-Erdene; K. D. Wilkinson; H. L. Ploegh; B. M. Kessler, *Chem. Biol.* **2002**, *9* (10), 1149–1159.
73. A. Borodovsky; H. Ovaa; W. J. Meester; E. S. Venanzi; M. S. Bogyo; B. G. Hekking; H. L. Ploegh; B. M. Kessler; H. S. Overkleeft, *ChemBioChem* **2005**, *6* (2), 287–291.
74. C. M. Overall; C. Lopez-Otin, *Nat. Rev. Cancer* **2002**, *2*, 657–672.
75. X. S. Peunte; L. M. Sanchez; C. M. Overall; C. Lopez-Otin, *Nat. Rev. Genet.* **2003**, *4*, 544–558.
76. A. Saghatelian; N. Jessani; A. Joseph; M. Humphrey; B. F. Cravatt, *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101* (27), 10000–10005.
77. E. A. Garbett; M. W. Reed; N. J. Brown, *Br. J. Cancer* **1999**, *81* (2), 287–293.
78. E. W. Chan; S. Chattopadhaya; R. C. Panicker; X. Huang; S. Q. Yao, *J. Am. Chem. Soc.* **2004**, *126* (44), 14435–14446.
79. M. Uttamchandani; J. Wang; J. Li; M. Hu; H. Sun; K. Y. Chen; K. Liu; S. Q. Yao, *J. Am. Chem. Soc.* **2007**, *129* (25), 7848–7858.
80. S. A. Sieber; S. Niessen; H. S. Hoover; B. F. Cravatt, *Nat. Chem. Biol.* **2006**, *2* (5), 274–281.
81. S. Minucci; P. G. Pelicci, *Nat. Rev. Cancer* **2006**, *6* (1), 38–51.
82. M. S. Finnin; J. R. Donigian; A. Cohen; V. M. Richon; R. A. Rifkind; P. A. Marks; R. Breslow; N. P. Pavletich, *Nature* **1999**, *401*, 188–193.
83. A. Vannini; C. Volpari; G. Filocamo; E. C. Casavola; M. Brunetti; D. Renzoni; P. Chakravarty; C. Paolini; R. De Francesco; P. Gallinari; C. Steinkuhler; Di S. Marco, *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101* (42), 15064–15069.
84. C. M. Salisbury; B. F. Cravatt, *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104* (4), 1171–1176.
85. C. M. Salisbury; B. F. Cravatt, *J. Am. Chem. Soc.* **2008**, *130* (7), 2184–2194.
86. D. J. Selkoe, *Trends Cell Biol.* **1998**, *8* (11), 447–453.
87. J. M. Travins; M. G. Bursavich; D. F. Veber; D. H. Rich, *Org. Lett.* **2001**, *3* (17), 2725–2728.
88. Y. M. Li; M. Xu; M. T. Lai; Q. Huang; J. L. Castro; J. DiMuzio-Mower; T. Harrison; C. Lellis; A. Nadin; J. G. Neduvelil; R. B. Register; M. K. Sardana; M. S. Shearman; A. L. Smith; X. P. Shi; K. C. Yin; J. A. Shafer; S. J. Gardell, *Nature* **2000**, *405*, 689–694.
89. L. M. Elphick; S. E. Lee; V. Gouverneur; D. J. Mann, *ACS Chem. Biol.* **2007**, *2* (5), 299–314.
90. G. Manning; D. B. Whyte; R. Martinez; T. Hunter; S. Sudarsanam, *Science* **2002**, *298* (5600), 1912–1934.
91. D. S. Krause; R. A. Van Etten, *N. Engl. J. Med.* **2005**, *353* (2), 172–187.
92. M. P. Wymann; R. Marone, *Curr. Opin. Cell. Biol.* **2005**, *17* (2), 141–149.
93. M. E. Noble; J. A. Endicott; L. N. Johnson, *Science* **2004**, *303* (5665), 1800–1805.
94. O. Fedorov; B. Marsden; V. Pogacic; P. Rellos; S. Muller; A. N. Bullock; J. Schwaller; M. Sundstrom; S. Knapp, *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104* (51), 20523–20528.
95. O. Fedorov; M. Sundstrom; B. Marsden; S. Knapp, *Drug Discov. Today* **2007**, *12* (9–10), 365–372.
96. M. P. Patricelli; A. K. Szardenings; M. Liyanage; T. K. Nomanbhoy; M. Wu; H. Weissig; A. Aban; D. Chun; S. Tanner; J. W. Kozarich, *Biochemistry* **2007**, *46* (2), 350–358.
97. Y. Liu; K. R. Shreder; W. Gai; S. Corral; D. K. Ferris; J. S. Rosenblum, *Chem. Biol.* **2005**, *12* (1), 99–107.

98. M. C. Yee; S. C. Fas; M. M. Stohlmeyer; T. J. Wandless; K. A. Cimprich, *J. Biol. Chem.* **2005**, *280* (32), 29053–29059.
99. J. A. Blair; D. Rauh; C. Kung; C. H. Yun; Q. W. Fan; H. Rode; C. Zhang; M. J. Eck; W. A. Weiss; K. M. Shokat, *Nat. Chem. Biol.* **2007**, *3* (4), 229–238.
100. M. S. Cohen; C. Zhang; K. M. Shokat; J. Taunton, *Science* **2005**, *308* (5726), 1318–1321.
101. M. S. Cohen; H. Hadjivassiliou; J. Taunton, *Nat. Chem. Biol.* **2007**, *3* (3), 156–160.
102. M. C. Hagenstein; J. H. Mussgnug; K. Lotte; R. Plessow; A. Brockhinke; O. Kruse; N. Sewald, *Angew. Chem. Int. Ed. Engl.* **2003**, *42*, 5635–5638.
103. N. K. Tonks, *Nat. Rev. Mol. Cell Biol.* **2006**, *7* (11), 833–846.
104. L. C. Lo; T. L. Pang; C. H. Kuo; Y. L. Chiang; H. Y. Wang; J. J. Lin, *J. Proteome Res.* **2002**, *1* (1), 35–40.
105. S. Kumar; B. Zhou; F. Liang; W. Q. Wang; Z. Huang; Z. Y. Zhang, *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101* (21), 7943–7948.
106. S. Kumar; B. Zhou; F. Liang; H. Yang; W. Q. Wang; Z. Y. Zhang, *J. Proteome Res.* **2006**, *5* (8), 1898–1905.
107. S. Liu; B. Zhou; H. Yang; Y. He; Z. X. Jiang; S. Kumar; L. Wu; Z. Y. Zhang, *J. Am. Chem. Soc.* **2008**, *130* (26), 8251–8260.
108. K. R. Shreder; Y. Liu; T. Nomanhboy; S. R. Fuller; M. S. Wong; W. Z. Gai; J. Wu; P. S. Leventhal; J. R. Lill; S. Corral, *Bioconjug. Chem.* **2004**, *15* (4), 790–798.
109. D. L. Zechel; S. G. Withers, *Acc. Chem. Res.* **2000**, *33* (1), 11–18.
110. C. S. Tsai; Y. K. Li; L. C. Lo, *Org. Lett.* **2002**, *4* (21), 3607–3610.
111. C. P. Lu; C. T. Ren; Y. N. Lai; S. H. Wu; W. M. Wang; J. Y. Chen; L. C. Lo, *Angew. Chem. Int. Ed. Engl.* **2005**, *44* (42), 6888–6892.
112. D. J. Vocadlo; C. R. Bertozzi, *Angew. Chem. Int. Ed. Engl.* **2004**, *43* (40), 5338–5342.
113. D. J. Vocadlo; H. C. Hang; E. J. Kim; J. A. Hanover; C. R. Bertozzi, *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100* (16), 9116–9121.
114. K. A. Stubbs; A. Scaffidi; A. W. Debowski; B. L. Mark; R. V. Stick; D. J. Vocadlo, *J. Am. Chem. Soc.* **2008**, *130* (1), 327–335.
115. O. Hekmat; Y. W. Kim; S. J. Williams; S. He; S. G. Withers, *J. Biol. Chem.* **2005**, *280* (42), 35126–35135.
116. I. G. Denisov; T. M. Makris; S. G. Sligar; I. Schlichting, *Chem. Rev.* **2005**, *105* (6), 2253–2277.
117. R. W. Brueggemeier; J. C. Hackett; E. S. Diaz-Cruz, *Endocr. Rev.* **2005**, *26* (3), 331–345.
118. J. C. Hackett; R. W. Brueggemeier; C. M. Hadad, *J. Am. Chem. Soc.* **2005**, *127* (14), 5224–5237.
119. A. T. Wright; B. F. Cravatt, *Chem. Biol.* **2007**, *14* (9), 1043–1051.
120. E. R. Vossenaar; A. J. Zendman; W. J. van Venrooij; G. J. Pruijn, *Bioessays* **2003**, *25* (11), 1106–1118.
121. Y. Luo; B. Knuckley; M. Bhatia; P. J. Pellechia; P. R. Thompson, *J. Am. Chem. Soc.* **2006**, *128* (45), 14468–14469.
122. G. C. Adam; B. F. Cravatt; E. J. Sorensen, *Chem. Biol.* **2001**, *8*, 81–95.
123. G. C. Adam; E. J. Sorensen; B. F. Cravatt, *Nat. Biotechnol.* **2002**, *20* (8), 805–809.
124. K. T. Barglow; B. F. Cravatt, *Chem. Biol.* **2004**, *11* (11), 1523–1531.
125. K. T. Barglow; B. F. Cravatt, *Angew. Chem. Int. Ed. Engl.* **2006**, *45* (44), 7408–7411.
126. E. Weerapana; G. M. Simon; B. F. Cravatt, *Nat. Chem. Biol.* **2008**, *4* (7), 405–407.
127. G. M. Cragg; D. J. Newman; K. M. Snader, *J. Nat. Prod.* **1997**, *60* (1), 52–60.
128. D. J. Newman; G. M. Cragg; K. M. Snader, *Nat. Prod. Rep.* **2000**, *17* (3), 215–234.
129. D. J. Newman; G. M. Cragg; K. M. Snader, *J. Nat. Prod.* **2003**, *66* (7), 1022–1037.
130. C. Drahl; B. F. Cravatt; E. J. Sorensen, *Angew. Chem. Int. Ed. Engl.* **2005**, *44* (36), 5788–5809.
131. N. O. Sykes; S. J. Macdonald; M. I. Page, *J. Med. Chem.* **2002**, *45* (13), 2850–2856.
132. M. J. Evans; A. Saghatelian; E. J. Sorensen; B. F. Cravatt, *Nat. Biotechnol.* **2005**, *23* (10), 1303–1307.
133. D. H. Kim; J. I. Park; S. J. Chung; J. D. Park; N. K. Park; J. H. Han, *Bioorg. Med. Chem.* **2002**, *10* (8), 2553–2560.
134. A. A. Tymiak; C. A. Culver; M. F. Malley; J. Z. Gougoutas, *J. Org. Chem.* **1985**, *50*, 5491–5495.
135. D. C. Aldridge; D. Giles; W. B. Turner, *J. Chem. Soc. Perkin Trans. 1* **1971**, *23*, 3888–3891.
136. R. L. Danheiser; J. S. Nowick, *J. Org. Chem.* **1991**, *56*, 1176–1185.
137. T. Böttcher; S. A. Sieber, *Angew. Chem. Int. Ed. Engl.* **2008**, *47* (24), 4600–4603.
138. Z. Wang; C. Gu; T. Colby; T. Shindo; R. Balamurugan; H. Waldmann; M. Kaiser; R. A. van der Hoorn, *Nat. Chem. Biol.* **2008**, *4* (9), 557–563.
139. P. Macheboeuf; C. Contreras-Martel; V. Job; O. Dideberg; A. Dessen, *FEMS Microbiol. Rev.* **2006**, *30* (5), 673–691.
140. C. T. Walsh, *Antibiotics: Actions, Origins, Resistance*; ASM Press: Washington, DC, 2003.
141. D. A. Preston; C. Y. Wu; L. C. Blaszczak; D. E. Seitz; N. G. Halligan, *Antimicrob. Agents Chemother.* **1990**, *34* (5), 718–721.
142. G. Zhao; T. I. Meier; S. D. Kahl; K. R. Gee; L. C. Blaszczak, *Antimicrob. Agents Chemother.* **1999**, *43* (5), 1124–1128.
143. I. Staub; S. A. Sieber, *J. Am. Chem. Soc.* **2008**, *130* (40), 13400–13409.
144. A. Arcaro; M. P. Wymann, *Biochem. J.* **1993**, *296* (Pt. 2), 297–301.
145. H. Yano; S. Nakanishi; K. Kimura; N. Hanai; Y. Saitoh; Y. Fukui; Y. Nonomura; Y. Matsuda, *J. Biol. Chem.* **1993**, *268* (34), 25846–25856.
146. V. Gupta; A. K. Ogawa; X. Du; K. N. Houk; R. W. Armstrong, *J. Med. Chem.* **1997**, *40* (20), 3199–3206.
147. L. Zhou; H. Yu; K. Chen, *Biomed. Environ. Sci.* **2002**, *15* (2), 166–171.
148. B. Sato; H. Nakajima; Y. Hori; M. Hino; S. Hashimoto; H. Terano, *J. Antibiot. (Tokyo)* **2000**, *53* (2), 204–206.
149. G. C. Adam; C. D. Vanderwal; E. J. Sorensen; B. F. Cravatt, *Angew. Chem. Int. Ed. Engl.* **2003**, *42*, 5480–5484.
150. A. L. MacKinnon; J. L. Garrison; R. S. Hegde; J. Taunton, *J. Am. Chem. Soc.* **2007**, *129* (47), 14560–14561.
151. N. Ramos-DeSimone; E. Hahn-Dantona; J. Sipley; H. Nagase; D. L. French; J. P. Quigley, *J. Biol. Chem.* **1999**, *274* (19), 13066–13076.
152. J. Lu; Y. J. Xiao; L. M. Baudhuin; G. Hong; Y. Xu, *J. Lipid. Res.* **2002**, *43* (3), 463–476.
153. J. A. Joyce; A. Baruch; K. Chehade; N. Meyer-Morse; E. Giraudo; F. Y. Tsai; D. C. Greenbaum; J. H. Hager; M. Bogyo; D. Hanahan, *Cancer Cell* **2004**, *5* (5), 443–453.
154. D. C. Greenbaum; A. Baruch; M. Grainger; Z. Bozdech; K. F. Medzihradszky; J. Engel; J. DeRisi; A. A. Holder; M. Bogyo, *Science* **2002**, *298* (5600), 2002–2006.
155. S. Arastu-Kapur; E. L. Ponder; U. P. Fonovic; S. Yeoh; F. Yuan; M. Fonovic; M. Grainger; C. I. Phillips; J. C. Powers; M. Bogyo, *Nat. Chem. Biol.* **2008**, *4* (3), 203–213.
156. D. Frees; S. N. Qazi; P. J. Hill; H. Ingmer, *Mol. Microbiol.* **2003**, *48* (6), 1565–1578.
157. T. Böttcher; S. A. Sieber, *J. Am. Chem. Soc.* **2008**, *130*, 14400–14401.

158. T. Böttcher; S. A. Sieber, *ChemBioChem* **2009**, *10*, 663–666.
159. H. C. Hang; J. Loureiro; E. Spooner; A. W. van der Velden; Y. M. Kim; A. M. Pollington; R. Maehr; M. N. Starnbach; H. L. Ploegh, *ACS Chem. Biol.* **2006**, *1* (11), 713–723.
160. N. Jessani; J. A. Young; S. L. Diaz; M. P. Patricelli; A. Varki; B. F. Cravatt, *Angew. Chem. Int. Ed. Engl.* **2005**, *44* (16), 2400–2403.
161. D. C. Greenbaum; W. D. Arnold; F. Lu; L. Hayrapetian; A. Baruch; J. Krumrine; S. Toba; K. Chehade; D. Bromme; I. D. Kuntz; M. Bogyo, *Chem. Biol.* **2002**, *9* (10), 1085–1094.
162. W. Li; J. L. Blankman; B. F. Cravatt, *J. Am. Chem. Soc.* **2007**, *129* (31), 9594–9595.
163. D. Leung; C. Hardouin; D. L. Boger; B. F. Cravatt, *Nat. Biotechnol.* **2003**, *21*, 687–691.
164. J. Zhang; R. E. Campbell; A. Y. Ting; R. Y. Tsien, *Nat. Rev. Mol. Cell Biol.* **2002**, *3* (12), 906–918.
165. R. N. Day; F. Schaufele, *Mol. Endocrinol.* **2005**, *19* (7), 1675–1686.
166. G. Blum; S. R. Mullins; K. Keren; M. Fonovic; C. Jedeszko; M. J. Rice; B. F. Sloane; M. Bogyo, *Nat. Chem. Biol.* **2005**, *1* (4), 203–209.
167. G. Blum; G. von Degenfeld; M. J. Merchant; H. M. Blau; M. Bogyo, *Nat. Chem. Biol.* **2007**, *3* (10), 668–677.

Biographical Sketches

Stephan A. Sieber studied chemistry at the universities of Marburg, Germany and Birmingham, England. He did his graduate studies, supported by a stipend of the National German National Academic Foundation, in the laboratories of Professor Marahiel, University of Marburg and Professor Christopher T. Walsh at Harvard Medical School, Boston, USA. In 2004, he obtained his Ph.D. and received the Friedrich Weygand award for his thesis work. Soon after his graduation he joined the group of Professor Benjamin F. Cravatt for his postdoctoral work for which he received a DFG stipend. In 2006 he started his independent research career at the University of Munich, Germany, funded by a DFG Emmy Noether grant.

Thomas Böttcher studied chemistry and biochemistry at the Ludwig-Maximilians-Universität (LMU) in Munich, Germany. He conducted his bachelor thesis at the Ruhr-Universität Bochum, Germany with Günter von Kiedrowski and received his B.Sc. in 2006 at the LMU. In the same year he joined the group of Stephan A. Sieber and started his Ph.D.

focused mainly on β-lactones as ABPP probes and highly specific ClpP inhibitors. Currently, he is a fellow of the German National Academic Foundation.

Isabell Staub studied chemistry at the Ludwig-Maximilians-University of Munich, Germany. She received her diploma degree under the guidance of Stephan Sieber in 2007 for her work on the synthesis and application of functionalized β-lactam antibiotics. In the same year she started her Ph.D. in the group of Stephan Sieber and continues her work on the application of probes based on β-lactam core structure in ABPP.

Ronald Orth studied biochemistry at the University of Bayreuth, Germany, where he completed his Diplom under the guidance of Carlo Unverzagt in 2006. After postgraduate studies with Sebastian Wesselborg at the University of Tübingen, Germany, he started his Ph.D. with Stephan Sieber in 2007 at the Ludwig-Maximilians-University in Munich, Germany. His thesis deals with the synthesis and application of various probes for ABPP.

9.18 Metabolic Studies Using the Retrobiosynthesis Concept – Theory, Technology, and Examples

Adelbert Bacher and Wolfgang Eisenreich, Technische Universität München, Garching, Germany

9.18.1 Isotopes in Metabolic Sciences

The heydays of metabolism research go long back to the twentieth century, with breathtaking discoveries of pathways such as glycolysis, the citric acid cycle, and photosynthesis that were later followed by the elucidation of the biosynthesis of complex metabolites such as vitamins, alkaloids, and antibiotics. That work, among other aspects, served as the basis for the elucidation and therapy of metabolic disease.

The emergence of molecular biology and, in its wake, the study of genomics, transcriptomics, and proteomics was accompanied by a decreased interest in metabolism.

More recently, there have been tendencies to complement the top-down approach where all aspects of living systems appear as the deployment of the genetic program by renewed efforts in the area of phenomenological study of metabolism including biosynthesis. Even the term ‘metabolomics’ has been introduced and may seem to confer a sense of renewed respectability and even urgency to the area, although the term is typically used to simply denote attempts to describe the chemical composition of living systems in quantitative terms.

In terms of methodology, the history of biosynthesis research is dominated by the interplay of natural products chemistry, enzymology, genetics, and the application of isotopes as tracers. Notably, isotope methods found their way into biosynthesis research almost immediately after deuterium and, with it, the physical principle of isotopology were discovered in the 1930s. Isotopes have had a tremendous impact on all aspects of biosynthesis research; a particularly notable example was the elucidation of carbon fixation in plant photosynthesis.[1]

The history of isotope technology in the area of biosynthesis research is essentially dominated by the emergence of physical methods for isotope production, detection, and quantification. Because deuterium was the first isotope that was discovered, pertinent biological studies started off in the domain of stable isotopes.[2,3]

Following the discovery of radioisotopes, interest moved rapidly in that direction.[4] The most important reason for that was the ease, sensitivity, and accuracy of radioisotope detection using relatively simple equipment such as a Geiger counter and, later on, liquid scintillation counters. Beta-emitting isotopes of hydrogen, carbon, sulfur, and phosphorus rapidly captured a commanding role in biosynthesis research. Only much later did stable isotopes get some of the limelight back when their observation and quantification was progressively improved by the emergent technologies of nuclear magnetic resonance (NMR) and mass spectrometry. Notably, however, both techniques, but especially NMR, come with hefty price tags for the required hardware.

For application as biosynthetic tracers, organic molecules can be labeled in single positions or in multiple positions using single or multiple isotope types. Experiments can even be conducted using complex mixtures of isotope-labeled compounds as tracers, but that approach has been used infrequently up until now. The chemical complexity of tracers can be minimal, starting with 1-carbon compounds such as $^{14}CO_2$ that has been used with immense success in the elucidation of carbon fixation in photosynthesis by higher plants.[1] On the other hand, there is no apparent limit to the potential chemical complexity of isotope-labeled tracer compounds.

Analysis of the isotope distribution in target compounds can be, and frequently been, limited to measuring bulk specific activity (in the case of radioisotopes) or the overall abundance (in the case of stable isotopes) of the tracer isotope. Alternatively, the topological distribution of the isotope label can be narrowed down by chemical degradation in the case of radioactive tracers. In case of stable isotopes, the topology of isotope distribution in the target molecule can be assessed in considerable detail by NMR spectrometry, mass spectrometry, or a combination of these methods.

9.18.2 Isotope Propagation in Metabolic Networks

The technology of isotope application for the study of problems in biosynthesis is typically viewed as simple and self-explanatory. When an isotope-labeled organic compound (typically designated as tracer) is found to be incorporated into a metabolite it is straightforward to assume that a metabolic connection must exist between the tracer and the target compound, and all that remains to be done is to construct a plausible sequence of chemical reaction steps explaining that connection. Hence, the application of isotopes in biosynthesis has been largely the domain of serendipity. In light of the success of this approach, there was not much impetus for theoretical treatment.

In contrast to that quasi-linear concept for the interpretation of isotope studies, it is worth noting that cells comprise at least hundreds (in bacteria) to thousands (in eukaryotes) of low-molecular-weight compounds. In their entirety, the low-molecular-weight components of cells constitute a complex metabolic network. Each specific metabolite can be viewed as a node in that network, and individual nodes are connected to other nodes by chemical transformations that are mediated by the vast number of enzymes (at least hundreds in bacteria and thousands in eukaryotes). In terms of network theory, cellular metabolism can be described by the so-called scale-free or small-world networks where any two nodes are connected to each other via a relatively low number of intermediary nodes. If any tracer is added to a biological system, the isotope label can spread from that source to various other nodes that can be viewed as sinks for the that label. A given source can contribute label to a variety of nodes; on the other hand, a given sink can acquire an isotope label from various sources. Clearly, the transfer of isotope label from any source to any sink proves the existence of at least one metabolic connection between these two nodes; however, it may (and frequently does) implicate the existence of multiple paths from source to sink. The aim of many typical isotope incorporation experiments, that is, to establish the metabolic route from source to sink, becomes a complex problem in light of the network topology of metabolism in cells and organisms.

The naive, quasi-one-dimensional approach to the interpretation of isotope studies can be justified in cases where the transfer from a given source to a given sink proceeds with high efficiency as indicated by high relative specific radioactivity or high relative abundance of the sink metabolite. In those cases, it appears plausible that source and sink are connected by at least one route via a small number of nodes; that route can then be construed as the pathway from source to sink, and the source compound can be addressed as the precursor of the sink compound in the conventional sense of the words. Frequently, however, the incorporation rates are only moderate or even low. Although such studies would have to be labeled routinely as inconclusive, researchers typically prefer to justify their efforts by a wide variety of assumptions designed to explain that a given source compound is the specific precursor (rather than one of various putative precursors) of the sink compound.

9.18.3 A Quasi-Steady-State Concept of Isotope Distribution

Stable as well as radioactive isotopes of carbon and hydrogen are natural constituents of living organisms. The stable isotopes, ^{2}H and ^{13}C occur naturally at abundances of 0.015 and 1.1%, respectively. Geophysical processes as well as metabolic transformations have a low degree of isotope selectivity, and the ^{2}H and ^{13}C content of organic matter therefore reflects its origin. Although these small differences are exploited for analytical purposes, for example, in order to determine the geographic or synthetic origin of biological materials, they are beyond the scope of this chapter.

The biosphere experiences a constant influx of radioactive isotopes, ^{3}H and ^{14}C, which is generated by nuclear processes in the atmosphere that are driven by cosmic radiation and washed down by atmospheric processes and incorporated into biomatter via $^{14}CO_2$ and $^{3}H_2O$. Like the stable isotopes described above, these are distributed almost at random in biomass of recent origin; measuring the ^{14}C content of a biomasses of ancient origin provides an important method of age determination that is not discussed in this chapter.

The close-to-random distribution of these isotopes can be perturbed experimentally by adding radioactive or stable isotope-labeled compounds to living cells or organisms. In fact, the specific enrichment of individual cellular components can be increased experimentally by many orders of magnitude by adding labeled metabolites. Such a perturbation triggers a relaxation process whereby the isotope label spreads to virtually all nodes of the metabolic network, albeit at rates that can again differ over a wide range. If these relaxation processes were to proceed for an infinite time in a closed system, the result should be a novel quasi-equilibrium distribution at an increased abundance level. In practice, however, isotope incorporation studies are conducted over relatively short periods that are insufficient in reaching a global steady-state distribution of the tracer.

9.18.4 A Rational Perturbation/Relaxation Strategy for Assessment of Metabolic Pathways

Based on the arguments considered above, a rigorous analysis of isotope flux should involve the appearance of isotope label at all or many nodes of the metabolic network as a function of time, after an initial perturbation of the quasi-steady-state distribution that was engineered by the addition of an isotope-labeled compound or a mixture of isotope-labeled compounds. The resulting data could then be translated into rates of isotope flux from a give source to many different potential sinks. A high flux rate from the source node to a given sink would translate into a relatively close precursor relationship. Any conclusions on biosynthetic pathways would be based on a comparative analysis of flux rates to and of labeling patterns at different nodes. The experimental setup could be described as a multidimensional perturbation/relaxation strategy.

In practical terms, such a comprehensive analysis is hardly possible. However, as shown in the examples given below, the comparison of label flux to a small number of metabolites or even to different substructure components of a given metabolite can provide more detailed information than the measurement of flux from a single source node to a single sink metabolite.

9.18.5 The Retrobiosynthesis Concept in Practical Terms

The building blocks of all metabolites, regardless of their structural complexity, are invariably derived from the metabolite pools of primary intermediary metabolism including the major anabolic and catabolic pathway in a given organism, such as the basic pathways for the biosynthesis and degradation of carbohydrates, carboxylic acids, amino acids, nucleic acid building blocks, and basic lipid compounds. In case of a structurally complex metabolite, it is obviously useful to determine the building blocks as a first step in biosynthetic studies. Regardless of the structure of the isotope-labeled precursor, the building blocks of complex metabolites can be narrowed down rapidly by a comparing the labeling patterns of the metabolite under study with the patterns of basic metabolites such as carbohydrates and amino acids from the same experiment. Based on available data on central intermediary metabolism, the labeling data of carbohydrates and amino acids also enable the reconstruction of other labeling patterns of central intermediary metabolism, such as dicarboxylic acids and keto acids. The principles of this retrodictive approach have been explained in detail elsewhere.[5] Essential correlations are summarized in **Figure 1**. As a specific example, consider the labeling patterns of aromatic amino acids in plants and in microorganisms endowed with the aromatic amino acid pathway. Because phenylalanine, tyrosine, and tryptophan are exclusively synthesized via the shikimate pathway, it is easily seen that the side chains of all three amino acids (carbon atoms a–c) must have an unequivocal biosynthetic relationship with the phosphoenol pyruvate pool (**Figure 2**). Moreover, the labeling patterns of phosphoenol

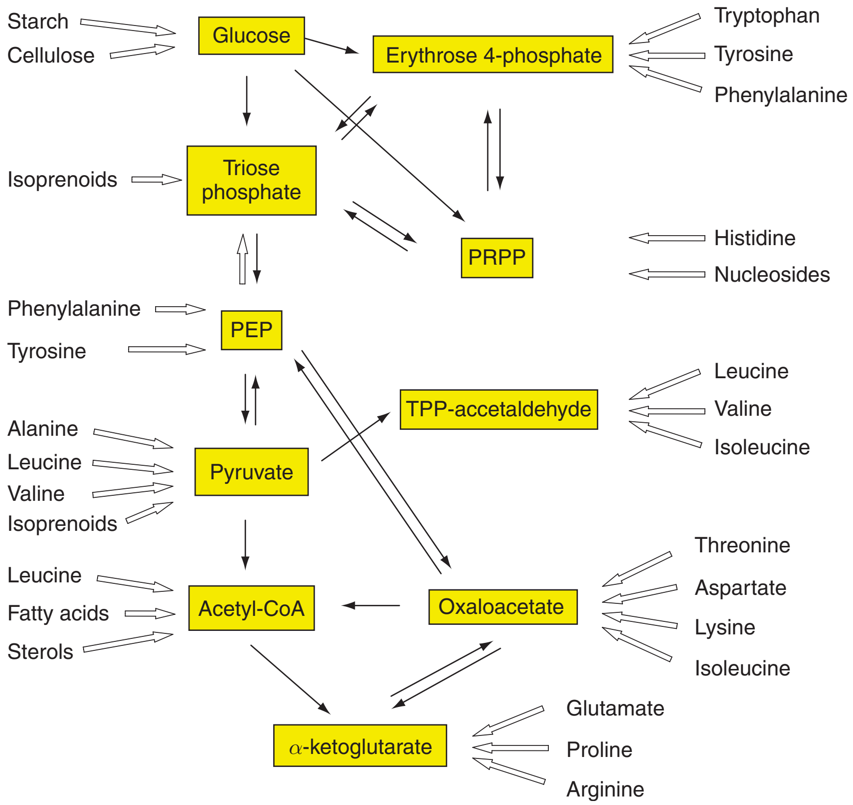


Figure 1 The retrobiosynthetic principle. Labeling patterns of central metabolic intermediates (shown in yellow boxes) are reconstructed from the labeling patterns of sink metabolites, such as protein-derived amino acids, storage metabolites (starch and lipids), cellulose, isoprenoids, or RNA-derived nucleosides. The reconstruction is symbolized by retro arrows following the principles of retrosynthesis in synthetic organic chemistry. The figure is based on known biosynthetic pathways of amino acids, starch, cellulose, nucleosides, and isoprenoids in plants. The profiles of the central metabolites can then be used for predictions of the labeling patterns of secondary metabolites. In comparison with the observed labeling patterns of the target compounds, hypothetical pathways can be falsified on this basis.

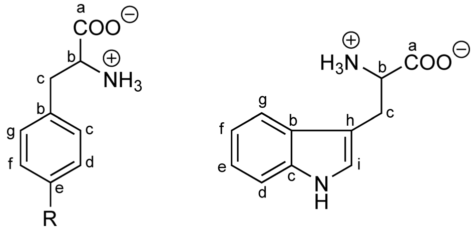


Figure 2 Atom mapping in aromatic amino acids. Atoms a–c are biosynthetically equivalent to atoms 1–3 in phosphoenol pyruvate, atoms d–g are equivalent to carbon atoms 1–4 of erythrose 4-phosphate, and atoms i–h are equivalent to carbon atoms 1–2 in ribose 5-phosphate.

pyruvate and erythrose 4-phosphate must be reflected in the benzenoid ring (carbon atoms b, c, and d–g, respectively). Finally, the carbon atoms 1 and 2 of ribose are reflected by carbon atoms i and h of the indole moiety in tryptophan (**Figure 2**).

Because the turnover of many proteins is relatively low, their amino acid constituents can be viewed as quasi-fossilized material. Subsequent to hydrolysis of cell protein, amino acids can be isolated relatively easily, and their isotope compositions can be determined. Deconvolution of their labeling patterns affords time-averaged labeling patterns for central intermediary metabolites such as acetyl-CoA, pyruvate, oxaloacetate, α-ketoglutarate, and glyceraldehyde phosphate on the basis of the known biosynthetic pathways of amino acids in the organism under study (**Figure 1**). The best conditions for this deconvolution process are found in auxotrophic organisms that biosynthesize all amino acids *de novo*, but even with heterotrophs, abundant information can still be obtained from those amino acids that are actually biosynthesized *de novo*. The labeling patterns of the reconstructed central metabolites can then serve as the basis for the deconvolution of the labeling patterns of complex metabolites to discover their basic building blocks. Similarly, the patterns of other sink metabolites, such as starch or cellulose in the plant example, lipids, and terpenoids can be used to reconstruct the labeling profiles in the respective precursors (**Figure 1**).

The approach described above has some resemblance with the concept of retrosynthesis. Briefly, in order to develop a strategy for the chemical synthesis of any complex organic structure, the target structure is dissected, on paper, by the disruption of bonds that can be generated by known chemical principles. This retroactive puzzle game is played until the target structure can be assembled from known building blocks by established chemical principles. The procedure, *per se*, is a time-honored concept in organic chemistry, but has been formalized and refined in recent decades under the heading of retrosynthesis.[6] The retrobiosynthetic concept, in turn, is an attempt to formalize concepts for the rapid discovery of the building blocks that are used in actual biosynthetic pathways.

9.18.6 Isotopologue Space

As described above, investigators using isotope incorporation strategies frequently found it sufficient to measure the overall specific incorporation rate for a given tracer compound into a given target molecule. More sophisticated approaches tried to locate the isotope label to specific positions of the target molecule, either in qualitative or in semiquantitative terms. In reality, it rarely happens that the tracer isotope is transferred to one specific position or any one group of specific positions in the target molecule. In reality, the complex interplay of catabolic and anabolic pathways in a given metabolic network enables the transfer of isotope from a given source node to a given sink node via more than one pathway, although the main flow may proceed via one specific main road.

It is now in order to discuss in detail the isotopologue composition of natural organic matter. As stated, organic matter invariably consists of about 98.9% ^{12}C, 1.1% ^{13}C, and traces of ^{14}C. The isotopes are distributed almost randomly to different molecular positions (again, biochemical reactions involved in the formation of biomatter have a low degree of isotope selectivity that can be exploited for analytical purposes but can be disregarded for the purpose of the present discussion that is focused on experiments where much larger perturbations of the quasi-equilibrium distribution of isotopes are generated by the addition of specific, isotope-labeled compounds). Natural organic matter is, on closer consideration, a mixture of many different isotopologues. For example, consider a compound with six carbon atoms such as glucose, the most abundant species representing about 93% consists exclusively of ^{12}C. Each of six isotopologues comprising one single ^{13}C atom accounts for about 1.1 mol%. Multiply labeled isotopologues are rare. Thus, each of the double-labeled isotopologues has an abundance of about 0.01 mol%, and the molecular species consisting exclusively of ^{13}C has an approximate abundance of about 10^{-10} mol% (about 1 ppt).

Considering only ^{12}C and ^{13}C, the number of isotopologues for any compound with n carbon atoms is 2^n. Total numbers of possible isotopologues are increased considerably if ^{2}H, ^{15}N, ^{17}O, and ^{18}O are taken into account. The entirety of all possible isotopologues of a given compound defines its isotopologue space.[7–9]

With tracer compounds carrying a single ^{13}C label in a single position at an abundance of close to 100%, the maximum perturbation of the natural isotopologue distribution is less than two orders of magnitude (the natural abundance of the respective isotopologue being 1.1% and the maximum possible value being less than 100%).

Much larger perturbations of the natural quasi-steady state of isotopologue distribution can be achieved with multiply labeled compounds. For example, since the natural abundance of [U-$^{13}C_6$]glucose is only about 1 ppt (see above), its abundance can be experimentally enhanced by more than 11 orders of magnitude by the addition of [U-$^{13}C_6$]glucose to the experimental system under study. The very large dynamic range of perturbations that can be generated with multiply labeled compounds is one, although not the only reason for their preferred use in perturbation/relaxation experiments.

An initial perturbation caused by the addition of one or more single-labeled or multiply labeled compounds will result in the consecutive perturbation of multiple metabolite pools as a consequence of the large number of metabolic transformations that are enabled by the enzyme catalysts of a given cell or organism. These processes constitute the relaxation process that need to be followed in quantitative terms in order to achieve an in-depth description of the relaxation process that has been triggered by the addition of the isotopic tracer(s) in an isotope incorporation experiment, as opposed to interpretations based on serendipity and educated guessing.

9.18.7 Quantitative Assessment of Isotopologue Abundance

The discussion of isotopologue space and its relevance in tracer experiments would be meaningless without robust technology to establish the isotopologue composition of molecules in sufficient detail. More specifically, the technology required is one that yields quantitative values for each isotopologue that is present in an experimental sample at a level exceeding a certain threshold. The required information is the isotopologue excess, above the level of the natural abundance of a given isotopologue, in a given experimental sample. For example, in tracer experiments with ^{13}C-labeled compounds, it is possible with present-day technology to reliably assess a 0.3 mol% isotopologue excess in case of isotopologues carrying a single ^{13}C atom and a 0.1 mol% isotopologue excess in case of multiply labeled isotopologues.

Radiolabeling experiments are frequently limited to the determination of relative specific activity in the target compound, thus providing no isotopologue information whatsoever. Even chemical degradation using elaborate degradation protocols can at best determine quantitative values for groups of isotopologues, but not individual isotopologues.

Advanced isotope quantification is best achieved by NMR spectroscopy using stable isotopes as tracers. Mass spectrometry can afford additional information. As a stand-alone technique, mass spectrometry can be used to determine groups of isotopologues, but it is poorly suited for measuring individual isotopologues in isotopologue mixtures.

Generally, NMR spectroscopy is the optimum approach for experiments using multiply ^{13}C-labeled tracer compounds. The technical details have been described in detail elsewhere.[9,10] Briefly, the presence of multiple ^{13}C atoms can result in signal splitting by scalar coupling. The ^{13}C^{13}C coupling constants are typically in the range of 35–50 Hz if the respective ^{13}C atoms are connected via a single bond. Coupling constants become progressively smaller with increasing numbers of covalent bonds between the respective ^{13}C atoms. Because the experimental limit for the resolution of coupling patterns is in the range of 1 Hz, the maximum range of observable ^{13}C^{13}C coupling can extend up to three to six covalent bonds, depending on molecular structure.

Groups of two connected ^{13}C atoms afford doublet signals, whereas larger numbers of adjacent ^{13}C atoms within the range of view can afford complex multiplets. The quantitative analysis of spectra comprising mixtures of ^{13}C isotopologues is well beyond the scope of this chapter. The best approach may be to study specific examples where the technique has been described in detail (see references quoted in Section 9.18.8).

9.18.8 Application of the Retrobiosynthesis Concept to Studies in Biosynthesis

During a period of more than six decades, isotope experiments have been employed for biosynthetic problems by countless biochemists. Aspects of the retrobiosynthetic concept as summarized above have been realized and exploited in the course of many studies.[11,12] The elements of the retrobiosynthetic concept, *per se*, are not novel. Instead, the concept is an attempt to sum up these elements in a coherent theory centered around perturbation/relaxation strategies in complex isotopologue space. In order to illustrate the history of the development and scope of the retrobiosynthetic concept, the examples described below are taken predominantly from the research of the present authors. This may be justified as the emphasis is on exemplification of the principles and techniques, rather than on specific pathways. It should be noted that the aim of this review is to illustrate the retrobiosynthetic method, not the comprehensive treatment of the biosynthetic pathways selected as examples. In other words, this review highlights certain episodes from the unfolding biosynthetic pathways under consideration. Research on many of these pathways has developed further, after the publication of the quoted articles that were designed as mere stepping stones and should be viewed as such. Notably, the entire scientific literature on certain topics that are addressed below comprises dozens or even hundreds of articles; for example, a database search for riboflavin biosynthesis retrieves more than 400 citations.

9.18.8.1 The Retrobiosynthetic Method Evolved in the Context of Studies on Riboflavin Biosynthesis

Work on the biosynthesis of riboflavin has been causal for the present authors' search for a unifying concept of isotope technology as set out in the retrobiosynthetic concept. Already in the 1940s, it had been noted that certain microorganisms can produce riboflavin in significant excess over their metabolic requirement. The pioneering finding by MacLaren that microbial flavinogenesis can be enhanced by adding purines to the culture medium was conducive to the purine hypothesis of riboflavin biosynthesis.[13] In due course, isotope incorporation studies by a number of research groups documented beyond doubt that all nitrogen atoms and all carbon atoms, with the exception of C-8, can be transferred from adenine and related purines to riboflavin (for review, see Plaut *et al.*[14]). However, it remained unknown which specific purine metabolite serves as the direct starting point of the riboflavin biosynthetic pathway. Because the typical nucleic acid building blocks are subject to efficient turnover via a multiplicity of salvage pathway, it was not easy to resolve the problem with additional isotope incorporation experiments.

9.18.8.1.1 Using nucleosides as reference metabolites – Structure of the purine precursor of riboflavin

The closest structural similarity with riboflavin in the group of naturally occurring purine derivatives exists with the xanthine family, which shares with riboflavin the carbonyl groups at the pyrimidine ring, whereas the substitution patterns of adenine and guanine are different from that of the vitamin. In order to restrict the metabolic flux of xanthine derivatives, experiments were conducted with a mutant of *Aerobacter* (*Klebsiella*) *aerogenes* that was unable to convert xanthosine monophosphate into guanosine monophosphate (GMP); this genetic defect could also be expected to restrict the conversion of xanthine monophosphate into compounds of the guanine series. Incorporation experiments were then performed using various ^{14}C-labeled purines, respectively.[15] In order to directly measure the interconversion of purines in these experiments, the transfer of radioactivity into biosynthetic xanthosine was measured, in addition to the measurement of label transfer into riboflavin. In that way, it was established that no radioactivity from ^{14}C-labeled adenine, hypoxanthine, or xanthine was incorporated into riboflavin, whereas all of these compounds were efficiently converted to xanthosine. On the other hand, [2-^{14}C]guanine was incorporated into riboflavin without significant dilution, whereas only traces of radioactive label were incorporated into xanthosine. The results left no doubt that the biosynthesis of the vitamin starts at the level of a guanine compound. This meant that the amino substituent present on the pyrimidine ring of guanine had to be replaced at some later step by a carbonyl group. This was later confirmed by the discovery of enzymes catalyzing the hydrolytic removal of the position 2 amino substituent from intermediates of the riboflavin pathway (for review, see Fischer and Bacher[16]). The crucial aspect of the experiments was the measurement of isotope transfer into xanthosine in order to gain a reference value from which the utilization of labeled purines could be gleaned unequivocally, despite the complex network of pathways for the *de novo* formation and the salvage of purines.

A similar strategy was used to analyze the origin of the ribityl side chain of the vitamin.[17] Stereochemical aspects suggested an origin from ribose. However, it was unknown whether the ribose side chain of the purine nucleoside or nucleotide precursor was carried all the way through to the vitamin or whether a ribose was later linked to with a heterocylic intermediate of the riboflavin pathway that was obtained from the initial committed purine precursor. Again, it was decided to limit the metabolic interconversion of purines by mutations. Specifically, a *Salmonella typhimurium* strain was used that was deficient of inosine monophosphate (IMP) dehydrogenase, purine nucleoside phosphorylase, and purine nucleotide pyrophosphorylase. As a consequence thereof, it was unable to grow on guanine as the sole purine source, whereas growth on guanosine was possible because the nucleoside could be converted into GMP by a resident nucleoside kinase.

The mutant incorporated [1′,2′,3′,4′,5′-^{13}C]guanosine into riboflavin and GMP (obtained by hydrolyzing cellular RNA) without dilution.[17] The isolated compounds were exclusively labeled in the ribityl and ribosyl side chains, respectively. Adenosine monophosphate (AMP) and cytidine monophosphate (CMP) (obtained by hydrolysis of cellular RNA) were not labeled, because ribose could not be diverted from the proffered guanosine to the general pentose phosphate pool due to the absence of the purine salvage pathway enzymes, purine nucleoside phosphorylase, and purine nucleotide pyrophosphorylase. Further control experiments confirmed that radioactivity from [2-^{14}C]guanosine was incorporated into riboflavin and GMP without dilution; the isolated compounds were exclusively labeled in the isoalloxazine and guanine moiety, respectively, and AMP and CMP were again unlabeled. In conjunction with the discovery[18] of an enzyme catalyzing the conversion of guanosine triphosphate (GTP) (**1**) into 2,5-diamino-6-ribosylamine-4(3*H*)-pyrimidinone phosphate (**2**) (**Figure 3**), these results showed conclusively that the first committed step in the biosynthesis starts from GTP, whose ribose side chain is transformed into the ribityl chain of the vitamin (**7**) (for review, see Fischer and Bacher[16]).

It may be relevant to note that a different strategy would have been possible for the isotope incorporation studies on the riboflavin side chain if present-day instrumentation had been available. Briefly, a modern version of the experiment could start from a dual-labeled guanosine derivative such as [1′-^{13}C$_1$,U-^{15}N$_4$]guanosine or [1′,2′,3′,4′,5′-^{13}C$_5$,8-^{15}N$_1$]guanosine. ^{13}C NMR spectroscopy of the resulting metabolites (riboflavin and nucleosides derived from RNA) would have been able to quantitatively assess the retention of the ribose chain of the proffered precursor on the basis of the ^{13}C^{15}N coupling pattern. In fact, such a strategy was in fact used later on to study the biosynthesis of molybdopterin (see below).

9.18.8.1.2 Retroanalysis by intramolecular pattern comparison – Origin of the xylene ring in riboflavin

Even prior to the elucidation of the first committed step of the riboflavin pathway, it had been shown that the benzenoid ring of riboflavin is assembled from two identical 4-carbon precursors. More specifically, the final step in the biosynthesis of the vitamin involves a dismutation of 6,7-dimethyl-8-ribityllumazine (**6**), where one of the substrate molecules serves as donor and the other as acceptor of a 4-carbon segment.[19,20] 6,7-Dimethyl-8-ribityllumazine, in turn, is formed in the penultimate step of the biosynthetic pathway from 5-amino-6-ribitylamino-4(3*H*)-pyrimidinedione (**3**), an intermediate that is obtained from the product of GTP cyclohydrolase II by a sequence of deamination, side chain reduction, and dephosphorylation (**Figure 3**). The nature of the 4-carbon precursor required for the formation of 6,7-dimethyl-8-ribityllumazine (**6**) from 5-amino-6-ribitylamino-4(3*H*)-pyrimidinedione (**3**) remained controversial for quite a long period, with working hypotheses including, but not limited to, tetroses, pentoses, and acetoin.

Because the ribityl side chain of the vitamin had already been shown to be derived from phosphoribosyl pyrophosphate via GTP (**1**), it could serve as an intramolecular reference for retrobiosynthetic analysis.

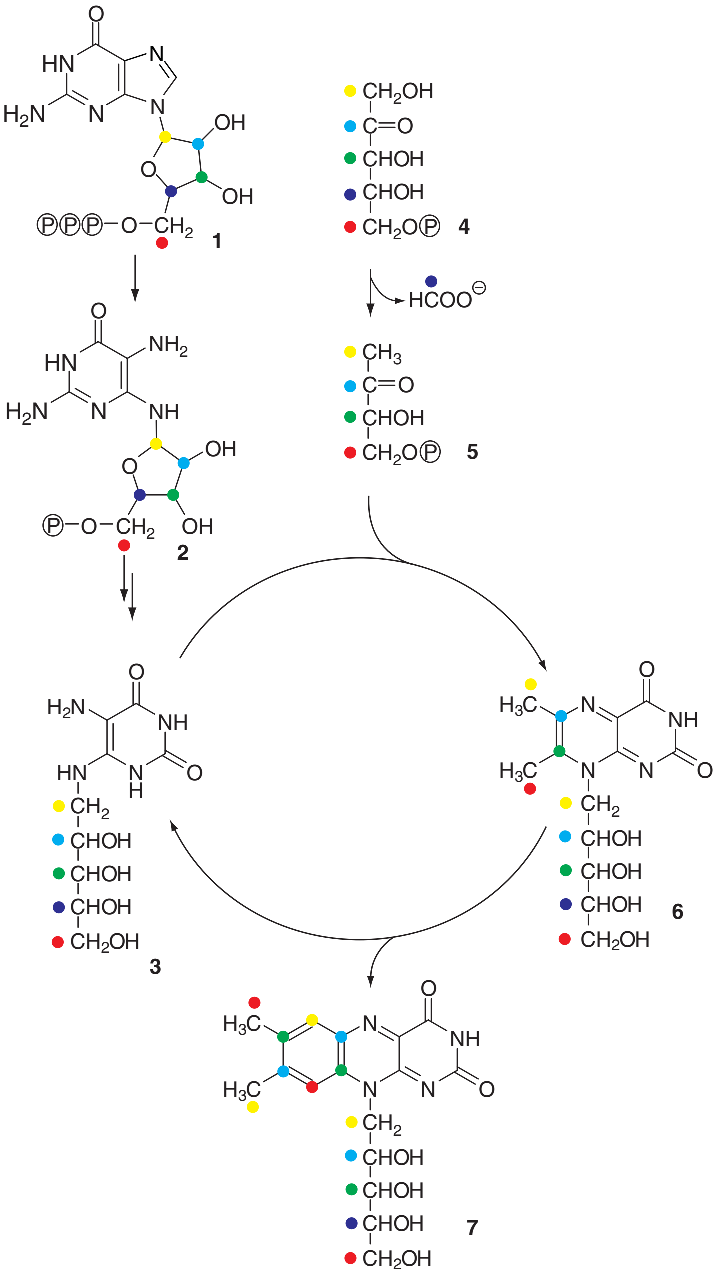


Figure 3 Biosynthesis of riboflavin (7). Biosynthetically equivalent atoms are indicated by colored dots. For details, see text.

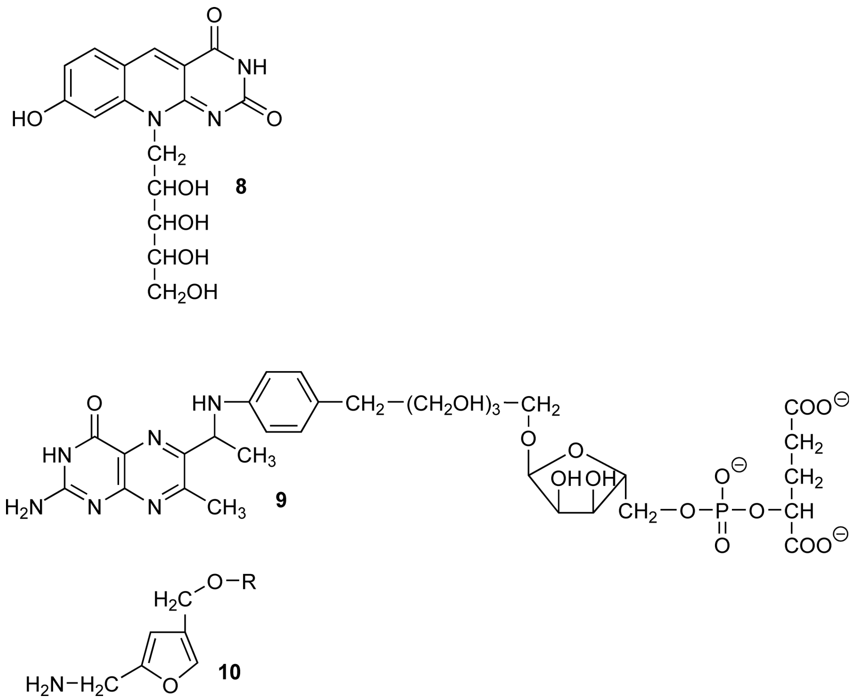
The experimental strategy involved feeding experiments with a variety of ^{13}C-labeled compounds, including acetate, glucose, glycerol, and ribose. Biosynthetic riboflavin was analyzed by ^{13}C NMR spectroscopy, affording the labeling patterns of the ribityl side chain and of the elusive 4-carbon precursor that could be extracted on basis of the known regioselectivity of riboflavin synthase – the terminal enzyme of the riboflavin pathway (notably, the two 4-carbon segments have inverse orientations in riboflavin). Each experiment in a relatively large series using different tracers and different labeling patterns of each respective precursor invariably showed a close correspondence between the labeling patterns of carbon atoms 1′ to 3′ of the ribityl side chain with three consecutive carbon atoms of the 4-carbon precursor.[21–25] Surprisingly, however, the fourth carbon atom of the 4-carbon precursor followed the labeling pattern of carbon 5′ of the ribityl side chain

(Figure 3). This unexpected result was best explained by the hypothesis that the 4-carbon precursor was derived from the pentose pool by an intramolecular rearrangement, resulting in the extrusion of carbon 4 and the ensuing linkage of C-5 to C-3 of the pentose precursor. The most direct evidence was obtained by an experiment using [1,3-$^{13}C_2$]glycerol as tracer. In the biosynthetic riboflavin from that fermentation, each of the two 4-carbon segments contained two ^{13}C atoms that were directly linked, whereas two ^{13}C atoms could be incorporated into the 3′ and 5′ position of the ribityl side chain of biosynthetic riboflavin, linked by a ^{12}C atom in position 4′. Retrosynthetic deconstruction suggested a sigmatropic rearrangement of ribulose 5-phosphate (4) with subsequent elimination of carbon atom 4 as a formate. With that information in hand, it was possible to identify the cognate enzyme, 3,4-dihydroxy-2-butanone 4-phosphate synthase affording 5.[26–28]

9.18.8.2 Small Precursor Molecules Such as Acetate Can Be Sufficient for Retroanalysis – Biosynthesis of Heterocyclic Coenzymes in Methanogenic Bacteria

Methanogenic bacteria can generate energy by the reduction of CO_2 or of simple carboxylic acids with elemental hydrogen. This unique metabolic strategy implicates the use of a number of special coenzymes occurring exclusively or predominantly in this taxonomic group. Whereas simple carboxylates such as acetate are avidly used by methanogens, the limited capacity for uptake of structurally complex precursors can be a problem for metabolic studies. Despite these limitations, ^{13}C-labeled acetate could be used to identify the building blocks for a number of methanogenic coenzymes by comparing it with nucleosides obtained by hydrolysis of RNA and amino acids obtained by hydrolysis of protein.

This approach was used to show that the biosynthesis of riboflavin in methanogenic bacteria uses the same building blocks as in eubacteria, fungi, and plants.[29] The pyrimidine ring of the deazaflavin chromophore (8, Figure 4) of the methanogenic coenzyme F420 could be shown to have a purine origin.[29,30] Later, it could be elaborated that the biosynthesis of deazaflavin branches off the riboflavin pathway at the level of 5-amino-6-ribitylamino-2,4(1*H*,3*H*)-pyrimidinedione (3, cf. Figure 3).[31] The carbocyclic ring of the deazaflavin


8

9

10

Figure 4 Structures of methanogenic coenzymes: (8), chromophore of the deazaflavin, coenzyme F420; (9), methanopterin; (10), furane moiety of methanofuran.

chromophore was shown to be derived from a shikimate pathway derivative.[29,30] Later, the origin of that ring could be narrowed down to 4-hydroxyphenylpyruvate.[31]

The same experimental approach enabled the building blocks of methanopterin (**9**, **Figure 4**).[32–34] Specifically, a hypothetical labeling pattern for the pteridine moiety could be easily predicted on the basis of the labeling patterns of the purine and ribose moiety and was found to agree with the observed labeling pattern. The ribose labeling pattern could also be used to confirm the hypothesis that the ribityl side chain of the ribitylaniline moiety is derived from the pentose pool. The aniline moiety could be shown to be derived from a shikimate dervivate, that is, 4-aminobenzoate, as the rules to predict the labeling of tetrose phosphate were already well established.

A closely similar approach enabled the identification of two triose phosphate moieties as the building blocks of the furane moiety in methanofuran (**10**).[35,36]

9.18.8.3 Identifying Basic Building Blocks in Difficult Cases – The Concept of Coherent Label Transfer

All complex metabolites are ultimately mosaics stitched together from basic building blocks derived from the central metabolite pools, including carbohydrates and carboxylic acids as well as amino acids and their respective biosynthetic precursors. In principle, it should be a simple task to identify at least the carbon skeleton of each respective building block by a systematic search for the largest units with contiguous ^{13}C labeling that can be contributed to the metabolite under study. If, for example, a block of four contiguous ^{13}C atoms can be contributed from any multiply labeled precursor proffered on a high background of unlabeled nutrients, it follows that the precursor supplying the respective structural motif comprised a minimum of four carbon atoms.

Numerous carbohydrates, carboxylic acids, and amino acids carrying multiple ^{13}C labels are now routinely available from commercial sources at acceptable prices. However, it is important to note that the intermediates of central intermediary metabolism are subject to rapid metabolic turnover in the densely interconnected central part of the metabolic network. More specifically, the isotopologue pattern of a proffered precursor can be modulated by passing through metabolic cycles involving the catabolic disruption of the proffered precursor into fragments and its subsequent regeneration from these fragments by anabolic processes. For example, animals rapidly convert proffered [U-^{13}C$_6$]glucose into complex isotopologue mixtures comprising about eight multiply ^{13}C-labeled glucose species in significant abundance.[37] Even so, the incorporation at significant abundance of a fragment of *n* consecutive ^{13}C atoms is sufficient to show that the specific precursor unit must have had a minimum size of *n* carbon atoms. Moreover, retrobiosynthetic comparison between the labeling patterns of different metabolites can be used to obtain estimates of the degree of coherence fragmentation by cyclic metabolic processes.

A simple way to avoid the disruption of label coherence by cyclic metabolic processes is the use of complex precursors from a relatively late stage of a biosynthetic pathway under study. However, this approach is subject to other limitations such as the failure to make the correct guess with regard to the structure of precursors, the failure of correctly guessed precursors to reach the intracellular site of the biosynthetic process under study,or the impossibility of preparing or otherwise obtaining a multilabeled specimen at an acceptable price and/or effort.

Despite its simple chemical structure, the plant metabolite chelidonic acid provides particular problems for isotope studies that result from the inherent mirror symmetry of the 7-carbon compound (**11**, **Figure 5**). Studies with [U-^{13}C$_6$]glucose and with a mixture of universally ^{13}C-labeled ribose and ribulose showed unequivocally that the 7-carbon skeleton of chelidonic acid could be retrobiosynthetically dissected into a 4-carbon and a 3-carbon piece, respectively (**Figure 5(a)**).[38] Interestingly, a $4+3$ pattern had been reported much earlier by authors who had interpreted their findings in terms of a shikimate origin for chelidonate (which appeared reasonable at the time because shikimate is assembled in $4+3$ pattern from erythrose phosphate and phosphoenol pyruvate).[39]

In closer detail, chelidonic acid biosynthesized from the universally ^{13}C-labeled precursors was obtained as a very complex mixture of isotopologues that nevertheless yielded to a rigorous isotopologue deconvolution affording the components shown in **Figure 5(a)**. We note in passing that multiple labeling results in breaking of

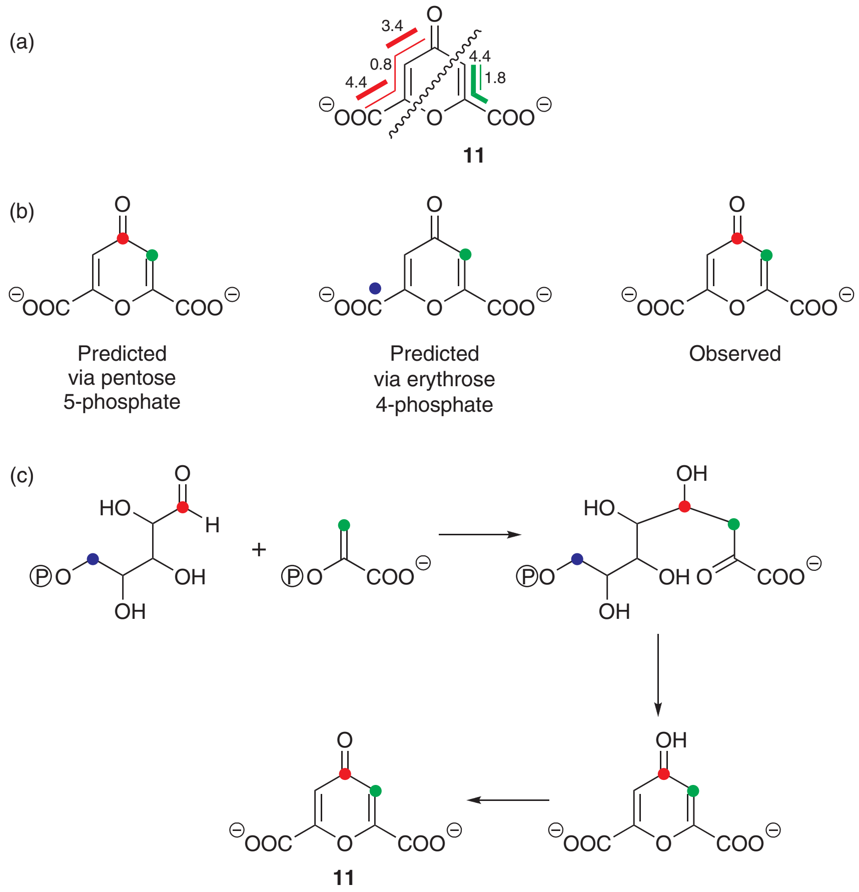


Figure 5 Retrobiosynthetic analyis of the biosynthesis of chelidonic acid. (a) Labeling patterns of chelidonic acid obtained from incorporation of [U-$^{13}C_6$]glucose diluted with unlabeled glucose into a cell culture of *Leucojum aestivum*.[38] Bold bars indicate multiple ^{13}C-labeled isotopologues. The numbers indicate mol% abundances. The colored dots indicate the biosynthetic dissection into two fragments. (b) Labeling patterns of chelidonic acid obtained from [1-^{13}C]glucose. The dots indicate highly ^{13}C-enriched carbon atoms. The patterns shown in the left and in the middle are based on predictions, and the right pattern was observed.[38] (c) Hypothetical biosynthetic pathway of chelidonic acid. The dots indicate atoms ^{13}C-enriched from [1-^{13}C]glucose.

the mirror symmetry and the resulting ^{13}C signal degeneration, in consequence of heavy isotope chemical shift variation. A technical discussion of the phenomenon can be found in the original publication.[38]

Simpler isotopologue mixtures were detected in chelidonic acid labeled from [1-^{13}C]glucose. Specifically, [4-$^{13}C_1$]- and [3-$^{13}C_1$]chelidonic acid were observed in high abundances (Figure 5(b)). The labeling patterns could now be compared to the shikimate labeling pattern obtained by retrobiosynthetic deconstruction of aromatic amino acids. Surprisingly, this comparison showed that the 4-carbon fragment of chelidonic acid does not reflect the four bottom carbon atoms of the pentose pool (as it should if it had been derived from erythrose 4-phosphate), but rather the four top carbon atoms of a pentulose (Figure 5(b)). Based on these findings, a biosynthetic pathway involving the condensation of a precursor obtained from the pentose pool with phosphoenol pyruvate was proposed (Figure 5(c)).

The biosynthesis of the coenzyme molybdopterin has been particulary resistant to biochemical exploration, even though genes and even proteins of the pathway had been known for a long time. It had been known beyond doubt that pyrimidine ring motifs in several biosynthetic heterocycles including riboflavin (7), deazaflavin (8), folic acid, and folic acid analogues are all biosynthetically derived from the pyrimidine ring of GTP (1); there was no doubt that this concept also applies for molybdoterin, a cofactor of certain oxidoreductases.[40,41]

In the folate pathway, the ribosyl moiety of GTP supplies the carbon atoms required for the formation of a second heterocyclic ring, thus affording the first committed intermediate, dihydroneopterin triphosphate. A motif with close structural similarity to the dihydroneopterin motif is a structural part of molybdopterin; the problem, however, is that the two carbon atoms required for the formation of the pyrazine ring plus the carbon atoms of the position 6 side chain add up to 6, and the ribosyl moiety of GTP is obviously insufficient to supply them all.

The study can serve as an example for the power of label coherence transfer analysis and was specifically designed to check whether the 'missing' carbon atom could be supplied by C-8 from the imidazole ring of the GTP precursor (**12**, **Figure 6**) by way of an intramolecular conversion. Experiments were conducted with an *Escherichia coli* strain that had been engineered for the overexpression of three open reading frames (*moaABC*) believed to be involved in the biosynthesis of the molybdopterin precursor, compound Z. [7-^{15}N$_1$,8-^{13}C$_1$]Guanine, [U-^{13}C$_6$]glucose, or a mixture of universally ^{13}C-labeled ribose and ribulose, were used as tracers on a background of unlabeled glucose serving as the major carbon source.[42]

Retrobiosynthetic analysis was based on a comparison between compound Z′ (**13**, **Figure 6**), an oxidation product of the pathway intermediate, compound Z, with the labeling patterns of four different nucleosides obtained by hydrolysis of RNA; thus, the labeling patterns of the ribosyl moiety, and thereby the labeling pattern of the pentose phosphate pool, was independently determined four times. Comparing the labeling patterns showed that carbon atoms 7, 6, 2′, 3′, and 4′ of compound Z′ could all be obtained from a single pentose moiety via an intramolecular process (**Figure 6**). On the other hand, it was also shown that N-5 and C-1′ of compound Z′ could be both obtained from [7-^{15}N$_1$,8-^{13}C$_1$]guanine, also via an intramolecular process. This rather unusual choreography of atoms translates into a hypothetical mechanism that turned out to be remarkably similar to the transformation of ribulose 5-phosphate into 3,4-diyhdroxy-2-butanone 4-phosphate in the pathway of riboflavin biosynthesis (see above).[42] The latter reaction involves the extrusion of a formate moiety from a pentose moiety, and the former reaction involves the incorporation of a formate unit into a carbohydrate unit. In both cases, the reactions proceed by way of intramolecular rearrangements.

It should be noted that the concept of label coherence transfer is also being addressed in the earlier as well as the following chapters in Section 9.18.8. This chapter, however, was written with the specific purpose of demonstrating how labeling coherence tracking in combination with retrobiosynthetic comparison can provide elegant solutions for problems that are not easily resolved by other techniques.

9.18.8.4 Identifying Branch Points in Complex Pathways – Studies on Shikimate Derivatives

The shikimate pathway of aromatic acid biosynthesis is the source of building blocks for a very wide number of natural products from microbial and plant kingdoms. Remarkably, it appears that this pathway is the unique source for the *de novo* synthesis of phenylalanine, tyrosine, and tryptophan. Although many shikimate-derived

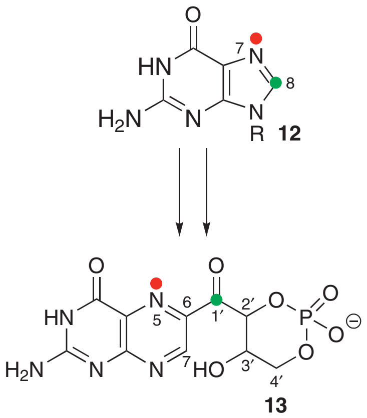


Figure 6 Biosynthesis of factor Z′, an oxidation product derived from an intermediate in molydopterin biosynthesis. Biosynthetically equivalent positions in factor Z′ and guanine precursors are indicated the colored. For details, see text.

natural products are obtained from these aromatic amino acids as precursors, others start from earlier intermediates of the shikimate biosynthetic pipeline.

The conversion of chorismate into any of the three proteinogenic aromatic amino acids is accompanied by the loss of its position 1 carboxylate group whose biosynthetic origin can be traced back to a phosphoenol pyruvate building block. The cotransfer of the carboxylic carbon into a downstream product, together with the other two carbons introduced into shikimate from the phosphoenol pyruvate precursor, provides compelling evidence for a preprephenate or preanthranilate branch point (and, most of the time, a prechorismate branch point) (**Figure 7**). An unequivocal and conceptually simple approach to distinguish between prechorismate and postchorismate branching could involve feeding of [U-$^{13}C_7$]shikimate or other appropriate shikimate orthologues. Clearly, however, this approach depends on the availability of an appropriate isotopologue and on the ability of shikimate to reach the site of biosynthesis inside the cells or organisms under study.

Alternatively, a wide variety of multiply labeled precursors including, but not limited to, carbohydrates, such as [U-$^{13}C_6$]glucose, can be used in a retrobiosynthetic approach, and the selection of the precursor can be based on criteria of both availability and on permeability aspects. The power of the retrobiosynthetic approach to distinguish between the prechorismate and postchorismate origin of shikimate-derived natural products can be illustrated by studying the biosynthesis of simple and complex secondary products from plants and fungi, which are briefly reported below.

Even a quick comparison between the isotopologue patterns of the benzenoid rings illustrated in **Figure 7** reflects the possible branch points in the phenylpropanoid pathway. Although the label coherence pattern does not extend beyond the aromatic ring in phenylpyruvate (**19**) or phenylalanine (**20**), it does so in the case of gallic acid (**21**),[43,44] hydroxyxanthones (**22**),[45] and amarogentin (**23**).[46] These extensions signify the undisrupted integration of a 3-carbon segment that reflects the phosphoenol pyruvate building block (**14**) of shikimate (**16**) and includes the carbon atom that stems from the position 1 carboxylic group of the early shikimate precursors, that is, the carbon atom that would have been lost in the transformation of chorismate (**17**) to prephenate (**18**) or arogenate.

It may be interesting to note that the biosynthesis of gallic acid (**21**), the universal precursor of tannic acids, had been addressed repeatedly by various noted natural products chemists, who came out with conflicting results.[47–52] It should also be noted that the information contained in the isotopologue profiles of the target compounds shown in **Figure 7** extends way beyond the question of a prechorismate or postchorismate origin of benzenoid rings. Thus, it was immediately apparent from the isotopologue pattern that ring A of the xanthone (**22**) has a polyketide origin; quite obviously, the biosynthesis involves the extension, by way of ester condensation of the carboxyl side chain originating from the carboxylate group carried over from shikimate. In case of amarogentin (**23**), extension of a 3-hydroxybenzoate by a polyketide sequence was also immediately obvious from the isotopologue deconvolution. Moreover, the isotopologue pattern showed that the building blocks for the bicyclic moiety in amarogentin originate predominantly (>95%) from the nonmevalonate pathway of terpenoid biosynthesis that is discussed in more detail in a subsequent chapter. Just for the record, the hexose unit present in amarogentin served as an internal label reference that permitted the reconstruction of the shikimate labeling.

9.18.8.5 Measuring Crosstalk between Pathways – Biosynthesis of Isoprenoids in Plants

Isoprenoids are the largest group of known natural products comprising at least 35 000 compounds. Pioneering work in the 1950s resulted in the elucidation of the mevalonate pathway for the biosynthesis of the two universal isoprenoid precursors isopentenyl diphosphate (IPP, **29**, **Figure 8**) and dimethylallyl diphosphate (DMAPP, **28**).

A second pathway for the biosynthesis of IPP and DMAPP eluded discovery until the 1990s but was then rapidly elucidated in considerable detail (deoxyxylulose pathway, **Figure 8**). These events have been reviewed repeatedly and will not be discussed at this point; suffice it to say that retrobiosynthetic analysis has been involved to some extent in the elucidation of the nonmevalonate pathway (specific examples are given in Eisenreich *et al.*[53]).

The single (and, hence, the largest) building block used in the mevalonate pathway is the acetyl moiety in acetyl-CoA (**30**, **Figure 8**); thus, labeling coherence cannot exceed the level of two contiguous ^{13}C atoms. More

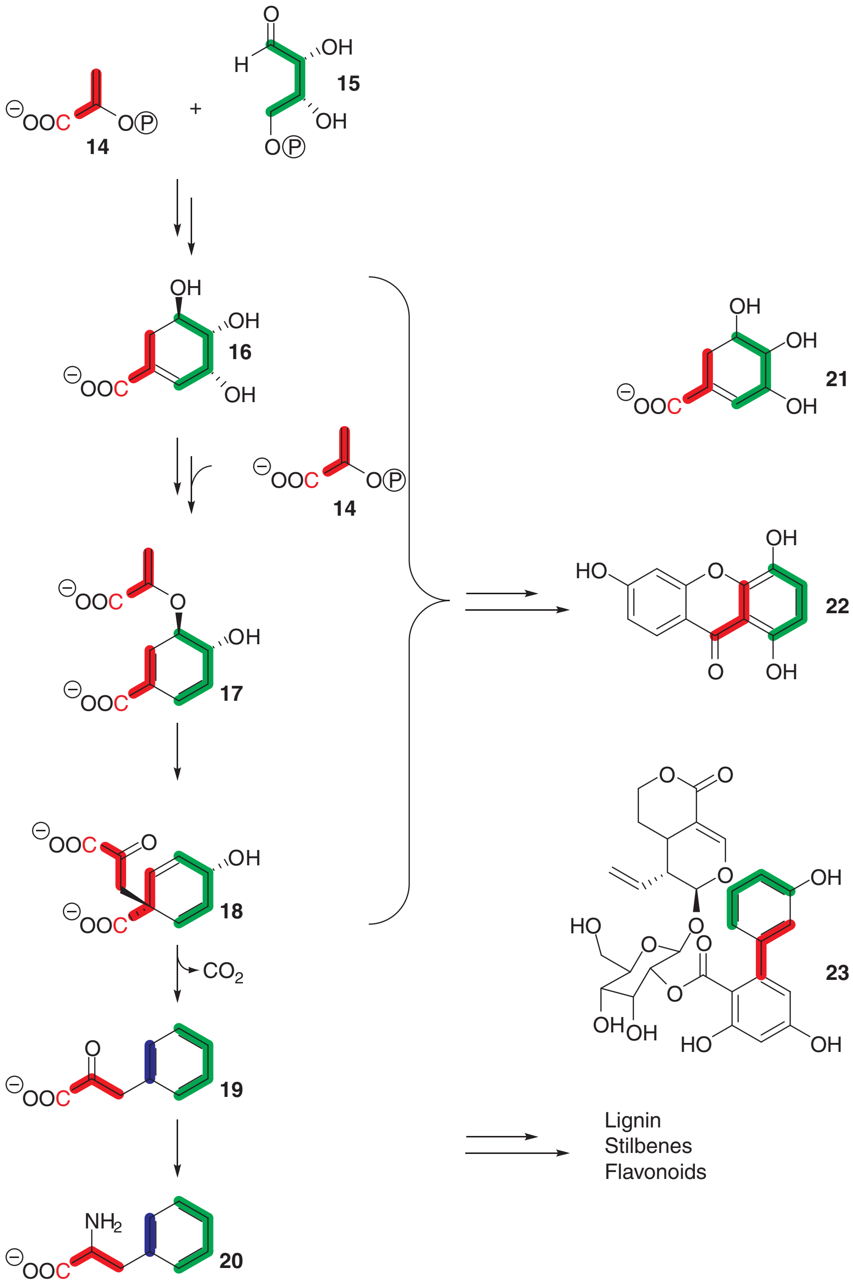


Figure 7 Biosynthesis of aromatic amino acids and products derived from phenylalanine or from intermediates of the shikimate pathway. Biosynthetically equivalent positions are indicated by colored bars. The atoms indicated by the blue bars are equivalent to atoms from phosphoenol pyruvate precursor followed by the loss of one carbon atom by decarboxylation.

specifically, the mevalonate pathway stitches three acetate units together by way of ester condensations, and one carbon is then sacrificed in order to yield the 5-carbon building blocks, IPP (29) and DMAPP (28) via mevalonate (31). By comparison, the nonmevalonate pathway uses two 3-carbon precursors, glyceraldehyde phosphate (25) and pyruvate (24). One carbon atom is also sacrificed to afford a 5-carbon skeleton. However, the labeling pattern of the glyceraldehyde phosphate precursor (25) is retained in the downstream products, although the labeling coherence becomes interrupted, but without being destroyed, by way of a strictly

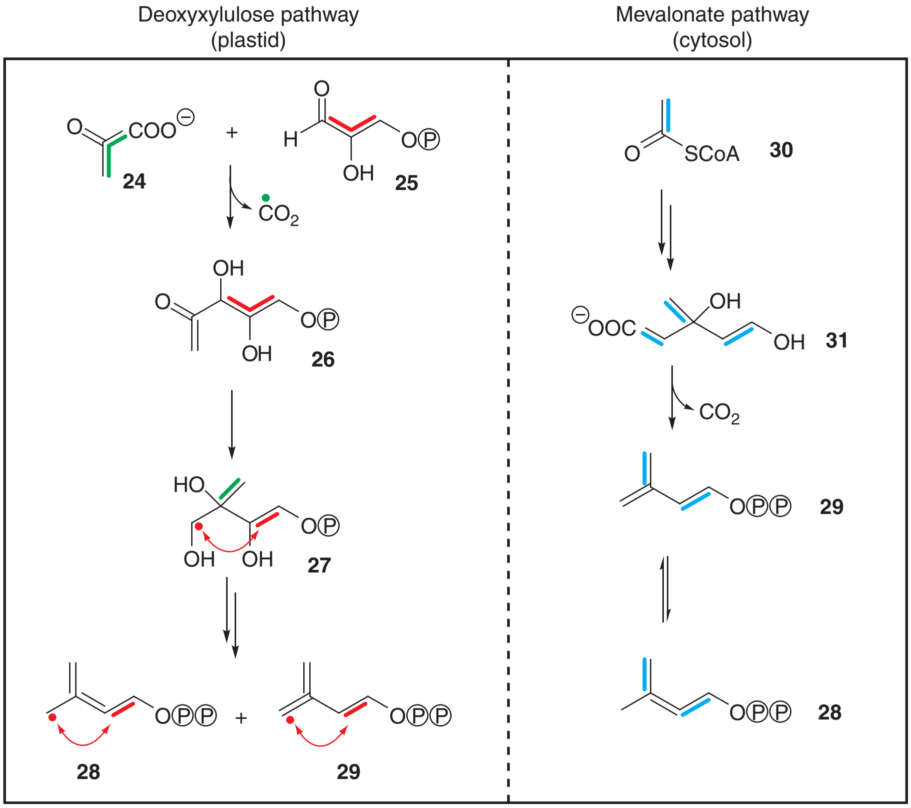


Figure 8 Biosynthesis of terpenoid precursors in plants. Colored bars indicate biosynthetically equivalent positions.

intramolecular rearrangement catalyzed by IspC protein.[54–56] Nevertheless, the presence of triple ^{13}C-labeled isoprenoids (caused by [U-$^{13}C_3$]glyceraldehyde phosphate precursors in the nonmevalonate pathway) can be detected by long-range ^{13}C-couplings via two or three bonds in many cases (for an example, see Eichinger *et al.*[57]).

It is now established knowledge that, with few exceptions, microorganisms use exclusively one of the two isoprenoid pathways. In cases where sufficient sequence data (and not necessarily a complete genome sequence) are available, the assignment of the respective organism to one of the isoprenoid pathways is easily possible on basis of bioinformatic analysis. In the absence of sufficient sequence data, the assignment can be performed experimentally using retrobiosynthetic technology (as an example, see Eichinger *et al.*[58]).

A far more complex situation arises in higher plants that use both the pathways in parallel.[53] With hindsight, it is even obvious that the belated discovery of the deoxyxylulose pathway can be traced to a significant extent to the very occurrence of both the pathways in plants. More specifically, due to metabolite exchange between the two pathways that is the subject of this chapter, it appears likely that labeled mevalonate can contribute at least some label to most if not all plant isoprenoids; hence, it was easy to jump to the conclusion – fallacious as we now know – that all plant isoprenoids are invariably biosynthesized from mevalonate.

Bioinformatic and biochemical data indicate that higher plants direct the proteins of the nonmevalonate pathway to the plastid compartment, due to N-terminal targeting sequences, whereas the mevalonate enzymes remain in the cytoplasm.[53] Similarly, enzymes catalyzing reaction steps located downstream from IPP and DMAPP are also subject to compartmentalization. However, compartmental separation of biosynthetic intermediates is apparently not absolute. On the contrary, a certain extent of transfer of intermediates between different compartments appears to be the rule rather than the exception. Hence, plant isoprenoids can and typically do contain 5-carbon building blocks of different (mevalonate or nonmevalonate) origin. This phenomenon has been termed as crosstalk that can vary over a relatively wide range.[57–62]

Although many details of isoprenoid pathway crosstalk in plants remain to be uncovered, it is at least possible to quantify the extent of crosstalk using retrobiosynthetic technology. Details of the technology are given in Schuhr *et al.*[63]

9.18.8.6 Using Complex Mixtures of Universally ^{13}C-Labeled Compounds as Precursors – Biosynthesis of Lipstatin

The almost universally applied standard approach in isotope labeling is the use of a single isotopologue of a single compound as a tracer (although readers may recall that we have presented examples where pentose/pentulose mixtures were used). In a rigorous deviation from the standard concept, the biosynthesis of lipstatin (**34**, **Figure 9**), whose tetrahydro derivative is marketed as an antiobesity drug (Xenical) was successfully addressed using a chloroform extract of algae grown with $^{13}CO_2$.[64] The blackish, oily material extracted from the labeled algae consisted of an unknown, but definitely large, number of different lipids and lipid derivatives and was added to fermentation cultures of *Streptomyces toxytricini* without any prior purification. Biosynthetic lipstatin was then isolated from the cultures and analyzed by NMR spectroscopy.

NMR spectra revealed that the highly lipophilic β lactone moiety of lipstatin consisted of two segments of uninterrupted sequences of 8 and 14 ^{13}C atoms, respectively (green and red bars in **Figure 9**). These had obviously been contributed from some long-chain aliphatic molecules in the crude algal extract. On the other hand, the formyl leucine residue present in lipstatin had incorporated two pairs of directly linked ^{13}C atoms (blue bars in **Figure 9**), and that pattern could be easily explained by catabolism of components in the proffered lipid mixture to the level of [$^{13}C_2$]acetyl-CoA, which had then contributed label to the pyruvate pool via the citric acid cycle affording [$^{13}C_2$]oxaloacetate. Notably, this is again an example for the application of the retro concept via intramolecular comparison of labeling patterns, in this case of the lipophilic β lactone moiety and the formyl leucine moiety of lipstatin.

The labeling pattern of the β lactone moiety could be easily explained by Claisen condensation of octanoyl-CoA (**33**) with 3-hydroxy-5,8-tetradecanoyl-CoA (**32**) obtained by β oxidation of linoleic acid, as opposed to polyketide-type biosynthesis from low-molecular-weight fragments.[64]

The proposed reaction mechanism received additional confirmation by experiments using chloroform extract of algae that had been grown in D_2O with $^{13}CO_2$ as the only carbon source.[65] Feeding the resulting [U-^{13}C,U-^{2}H]lipid mixture on a background of unlabeled sunflower oil to *S. toxytricini* afforded lipstatin whose β lactone moiety contained three hydrogen atoms that were not derived from the proffered lipid but from solvent (as a technical note, it should be said that in this case the ^{13}C labeling served only the purpose to increase the sensitivity of ^{2}H detection via coupling to ^{13}C). Confirmatory evidence was obtained by the

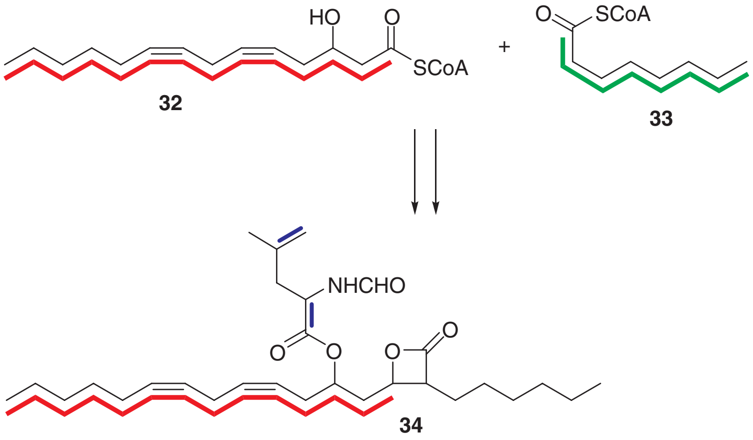


Figure 9 Biosynthesis of lipstatin (**34**). Multiple ^{13}C-labeled isotopologues detected in lipstatin from a crude mixture of universal ^{13}C-labeled algal lipid.[63] For details, see text.

biosynthesis of lipstatin in D_2O with unlabeled sunflower oil as carbon source. Again, the leucine moiety of lipstatin could be used as an intramolecular reference.

The findings obtained with the crude lipid mixtures were later independently confirmed by incorporation studies with specific fatty acid precursors (more specifically, (5*Z*,8*Z*)-[10,11,12,12-^{2}H]tetradeca-5,8-dienoic acid, a mixture of [2,2-$^2H_2$]- and [8,8,8-$^2H_3$]octanoic acid, [3,10,11,12-^{2}H]-(3*S*,5*Z*,8*Z*)-3-hydroxytetradeca-5,8-dienoic acid, and [7,8-$^2H_2$]hexylmalonate, respectively).[66]

9.18.8.7 *In Situ* Generation of Coherently Labeled Precursor Populations – Analyzing the Biosynthesis of Plant Metabolites via Photosynthetic Pulse/Chase Labeling with $^{13}CO_2$

Pulse labeling with $^{14}CO_2$ was a breakthrough technology for the Nobel Prize winning work on photosynthetic carbon fixation in plants.[1] Later, the technology was transferred to labeling work with $^{13}CO_2$, published in the 1970s.[67–69] In principle, photosynthetic pulse labeling with $^{13}CO_2$ should afford a population of coherently ^{13}C-labeled carbohydrates, and a consecutive chase phase driven by photosynthesis in normal atmosphere should enable relaxation in a similar way as described above for labeling studies with ^{13}C-labeled organic tracer compounds. The techniques for quantitative assessment of individual isotopologues in isotopologue mixtures by systematic deconvolution of NMR spectra have made significant progress, and two recent papers have reported on the biosynthesis of nicotine and of hermidin in plants using $^{13}CO_2$ as a precursor.[70,71] The technology may become relevant for cases where other approaches are not possible. In any case, it is a technique enabling the study of experimentally unperturbed whole plants under strictly physiological conditions.

9.18.9 Conclusion and Outlook

The introduction of a general ^{13}C-labeled carbon source into the metabolic networks of organisms generates local perturbations that spread through the entire networks by metabolic reactions. The process affords highly specific isotopologue profiles in many metabolites, which can be assessed by quantitative NMR spectroscopy and/or mass spectrometry as a function of time. On the basis of retrobiosynthetic analysis, information about metabolic pathways under *in vivo* conditions is obtained. The many examples establish isotopologue profiling and the retrobiosynthetic approach as a welcome additional method to transcriptomic, proteomic, and metabolomic analyses in order to understand what happens in living organisms.

Glossary

isotopologues Isotopologues are molecular species of a given chemical compound that differ in isotope composition (number of isotopic substitutions), whereas isotopomers differ in the position of isotopic substitution(s) but not in the net isotope composition. In that sense, sets of isotopomers are subsets of the isotopologue superset.

References

1. J. A. Bassham; K. Shibata; K. Steenberg; J. Bourdon; M. Calvin, *J. Am. Chem. Soc.* **1956**, *78*, 4120–4124.
2. E. W. Washburn; E. R. Smith, *Science* **1934**, *79*, 188–189.
3. O. Reitz; K. F. Bonhoeffer, *Naturwissenschaften* **1934**, *22*, 744.
4. C. C. Butler, *Proc. Conf. Nucl. Chem.* **1947**, 159–166.
5. A. Bacher; C. Rieder; D. Eichinger; D. Arigoni; G. Fuchs; W. Eisenreich, *FEMS Microbiol. Rev.* **1998**, *22*, 567–598.
6. E. J. Corey; X.-M. Cheng, *The Logic of Chemical Synthesis*; Wiley: New York, 1995.
7. A. Bacher; W. Eisenreich, Application of Isotopes in Investigating Biosynthetic Pathways. In *Proceedings of the 7th International Symposium*, Dresden, Germany, 18–22 June, 2001.
8. C. Ettenhuber; T. Radykewicz; W. Kofer; H.-U. Koop; A. Bacher; W. Eisenreich, *Phytochemistry* **2005**, *66*, 323–335.
9. W. Eisenreich; A. Bacher, *Phytochemistry* **2007**, *68*, 2799–2815.

10. B. Schneider, *Prog. Nucl. Magn. Reson. Spectosc.* **2007**, *51*, 155–198.
11. J. A. Gudgeon; J. S. E. Holker; T. J. Simpson, *J. Chem. Soc. Chem. Commun.* **1974**, 636–638.
12. T. J. Simpson, *Top. Curr. Chem.* **1998**, *195*, 1–48.
13. J. A. MacLaren, *J. Bacteriol.* **1952**, *63*, 233–241.
14. G. W. E. Plaut; C. M. Smith; W. L. Alworth, *Annu. Rev. Biochem.* **1974**, *43*, 899–922.
15. A. Bacher; B. Mailänder, *J. Biol. Chem.* **1973**, *248*, 6227–6231.
16. M. Fischer; A. Bacher, *Nat. Prod. Rep.* **2005**, *22*, 324–350.
17. B. Mailänder; A. Bacher, *J. Biol. Chem.* **1976**, *251*, 3623–3628.
18. F. Foor; G. M. Brown, *J. Biol. Chem.* **1975**, *250*, 3545–3551.
19. G. W. Plaut, *J. Biol. Chem.* **1963**, *238*, 2225–2243.
20. H. Wacker; R. A. Harvey; C. H. Winestock; G. W. Plaut, *J. Biol. Chem.* **1964**, *239*, 3493–3497.
21. A. Bacher; Q. Le Van; P. J. Keller; H. G. Floss, *J. Biol. Chem.* **1983**, *258*, 13431–13437.
22. P. J. Keller; Q. Le Van; A. Bacher; J. F. Kozlowski; H. G. Floss, *J. Am. Chem. Soc.* **1983**, *105*, 2505–2507.
23. H. G. Floss; Q. Le Van; P. J. Keller; A. Bacher, *J. Am. Chem. Soc.* **1983**, *105*, 2493–2494.
24. A. Bacher; Q. Le Van; P. J. Keller; H. G. Floss, *J. Am. Chem. Soc.* **1985**, *107*, 6380–6385.
25. Q. Le Van; P. J. Keller; D. H. Bown; H. G. Floss; A. Bacher, *J. Bacteriol.* **1985**, *162*, 1280–1284.
26. R. Volk; A. Bacher, *J. Biol. Chem.* **1990**, *265*, 19479–19485.
27. R. Volk; A. Bacher, *J. Biol. Chem.* **1991**, *266*, 20610–20618.
28. G. Richter; R. Volk; C. Krieger; H. W. Lahm; Ü. Röthlisberger; A. Bacher, *J. Bacteriol.* **1992**, *174*, 4050–4056.
29. W. Eisenreich; B. Schwarzkopf; A. Bacher, *J. Biol. Chem.* **1991**, *266*, 9622–9623.
30. Q. Le Van; B. Schwarzkopf; A. Bacher; P. J. Keller; S. Lee; H. G. Floss, *J. Am. Chem. Soc.* **1985**, *107*, 8300–8301.
31. B. Reuke; S. Korn; W. Eisenreich; A. Bacher, *J. Bacteriol.* **1992**, *174*, 4042–4049.
32. P. J. Keller; H. G. Floss; Q. Le Van; B. Schwarzkopf; A. Bacher, *J. Am. Chem. Soc.* **1986**, *108*, 344–345.
33. B. Schwarzkopf; B. Reuke; A. Kiener; A. Bacher, *Arch. Microbiol.* **1990**, *153*, 259–263.
34. W. Eisenreich; A. Bacher, *Pteridines* **1994**, *5*, 8–17.
35. W. Eisenreich; B. Schwarzkopf; Q. Le Van; P. J. Keller; A. Bacher, *J. Chem. Soc. Chem. Commun.* **1988**, 1294–1296.
36. W. Eisenreich; A. Bacher, *J. Biol. Chem.* **1992**, *267*, 17574–17580.
37. W. Eisenreich; C. Ettenhuber; R. Laupitz; C. Theus; A. Bacher, *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 6764–6769.
38. Z.-W. Shen; U. Fisinger; A. Poulev; W. Eisenreich; I. Werner; E. Pleiner; A. Bacher; M. H. Zenk, *Phytochemistry* **2001**, *57*, 33–42.
39. M. J. Malcolm; J. R. Gear, *Can. J. Biochem.* **1971**, *49*, 412–416.
40. R. B. Irby; W. Adair; W. Lee, Jr., *J. Biol. Chem.* **1994**, *269*, 23981–23987.
41. M. M. Wübbens; K. V. Rajagopalan, *J. Biol. Chem.* **1995**, *270*, 1082–1087.
42. C. Rieder; W. Eisenreich; J. O'Brien; G. Richter; E. Götze; P. Boyle; S. Blanchard; A. Bacher; H. Simon, *Eur. J. Biochem.* **1998**, *255*, 24–36.
43. I. Werner; A. Bacher; W. Eisenreich, *J. Biol. Chem.* **1997**, *272*, 25474–25482.
44. W. Eisenreich; I. Werner; A. Bacher, *NMR Microbiol.* **2000**, 381–409.
45. C.-Z. Wang; U. H. Maier; M. Keil; M. H. Zenk; A. Bacher; F. Rohdich; W. Eisenreich, *Eur. J. Biochem.* **2003**, *270*, 2950–2958.
46. C.-Z. Wang; U. H. Maier; W. Eisenreich; P. Adam; I. Obersteiner; M. Keil; A. Bacher; M. H. Zenk, *Eur. J. Org. Chem.* **2001**, 1459–1465.
47. M. H. Zenk, *Z. Naturforsch.* **1964**, *19b*, 83–84.
48. E. E. Conn; T. Swain, *Chem. Ind.* **1961**, 592–593.
49. E. Haslam; R. D. Haworth; P. F. Knowles, *J. Chem. Soc.* **1961**, 1854–1859.
50. D. Cornthwaite; E. Haslam, *J. Chem. Soc.* **1965**, 3008–3011.
51. P. M. Dewick; E. Haslam, *Chem. Commun.* **1968**, 673–675.
52. P. M. Dewick; E. Haslam, *Biochem. J.* **1969**, *113*, 537–542.
53. W. Eisenreich; A. Bacher; D. Arigoni; F. Rohdich, *Cell. Mol. Life Sci.* **2004**, *61*, 1401–1426.
54. S. Takahashi; T. Kuzuyama; H. Watanabe; H. Seto, *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 9879–9884.
55. J. F. Hoeffler; D. Tritsch; C. Grosdemange-Billiard; M. Rohmer, *Eur. J. Biochem.* **2002**, *269*, 4446–4457.
56. S. Lauw; V. Illarionova; A. Bacher; F. Rohdich; W. Eisenreich, *FEBS J.* **2008**, *275*, 4060–4073.
57. D. Eichinger; A. Bacher; M. H. Zenk; W. Eisenreich, *Phytochemistry* **1999**, *51*, 223–236.
58. W. Eisenreich; A. Bacher; A. Berry; W. Bretzel; M. Hümbelin; R. Lopez-Ulibarri; A. F. Mayer; A. Yeliseev, *J. Org. Chem.* **2002**, *67*, 871–875.
59. O. Laule; A. Furholz; H.-S. Chang; T. Zhu; X. Wang; P. B. Heifetz; W. Gruissem; M. Lange, *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 6866–6871.
60. A. Hemmerlin; J. F. Hoeffler; O. Meyer, *J. Biol. Chem.* **2003**, *278*, 26666–26676.
61. M. Rodriguez-Concepcion; O. Fores; J. F. Martinez-Garcia; V. Gonzalez; M. A. Phillips; A. Ferrer; A. Boronat, *Plant Cell* **2004**, *16*, 144–156.
62. D. Hampel; A. Mosandl; M. Wüst, *Phytochemistry* **2005**, *66*, 305–311.
63. C. A. Schuhr; T. Radykewicz; S. Sagner; C. Latzel; M. H. Zenk; D. Arigoni; A. Bacher; F. Rohdich; W. Eisenreich, *Phytochem. Rev.* **2003**, *2*, 3–16.
64. W. Eisenreich; E. Kupfer; W. Weber; A. Bacher, *J. Biol. Chem.* **1997**, *272*, 867–874.
65. M. Goese; W. Eisenreich; E. Kupfer; W. Weber; A. Bacher, *J. Biol. Chem.* **2000**, *275*, 21192–21196.
66. M. Goese; W. Eisenreich; E. Kupfer; P. Stohler; W. Weber; H. G. Leuenberger; A. Bacher, *J. Org. Chem.* **2001**, *66*, 4673–4678.
67. J. Schaefer; E. O. Stejskal; C. F. Beard, *Plant Physiol.* **1975**, *55*, 1048–1053.
68. J. Schaefer; L. D. Kier; E. O. Stejskal, *Plant Physiol.* **1980**, *65*, 254–259.
69. C. R. Hutchinson; M. T. Hsia; C. R. A. Stephen, *J. Am. Chem. Soc.* **1976**, *98*, 6006–6011.
70. W. Römisch-Margl; N. Schramek; T. Radykewicz; C. Ettenhuber; E. Eylert; C. Huber; L. Römisch-Margl; C. Schwarz; M. Dobner; N. Demmel; B. Winzenhörlein; A. Bacher; W. Eisenreich, *Phytochemistry* **2007**, *68*, 2273–2289.
71. E. Ostrozhenkova; E. Eylert; N. Schramek; A. Golan-Goldhirsh; A. Bacher; W. Eisenreich, *Phytochemistry* **2007**, *68*, 2816–2824.

Biographical Sketches

Adelbert Bacher served as associate professor of microbiology at the University of Frankfurt and as professor of organic chemistry and biochemistry at the Technical University of Munich.

Wolfgang Eisenreich studied chemistry at the Technical University of Munich (TUM), Germany, graduating in 1987. During his doctoral studies at the TUM (1988–90), he elucidated the biosynthetic pathways of methanogenic cofactors. For his work, he received the Hans Fischer Prize in 1991. Since 1991, Wolfgang Eisenreich established an independent group at the TUM, focusing on the analysis of metabolism in various organisms including plants. Using stable isotope-labeled precursors and using the 'retrobiosynthetic' concept of data interpretation, the pathways of many plant metabolites (e.g., isoprenoids) were identified and quantified on the basis of NMR spectroscopy. Part of this work was awarded the Pierre Fabre Prize by the Phytochemical Society of Europe in 2002.

9.19 Bacterial Protein Overexpression Systems and Strategies

C. Kinsland, Cornell University, Ithaca, NY, USA

9.19.1 Introduction

Most proteins are not present in high enough levels to be economically purified in useful quantities from their natural hosts. Luckily, in the decade since the last review in *Comprehensive Natural Products Chemistry* by Pickert and Miller[1] several new technologies and techniques to enable researchers to clone, overexpress, and purify proteins from a variety of hosts have been developed. In many cases, all the necessary components can be purchased in prepackaged kits, which require only the input gene sequence in some PCR-ready format as a starting material. However, different companies have taken different approaches. Unfortunately, for researchers just entering the area, this means that there is a confusing array of different techniques and strategies, which all have the same final goal in mind. Sifting through the possibilities and choosing a cloning, expression, and purification strategy becomes a daunting task, made more formidable by the looming shadow of making an unproductive choice. For researchers well versed in cloning, expression, and purification, the problem is slightly different. For these seasoned veterans, the problem is deciding if any of the newer techniques are sufficiently 'better' (however that can be judged) than the techniques entrenched in the laboratory. Most people are loathe to replace a technique that they understand, that works for them, and that they know how to troubleshoot; therefore, a new technique needs to offer a significant improvement to be adopted. In recent years, there have been a few disruptive technologies introduced that have both facilitated the entry of neophytes into the area and enticed experienced researchers away from their well-tested laboratory standard operating procedures.

This review will cover several aspects of recombinant protein production in *Escherichia coli* from initial planning through to purified protein. The focus of the review will be on overexpression in *E. coli* as it is a reasonable starting point for any project. *E. coli* has several advantages over other systems: genetic manipulations are straightforward, the cultures are facile to grow, media is inexpensive, and many recombinant proteins express to high levels in soluble and active form. This review is geared toward providing an overview of the methods available and focuses on providing an outline of a reasonable starting strategy and options and rescue strategies for the more difficult projects (see **Figure 1**). Every protein is different and what works for one will likely not work for another; however, some methods have a higher success rate than others and have become the standard strategy for initial studies in a majority of laboratories.

Escherichia coli does not have the necessary machinery to incorporate many post-translational modifications that are common in proteins from higher organisms (glycosylation, phosphorylation, acylation) and disulfide bond formation can be problematic. These limitations can cause proteins expressed in *E. coli* to be misfolded or otherwise inactivated. Therefore, it may be necessary to use another production system (baculovirus, mammalian cells, or *in vitro* methods) for some proteins. However, perhaps surprisingly given its simplicity, *E. coli* has proven to be capable of producing a wide range of classes of proteins. It is incapable of performing some post-translational modifications and the probability of success drops as the molecular weight of the target protein increases.[2] Data from the structural genomics consortium indicate that 50% of bacterial proteins and 10% of eukaryotic proteins can be produced in *E. coli* in soluble form.[2] Possible causes of failure to express well in *E. coli* include toxicity of the target protein, codon bias problems, poor protein folding, proteolysis of the expressed protein, and mRNA problems (instability and hairpin formation). Despite the predicted high failure rate, the low cost and convenience of *E. coli* make it a reasonable starting point for attempting to obtain a recombinant protein. The investment of time and resources it takes to test expression in *E. coli* is quite minimal compared to other systems.

9.19.2 Project Design

In planning a protein production project, there are a number of factors to be taken into consideration. What is the intended final use of the protein (enzymatic assays, antibody generation, crystallography, etc.)? Does the protein need to be active and correctly folded? Is the protein expected to have posttranslational modifications (phosphorylation, glycosylation, site-specific proteolysis, etc.)? How much protein is going to be needed? Will any fusions need to be removed? Is the protein expected to incorporate cofactors? Is the protein expected to

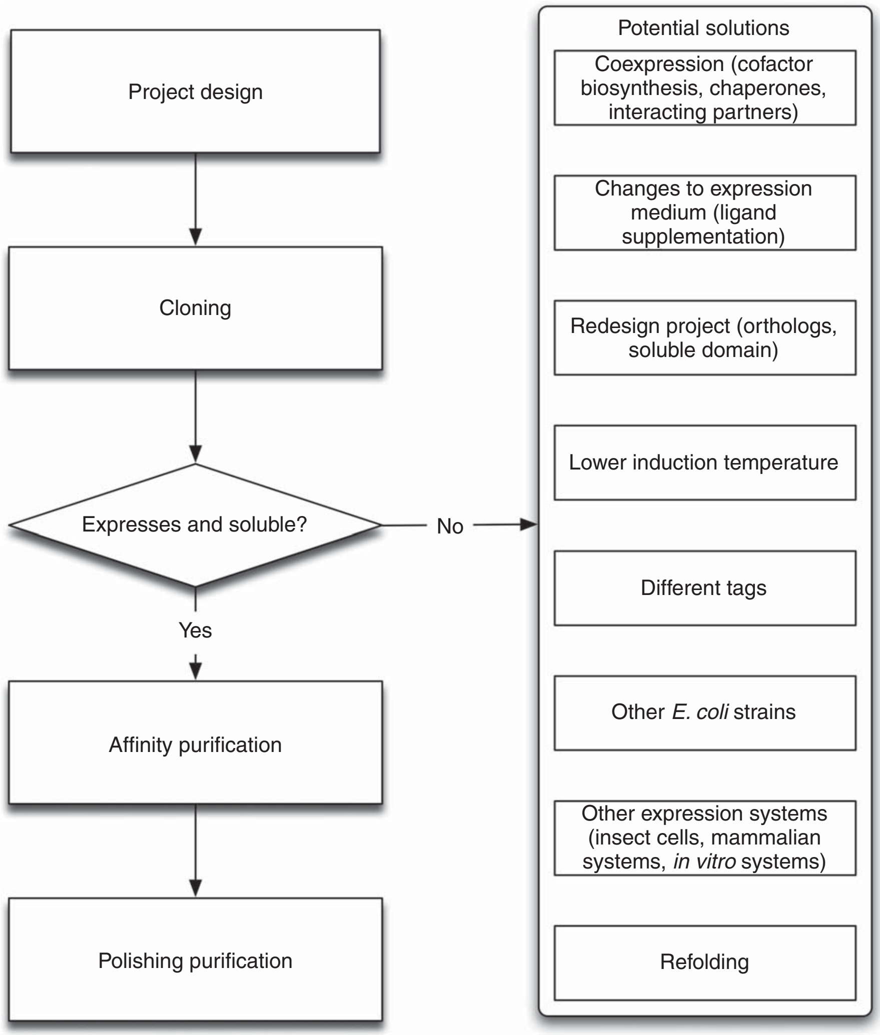

Project design
Cloning
Expresses and soluble?
No
Yes
Affinity purification
Polishing purification
Potential solutions
Coexpression (cofactor biosynthesis, chaperones, interacting partners)
Changes to expression medium (ligand supplementation)
Redesign project (orthologs, soluble domain)
Lower induction temperature
Different tags
Other *E. coli* strains
Other expression systems (insect cells, mammalian systems, *in vitro* systems)
Refolding

Figure 1 A flow chart for a recombinant protein project in *Escherichia coli.*

form a heterodimer or associate with another protein? Which expression system should be used? Which vector? Should a fusion partner be added to facilitate purification and/or expression and solubility? Should the entire protein be expressed, or would some fragment of the protein be more easily obtainable and still be sufficient for the needs of the research? What purification strategy should be followed?

9.19.2.1 Target Selection

Once a protein of interest has been identified based on the scientific goals of the project several choices must be made. Seemingly insignificant changes to the amino acid sequence or minor variations on the boundaries of the polypeptide can result in radically different expression and solubility levels and have huge impact on the stability and performance of the protein. Despite huge efforts to analyze the data on protein expression and purification being generated by the high-throughput efforts of the structural genomics consortia, predicting protein behavior from amino acid sequence remains difficult.[3–5] To increase the odds of obtaining a well-behaved protein construct, it is frequently useful to screen proteins from different organisms with the same function (orthologues) if the project goals do not rely on the protein from a specific organism.[6] For example, in our laboratory, it is common to clone 3–5 orthologues for proteins that are destined for structural studies. Despite high sequence homology, the proteins will often demonstrate radically different expression and solubility levels and dramatically variable behaviors upon purification and crystallization screening. It is commonly thought that proteins from thermophiles will be better behaved and more amenable to structural studies because they are predicted to have fewer disordered regions and a higher proportion of surface salt bridges.[7–9] This belief seems to arise predominantly from anecdotal evidence. In comparing a series of 68 homologous proteins from *E. coli* and *Thermotoga maritima*, Savchenko *et al.* did not find a distinct advantage to the use of thermophile-derived proteins. However, they did find that the use of orthologues increased the probability of obtaining a protein sample suitable for structural studies.[6]

A recent overview of the experience of the Protein Structure Initiative (PSI) with over 100 000 proteins pointed out that 55% of all targets were of bacterial origin because of the easier access to the genes and the generally more tractable nature of their proteins.[10] However, they did note that, of eukaryotes, *Homo sapiens*, *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, and *Mus musculus* proteins have been most successful. They also noted that parasites (*Plasmodium* sp., *Trypanosoma* sp., and *Leishmania* sp.), *Drosophila melanogaster*, and *Caenorhabditis elegans* were particularly difficult both to clone and to express in soluble form. Therefore, it may be beneficial to avoid these species when choosing an orthologue set for screening if possible given the project goals.

9.19.2.2 Obtaining DNA for the Gene of Interest

Once a protein of interest is chosen for study the next step is to obtain the target DNA for the gene of interest. Most commonly, DNA containing the gene of interest is used as a template in a polymerase chain reaction (PCR) to obtain sufficient quantities for cloning. There are a variety of resources available for obtaining DNA from both prokaryotic and eukaryotic organisms. For bacterial sources, genomic DNA is readily available from the American Type Culture Collection (ATCC; http://www.atcc.org/) and the DSMZ (http://www.dsmz.de/index.htm). It is important to remember that eukaryotic genes often contain introns, which makes obtaining DNA for expression in *E. coli* slightly more difficult for eukaryotic projects. Genomic DNA can be used as a PCR template, but *E. coli* will not correctly process the introns out, so they must be removed at some point during the cloning stage. This can be done during the PCR stages of a cloning project, but it can be tedious if there are several introns and it is more straightforward to use DNA template free of introns. Typically, this means using cDNA or a cDNA library as a template. The Mammalian Gene Collection (MGC)[11–13] maintains a vast repository of full-length, sequence-verified cDNA from humans, mice, and rats and their clones are generally of high quality. For other organisms, cDNA or cDNA libraries can be purchased from a number of commercial vendors or obtained from other government-funded efforts.[14]

If no suitable source of template DNA can be found, total gene synthesis is a viable option. The cost of gene synthesis has been decreasing rapidly over the last few years and, in fact, gene synthesis is becoming a common choice even when a suitable template is readily available. There are several advantages to synthesizing the gene for a protein of interest instead of relying on a natural DNA source as a template. First, total gene synthesis allows the gene coding to be matched to the preferences of the host organism that is going to be used for recombinant protein production (see Section 9.19.5.4.1). Second, any restriction sites can be added or removed easily during the synthesis to facilitate cloning, truncation, domain removal, and other alterations to the full-length sequence. Third, attempts to increase expression by mitigating the impact of mRNA secondary structure, eliminating repetitive sequences, and optimization of other characteristics of the gene sequence through sequence design can be easily incorporated. There is also evidence that codon pairs encode signals that control elongation rate, introducing pauses in translation that may encourage proper folding.[15] Finally, it is easy to plan for and create multiple variants or randomized libraries of the same gene during gene synthesis.

There are several published protocols for gene synthesis[16–23] but they all rely on the availability of high-quality oligonucleotides and proofreading DNA polymerases. Although gene synthesis from oligonucleotides is fairly facile, many laboratories outsource to a commercial vendor, of which there are several.

9.19.2.3 Protein Domains

The activity of interest (enzymatic catalysis or ligand binding, for example) often resides in a discrete domain of a full-length protein and it is often easier to obtain the active domain alone than to express and purify the full-length protein. Choosing the limits of the domain and deciding on the exact boundaries for the protein to be expressed is not a straightforward process. However, there are some useful computational tools for helping to predict secondary structure and degree of disorder.[24–36] Generally, it is suggested that predicted membrane-spanning regions be eliminated, predicted secondary structural elements and globular domains be left intact, and low-complexity regions and hydrophobic residues be avoided at the termini.[37] Unfortunately, even seemingly minor changes to the N- and C-terminal boundaries (a few amino acids) can have a dramatic affect on solubility and aggregation behavior.[38] For this reason, it is pragmatic to test several constructs with different termini.[39] For example, one suggestion derived from experience with 10 000 proteins is to test 10 distinct constructs of the targeted domain (one full-length, nine from a combination of three different C-termini and three different N-termini).[37] In one trial of 400 human protein domains the use of multiple constructs with different termini nearly doubled the probability of obtaining a soluble protein and more than quadrupled the odds of achieving well-diffracting crystals.[40] Given the difficulty of predicting appropriate boundaries, high-throughput methods of searching for soluble domains are often employed (see Section 9.19.5.6).

9.19.3 *Escherichia Coli* Expression

9.19.3.1 Vector Basics

When expressing a protein in *E. coli*, there are two parts to the system: the vector (plasmid) incorporating the DNA for the protein of interest and the host cell for that vector. There are hundreds of commercially available plasmids for protein expression in *E. coli* and dozens of potential host strains; therefore, potential combinations are innumerable. However, the vectors all share a basic structure, although they vary in the details (see **Figure 2**). All expression vectors have an origin of replication (ORI), the location from which DNA replication begins. The ORI determines the copy number of the plasmid and most expression vectors are medium-copy (15–60 per cell, often pMB1/ColE1 derived). Vectors also contain an antibiotic resistance marker that is used for selection and maintenance. Additionally, some vectors rely on nonantibiotic-driven positive selection for maintenance and stabilization.[41] Expression vectors generally utilize a strong promoter region, under the control of a repressor, which is often expressed in trans. This means that expression is generally kept off, or low, by the presence of the repressor protein, which binds to the promoter region and thwarts transcription. Protein production is typically enabled by adding a small molecule that binds to the repressor protein and causes its dissociation from the promoter (see **Figure 3**). This clears the path for RNA polymerase and enables transcription to start. There are other mechanisms of controlling expression, but this general scheme is valid for the most commonly used promoters. Promoters vary in their strength, repressors vary in their efficacy, and the tunability of the possible promoter/repressor combinations covers a wide range.

9.19.3.1.1 Promoters

Ideally, a promoter for expression of recombinant protein in *E. coli* would have the following characteristics: (1) drive high-level protein production by directing efficient transcription, (2) be tightly regulated to reduce leaky expression, thereby minimizing metabolic burdens and toxic effects, (3) have variable induction for the fine-tuning of expression levels, (4) be inducible by an inexpensive chemical or by straightforward change of growth conditions, and (5) be readily available. Several promoters are available and used with some frequency to drive recombinant protein production: *lac*, *trp*, *tac*, *ara*, *T*7, and others.[42] The availability of a diversity of promoters with different inducing agents allows the independent regulation of expression of two or more proteins (commonly used for the coexpression of folding chaperones, for example).

The most commonly used promoter in *E. coli* for recombinant protein production is the hybrid *T7lac* promoter (present in many pET plasmids from EMD/Novagen, Madison, WI, for example), which has the strong phage T7 late promoter followed by the *lac* operator. In this system, LacI binds to the operator region in

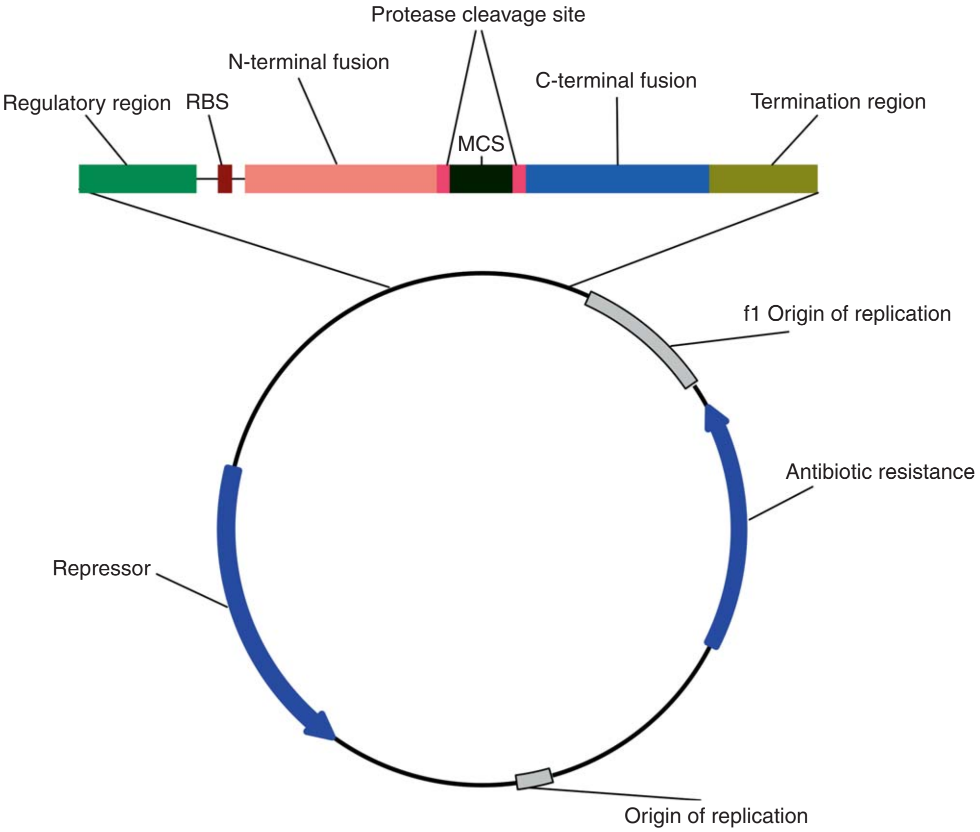


Figure 2 A general overexpresion vector. Most *Escherichia coli* vectors for recombinant protein production will have most of these features. The N- and C-terminal fusions and associated protease cleavage sites are optional. The multiple cloning site (MCS) may be replaced by DNA sequences for site-specific recombination in some vectors.

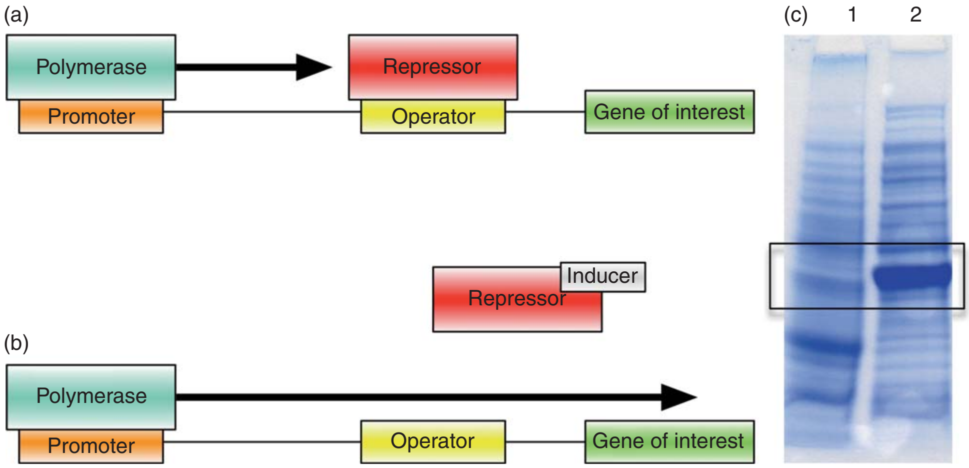


Figure 3 General structure of a promoter/repressor system in *Escherichia coli*. (a) In the absence of inducing agent, the repressor protein binds to the operator sequence and impedes the progress of RNA polymerase, halting transcription. (b) When the inducer is added, it binds to the repressor protein and causes its dissociation from the operator, clearing the path for the RNA polymerase to transcribe through the gene of the protein of interest. (c) Lane 1: total cellular protein in the absence of inducer. Lane 2: total cellular protein 3 h after the addition of inducing agent. Production of the protein of interest can be seen in the boxed area.

the absence of lactose or the lactose analogue isopropyl-β-D-thiogalactoside (IPTG) and prevents transcription. Expression is induced by adding lactose or IPTG, which binds to the repressor and causes its dissociation from the operator region, allowing the T7 RNA polymerase to proceed with transcription. The T7 RNA polymerase needs to be supplied either by infection with bacteriophage CE6 or, far more commonly, by use of a host strain containing a prophage (λDE3) encoding the polymerase under the control of the IPTG-inducible *lac*UV5 promoter.[43] An alternative to the manual intervention required for addition of lactose or IPTG is to induce protein production by the use of autoinduction medium, which enables unattended induction of transcription (see Section 9.19.5.3.2).[44]

9.19.3.1.2 Fusions

In addition to putting a strong promoter and control elements in front of the open reading frame (ORF) of interest in order to drive high-level expression, many expression vectors are designed to add amino acids to the protein of interest. These accessory amino acids (fusion tags or fusion proteins, depending on length and source) are added to enable facile purification by affinity chromatography, increase expression levels, increase solubility, target the protein to the periplasmic space, and/or enable detection (by fluorescence, antibody-based methods, or enzymatic activity, for example).[45] Fusions may be added to either the N- or C-terminus of a protein and are often separated from the protein of interest by amino acids that will enable proteolysis to separate the two. Multiple fusions are often attached with a common construction being a small tag that enables affinity purification and a larger tag to enhance solubility. Tags are sometimes put at both ends of a protein, often to enable tandem affinity purification (TAP).[46–50] Examples of typical protein–fusion combinations are shown in **Figure 4**.

9.19.3.2 Small-Scale Expression Testing

One benefit of the simplicity of the *E. coli* expression system is the ability to test dozens or hundreds of variables in parallel in small volumes (0.5–10 ml)[51,52] or even directly in colonies on a plate (see Section 9.19.5.6).[53–55] This allows the assessment of the relative utility of constructs, fusion tags, growth and induction conditions,

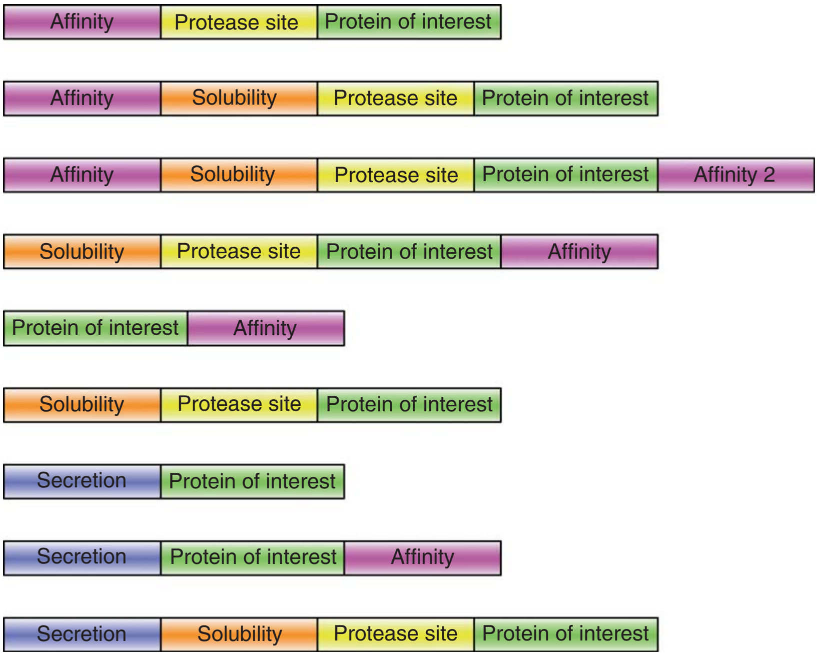


Figure 4 Example fusion constructs that may be used for recombinant protein production. Fusions for affinity purification, solubility enhancement, or secretion may be added. Other organizations or combinations not shown here may also be constructed.

host strains, coexpressed proteins, and media composition in a highly parallel fashion. Many laboratories use expression in 96-well, deep-well plates followed by antibody detection of an encoded tag.[52,56] Antibody-based detection protocols have limited throughput; consequently, several reporter assays have been developed. For example, methods based on detection of the enzymatic activity of a protein fused to the C-terminus of the protein of interest have been reported. Often, these types of reporter assays are employed when screening a library of constructs, typically truncations of a given protein. Another common approach to screening expression condition enlists the assistance of robotic methods that have been developed for the high-throughput affinity purification of proteins by immobilized metal affinity chromatography (IMAC) and glutathione *S*-transferase (GST). For example, the Oxford Protein Production Facility utilized expression in 2.5 ml of media in 24-well, deep-well plates followed by robotic IMAC purification and sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS–PAGE) analysis to screen for constructs producing soluble protein.[57] In fact, even laboratories that lack expensive liquid-handling systems can utilize these protocols to test multiple conditions in parallel. Using spin columns, one person can easily run an IMAC purification protocol for 24 samples in under an hour. With a 96-well plate, a single worker can process 96 samples in the same amount of time. Analyzing postpurification gives more information about the quality, yield, and behavior of the protein, although it does limit throughput.

9.19.4 Cloning

9.19.4.1 Introduction

The choice of cloning method is driven by several factors, including cost, throughput level, personal preference, and laboratory culture. A laboratory that makes a handful of clones per year will weigh consumable costs very differently than a high-throughput effort making thousands of clones per year. As in any field, some will be loathe to abandon a 'tried and true' method in favor of a newer methodology, particularly if the new method involves significant re-training or monetary investment. In many cases, laboratories have spent years amassing a library of different vectors appropriate for a particular cloning method and it would make little sense for them to switch strategies. This review will give a brief overview of the three most commonly used cloning methods: restriction based, ligation independent, and recombination based. Each method has advantages and disadvantages regarding cost, efficiency, robustness, and the ease of transferring the gene into different expression vectors (to test different fusion proteins or to change expression host organism, for example). All three are used in both small laboratories and in high-throughput settings; however, the restriction-based method is the least frequently used in high-throughput laboratories. A recent review by Hartley describes the available cloning methods in great detail, including some more novel and less frequently used methods.[58]

In general, all cloning methods begin with PCR to amplify the DNA containing the gene of interest and append sequences to the termini that allow the gene to be specifically placed in the correct context of an expression vector. The PCR primers will contain sequence at their 5′ ends that is complementary to the gene being amplified (15–20 base pairs of complementary sequence is standard). Then, at the 3′ end of the primers whatever sequence is needed for the next cloning steps is introduced. During the PCR process, the entire sequence of the oligonucleotide primer is incorporated into the final product (see **Figure 5**). The sequence and length of the accessory sequence appended varies with the cloning method being used.

9.19.4.2 Restriction-Based Cloning

The first recombinant DNA molecule was created in 1972[59] and since then the vast majority of recombinant DNA construction has utilized digestion with restriction enzymes followed by enzymatic ligation into the vector of choice. Most commercially available plasmids are designed with multiple cloning sites (MCSs) containing DNA recognition sequence for a variety of infrequently cutting restriction enzymes (see **Figure 6**). Typically, appropriate restriction sites are added to the ends of the DNA containing the gene of interest by PCR and both the vector and DNA to be inserted are treated with the restriction enzyme or

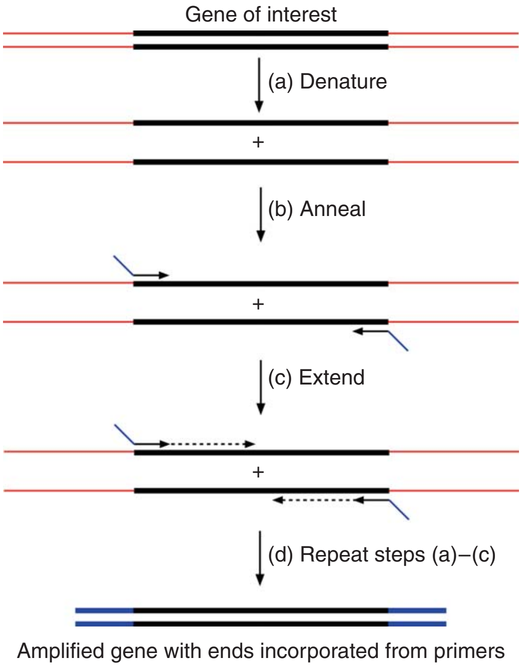


Figure 5 Gene amplification by PCR. (a) The double-stranded DNA template is denatured to single-stranded DNA by elevated temperature. (b) Oligonucleotide primers complementary to the DNA of interest anneal to the single-stranded DNA at a reduced temperature. (c) DNA polymerase synthesizes new DNA strands, extending outward from the primers. (d) The cycle is repeated, typically 25–30 times.

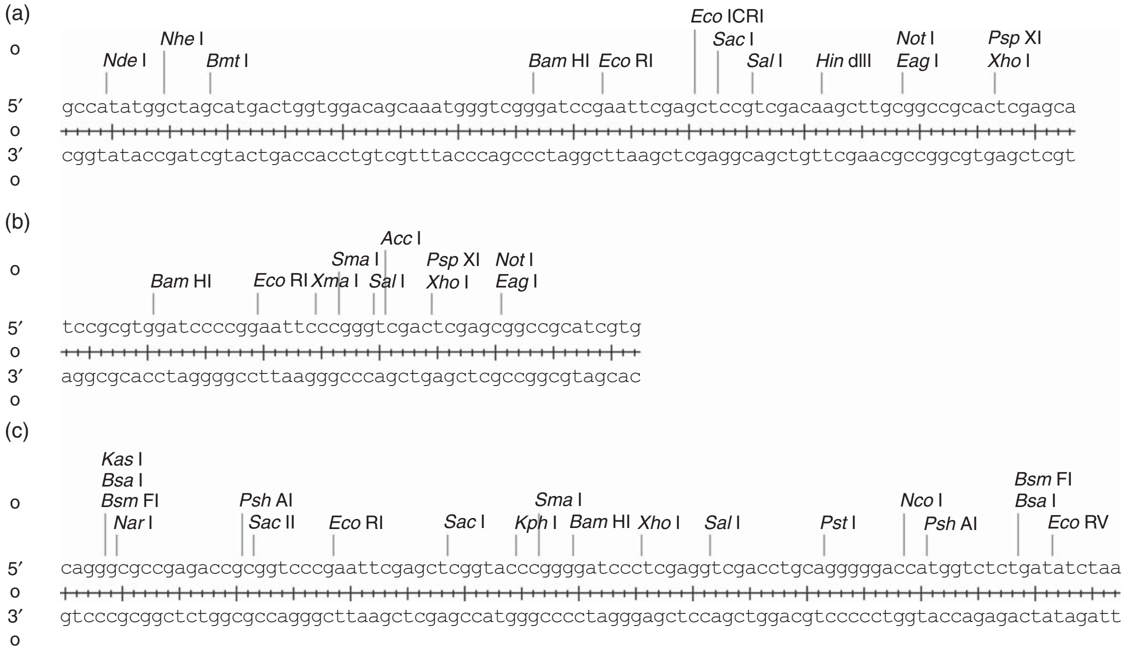

(a)
Eco ICRI
Nhe I
Sac I
Not I
Psp XI
Nde I
Bmt I
Bam HI
Eco RI
Sal I
Hin dIII
Eag I
Xho I
5′ gccatatggctagcatgactggtggacagcaaatgggtcgggatccgaattcgagctccgtcgacaagcttgcggccgcactcgagca
3′ cggtataccgatcgtactgaccacctgtcgtttacccagccctaggcttaagctcgaggcagctgttcgaacgccggcgtgagctcgt
(b)
Acc I
Sma I
Psp XI
Not I
Bam HI
Eco RI
Xma I
Sal I
Xho I
Eag I
5′ tccgcgtggatccccggaattcccgggtcgactcgagcggccgcatcgtg
3′ aggcgcacctaggggccttaagggcccagctgagctcgccggcgtagcac
(c)
Kas I
Bsa I
Bsm FI
Nar I
Psh AI
Sac II
Eco RI
Sac I
Sma I
Kph I
Bam HI
Xho I
Sal I
Pst I
Nco I
Psh AI
Bsm FI
Bsa I
Eco RV
5′ cagggcgccgagaccgcggtcccgaattcgagctcggtacccggggatccctcgaggtcgacctgcaggggaccatggtctctgatatctaa
3′ gtcccgcggctctggcgccagggcttaagctcgagccatgggcccctagggagctccagctggacgtcccctggtaccagagactatagatt

Figure 6 Different multiple cloning sites in different vectors. (a) MCS from pET-28a (Novagen, Madison, WI). (b) MCS from pGEX-4T-1 (GE Healthcare, Piscataway, NJ). (c) pASK-IBA6 (IBA, Germany).

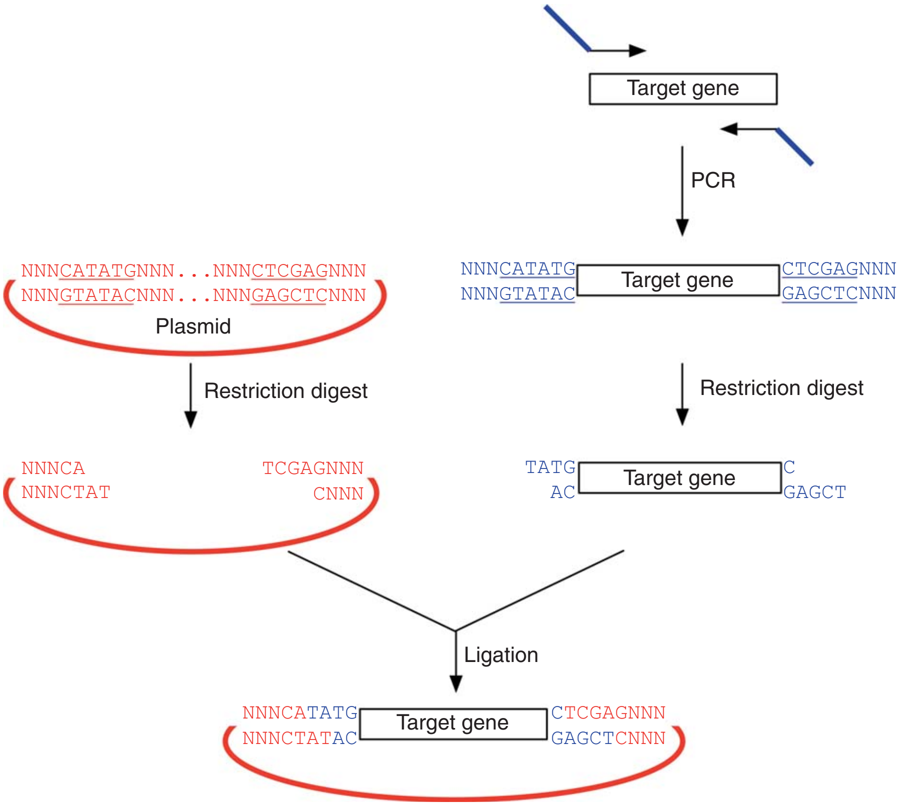


Figure 7 Cloning by restriction digest and ligation. Both the PCR-amplified gene of interest and the target vector are digested with the same restriction enzymes (or with enzymes that produce compatible ends). The short section of overhanging sequence (called 'sticky ends') are complementary to each other. The digested DNAs are mixed, treated with ligase, and transformed into a suitable *Escherichia coli* host strain to produce a new recombinant DNA molecule with the gene of interest specifically inserted into a host vector.

enzymes. Frequently, the vector is further treated with a phosphatase to remove the 5′ phosphates to reduce the background of vector religation resulting from incomplete digestion. Finally, the fragments are ligated together and transformed into competent *E. coli* (see **Figure 7**). For most laboratories, restriction-based cloning has sufficient throughput and is flexible enough to accommodate their cloning needs.

Restriction-based cloning is the most widely used of the cloning methods available and a large number of vectors are available with various MCSs. Vectors designed for use with restriction-based cloning methods are available for every commonly attached fusion protein, for every expression system, and with the full range of promoters. However, because these vectors have differing lineages and are available from different sources, they often carry different MCSs (see **Figure 6**). This can make testing several fusion proteins, for example, quite arduous as the whole cloning protocol, from PCR onward, needs to be repeated in order to incorporate restriction sites for each vector to be tried. This incompatibility of vector MCS is the biggest drawback of using restriction-based cloning. Many laboratories circumvent this problem by making their own set of vectors for the fusions, promoters, etc. that they commonly test with compatible MCSs.[60] For example, in our laboratory, we have a set of vectors for attaching MBP, GST, His tag, secretion signals, small ubiquitin-like modifier (SUMO), and a few others, all with the same MCS. Once cloned into one vector, the gene can be moved to any other vector by digestion and ligation in a one-day procedure with high efficiency. We also make an effort to clone all of our genes with the same two restriction enzymes, which facilitates parallelization and decreases failures due to human error. However, another drawback of the restriction-based cloning method is that extra restriction sites in the coding region of the gene lead to cleavage of the target DNA when the restriction enzyme is added. Problematic restriction sites can be removed by single overlap extension–PCR (SOE–PCR)[61] during the cloning process; however, it is not always facile to do so. Vectors that use extremely rare restriction sites (Flexi Vectors, Promega, Madison, WI) have been developed to avoid the issue of extra restriction sites in the target gene.[62] These vectors have been utilized in high-throughput structural genomics efforts.[63,64]

9.19.4.3 Recombination-Based Cloning Methods

The recombination-based cloning methods work by swapping DNA between two vectors or between a vector and a linear DNA fragment at points of sequence homology. Two site-specific recombination methods predominate:[65] the Gateway system (Invitrogen, Madison, WI), which transfers ORFs by λ-att recombination at the 25-bp *att*B site,[66] and the Creator system (Clontech, Madison, WI), which utilizes Cre-*loxP* recombination of the 34-bp lox-P site upstream of the ORF.[67] Conceptually, the recombination regions can be thought of as large restriction sites and the recombination reaction as a concerted digestion and ligation. The long length of the region of homology gives the reactions high specificity. The InFusion method (ClonTech) is slightly different in that it uses homologous recombination instead of site-specific recombination. This allows the use of any vector, but, like the two site-specific methods, it requires fairly long extensions on the ends of the PCR product.

Recombinational cloning has the advantage of being readily adapted to high-throughput settings and simplifying the transfer of an ORF into many different expression vectors without the issues of MCS matching and inconvenient restrictions sites that plague restriction/ligation-based cloning. For these reasons, many high-throughput structural genomics centers rely on recombinational cloning[68,69] and libraries of human ORFs are available in recombination-ready vectors.[14]

9.19.4.4 Ligation-Independent Cloning

Ligation-independent cloning–PCR (LIC–PCR) uses T4 DNA polymerase in the presence of a single deoxyribonucleotide to produce 12–15 bp overhangs on a PCR product.[70–72] The recipient vector is similarly treated so that both PCR product and vector contain long complementary overhangs. This is similar to the short sticky ends produced by restriction digest. However, the length (12–15 bases) means that the extensions anneal strongly enough to survive transformation into *E. coli* without the need for *in vitro* ligation. Once in the host, cellular repair enzymes close the nicks. LIC–PCR is used in several high-throughput settings.[69]

9.19.5 Methods to Improve Expression and Solubility

Unfortunately, the majority of proteins will not express well and in soluble form with an N-terminal His tag in *E. coli*. The high levels of mRNA produced by the strong promoters used for overexpression can lead to ribosome destruction and cell death. The high rate of protein synthesis and the accumulation of elevated levels of the recombinant protein can lead to aggregation and precipitation of the protein in the form of inclusion bodies, insoluble masses of aggregated protein. These bodies can accumulate in the cytoplasm or periplasm (if the protein has been engineered to secrete). There are several contributing factors to protein misfolding in recombinant overexpression systems. The *E. coli* cytoplasm is crowded with proteins (macromolecule concentration can reach 300–400 mg ml^{-1})[73,74] and the tight coupling of transcription and translation means that a complete protein chain can leave the ribosome every 35 s.[75] The crowded environment and rapid protein production makes proper folding difficult. Small, quickly folding proteins can often achieve the correct conformation without assistance; however, large, complex, multidomain proteins often require interactions with chaperones or other proteins (e.g., folding catalysts such as prolyl isomerases) in order to fold properly. However, the use of strong promoters and robust induction may lead to protein production rates that can swamp the available cellular folding machinery. Additionally, the high-level production of overexpressed protein (50% or more of total protein in some cases) can cause the rate of protein aggregation to be faster than proper protein folding. For many heterologously expressed proteins, the inability of *E. coli* to supply necessary post-translational modifications (such as glycosylation and phosphorylation) also contributes to misfolding and the formation of inclusion bodies. The strongly reducing cytoplasm of wild-type *E. coli* disfavors the formation of disulfide bonds, which can lead to aggregation of disulfide bond-containing proteins.

Inclusion bodies are typically composed of 80–95% the protein being overproduced with contaminants being composed of outer membrane proteins, ribosomal proteins, phospholipids, nucleic acids, and, in some cases, folding modulators.[76–78] Proteins sequestered in inclusion bodies are resistant to proteolysis. The high purity of target protein and the protection from proteases makes inclusion bodies an attractive source for recombinant proteins provided that refolding conditions can be found (see Section 9.19.5.7) or that the end use does not require properly folded protein. In fact, for some highly proteolytically labile polypeptides, it can be advantageous to deliberately drive the protein to inclusion bodies by fusing it to the insoluble KSI tag.[79,80] Short polypeptides that may not have much secondary structure are particularly prone to proteolysis and can benefit from being the protection that inclusion bodies afford. Additionally, peptides that are toxic to *E. coli* may be produced by deliberately sequestering them in inclusion bodies.

Luckily, there are a number of rescue strategies available that may enable sufficient quantities of protein to be obtained without having to switch to a more costly expression system. Baneyx and Mujacic[81] reviewed the mechanisms of protein folding in *E. coli* and strategies to obtain correctly folded recombinant proteins in great detail. It is reasonably facile and inexpensive to test multiple cell strains, expression conditions, and fusion partners; therefore, it makes sense to screen several *E. coli* conditions before moving to an alternate system.

9.19.5.1 Induction Temperature

In our laboratory, reduced temperature at induction is the first thing attempted to increase the solubility of an expressed protein. Frequently, reducing the induction temperature from 37 to 15–20 °C can change the solubility level of an overproduced protein from undetectable to adequate.[82] Reduced temperature slows both transcription and translation rates and reduces the strength of hydrophobic interactions. However, below 15 °C many promoter systems are inefficient. For these very low temperatures, use of cold-inducible promoters can be advantageous as most commonly used promoter systems work with reduced efficiency at decreased temperatures.[83–85] Systems based on the promoter from the major *E. coli* cold-shock gene *cspA* have been described and have demonstrated their utility with membrane-associated proteins as well as those prone to proteolysis.[86,87]

9.19.5.2 Fusion Tags

Addition of a fusion protein that encourages proper folding to the N-terminus of the protein of interest is a frequently successful strategy for obtaining soluble protein. Dozens of proteins have been described in the literature as increasing the solubility of proteins fused to their C-termini; however, only a handful of fusion proteins have achieved common usage[88] (see **Table 1**). The most commonly used fusions for solubility enhancement are GST,[89] maltose-binding protein (MBP),[90,91] NusA,[92] thioredoxin (TRX),[93] and ubiquitin or ubiquitin-like proteins.[94–97] There have been several studies of the relative efficacy of these fusions for the enhancement of solubility.[92,98–104] Within each study, different proteins were solubilized by different fusions and no one fusion emerged as a panacea. The different studies found different rankings for the efficacy of the fusions. The inconsistency from study to study is probably explained by the different proteins used in each trial and serves as a reminder that the optimal conditions for each protein will be different. Therefore, it is common to test several different fusions for their ability to encourage soluble expression of the protein of interest.

The mechanism by which fusion proteins enhance expression and solubility remains unclear. The molten globule hypothesis is that the fusion tag acts as a nucleus of folding.[105–107] Several studies have been performed with MBP to elucidate the mechanisms by which it promotes the solubility of fusion partners.[108–110]

It is not uncommon to discover that a protein expressed in soluble form by attachment of a fusion partner is not correctly folded. Therefore, it is recommended that additional measurements (ligand binding, enzymatic activity, circular dichroism (CD) spectropolarimetry, or others) be taken to determine if the protein is correctly folded. Additionally, many of the solubility-enhancing fusions are large, which decreases the overall yield of the protein of interest.[111]

9.19.5.2.1 Periplasmic localization

An option for allowing disulfide bond formation in heterologously expressed protein is to export the protein to the periplasmic space. This is typically achieved by fusing a signal sequence to the N-terminus of the protein (the signal sequence of PelB or OmpA is commonly used) or by fusion to ecotin[112] or other periplasmically localized proteins.[45,113,114] The *E. coli* periplasm is home to DsbA (a thiol oxidant) and DsbC (a disulfide bond isomerase)[115] both of which act to encourage proper disulfide bond formation. There are also a number of chaperones and other proteins involved in protein folding that are specific to the periplasmic space. There are other advantages to periplasmic localization of heterologous proteins: (1) removal of the signal peptide by leader peptidases allows the production of an authentic N-terminus, (2) there are fewer proteases in the periplasm than in the cytoplasm and they tend to have specific substrates, and (3) the periplasm contains fewer proteins than the cytoplasm and they can be released specifically by osmotic shock or other methods, simplifying purification.[81,116–120] The mechanisms of export to the periplasm and the cellular machinery associated with both export and folding have been well reviewed.[81,121] The engineering of strains with improved characteristics for periplasmic localization of recombinant proteins is ongoing.[122–125]

9.19.5.3 Growth Medium

9.19.5.3.1 Inducer titration

Reducing the synthesis rate of a target protein is a common approach to improving folding.[126] Typically, this is achieved by reducing the induction temperature (see Section 9.19.5.1), changing to a weaker promoter,[127–129] or by decreasing the concentration of inducing agent.[130–132] Promoters utilizing control elements of the *lac* operon (e.g., *tac*, *trc*, or the hybrid T7*lac* promoters) can be partially induced by IPTG concentrations below 100 μmol l^{-1} (1 mmol l^{-1} is a typical concentration for full induction). However, heterogeneity of expression of the *lacY* permease can lead to varying doses of IPTG within the population of cells in the culture; therefore, it may be beneficial to use a *lacY*-deficient strain (such as the Tuner strains from Novagen, Madison, WI).[133–135] In the absence of permease all the cells of the culture receive a dose of IPTG dependent only on diffusion across the membrane. The *araB* (pBAD) promoter is often considered more easily titrated than the *lac* promoters, but it is off/on in wild-type cells. In order to assess the impact of partial induction by subsaturating concentrations of inducer (arabinose) it is necessary to use host strains that have been engineered to constitutively transport arabinose by expression of the AraE transporter.[136,137] However, it is possible to obtain a linear relationship between expression level and inducer concentration over two orders of magnitude.[138–140]

9.19.5.3.2 Auto induction

Media that allows unattended induction of protein production has been developed.[44] This medium removes the requirement for monitoring of culture density and human intervention to add inducing agent at the correct growth stage. Additionally, significantly greater cell mass is obtained from the same culture volume than is achievable in shake flasks with standard media formulations. A recent study investigated the effects of different promoter and medium compositions on protein expression and presented media changes that increased protein production levels for a variety of proteins.[141]

9.19.5.3.3 Ligand supplementation

Ligand binding to the active site of an enzyme or receptor frequently results in protein stabilization by polar and hydrophobic interactions. This change in protein stability is utilized in the thermal shift assay, which detects changes in the midpoint transition temperature upon thermal denaturation of the protein.[142,143] Binding of the ligand stabilizes the protein and results in an increase in the transition temperature. This protein stabilization effect can be used to increase the stability and solubility of heterologously expressed proteins.[144] For example, Elleby *et al.* found that inclusion of a small protein ligand during induction increased the level of soluble protein for their target more than 100-fold.[145] In a larger trial, 500 constructs representing 65 distinct genes, all of the short-chain dehydrogenases/reductases family, were screened for soluble expression in the presence and absence of carbenoxolone, a compound known to bind to several members of the family. Soluble expression was increased from 70 constructs to 86 constructs in the presence of carbenoxolone; in most cases the enhancement of soluble production was from two- to three-fold, in some the increase was from undetectable to mg l^{-1} levels.[146] The beneficial effects of ligand supplementation have also been seen in the baculovirus expression system.[147,148]

It is hypothesized that the ligand can bind to a partially folded intermediate and stabilize it, hence encouraging proper folding and discouraging aggregation. In order to implement this method, it is necessary to have some ligand information. However, it may be possible, as in the carbenoxolone study, to utilize general inhibitors that demonstrate binding activity against several members of a protein family.

9.19.5.4 *Escherichia coli* Host Strain

There are many commercially available *E. coli* cell strains for the heterologous expression of recombinant proteins. By far the most common promoter system is the T7 expression system and the most commonly used expression host is BL21(DE3). If one is using a promoter that is recognized by the *E. coli* RNA polymerase, any cell strain can be used. Frequently useful are cell strains that supplement rare codons and strains that alter the ability of *E. coli* to produce disulfide bonds in the cytoplasm. However, there are cell strains that have been developed for their ability to express toxic proteins[149,150] and a wide range of other strains that may be useful for any particular protein.

9.19.5.4.1 Codon supplementation

The 20 amino acids plus stop codons are coded for by 64 different triplet codons. This means that there is degeneracy in the coding. Some amino acids are only coded for by one possible triplet (methionine by ATG and tryptophan by TGG) or two possible triplets. However, some amino acids can be coded for by four or six possible codons. Coding is nonrandom and organisms exhibit codon bias.[151,152] *E. coli* exhibits distinct preferential codon usage for highly expressed (Class II) proteins.[153] Therefore, in heterologous expression systems, it is not unusual to have a gene that utilizes a very different set of codons than the host organism. This can cause problems during translation because the intracellular pool of tRNA is biased toward the usage of the host organism. In high-level expression of a gene that uses a different set of codons, the pool of a codon that is rarely used by the host can be depleted, causing pauses and stalling of translation as the ribosome waits for tRNA. There are two good solutions to this problem. The first is to mutate or synthesize the gene of interest to alter to the coding to match the host organism. Using the *E. coli* preferred coding is common when genes are synthesized and several proteins have been successfully produced by mutating rare codons to more preferred ones.[154–162] Another option is to express the protein in a host strain incorporating a plasmid that provides higher levels of the rare codons, such as the Rosetta strains (Novagen, Madison, WI) or the BL21-CodonPlus strains (Stratagene, La Jolla, CA).[163–168] Both methods work with roughly equivalent efficacy, with the codon supplementation methods being easier to implement and less expensive.[169–171] However, synthesizing codon-optimized genes allows improvement of other aspects of the gene (such as mRNA secondary structure). Additionally, utilization of the plasmids for tRNA supplementation may limit the ability to perform coexpression experiments.

9.19.5.4.2 Disulfide bond formation

In the *E. coli* cytoplasm the thioredoxin system (thioredoxin reductase and two thioredoxins) and the glutaredoxin system (glutathione reductase, glutathione, and three glutaredoxins) use the reducing potential of NADPH to reduce disulfide bonds.[115] The combined action of these two systems disallows the formation of stable disulfide bonds in the *E. coli* cytoplasm. Many eukaryotic proteins contain disulfide bridges whose formation is integral to correct folding. The first strain engineered to allow disulfide bonds in the cytoplasm contained a disrupted *trxB*,[172] which actually converted TrxA and TrxC from reductases to oxidases by causing them to accumulate in their disulfide bonded form.[173] Strains containing the *trxB* mutation were shown to allow correct disulfide bond formation and increase soluble production of some recombinant proteins.[174,175] The cytoplasm can be made more oxidizing additionally eliminating the activity of the glutathione pathway by creating null mutations in *gor* or *gshA*;[176] however, the double mutant displays impeded aerobic growth. Further work isolated suppressor strains that allowed good growth of the double mutants and these strains have demonstrated their utility for the production of disulfide-bonded heterologous proteins.[114, 177–186]

9.19.5.5 Coexpression

By using plasmids with different antibiotic resistances and different origins of replication, more than one plasmid can be transformed into an *E. coli* host cell. This allows the production of different proteins at the same time or, by using different promoter systems, nonsimultaneously within the same cell. There are also plasmids with more than one cloning site and promoter (such as the Duet plasmids from Novagen, Madison, WI). These systems are often used to coexpress two or more proteins.

9.19.5.5.1 Chaperones

It is often assumed that the amino acid sequence of a protein contains all the information needed for correct folding into the final tertiary structure. However, folding pathways display a great deal of complexity. Many proteins need to interact with one or more foldases or chaperones in order to fold correctly. High-level overexpression of a protein can titrate the available cellular pool of folding modulators. Coexpression of common chaperone proteins is typically affected by use of a separate expression plasmid using a different promoter system. This allows nonsimultaneous induction of chaperone and protein of interest. Typically, the chaperone is induced before the target protein to insure that the chaperones are present at the start of target protein synthesis. Several plasmids have been described in the literature for chaperone coexpression.[187,188]

Stratagene introduced a cell strain designed to facilitate expression at very reduced temperatures (down to 10 °C). The ArcticExpress cell strain coexpresses two cold-shock chaperones (Cpn60 and Cpn10 from the psychrophilic bacterium *Oleispira antarctica*) and allows expression down to 10 °C. One caveat with the strain is that the chaperones can be difficult to separate from the protein of interest. However, a recent report demonstrates that incubation with $MgCl_2$/ATP/KCl removes the copurifying chaperonin,[189] which should be independent of the protein target because it acts directly on the chaperonin system.

9.19.5.5.2 Interacting partners

Many proteins, particularly those of eukaryotic origin, require an interacting protein partner for correct folding and stability.[190–193] Often, these proteins contain intrinsically disordered domains that mediate the protein interaction. Frequently, proteins with unstructured domains cannot be expressed solubly in *E. coli.* In some cases, coexpression of the interacting protein has improved the stability and solubility of the protein of interest.[194–197] As an example, the yield of inducible nitric oxide synthase (iNOS) was 20–25 times greater when coexpressed with its natural partner, calmodulin, than in the absence of calmodulin. Additionally, soluble iNOS could be isolated when expressed without calmodulin but it was deficient in heme and flavins and almost completely inactive.[198] Methods for coexpression in both prokaryotic and eukaryotic hosts have been recently reviewed.[199]

9.19.5.5.3 Accessory proteins

In some cases, the protein of interest requires the assistance of accessory proteins to correctly assemble an active site structure. For example, iron–sulfur cluster-containing proteins are typically expressed in their apo forms in *E. coli.* Coexpression of a set of proteins that assists in the delivery of sulfur to these clusters can allow the protein to be obtained with its cluster intact and active. Correct cluster formation can stabilize the protein and increase the soluble yield.[200–202]

9.19.5.6 Project Redesign

In addition to changing the expression strain, media, induction conditions, fusion proteins, and other experimental parameters previously discussed, it is sometimes beneficial to actually make changes to the coding sequence of the protein of interest. These alterations can be solely at the DNA level (silent mutations), which may affect mRNA folding and stability. Changes to charged surface residues can affect solubility.[203,204] Other point mutations can have large and largely unpredictable effects on the stability of the protein, although there are algorithms for rationally design of point mutations to increase stability.[205,206] Finally, truncated versions of the full-length protein can be investigated in the search for a well-behaved polypeptide.

Generally, changes at the 5′ end of the gene have the largest effect on expression levels. For example, alteration of the mRNA sequence by silent mutagenesis can reduce deleterious mRNA secondary structure formation and increase expression.[207] The amino acid sequence at the N-terminus of the protein can have a large effect on heterologous expression.[208–210] For example, NGG codons at positions + 2, + 3, and + 5 lower expression levels in *E. coli* at the translational level. It is thought that the low expression level is not caused by mRNA secondary structure or depletion of the intracellular tRNA pool.[211]

There are high-throughput screening methods for searching for proteins of altered structure with increased solubility. Many of these methods were developed for searching for soluble domains of full-length proteins by random truncation; however, the methods are equally applicable to screening for other types of alterations such as mutations arising from error-prone PCR, DNA shuffling, or other methods. Typically, these rely on appending a fusion partner to the C-terminus of the protein and monitoring the activity of the C-terminal fusion partner. The assumption is that constructs that result in higher levels of expressed soluble protein will also demonstrate higher activity of the C-terminal partner. Several different reporters of soluble expression have been described. These can be grouped into those that lead to a color,[212] or fluorescence[54,213–215] and those that confer a growth advantage on clones producing soluble protein, typically by resistance to an antibiotic.[216–218] Generally, the color and fluorescence assays are useful for smaller libraries with nonrandom insert cloning. Random insert cloning leads to only 1 in 18 of the clones being in frame and in the correct orientation, which necessitates larger library sizes. These large libraries are more quickly screened with one of the dominant genetic markers.[219] It can be beneficial to make a nested set of constructs as differences of a few amino acids at either end can vastly impact protein expression, solubility, stability, and performance.

9.19.5.7 Refolding

Many consider refolding to be a method of last resort. However, the high purity of protein in inclusion bodies and advances in refolding screening technologies have made refolding an attractive option.[220] Good yields of high-quality protein can be obtained from refolding strategies. Generally, the insoluble protein is solubilized in a chaotropic agent (high concentrations of urea or guanidine are most commonly used) and then refolded by buffer exchange. Disulfide exchange agents are added as needed and folding chaperones or other enzymes can be added as needed to effect proper folding. There are protocols for refolding of proteins on columns[221–223] and at high pressures.[224–226] Refolding is also commonly performed by dialysis or rapid dilution. There are several good reviews of protein refolding techniques.[227–230]

9.19.6 Affinity Protein Purification

9.19.6.1 Introduction

Even seemingly insignificant changes in amino acid sequence can lead to large perturbations in protein behavior. It is quite straightforward to perform DNA manipulations in a high-throughput effort as DNA behavior is largely sequence independent; conversely, it is quite difficult to predict, *de novo*, the behavior of a protein based solely on primary amino acid sequence. However, the introduction of affinity sequences enables the use of standardized methods for the purification of proteins. With affinity fusion tags, a fairly general purification scheme can be utilized that will serve as an excellent starting point for many proteins. A comparative study of the efficacy and cost of some common affinity tags is available.[231]

9.19.6.2 Affinity Methods

Affinity tags allow the recombinant protein to be separated from host cell proteins by virtue of a specific binding interaction to an immobilized ligand. A general scheme for an affinity purification is shown in **Figure 8**. Greater than 1000-fold purification can be achieved in a single pass of a sample through an affinity column. Often, the purity level obtained after a single affinity purification step is sufficient for the intended use of the target protein and no further manipulations are necessary. In the ideal case, the fusion tag would be small to minimize the likelihood of interfering with the activity and/or structure of the protein of interest. Additionally,

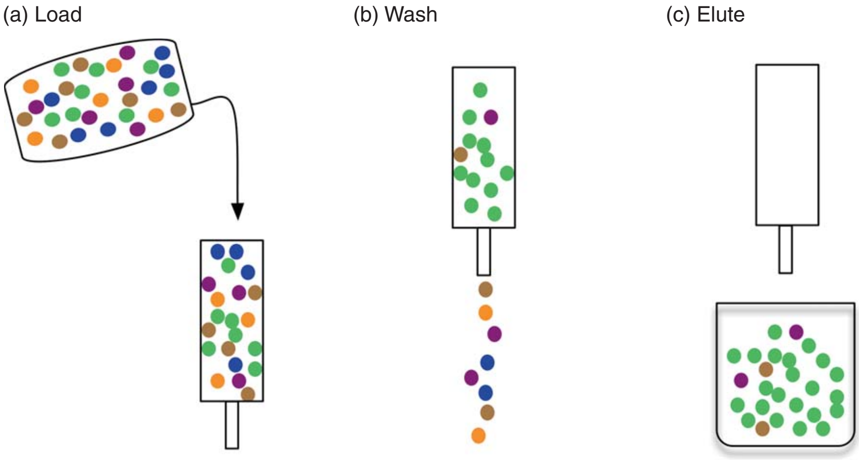


Figure 8 Affinity purification overview. (a) A mixture of cellular proteins, including the protein of interest with an affinity tag attached (green) is loaded onto a solid-support resin chemically modified to bind the affinity tag. (b) The solid support is washed and the majority of proteins not specifically bound to the resin are removed. (c) The protein of interest is eluted by interfering with the affinity interaction (commonly by competition with a small molecule).

the ideal affinity tag would bind tightly and rapidly to an inexpensive, high-capacity resin and no (or few) host cell proteins would bind. Finally, elution would be simple, specific, and utilize inexpensive reagents without resorting to unusual buffer conditions that might be deleterious to the target protein. Dozens of fusion tags enabling affinity purification have been reported in the literature (Table 1). However, there are only a few that have been widely adopted. Many fusions are also used as reporters, enabling facile, sensitive detection of the fusion protein. Still others have the potential to increase the solubility of the expressed fusion.

9.19.6.2.1 Immobilized metal affinity chromatography

The His tag[248] is by far the most popular affinity tag for purification of recombinant proteins. Typically, the tag is composed of 6–10 consecutive histidines at either terminus of the protein of interest, often separated by a protease-cleavage site. The presence of a His tag enables the use of IMAC for purification. IMAC is a rapid

Table 1 Common fusion tags for recombinant protein production

Fusion	*Size (kDa)*	*Solubility (S), Affinity (A), or Detection (D)*	*Reference(s)*
Maltose-binding protein (MBP)	40	S, A	102, 232–234
Glutathione *S*-transferase (GST)	26	S, A	235, 236
Ubiquitin or ubiquitin-like (SUMO, Smt3, Ub, HUE)	11–15	S	94, 95, 98, 237–240
NusA	55	S	232
DsbA	23	S	241
Thioredoxin	12	S	102
GB1	7	S	99, 242
Cellulose-binding domain	5	A	243–248
His tag	1	A	249, 250
FLAG	1	A, D	251–255
Calmodulin-binding peptide	2	A	256–259
Biotin affinity peptide (AviTag, Bioease)	2	A, D	260–264
StrepII	2	A, D	265, 266
Green fluorescent protein (GFP)	27	D	215, 267–269
Halo tag	33	A, D	64, 270–272
SNAP	24	D	273

affinity method based on binding of the adjacent histidines of the His tag to an immobilized divalent metal ion (nickel is the most frequently used, followed by cobalt, but other divalent metals are also used). The His tag binds to the metal at neutral to slightly basic pH (pH 7.5–8 is typical) and the protein can be eluted by lowering the pH to 4–5, stripping the metal from the polymeric support with high concentrations of ethylenediaminetetraacetic acid (EDTA) or, most commonly, by competition with imidazole. The method is quick, inexpensive, and straightforward, which has led to widespread adoption. According to Derewenda 90% of deposited crystal structures are derived from recombinantly produced proteins, with almost 60% using a His tag for purification.[274]

The His tag is small (6–10 amino acids) and often does not require removal before use of the recombinant protein. In fact, although many crystallographers cleave the tag before screening for crystallization conditions,[274,275] Carson *et al.* showed that the tags generally had no significant effect on the structure of the attached protein.[276] However, the tag may still alter solubility or aggregation state of the purified protein. An additional benefit of the His tag is that its affinity for divalent metal cations is not dependent on a particular protein fold or secondary structure; therefore, IMAC can be performed under denaturing conditions ($6\,\mathrm{mol\,l^{-1}}$ guanidine hydrochloride or $8\,\mathrm{mol\,l^{-1}}$ urea, for example). This can be useful when the protein cannot be expressed in a soluble form and refolding is being pursued (see Section 9.19.5.7).

9.19.6.2.2 Glutathione S-transferase

GST has been used as an affinity fusion since 1988.[89] In many cases, GST has also been shown to enhance the solubility of the protein to which it is fused. GST was one of the first widely adopted fusion tags for affinity purification and it has been used for the purification of innumerable proteins. Generally, GST fusions are separated from the protein of interest by a protease recognition site (thrombin or Factor Xa in the original vectors). In more recent years, a fusion of the human rhinovirus 3c protease and GST has been marketed for cleavage of appropriately constructed GST fusions (PreScission protease, GE Healthcare, Piscataway, NJ).[277,278] Plasmids for expression of the GST-3c fusion are also widely distributed in the academic community. An advantage of using the fused protease is the facile removal of the protease and reaction byproducts by subtractive affinity chromatography (see Section 9.19.6.3.4).

9.19.6.3 Tag Removal

Although the introduction of fusion tags has many potential benefits for the production and purification of recombinant proteins, they may interfere with the final use of the protein; therefore, they often need to be removed. Many plasmids are designed to introduce a protease recognition site between an introduced tag and the protein of interest. Commonly used proteases and their recognition sites are shown in **Table 2**. Generally,

Table 2 Tag removal enzymes

Protease	*Recognition sequence*	*Reference(s)*
Factor Xa	I-E/D-G-R-!	279
Enterokinase	D-D-D-D-K-!	255
Genenase	P-G-A-A-H-Y-!	280, 281
Furin	R-X-K/R-R-!	282, 283
Thrombin	L-V-P-R-!-G-S	279, 284–286
Tobacco etch virus (TEV)	E-N-L-Y-F-Q-!-G/S	287–290
Tobacco vein mottling virus (TVMV)	E-T-V-R-F-Q-!-S	291
Casp4	L-E-V-D-!	292
Casp6	V-E-I-D-!	293, 294
3c (PreScission)	L-E-V-L-F-Q-!-G-P	277, 278, 295
Ulp1	Specific to SUMO tag	94, 296
SenP2	Specific to SUMO tag	94, 296
Intein processing	Specific autocleavage	297–300
TAGZyme system	Exoproteolysis	301–303

Enzymes cut at the position of the '!'.

the fusion protein containing the protein of interest is purified and then cleaved by endoproteolysis. Post cleavage, it is necessary to remove the cleaved fusion tag, the protease, and any uncleaved fusion protein. For this reason, it is beneficial to have an engineered protease that incorporates a handle to enable facile removal. For example, thrombin can be purchased in a biotinylated form, which enables the removal of thrombin from the proteolysis reaction mixture by binding to immobilized streptavidin. Two proteases in particular have become extremely popular. Both the tobacco etch virus (TEV) protease and the rhinovirus 3c protease are easy to produce and purify in-house and many laboratories have expression constructs for these enzymes. Other enzymes can remove specific tags and leave no unwanted amino acids behind.

9.19.6.3.1 TEV protease

TEV protease has become one of the most popular proteases for use in separating recombinant proteins from their attached fusion partners because of its stringent sequence specificity and activity over a broad temperature range. The protein can be recombinantly produced in *E. coli* with a His tag for easy purification and subsequent removal from reaction mixtures.[290,304,305] The wild-type enzyme suffers from autoproteolysis, which leads to a truncated form with vastly reduced activity; therefore, specific mutants are used that stabilize the protein and increase its catalytic efficiency.[288] The enzyme is well studied and its crystal structure has been solved.[287,289,306] An additional benefit of TEV protease is that many amino acids are tolerated in the P1′ position of its recognition site; however, efficiency decreases for some amino acids and proline abolishes cleavage.[287] Since the P1′ amino acid will remain on the protein of interest when N-terminal fusions are cleaved off, this does allow some flexibility to match the naturally occurring N-terminus post cleavage.

9.19.6.3.2 Rhinovirus 3c protease

Like TEV protease, the rhinovirus 3c protease is widely used and plasmids for its production are readily available.[277,278] The 3c protease is marketed with a GST tag by GE Healthcare and several of the frequently used commercial vectors for production of protein with GST tag carry the 3c protease recognition site.

9.19.6.3.3 Other tag removal systems

In addition to the commonly used TEV and 3c proteases, several other proteases are commonly used to remove fusion tags (see **Table 2**). There are also other tag removal systems that do not rely on proteases. For example, the intein system is an autocatalytic cleavage of the produced fusion protein.[297] For SUMO fusion proteins there are several hydrolases available for tag removal. Because the SUMO hydrolases recognize the whole fold of the tag and not just an amino acid sequence, they are highly specific and not prone to unintended cleavage.[94,296]

9.19.6.3.4 Subtractive affinity chromatography

By adding the same affinity tag onto the protease as is on the protein of interest, the protease, cleaved tag, and uncleaved fusion can be separated from the protein of interest in one step as all three will bind to the same affinity resin, while the protein of interest, now devoid of its affinity tag, no longer binds. This is sometimes called a ‘subtractive’ chromatography step as the protein of interest flows through the column without binding and the impurities bind and are removed. An additional benefit of a subtractive chromatography step is that, frequently, impurities that are copurified in the original affinity chromatography step will bind to the column again. Then, because no competitive elution reagent is used, they remain bound and cease to contaminate the protein of interest. An example of subtractive IMAC is shown in **Figure 9**. In lane 1, uncleaved fusion protein, cleaved protein of interest, and the cleaved tag are all visible. After subtractive IMAC, the cleaved protein of interest is nearly homogeneous. As noted above, TEV protease is commonly available with a His tag and the Rhinovirus 3c protease is commercially available as a GST fusion, although an expression vector for a His tag version is also widely circulated. The SUMO hydrolases are also commercially available with His tags (LifeSensors, Malvern, PA).

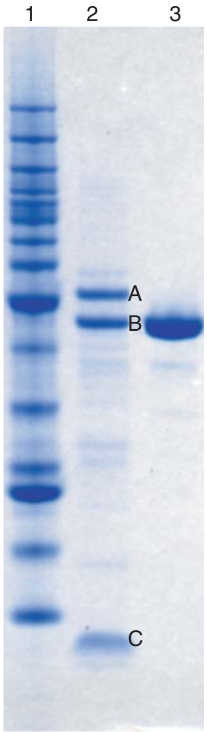


Figure 9 Subtractive affinity chromatography. Lane 1: Molecular weight markers. Lane 2: Partially proteolized fusion protein. Lane 3: Purified protein after subtractive affinity chromatography. A: Full-length fusion protein. B: Protein of interest cleaved from fusion tag. C: Cleaved fusion tag. The full-length fusion protein and cleaved fusion tag (fragment A and C) are both bound to the affinity column. Fragment B (the protein of interest) did not bind to the affinity column and was collected in the flow through fraction.

9.19.7 Conclusions

A high proportion of targets can be obtained from *E. coli* in yields suitable for many uses. Progress continues to be made in engineering of cell strains, fusion technologies, growth medium alterations, proteases, and other variables of recombinant protein production in *E. coli.* The advantages of rapid screening in small volumes, low cost, and ease of manipulation make *E. coli* a very attractive starting point for any protein production project. Although the range of options to be explored can seem overwhelming, many variables can be assayed in parallel, increasing the odds of finding a set of conditions that will enable the production of adequate amounts of the protein of interest.

Abbreviations

CD	circular dichroism
cDNA	complementary DNA
DNA	deoxyribonucleic acid
GST	glutathione *S*-transferase
IMAC	immobilized metal affinity chromatography
IPTG	isopropyl-β-D-thiogalactoside
LIC–PCR	ligation-independent cloning–polymerase chain reaction
MBP	maltose-binding protein
MCS	multiple cloning site
MGC	Mammalian Gene Collection

mRNA	messenger ribonucleic acid
ORF	open reading frame
ORI	origin of replication
PCR	polymerase chain reaction
RNA	ribonucleic acid
SDS–PAGE	sodium dodecyl sulfate–polyacrylamide gel electrophoresis
SOE–PCR	single overlap extension–polymerase chain reaction
SUMO	small ubiquitin-like modifier
TAP	tandem affinity purification
TEV	tobacco etch virus
TRX	thioredoxin
TVMV	tobacco vein mottling virus

References

1. J. D. Pickert; B. L. Miller. Cloning as a Tool for Organic Chemists. In *Comprehensive Natural Products Chemistrty*; E. T. Kool, Ed.; Elsevier Science: Amsterdam, 1999; Vol. 7, pp 643–674.
2. P. Braun; J. LaBaer, *Trends Biotechnol.* **2003**, *21*, 383–388.
3. J. M. Canaves; R. Page; I. A. Wilson; R. C. Stevens, *J. Mol. Biol.* **2004**, *344*, 977–991.
4. M. Linial; G. Yona, *Prog. Biophys. Mol. Biol.* **2000**, *73*, 297–320.
5. C. S. Goh; N. Lan; S. M. Douglas; B. Wu; N. Echols; A. Smith; D. Milburn; G. T. Montelione; H. Zhao; M. Gerstein, *J. Mol. Biol.* **2004**, *336*, 115–130.
6. A. Savchenko; A. Yee; A. Khachatryan; T. Skarina; E. Evdokimova; M. Pavlova; A. Semesi; J. Northey; S. Beasley; N. Lan; R. Das; M. Gerstein; C. H. Arrowmith; A. M. Edwards, *Proteins* **2003**, *50*, 392–399.
7. L. Wesson; D. Eisenberg, *Protein Sci.* **1992**, *1*, 227–235.
8. S. Chakravarty; R. Varadarajan, *FEBS Lett.* **2000**, *470*, 65–69.
9. S. Chakravarty; R. Varadarajan, *Biochemistry* **2002**, *41*, 8152–8161.
10. B. G. Fox; C. Goulding; M. G. Malkowski; L. Stewart; A. Deacon, *Nat. Methods* **2008**, *5*, 129–132.
11. D. S. Gerhard; L. Wagner; E. A. Feingold; C. M. Shenmen; L. H. Grouse; G. Schuler; S. L. Klein; S. Old; R. Rasooly; P. Good; M. Guyer; A. M. Peck; J. G. Derge; D. Lipman; F. S. Collins; W. Jang; S. Sherry; M. Feolo; L. Misquitta; E. Lee; K. Rotmistrovsky; S. F. Greenhut; C. F. Schaefer; K. Buetow; T. I. Bonner; D. Haussler; J. Kent; M. Kiekhaus; T. Furey; M. Brent; C. Prange; K. Schreiber; N. Shapiro; N. K. Bhat; R. F. Hopkins; F. Hsie; T. Driscoll; M. B. Soares; T. L. Casavant; T. E. Scheetz; M. J. Brown-stein; T. B. Usdin; S. Toshiyuki; P. Carninci; Y. Piao; D. B. Dudekula; M. S. Ko; K. Kawakami; Y. Suzuki; S. Sugano; C. E. Gruber; M. R. Smith; B. Simmons; T. Moore; R. Waterman; S. L. Johnson; Y. Ruan; C. L. Wei; S. Mathavan; P. H. Gunaratne; J. Wu; A. M. Garcia; S. W. Hulyk; E. Fuh; Y. Yuan; A. Sneed; C. Kowis; A. Hodgson; D. M. Muzny; J. McPherson; R. A. Gibbs; J. Fahey; E. Helton; M. Ketteman; A. Madan; S. Rodrigues; A. Sanchez; M. Whiting; A. Madari; A. C. Young; K. D. Wetherby; S. J. Granite; P. N. Kwong; C. P. Brinkley; R. L. Pearson; G. G. Bouffard; R. W. Blakesly; E. D. Green; M. C. Dickson; A. C. Rodriguez; J. Grimwood; J. Schmutz; R. M. Myers; Y. S. Butterfield; M. Griffith; O. L. Griffith; M. I. Krzywinski; N. Liao; R. Morin; D. Palmquist, *Genome Res.* **2004**, *14*, 2121–2127.
12. R. L. Strausberg; E. A. Feingold; L. H. Grouse; J. G. Derge; R. D. Klausner; F. S. Collins; L. Wagner; C. M. Shenmen; G. D. Schuler; S. F. Altschul; B. Zeeberg; K. H. Buetow; C. F. Schaefer; N. K. Bhat; R. F. Hopkins; H. Jordan; T. Moore; S. I. Max; J. Wang; F. Hsieh; L. Diatchenko; K. Marusina; A. A. Farmer; G. M. Rubin; L. Hong; M. Stapleton; M. B. Soares; M. F. Bonaldo; T. L. Casavant; T. E. Scheetz; M. J. Brownstein; T. B. Usdin; S. Toshiyuki; P. Carninci; C. Prange; S. S. Raha; N. A. Loquellano; G. J. Peters; R. D. Abramson; S. J. Mullahy; S. A. Bosak; P. J. McEwan; K. J. McKernan; J. A. Malek; P. H. Gunaratne; S. Richards; K. C. Worley; S. Hale; A. M. Garcia; L. J. Gay; S. W. Hulyk; D. K. Villalon; D. M. Muzny; E. J. Sodergren; X. Lu; R. A. Gibbs; J. Fahey; E. Helton; M. Ketteman; A. Madan; S. Rodrigues; A. Sanchez; M. Whiting; A. C. Young; Y. Shevchenko; G. G. Bouffard; R. W. Blakesley; J. W. Touchman; E. D. Green; M. C. Dickson; A. C. Rodriguez; J. Grimwood; J. Schmutz; R. M. Myers; Y. S. Butterfield; M. I. Krzywinski; U. Skalska; D. E. Smailus; A. Schnerch; J. E. Schein; S. J. Jones; M. A. Marra, *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 16899–16903.
13. R. L. Strausberg; E. A. Feingold; R. D. Klausner; F. S. Collins, *Science* **1999**, *286*, 455–457.
14. G. Temple; P. Lamesch; S. Milstein; D. E. Hill; L. Wagner; T. Moore; M. Vidal, *Hum. Mol. Genet.* **2006**, *15* (Spec. No. 1), R31–R43.
15. B. Irwin; J. D. Heck; G. W. Hatfield, *J. Biol. Chem.* **1995**, *270*, 22801–22806.
16. X. Gao; P. Yo; A. Keith; T. J. Ragan; T. K. Harris, *Nucleic Acids Res.* **2003**, *31*, e143.
17. D. M. Hoover; J. Lubkowski, *Nucleic Acids Res.* **2002**, *30*, e43.
18. S. J. Kodumal; K. G. Patel; R. Reid; H. G. Menzella; M. Welch; D. V. Santi, *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 15573–15578.
19. K. Majumder, *Gene* **1992**, *110*, 89–94.
20. J. M. Rouillard; W. Lee; G. Truan; X. Gao; X. Zhou; E. Gulari, *Nucleic Acids Res.* **2004**, *32*, W176–W180.
21. C. Withers-Martinez; E. P. Carpenter; F. Hackett; B. Ely; M. Sajid; M. Grainger; M. J. Blackman, *Protein Eng.* **1999**, *12*, 1113–1120.
22. A. S. Xiong; Q. H. Yao; R. H. Peng; X. Li; H. Q. Fan; Z. M. Cheng; Y. Li, *Nucleic Acids Res.* **2004**, *32*, e98.

23. L. Young; Q. Dong, *Nucleic Acids Res.* **2004**, *32*, e59.
24. K. Ginalski; A. Elofsson; D. Fischer; L. Rychlewski, *Bioinformatics* **2003**, *19*, 1015–1018.
25. J. J. Ward; L. J. McGuffin; K. Bryson; B. F. Buxton; D. T. Jones, *Bioinformatics* **2004**, *20*, 2138–2139.
26. L. J. McGuffin; K. Bryson; D. T. Jones, *Bioinformatics* **2000**, *16*, 404–405.
27. J. Liu; H. Hegyi; T. B. Acton; G. T. Montelione; B. Rost, *Proteins* **2004**, *56*, 188–200.
28. Z. R. Yang; R. Thomson; P. McNeil; R. M. Esnouf, *Bioinformatics* **2005**, *21*, 3369–3376.
29. L. Jaroszewski; L. Rychlewski; Z. Li; W. Li; A. Godzik, *Nucleic Acids Res.* **2005**, *33*, W284–W288.
30. L. Rychlewski; L. Jaroszewski; W. Li; A. Godzik, *Protein Sci.* **2000**, *9*, 232–241.
31. Z. Obradovic; K. Peng; S. Vucetic; P. Radivojac; C. J. Brown; A. K. Dunker, *Proteins* **2003**, *53* (Suppl. 6), 566–572.
32. Z. Obradovic; K. Peng; S. Vucetic; P. Radivojac; A. K. Dunker, *Proteins* **2005**, *61* (Suppl. 7), 176–182.
33. K. Peng; P. Radivojac; S. Vucetic; A. K. Dunker; Z. Obradovic, *BMC Bioinformatics* **2006**, *7*, 208.
34. K. Peng; S. Vucetic; P. Radivojac; C. J. Brown; A. K. Dunker; Z. Obradovic, *J. Bioinform. Comput. Biol.* **2005**, *3*, 35–60.
35. S. Vucetic; Z. Obradovic; V. Vacic; P. Radivojac; K. Peng; L. M. Iakoucheva; M. S. Cortese; J. D. Lawson; C. J. Brown; J. G. Sikes; C. D. Newton; A. K. Dunker, *Bioinformatics* **2005**, *21*, 137–140.
36. L. M. Iakoucheva; A. K. Dunker, *Structure* **2003**, *11*, 1316–1317.
37. S. Graslund; P. Nordlund; J. Weigelt; B. M. Hallberg; J. Bray; O. Gileadi; S. Knapp; U. Oppermann; C. Arrowsmith; R. Hui; J. Ming; S. dhe-Paganon; H. W. Park; A. Savchenko; A. Yee; A. Edwards; R. Vincentelli; C. Cambillau; R. Kim; S. H. Kim; Z. Rao; Y. Shi; T. C. Terwilliger; C. Y. Kim; L. W. Hung; G. S. Waldo; Y. Peleg; S. Albeck; T. Unger; O. Dym; J. Prilusky; J. L. Sussman; R. C. Stevens; S. A. Lesley; I. A. Wilson; A. Joachimiak; F. Collart; I. Dementieva; M. I. Donnelly; W. H. Eschenfeldt; Y. Kim; L. Stols; R. Wu; M. Zhou; S. K. Burley; J. S. Emtage; J. M. Sauder; D. Thompson; K. Bain; J. Luz; T. Gheyi; F. Zhang; S. Atwell; S. C. Almo; J. B. Bonanno; A. Fiser; S. Swaminathan; F. W. Studier; M. R. Chance; A. Sali; T. B. Acton; R. Xiao; L. Zhao; L. C. Ma; J. F. Hunt; L. Tong; K. Cunningham; M. Inouye; S. Anderson; H. Janjua; R. Shastry; C. K. Ho; D. Wang; H. Wang; M. Jiang; G. T. Montelione; D. I. Stuart; R. J. Owens; S. Daenke; A. Schutz; U. Heinemann; S. Yokoyama; K. Bussow; K. C. Gunsalus, *Nat. Methods* **2008**, *5*, 135–146.
38. H. E. Klock; E. J. Koesema; M. W. Knuth; S. A. Lesley, *Proteins* **2008**, *71*, 982–994.
39. W. Peti; R. Page, *Protein Expr. Purif.* **2007**, *51*, 1–10.
40. S. Graslund; J. Sagemark; H. Berglund; L. G. Dahlgren; A. Flores; M. Hammarstrom; I. Johansson; T. Kotenyova; M. Nilsson; P. Nordlund; J. Weigelt, *Protein Expr. Purif.* **2008**, *58*, 210–221.
41. C. Young-Jun; W. Tsung-Tsan; B. H. Lee, *Crit. Rev. Biotechnol.* **2002**, *22*, 225–244.
42. M. J. Weickert; D. H. Doherty; E. A. Best; P. O. Olins, *Curr. Opin. Biotechnol.* **1996**, *7*, 494–499.
43. F. W. Studier; B. A. Moffatt, *J. Mol. Biol.* **1986**, *189*, 113–130.
44. F. W. Studier, *Protein Expr. Purif.* **2005**, *41*, 207–234.
45. S. Nallamsetty; B. P. Austin; K. J. Penrose; D. S. Waugh, *Protein Sci.* **2005**, *14*, 2964–2971.
46. A. J. McCluskey; G. M. Poon; J. Gariepy, *Protein Sci.* **2007**, *16*, 2726–2732.
47. J. Geng; R. P. Carstens, *Protein Expr. Purif.* **2006**, *48*, 142–150.
48. O. Puig; F. Caspary; G. Rigaut; B. Rutz; E. Bouveret; E. Bragado-Nilsson; M. Wilm; B. Seraphin, *Methods* **2001**, *24*, 218–229.
49. S. Honey; B. L. Schneider; D. M. Schieltz; J. R. Yates; B. Futcher, *Nucleic Acids Res.* **2001**, *29*, E24.
50. C. J. Gloeckner; K. Boldt; A. Schumacher; R. Roepman; M. Ueffing, *Proteomics* **2007**, *7*, 4228–4234.
51. Y. Peleg; T. Unger, *Methods Mol. Biol.* **2008**, *426*, 197–208.
52. R. Vincentelli; S. Canaan; J. Offant; C. Cambillau; C. Bignon, *Anal. Biochem.* **2005**, *346*, 77–84.
53. S. Cabantous; G. S. Waldo, *Nat. Methods* **2006**, *3*, 845–854.
54. S. Cabantous; Y. Rogers; T. C. Terwilliger; G. S. Waldo, *PLoS ONE* **2008**, *3*, e2387.
55. T. Cornvik; S. L. Dahlroth; A. Magnusdottir; M. D. Herman; R. Knaust; M. Ekberg; P. Nordlund, *Nat. Methods* **2005**, *2*, 507–509.
56. R. K. Knaust; P. Nordlund, *Anal. Biochem.* **2001**, *297*, 79–85.
57. N. S. Berrow; D. Alderton; S. Sainsbury; J. Nettleship; R. Assenberg; N. Rahman; D. I. Stuart; R. J. Owens, *Nucleic Acids Res.* **2007**, *35*, e45.
58. J. L. Hartley, *Curr. Opin. Biotechnol.* **2006**, *17*, 359–366.
59. D. A. Jackson; R. H. Symons; P. Berg, *Proc. Natl. Acad. Sci. U.S.A.* **1972**, *69*, 2904–2909.
60. L. D. Cabrita; W. Dai; S. P. Bottomley, *BMC Biotechnol.* **2006**, *6*, 12.
61. S. N. Ho; H. D. Hunt; R. M. Horton; J. K. Pullen; L. R. Pease, *Gene* **1989**, *77*, 51–59.
62. P. G. Blommel; P. A. Martin; R. L. Wrobel; E. Steffen; B. G. Fox, *Protein Expr. Purif.* **2006**, *47*, 562–570.
63. R. O. Frederick; L. Bergeman; P. G. Blommel; L. J. Bailey; J. G. McCoy; J. Song; L. Meske; C. A. Bingman; M. Riters; N. A. Dillon; J. Kunert; J. W. Yoon; A. Lim; M. Cassidy; J. Bunge; D. J. Aceti; J. G. Primm; J. L. Markley; G. N. Phillips, Jr.; B. G. Fox, *J. Struct. Funct. Genomics* **2007**, *8*, 153–166.
64. T. Nagase; H. Yamakawa; S. Tadokoro; D. Nakajima; S. Inoue; K. Yamaguchi; Y. Itokawa; R. F. Kikuno; H. Koga; O. Ohara, *DNA Res.* **2008**, *15*, 137–149.
65. G. Marsischky; J. LaBaer, *Genome Res.* **2004**, *14*, 2020–2028.
66. A. Landy, *Annu. Rev. Biochem.* **1989**, *58*, 913–949.
67. J. Park; J. Labaer, *Curr. Protoc. Mol. Biol.* **2006**, *Chapter 3*, Unit 3. 20.
68. I. Hunt, *Protein Expr. Purif.* **2005**, *40*, 1–22.
69. P. M. Alzari; H. Berglund; N. S. Berrow; E. Blagova; D. Busso; C. Cambillau; V. Campanacci; E. Christodoulou; S. Eiler; M. J. Fogg; G. Folkers; A. Geerlof; D. Hart; A. Haouz; M. D. Herman; S. Macieira; P. Nordlund; A. Perrakis; S. Quevillon-Cheruel; F. Tarandeau; H. van Tilbeurgh; T. Unger; M. P. Luna-Vargas; M. Velarde; M. Willmanns; R. J. Owens, *Acta Crystallogr. D Biol. Crystallogr.* **2006**, *62*, 1103–1113.
70. C. Aslanidis; P. J. de Jong, *Nucleic Acids Res.* **1990**, *18*, 6069–6074.
71. C. Aslanidis; P. J. de Jong; G. Schmitz, *PCR Methods Appl.* **1994**, *4*, 172–177.
72. R. S. Haun; I. M. Serventi; J. Moss, *Biotechniques* **1992**, *13*, 515–518.
73. R. J. Ellis; A. P. Minton, *Biol. Chem.* **2006**, *387*, 485–497.
74. R. J. Ellis; A. P. Minton, *Nature* **2003**, *425*, 27–28.

75. G. H. Lorimer, *FASEB J.* **1996**, *10*, 5–9.
76. G. Georgiou; P. Valax, *Meth. Enzymol.* **1999**, *309*, 48–58.
77. P. Valax; G. Georgiou, *Biotechnol. Prog.* **1993**, *9*, 539–547.
78. M. M. Carrio; J. L. Corchero; A. Villaverde, *FEMS Microbiol. Lett.* **1998**, *169*, 9–15.
79. D. X. Zhao; Z. C. Ding; Y. Q. Liu; Z. X. Huang, *Protein Expr. Purif.* **2007**, *53*, 232–237.
80. A. Majerle; J. Kidric; R. Jerala, *J. Biomol. NMR* **2000**, *18*, 145–151.
81. F. Baneyx; M. Mujacic, *Nat. Biotechnol.* **2004**, *22*, 1399–1408.
82. Q. M. Sun; L. Cao; L. Fang; C. Chen; J. Dai; L. L. Chen; Z. C. Hua, *Protein Expr. Purif.* **2005**, *39*, 288–295.
83. J. A. Vasina; F. Baneyx, *Protein Expr. Purif.* **1997**, *9*, 211–218.
84. J. A. Vasina; F. Baneyx, *Appl. Environ. Microbiol.* **1996**, *62*, 1444–1447.
85. G. Qing; L. C. Ma; A. Khorchid; G. V. Swapna; T. K. Mal; M. M. Takayama; B. Xia; S. Phadtare; H. Ke; T. Acton; G. T. Montelione; M. Ikura; M. Inouye, *Nat. Biotechnol.* **2004**, *22*, 877–882.
86. M. Mujacic; K. W. Cooper; F. Baneyx, *Gene* **1999**, *238*, 325–332.
87. J. A. Vasina; M. S. Peterson; F. Baneyx, *Biotechnol. Prog.* **1998**, *14*, 714–721.
88. K. Terpe, *Appl. Microbiol. Biotechnol.* **2003**, *60*, 523–533.
89. D. B. Smith; K. S. Johnson, *Gene* **1988**, *67*, 31–40.
90. H. Bedouelle; P. Duplay, *Eur. J. Biochem.* **1988**, *171*, 541–549.
91. C. di Guan; P. Li; P. D. Riggs; H. Inouye, *Gene* **1988**, *67*, 21–30.
92. G. D. Davis; C. Elisee; D. M. Newham; R. G. Harrison, *Biotechnol. Bioeng.* **1999**, *65*, 382–388.
93. E. R. LaVallie; E. A. DiBlasio; S. Kovacic; K. L. Grant; P. F. Schendel; J. M. McCoy, *Biotechnology (N.Y.)* **1993**, *11*, 187–193.
94. M. P. Malakhov; M. R. Mattern; O. A. Malakhova; M. Drinker; S. D. Weeks; T. R. Butt, *J. Struct. Funct. Genomics* **2004**, *5*, 75–86.
95. X. Zuo; S. Li; J. Hall; M. R. Mattern; H. Tran; J. Shoo; R. Tan; S. R. Weiss; T. R. Butt, *J. Struct. Funct. Genomics* **2005**, *6*, 103–111.
96. X. Zuo; M. R. Mattern; R. Tan; S. Li; J. Hall; D. E. Sterner; J. Shoo; H. Tran; P. Lim; S. G. Sarafianos; L. Kazi; S. Navas-Martin; S. R. Weiss; T. R. Butt, *Protein Expr. Purif.* **2005**, *42*, 100–110.
97. T. R. Butt; S. C. Edavettal; J. P. Hall; M. R. Mattern, *Protein Expr. Purif.* **2005**, *43*, 1–9.
98. J. G. Marblestone; S. C. Edavettal; Y. Lim; P. Lim; X. Zuo; T. R. Butt, *Protein Sci.* **2006**, *15*, 182–189.
99. M. Hammarstrom; N. Hellgren; S. van Den Berg; H. Berglund; T. Hard, *Protein Sci.* **2002**, *11*, 313–321.
100. P. Braun; Y. Hu; B. Shen; A. Halleck; M. Koundinya; E. Harlow; J. LaBaer, *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 2654–2659.
101. Y. P. Shih; W. M. Kung; J. C. Chen; C. H. Yeh; A. H. Wang; T. F. Wang, *Protein Sci.* **2002**, *11*, 1714–1719.
102. M. R. Dyson; S. P. Shadbolt; K. J. Vincent; R. L. Perera; J. McCafferty, *BMC Biotechnol.* **2004**, *4*, 32.
103. V. De Marco; G. Stier; S. Blandin; A. de Marco, *Biochem. Biophys. Res. Commun.* **2004**, *322*, 766–771.
104. C. M. Guzzo; D. C. Yang, *Protein Expr. Purif.* **2007**, *54*, 166–175.
105. S. W. Englander, *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 213–238.
106. T. E. Creighton, *Curr. Biol.* **1997**, *7*, R380–R383.
107. T. E. Creighton, *Trends Biochem. Sci.* **1997**, *22*, 6–10.
108. J. D. Fox; R. B. Kapust; D. S. Waugh, *Protein Sci.* **2001**, *10*, 622–630.
109. S. Nallamsetty; D. S. Waugh, *Biochem. Biophys. Res. Commun.* **2007**, *364*, 639–644.
110. K. M. Routzahn; D. S. Waugh, *J. Struct. Funct. Genomics* **2002**, *2*, 83–92.
111. M. Hammarstrom; E. A. Woestenenk; N. Hellgren; T. Hard; H. Berglund, *J. Struct. Funct. Genomics* **2006**, *7*, 1–14.
112. A. Malik; R. Rudolph; B. Sohling, *Protein Expr. Purif.* **2006**, *47*, 662–671.
113. J. Winter; P. Neubauer; R. Glockshuber; R. Rudolph, *J. Biotechnol.* **2001**, *84*, 175–185.
114. M. W. Larsen; U. T. Bornscheuer; K. Hult, *Protein Expr. Purif.* **2008**, *62*, 90–97.
115. D. Ritz; J. Beckwith, *Annu. Rev. Microbiol.* **2001**, *55*, 21–48.
116. J. R. Allen; A. Y. Patkar; T. C. Frank; F. A. Donate; Y. C. Chiu; J. E. Shields; M. E. Gustafson, *Biotechnol. Prog.* **2007**, *23*, 1163–1170.
117. C. Huang; G. Ren; H. Zhou; C. C. Wang, *Protein Expr. Purif.* **2005**, *42*, 173–177.
118. M. Svensson; I. Svensson; S. O. Enfors, *Appl. Microbiol. Biotechnol.* **2005**, *67*, 345–350.
119. I. Huys; K. Dyason; E. Waelkens; F. Verdonck; J. van Zyl; J. du Plessis; G. J. Muller; J. van der Walt; E. Clynen; L. Schoofs; J. Tytgat, *Eur. J. Biochem.* **2002**, *269*, 1854–1865.
120. J. Robbens; A. Raeymaekers; L. Steidler; W. Fiers; E. Remaut, *Protein Expr. Purif.* **1995**, *6*, 481–486.
121. J. H. Choi; S. Y. Lee, *Appl. Microbiol. Biotechnol.* **2004**, *64*, 625–635.
122. M. Schlapschy; S. Grimm; A. Skerra, *Protein Eng. Des. Sel.* **2006**, *19*, 385–390.
123. M. Kraft; U. Knupfer; R. Wenderoth; A. Kacholdt; P. Pietschmann; B. Hock; U. Horn, *Appl. Microbiol. Biotechnol.* **2007**, *76*, 1413–1422.
124. Z. Ignatova; A. Mahsunah; M. Georgieva; V. Kasche, *Appl. Environ. Microbiol.* **2003**, *69*, 1237–1245.
125. S. Y. Li; B. Y. Chang; S. C. Lin, *J. Biotechnol.* **2006**, *122*, 412–421.
126. T. Schultz; L. Martinez; A. de Marco, *Microb. Cell Fact.* **2006**, *5*, 28.
127. L. J. Bailey; N. L. Elsen; B. S. Pierce; B. G. Fox, *Protein Expr. Purif.* **2008**, *57*, 9–16.
128. J. Yin; J. Lin; W. Li; D. I. Wang, *J. Biotechnol.* **2003**, *100*, 181–191.
129. M. R. Mayer; T. A. Dailey; C. M. Baucom; J. L. Supernak; M. C. Grady; H. E. Hawk; H. A. Dailey, *J. Struct. Funct. Genomics* **2004**, *5*, 159–165.
130. I. Rabhi-Essafi; A. Sadok; N. Khalaf; D. M. Fathallah, *Protein Eng. Des. Sel.* **2007**, *20*, 201–209.
131. P. Turner; O. Holst; E. N. Karlsson, *Protein Expr. Purif.* **2005**, *39*, 54–60.
132. A. M. Sanden; M. Bostrom; K. Markland; G. Larsson, *Biotechnol. Bioeng.* **2005**, *90*, 239–247.
133. D. Tielker; F. Rosenau; K. M. Bartels; T. Rosenbaum; K. E. Jaeger, *Biotechniques* **2006**, *41*, 327–332.
134. S. Picaud; M. E. Olsson; P. E. Brodelius, *Protein Expr. Purif.* **2007**, *51*, 71–79.
135. C. Bertoldo; M. Armbrecht; F. Becker; T. Schafer; G. Antranikian; W. Liebl, *Appl. Environ. Microbiol.* **2004**, *70*, 3407–3416.
136. A. Khlebnikov; K. A. Datsenko; T. Skaug; B. L. Wanner; J. D. Keasling, *Microbiology* **2001**, *147*, 3241–3247.
137. R. M. Morgan-Kiss; C. Wadler; J. E. Cronan, Jr., *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 7373–7377.

138. D. A. Siegele; J. C. Hu, *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 8168–8172.
139. P. Belin; J. Dassa; P. Drevet; E. Lajeunesse; A. Savatier; J. C. Boulain; A. Menez, *Protein Eng. Des. Sel.* **2004**, *17*, 491–500.
140. L. M. Guzman; D. Belin; M. J. Carson; J. Beckwith, *J. Bacteriol.* **1995**, *177*, 4121–4130.
141. P. G. Blommel; K. J. Becker; P. Duvnjak; B. G. Fox, *Biotechnol. Prog.* **2007**, *23*, 585–598.
142. F. H. Niesen; H. Berglund; M. Vedadi, *Nat. Protoc.* **2007**, *2*, 2212–2221.
143. M. Vedadi; F. H. Niesen; A. Allali-Hassani; O. Y. Fedorov; P. J. Finerty, Jr.; G. A. Wasney; R. Yeung; C. Arrowsmith; L. J. Ball; H. Berglund; R. Hui; B. D. Marsden; P. Nordlund; M. Sundstrom; J. Weigelt; A. M. Edwards, *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 15835–15840.
144. A. M. Hassell; G. An; R. K. Bledsoe; J. M. Bynum; H. L. Carter, 3rd; S. J. Deng; R. T. Gampe; T. E. Grisard; K. P. Madauss; R. T. Nolte; W. J. Rocque; L. Wang; K. L. Weaver; S. P. Williams; G. B. Wisely; R. Xu; L. M. Shewchuk, *Acta Crystallogr. D Biol. Crystallogr.* **2007**, *63*, 72–79.
145. B. Elleby; S. Svensson; X. Wu; K. Stefansson; J. Nilsson; D. Hallen; U. Oppermann; L. Abrahmsen, *Biochim. Biophys. Acta* **2004**, *1700*, 199–207.
146. V. Hozjan; K. Guo; X. Wu; U. Oppermann, *Expert Rev. Proteomics* **2008**, *5*, 137–143.
147. A. Strauss; G. Fendrich; M. A. Horisberger; J. Liebetanz; B. Meyhack; J. M. Schlaeppi; R. Schmitz, *Protein Expr. Purif.* **2007**, *56*, 167–176.
148. L. E. Pyle; P. Barton; Y. Fujiwara; A. Mitchell; N. Fidge, *J. Lipid Res.* **1995**, *36*, 2355–2361.
149. L. Dumon-Seignovert; G. Cariot; L. Vuillard, *Protein Expr. Purif.* **2004**, *37*, 203–206.
150. B. Miroux; J. E. Walker, *J. Mol. Biol.* **1996**, *260*, 289–298.
151. Y. Nakamura; T. Gojobori; T. Ikemura, *Nucleic Acids Res.* **2000**, *28*, 292.
152. E. G. Shpaer, *Protein Seq. Data Anal.* **1989**, *2*, 107–110.
153. A. Henaut; A. Danchin. Analysis and Predictions from *Escherichia coli* Sequences. In *Escherichia coli and Salmonella typhimurium Cellular and Molecular Biology*; F. Neidhardt, R. Curtiss, III, J. Ingraham, E. Lin, B. Low, B. Magasanik, W. Resnikoff, M. Riley, M. Schaechter, H. Umbarger, Eds.; American Society for Microbiology: Washington, D.C.; 1996; Vol. 2, pp 2047–2066.
154. T. Deng, *FEBS Lett.* **1997**, *409*, 269–272.
155. T. P. Holler; S. K. Foltin; Q. Z. Ye; D. J. Hupe, *Gene* **1993**, *136*, 323–328.
156. M. J. Pikaart; G. Felsenfeld, *Protein Expr. Purif.* **1996**, *8*, 469–475.
157. R. S. Hale; G. Thompson, *Protein Expr. Purif.* **1998**, *12*, 185–188.
158. I. Hubatsch; M. Ridderstrom; B. Mannervik, *Biochem. J.* **1998**, *330* (Pt. 1), 175–179.
159. K. K. Zimmerman; J. D. Scholten; C. C. Huang; C. A. Fierke; D. J. Hupe, *Protein Expr. Purif.* **1998**, *14*, 395–402.
160. L. Feng; W. W. Chan; S. L. Roderick; D. E. Cohen, *Biochemistry* **2000**, *39*, 15399–15409.
161. S. Klompus; G. Solomon; A. Gertler, *Protein Expr Purif* **2008**, *62*, 199–205.
162. D. V. Krishna Rao; C. Tulasi Ramu; J. Venkateswara Rao; M. Lakshmi Narasu; A. K. Bhujanga Rao, *Appl. Biochem. Biotechnol.* **2008**.
163. W. Fu; J. Lin; P. Cen, *Appl. Microbiol. Biotechnol.* **2007**, *75*, 777–782.
164. M. Aminian; S. Sivam; C. W. Lee; S. A. Halperin; S. F. Lee, *Protein Expr. Purif.* **2007**, *51*, 170–178.
165. I. Martinez-Martinez; J. Navarro-Fernandez; J. D. Lozada-Ramirez; F. Garcia-Carmona; A. Sanchez-Ferrer, *Biotechnol. Prog.* **2006**, *22*, 647–652.
166. A. V. Ivanov; A. N. Korovina; V. L. Tunitskaya; D. A. Kostyuk; V. O. Rechinsky; M. K. Kukhanova; S. N. Kochetkov, *Protein Expr. Purif.* **2006**, *48*, 14–23.
167. E. M. Diallo; D. L. Thompson; R. J. Koenig, *Protein Expr. Purif.* **2005**, *40*, 292–298.
168. S. Dabrowski; B. Kiaer Ahring, *Protein Expr. Purif.* **2003**, *31*, 72–78.
169. N. A. Burgess-Brown; S. Sharma; F. Sobott; C. Loenarz; U. Oppermann; O. Gileadi, *Protein Expr. Purif.* **2008**, *59*, 94–102.
170. Y. P. Chuan; L. H. Lua; A. P. Middelberg, *J. Biotechnol.* **2008**, *134*, 64–71.
171. A. H. Choi; M. Basu; M. M. McNeal; J. A. Bean; J. D. Clements; R. L. Ward, *Protein Expr. Purif.* **2004**, *38*, 205–216.
172. A. I. Derman; W. A. Prinz; D. Belin; J. Beckwith, *Science* **1993**, *262*, 1744–1747.
173. E. J. Stewart; F. Aslund; J. Beckwith, *EMBO J.* **1998**, *17*, 5543–5550.
174. E. L. Schneider; J. G. Thomas; J. A. Bassuk; E. H. Sage; F. Baneyx, *Nat. Biotechnol.* **1997**, *15*, 581–585.
175. P. E. Molloy; W. J. Harris; G. Strachan; C. Watts; C. Cunningham, *Mol. Immunol.* **1998**, *35*, 73–81.
176. W. A. Prinz; F. Aslund; A. Holmgren; J. Beckwith, *J. Biol. Chem.* **1997**, *272*, 15661–15667.
177. P. H. Bessette; F. Aslund; J. Beckwith; G. Georgiou, *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 13703–13708.
178. P. Jurado; D. Ritz; J. Beckwith; V. de Lorenzo; L. A. Fernandez, *J. Mol. Biol.* **2002**, *320*, 1–10.
179. H. Nakazawa; K. Okada; R. Kobayashi; T. Kubota; T. Onodera; N. Ochiai; N. Omata; W. Ogasawara; H. Okada; Y. Morikawa, *Appl. Microbiol. Biotechnol.* **2008**, *81*, 681–689.
180. A. Couvineau; J. C. Robert; T. Ramdani; J. J. Lacapere; C. Rouyer-Fessard; M. Laburthe, *J. Mol. Neurosci.* **2008**, *36*, 249–253.
181. C. Drees; C. A. Sturmer; H. M. Moller; G. Fritz, *Protein Expr. Purif.* **2008**, *59*, 47–54.
182. M. Kumano-Kuramochi; Q. Xie; Y. Sakakibara; S. Niimi; K. Sekizawa; S. Komba; S. Machida, *J. Biochem.* **2008**, *143*, 229–236.
183. D. Li; G. Xu; Y. Xu; T. Wu; J. Shen; H. Shu, *Biotechnol. Lett.* **2007**, *29*, 1363–1368.
184. D. Bottcher; E. Brusehaber; K. Doderer; U. T. Bornscheuer, *Appl. Microbiol. Biotechnol.* **2007**, *73*, 1282–1289.
185. S. Xiong; Y. F. Wang; X. R. Ren; B. Li; M. Y. Zhang; Y. Luo; L. Zhang; Q. L. Xie; K. Y. Su, *World J. Gastroenterol.* **2005**, *11*, 1077–1082.
186. E. A. Kersteen; J. J. Higgin; R. T. Raines, *Protein Expr. Purif.* **2004**, *38*, 279–291.
187. M. P. Castanie; H. Berges; J. Oreglia; M. F. Prere; O. Fayet, *Anal. Biochem.* **1997**, *254*, 150–152.
188. K. Nishihara; M. Kanemori; M. Kitagawa; H. Yanagi; T. Yura, *Appl. Environ. Microbiol.* **1998**, *64*, 1694–1699.
189. R. E. Joseph; A. H. Andreotti, *Protein Expr. Purif.* **2008**, *60*, 194–197.
190. A. K. Dunker; M. S. Cortese; P. Romero; L. M. Iakoucheva; V. N. Uversky, *FEBS J.* **2005**, *272*, 5129–5148.
191. Y. Cheng; T. LeGall; C. J. Oldfield; J. P. Mueller; Y. Y. Van; P. Romero; M. S. Cortese; V. N. Uversky; A. K. Dunker, *Trends Biotechnol.* **2006**, *24*, 435–442.

192. L. M. Iakoucheva; C. J. Brown; J. D. Lawson; Z. Obradovic; A. K. Dunker, *J. Mol. Biol.* **2002**, *323*, 573–584.
193. A. K. Dunker; C. J. Brown; J. D. Lawson; L. M. Iakoucheva; Z. Obradovic, *Biochemistry* **2002**, *41*, 6573–6582.
194. N. H. Tolia; L. Joshua-Tor, *Nat. Methods* **2006**, *3*, 55–64.
195. C. Romier; M. Ben Jelloul; S. Albeck; G. Buchwald; D. Busso; P. H. Celie; E. Christodoulou; V. De Marco; S. van Gerwen; P. Knipscheer; J. H. Lebbink; V. Notenboom; A. Poterszman; N. Rochel; S. X. Cohen; T. Unger; J. L. Sussman; D. Moras; T. K. Sixma; A. Perrakis, *Acta Crystallogr. D Biol. Crystallogr.* **2006**, *62*, 1232–1242.
196. J. I. Pons; S. Rodriguez; C. Madrid; A. Juarez; J. M. Nieto, *Protein Expr. Purif.* **2004**, *35*, 293–297.
197. J. Chiu; D. Tillett; P. E. March, *Protein Expr. Purif.* **2006**, *47*, 264–272.
198. C. Wu; J. Zhang; H. Abu-Soud; D. K. Ghosh; D. J. Stuehr, *Biochem. Biophys. Res. Commun.* **1996**, *222*, 439–444.
199. A. Perrakis; C. Romier, *Methods Mol. Biol.* **2008**, *426*, 247–256.
200. A. Chatterjee; Y. Li; Y. Zhang; T. L. Grove; M. Lee; C. Krebs; S. J. Booker; T. P. Begley; S. E. Ealick, *Nat. Chem. Biol.* **2008**, *4*, 758–765.
201. R. M. Cicchillo; L. Tu; J. A. Stromberg; L. M. Hoffart; C. Krebs; S. J. Booker, *J. Am. Chem. Soc.* **2005**, *127*, 7310–7311.
202. R. M. Cicchillo; K. H. Lee; C. Baleanu-Gogonea; N. M. Nesbitt; C. Krebs; S. J. Booker, *Biochemistry* **2004**, *43*, 11770–11781.
203. L. K. Mosavi; Z. Y. Peng, *Protein Eng.* **2003**, *16*, 739–745.
204. R. B. Greaves; J. Warwicker, *BMC Struct. Biol.* **2007**, *7*, 18.
205. L. D. Cabrita; D. Gilis; A. L. Robertson; Y. Dehouck; M. Rooman; S. P. Bottomley, *Protein Sci.* **2007**, *16*, 2360–2367.
206. J. M. Kwasigroch; D. Gilis; Y. Dehouck; M. Rooman, *Bioinformatics* **2002**, *18*, 1701–1702.
207. R. Cebe; M. Geiser, *Protein Expr. Purif.* **2006**, *45*, 374–380.
208. S. P. Sati; S. K. Singh; N. Kumar; A. Sharma, *Eur. J. Biochem.* **2002**, *269*, 5259–5263.
209. S. S. Orchard; H. Goodrich-Blair, *Microb. Cell Fact.* **2005**, *4*, 22.
210. K. H. Kim; J. K. Yang; G. S. Waldo; T. C. Terwilliger; S. W. Suh, *Methods Mol. Biol.* **2008**, *426*, 187–195.
211. E. I. Gonzalez de Valdivia; L. A. Isaksson, *Nucleic Acids Res.* **2004**, *32*, 5198–5205.
212. W. C. Wigley; R. D. Stidham; N. M. Smith; J. F. Hunt; P. J. Thomas, *Nat. Biotechnol.* **2001**, *19*, 131–136.
213. G. S. Waldo; B. M. Standish; J. Berendzen; T. C. Terwilliger, *Nat. Biotechnol.* **1999**, *17*, 691–695.
214. M. Hedhammar; M. Stenvall; R. Lonneborg; O. Nord; O. Sjolin; H. Brismar; M. Uhlen; J. Ottosson; S. Hober, *J. Biotechnol.* **2005**, *119*, 133–146.
215. S. Cabantous; T. C. Terwilliger; G. S. Waldo, *Nat. Biotechnol.* **2005**, *23*, 102–107.
216. A. C. Fisher; W. Kim; M. P. DeLisa, *Protein Sci.* **2006**, *15*, 449–458.
217. K. L. Maxwell; A. K. Mittermaier; J. D. Forman-Kay; A. R. Davidson, *Protein Sci.* **1999**, *8*, 1908–1911.
218. J. W. Liu; Y. Boucher; H. W. Stokes; D. L. Ollis, *Protein Expr. Purif.* **2006**, *47*, 258–263.
219. D. J. Hart; F. Tarendeau, *Acta Crystallogr. D Biol. Crystallogr.* **2006**, *62*, 19–26.
220. H. Lilie; E. Schwarz; R. Rudolph, *Curr. Opin. Biotechnol.* **1998**, *9*, 497–501.
221. N. Oganesyan; S.-H. Kim; R. Kim, *PharmaGenomics* **2004**, *4*, 22–26.
222. M. Li; Z. G. Su; J. C. Janson, *Protein Expr. Purif.* **2004**, *33*, 1–10.
223. S. M. Yin; Y. Zheng; P. Tien, *Protein Expr. Purif.* **2003**, *32*, 104–109.
224. M. B. Seefeldt; J. Ouyang; W. A. Froland; J. F. Carpenter; T. W. Randolph, *Protein Sci.* **2004**, *13*, 2639–2650.
225. R. J. St John; J. F. Carpenter; C. Balny; T. W. Randolph, *J. Biol. Chem.* **2001**, *276*, 46856–46863.
226. R. J. St John; J. F. Carpenter; T. W. Randolph, *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 13029–13033.
227. A. Jungbauer; W. Kaar, *J. Biotechnol.* **2007**, *128*, 587–596.
228. W. Swietnicki, *Curr. Opin. Biotechnol.* **2006**, *17*, 367–372.
229. K. Tsumoto; M. Umetsu; I. Kumagai; D. Ejima; J. S. Philo; T. Arakawa, *Biotechnol. Prog.* **2004**, *20*, 1301–1308.
230. A. P. Middelberg, *Trends Biotechnol.* **2002**, *20*, 437–443.
231. J. J. Lichty; J. L. Malecki; H. D. Agnew; D. J. Michelson-Horowitz; S. Tan, *Protein Expr. Purif.* **2005**, *41*, 98–105.
232. S. Nallamsetty; D. S. Waugh, *Protein Expr. Purif.* **2006**, *45*, 175–182.
233. K. D. Pryor; B. Leiting, *Protein Expr. Purif.* **1997**, *10*, 309–319.
234. R. B. Kapust; D. S. Waugh, *Protein Sci.* **1999**, *8*, 1668–1674.
235. J. V. Frangioni; B. G. Neel, *Anal. Biochem.* **1993**, *210*, 179–187.
236. J. C. Swaffield; S. A. Johnston, *Curr. Protoc. Mol. Biol.* **2001**, *Chapter 20*, Unit 20. 2.
237. D. Hasenwinkle; E. Jervis; O. Kops; C. Liu; G. Lesnicki; C. A. Haynes; D. G. Kilburn, *Biotechnol. Bioeng.* **1997**, *55*, 854–863.
238. C. D. Lee; H. C. Sun; S. M. Hu; C. F. Chiu; A. Homhuan; S. M. Liang; C. H. Leng; T. F. Wang, *Protein Sci.* **2008**, *17*, 1241–1248.
239. R. T. Baker, *Curr. Opin. Biotechnol.* **1996**, *7*, 541–546.
240. R. T. Baker; A. M. Catanzariti; Y. Karunasekara; T. A. Soboleva; R. Sharwood; S. Whitney; P. G. Board, *Meth. Enzymol.* **2005**, *398*, 540–554.
241. H. Huang; Y. Zhao; G. Yi-ru, *Chin. Med. J.* **2004**, *117*, 286–290.
242. P. Zhou; A. A. Lugovskoy; G. Wagner, *J. Biomol. NMR* **2001**, *20*, 11–14.
243. I. Levy; O. Shoseyov, *Biotechnol. Adv.* **2002**, *20*, 191–213.
244. E. Shpigel; A. Goldlust; A. Eshel; I. K. Ber; G. Efroni; Y. Singer; I. Levy; M. Dekel; O. Shoseyov, *Biotechnol. Appl. Biochem.* **2000**, *31* (Pt. 3), 197–203.
245. E. Ong; N. R. Gilkes; R. C. Miller, Jr.; A. J. Warren; D. G. Kilburn, *Enzyme Microb. Technol.* **1991**, *13*, 59–65.
246. M. Saleemuddin, *Adv. Biochem. Eng. Biotechnol.* **1999**, *64*, 203–226.
247. Y. Xu; F. C. Foong, *J. Biotechnol.* **2008**, *135*, 319–325.
248. M. C. Smith; T. C. Furman; T. D. Ingolia; C. Pidgeon, *J. Biol. Chem.* **1988**, *263*, 7211–7215.
249. P. Hengen, *Trends Biochem. Sci.* **1995**, *20*, 285–286.
250. B. P. Chen; T. Hai, *Gene* **1994**, *139*, 73–75.
251. A. Knappik; A. Pluckthun, *Biotechniques* **1994**, *17*, 754–761.
252. R. G. Chubet; B. L. Brizzard, *Biotechniques* **1996**, *20*, 136–141.
253. J. W. Slootstra; D. Kuperus; A. Pluckthun; R. H. Meloen, *Mol. Divers.* **1997**, *2*, 156–164.

254. M. Schuster; E. Wasserbauer; A. Einhauer; C. Ortner; A. Jungbauer; F. Hammerschmid; G. Werner, *J. Biomol. Screen.* **2000**, *5*, 89–97.
255. A. Einhauer; A. Jungbauer, *J. Biochem. Biophys. Methods* **2001**, *49*, 455–465.
256. N. B. Pestov; J. Rydstrom, *Nat. Protoc.* **2007**, *2*, 198–202.
257. W. Klein, *Methods Mol. Biol.* **2003**, *205*, 79–97.
258. S. Melkko; D. Neri, *Methods Mol. Biol.* **2003**, *205*, 69–77.
259. P. Vaillancourt; C. F. Zheng; D. Q. Hoang; L. Breister, *Meth. Enzymol.* **2000**, *326*, 340–362.
260. J. E. Cronan, Jr., *J. Biol. Chem.* **1990**, *265*, 10327–10333.
261. S. A. Lesley; D. J. Groskreutz, *J. Immunol. Methods* **1997**, *207*, 147–155.
262. P. A. Smith; B. C. Tripp; E. A. DiBlasio-Smith; Z. Lu; E. R. LaVallie; J. M. McCoy, *Nucleic Acids Res.* **1998**, *26*, 1414–1420.
263. A. Tirat; F. Freuler; T. Stettler; L. M. Mayr; L. Leder, *Int. J. Biol. Macromol.* **2006**, *39*, 66–76.
264. S. S. Ashraf; R. E. Benson; E. S. Payne; C. M. Halbleib; H. Gron, *Protein Expr. Purif.* **2004**, *33*, 238–245.
265. I. P. Korndorfer; A. Skerra, *Protein Sci.* **2002**, *11*, 883–893.
266. T. G. Schmidt; A. Skerra, *Nat. Protoc.* **2007**, *2*, 1528–1535.
267. L. Liu; J. Spurrier; T. R. Butt; J. E. Strickler, *Protein Expr. Purif.* **2008**, *62*, 21–28.
268. B. Coutard; M. Gagnaire; A. A. Guilhon; M. Berro; S. Canaan; C. Bignon, *Protein Expr. Purif.* **2008**, *61*, 184–190.
269. M. A. Coleman; V. H. Lao; B. W. Segelke; P. T. Beernink, *J. Proteome Res.* **2004**, *3*, 1024–1032.
270. T. Hata; M. Nakayama, *J. Biochem. Biophys. Methods* **2007**, *70*, 679–682.
271. M. Urh; D. Hartzell; J. Mendez; D. H. Klaubert; K. Wood, *Methods Mol. Biol.* **2008**, *421*, 191–209.
272. G. V. Los; K. Wood, *Methods Mol. Biol.* **2007**, *356*, 195–208.
273. L. Iversen; N. Cherouati; T. Berthing; D. Stamou; K. L. Martinez, *Langmuir* **2008**, *24*, 6375–6381.
274. Z. S. Derewenda, *Methods* **2004**, *34*, 354–363.
275. D. S. Waugh, *Trends Biotechnol.* **2005**, *23*, 316–320.
276. M. Carson; D. H. Johnson; H. McDonald; C. Brouillette; L. J. Delucas, *Acta Crystallogr. D Biol. Crystallogr.* **2007**, *63*, 295–301.
277. P. A. Walker; L. E. Leong; P. W. Ng; S. H. Tan; S. Waller; D. Murphy; A. G. Porter, *Biotechnology (N.Y.)* **1994**, *12*, 601–605.
278. M. G. Cordingley; P. L. Callahan; V. V. Sardana; V. M. Garsky; R. J. Colonno, *J. Biol. Chem.* **1990**, *265*, 9062–9065.
279. R. J. Jenny; K. G. Mann; R. L. Lundblad, *Protein Expr. Purif.* **2003**, *31*, 1–11.
280. D. Zhou; P. Yuen; D. Chu; V. Thon; S. McConnell; S. Brown; A. Tsang; M. Pena; A. Russell; J. F. Cheng; A. M. Nadzan; M. S. Barbosa; J. R. Dyck; G. D. Lopaschuk; G. Yang, *Protein Expr. Purif.* **2004**, *34*, 261–269.
281. R. L. Ward; M. A. Clark; J. Lees; N. J. Hawkins, *J. Immunol. Methods* **1996**, *189*, 73–82.
282. G. Vidricaire; J. B. Denault; R. Leduc, *Biochem. Biophys. Res. Commun.* **1993**, *195*, 1011–1018.
283. H. J. Jin; M. A. Dunn; D. Borthakur; Y. S. Kim, *Protein Expr. Purif.* **2004**, *35*, 1–10.
284. K. L. Guan; J. E. Dixon, *Anal. Biochem.* **1991**, *192*, 262–267.
285. M. G. Bolyard; S. T. Lord, *Blood* **1989**, *73*, 1202–1206.
286. G. Forsberg; B. Baastrup; H. Rondahl; E. Holmgren; G. Pohl; M. Hartmanis; M. Lake, *J. Protein Chem.* **1992**, *11*, 201–211.
287. R. B. Kapust; J. Tozser; T. D. Copeland; D. S. Waugh, *Biochem. Biophys. Res. Commun.* **2002**, *294*, 949–955.
288. R. B. Kapust; J. Tozser; J. D. Fox; D. E. Anderson; S. Cherry; T. D. Copeland; D. S. Waugh, *Protein Eng.* **2001**, *14*, 993–1000.
289. A. K. Mohanty; C. R. Simmons; M. C. Wiener, *Protein Expr. Purif.* **2003**, *27*, 109–114.
290. T. D. Parks; E. D. Howard; T. J. Wolpert; D. J. Arp; W. G. Dougherty, *Virology* **1995**, *210*, 194–201.
291. H. Y. Yoon; D. C. Hwang; K. Y. Choi; B. D. Song, *Mol. Cells* **2000**, *10*, 213–219.
292. H. R. Stennicke; G. S. Salvesen, *Methods* **1999**, *17*, 313–319.
293. P. K. Purbey; P. C. Jayakumar; P. D. Deepalakshmi; M. S. Patole; S. Galande, *Biotechniques* **2005**, *38*, 360, 362, 364 passim.
294. H. R. Stennicke; G. S. Salvesen, *J. Biol. Chem.* **1997**, *272*, 25719–25723.
295. J. A. Knott; D. C. Orr; D. S. Montgomery; C. A. Sullivan; A. Weston, *Eur. J. Biochem.* **1989**, *182*, 547–555.
296. D. Reverter; C. D. Lima, *Structure* **2004**, *12*, 1519–1531.
297. S. Chong; F. B. Mersha; D. G. Comb; M. E. Scott; D. Landry; L. M. Vence; F. B. Perler; J. Benner; R. B. Kucera; C. A. Hirvonen; J. J. Pelletier; H. Paulus; M. Q. Xu, *Gene* **1997**, *192*, 271–281.
298. M. W. Southworth; K. Amaya; T. C. Evans; M. Q. Xu; F. B. Perler, *Biotechniques* **1999**, *27*, 110–114, 116, 118–120.
299. T. C. Evans, Jr.; M. Q. Xu, *Biopolymers* **1999**, *51*, 333–342.
300. D. W. Wood; V. Derbyshire; W. Wu; M. Chartrain; M. Belfort; G. Belfort, *Biotechnol. Prog.* **2000**, *16*, 1055–1063.
301. J. Arnau; C. Lauritzen; G. E. Petersen; J. Pedersen, *Methods Mol. Biol.* **2008**, *421*, 229–243.
302. J. Arnau; C. Lauritzen; J. Pedersen, *Nat. Protoc.* **2006**, *1*, 2326–2333.
303. J. Arnau; C. Lauritzen; G. E. Petersen; J. Pedersen, *Protein Expr. Purif.* **2006**, *48*, 1–13.
304. L. J. Lucast; R. T. Batey; J. A. Doudna, *Biotechniques* **2001**, *30*, 544–546, 548, 550 passim.
305. L. Fang; K. Z. Jia; Y. L. Tang; D. Y. Ma; M. Yu; Z. C. Hua, *Protein Expr. Purif.* **2007**, *51*, 102–109.
306. J. Phan; A. Zdanov; A. G. Evdokimov; J. E. Tropea; H. K. Peters, III; R. B. Kapust; M. Li; A. Wlodawer; D. S. Waugh, *J. Biol. Chem.* **2002**, *277*, 50564–50572.

Biographical Sketch

Dr. C. Kinsland obtained a B.S. in chemistry and a B.A. in French literature from the University of Southwestern Louisiana and an M.S. and Ph.D. in bio-organic chemistry from Cornell University. Since 2001 she has directed the Protein Production and Characterization core facility at Cornell University.

9.20 Directed Evolution of Enzymes

Colin J. Jackson, CSIRO Entomology, Black Mountain, ACT, Australia

Elizabeth M. J. Gillam, University of Queensland, St. Lucia, QLD, Australia

David L. Ollis, Australian National University, Canberra, ACT, Australia

9.20.1 Introduction

By controlling the evolution of organisms, humans have been ‘directing’ evolution for thousands of years and the results are easy to see – the everyday domestic dog, for example. People have bred dogs for a number of purposes. This has been accomplished in an iterative process; dogs that displayed desirable traits were bred, their best progeny were bred, and the process repeated. Since the 1980s, with the advent of modern molecular biology, people have been able to direct the evolution of individual genes and the enzymes/proteins they encode. Much like the breeding of dogs, a common ancestor can be ‘bred’ in an iterative procedure for our needs. An ancestral dog species has given rise to dogs suited to a variety of purposes, such as hunting (bloodhound), racing (greyhound), fighting (pitbull), guarding (doberman), pest control (terriers), and even toys (chihuahuas). This is not dissimilar to the process of evolving molecules, in which a hydrolytic enzyme can be evolved to catalyze the hydrolysis of lactones, phosphotriesters, thiolactones, or esters (**Figure 1**). However, with existing technology there are limits to how much can be achieved with molecular evolution. Just as it is not possible to evolve a dog into a cat, not all molecular transformations are possible, such as

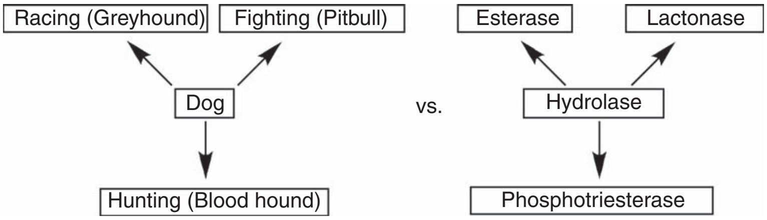


Figure 1 The comparison between dog breeding and directed molecular evolution. Over many generations, humans have directed the evolution of dogs to suit certain needs. Likewise, over several generations of directed molecular evolution, humans can direct the evolution of a hydrolase in different directions.

transforming a reductase into an efficient DNA polymerase. Thus, it is important to be aware of the potential and limitations of the techniques that are encompassed by the term 'directed molecular evolution'. Before we progress to this issue, however, we will first deal with a few preliminary matters.

The intention of this chapter is to give a broad overview of directed molecular evolution and by doing so promote interest in the field. We attempt to explain why one would want to evolve enzymes, how this might be done, and give some idea of the limitations of the available technology. We do not attempt to provide an up-to-date summary of the latest technical developments – that is better left to highly specialized review publications. Likewise, we cannot hope to cover all the branches of molecular evolution; instead we try to focus on those areas that we feel have the most relevance to chemists, natural product chemists in particular. In addition, we direct the reader to reviews and important original studies that cover material that may (or may not) become more important in the future.

The comparison of classical animal husbandry with molecular breeding is useful in many respects, but it breaks down in at least three ways. First, in animal breeding, the genetic material is never separated from the organism – this is not necessarily the case with molecular breeding. We manipulate DNA in one test tube and in the final stages we seek changes in the properties of the corresponding purified protein in another test tube. During most of this process, the genotype (DNA) and the phenotype (protein) are contained in a single 'vessel' – usually a bacterium. This may sound like a trivial point, but it is one that places significant restrictions on what can be achieved in the laboratory – a point we take up later. Second, classical genetics deals with fitness at the level of the organism – this may be conferred by changes in a single gene/protein or may be due to a multitude of changes spread over the genome. In studies with higher organisms, there is little control of how genetic diversity arises and the principal concern is with the results of selection. In simpler laboratory systems, like bacterial genetics, the experimenter has some control over genetic diversity and may use this to identify a gene that is responsible for an effect or activity. In molecular evolution, the experimenter has a considerable level of control over genetic diversity and the ability to focus it on one gene or perhaps even on one section of the gene. As a result, the mutation rate used in test tube evolution is much higher than is observed in nature. Finally, the classical animal breeder cannot make an organism, but the molecular biologist can make the DNA for a gene and express it in *Escherichia coli* (or some other suitable host) to give a protein of any desired sequence. This advantage is not without its limitations however. Although we can make genes and express proteins, we are restricted in our ability to predict how an amino acid sequence will fold. Relatively small proteins can be designed to fold in a particular manner[1,2] but the ability to design large proteins with a specified fold is some way off. Alternatively, one could take the gene for an existing protein and modify it, using site-specific mutagenesis, to produce a new protein. In this way, one could redesign a protein for some new purpose or improve its physical properties. Examples of the success of this approach can be found in the literature.[3] This approach requires a detailed knowledge of the structure and in the case of an enzyme a detailed knowledge of the mechanism – while it is an aesthetically pleasing approach, a quick search of the literature will demonstrate that evolutionary approaches are far more popular and productive. It should be noted that design and evolutionary methods are not mutually exclusive – the two approaches can be combined in a synergistic manner.[4]

9.20.1.1 Directed Molecular Evolution can be Applied in Many Ways

It may now be time to address the question 'what is directed molecular evolution?' The word 'directed' implies some purpose – in our case the purpose is of relevance to mankind and not the organism from which the enzyme was obtained. The term 'evolution' usually implies Darwinian evolution – the generation of genetic diversity followed by selection to produce a superior product. The process, at the molecular level, is summarized in **Figure 2**.

Three points are worth noting. First, the success of the evolutionary process will depend on the number of variants that can be tested. Second, in most cases, the process is iterative so that significant change usually occurs over a number of generations – it is the stepwise addition of a number of small effects. Third, most changes are negative and the effect of the selection process is, in part, to eliminate undesirables from the population. These points have considerable relevance to molecular evolution, but natural evolution and laboratory evolution differ in important ways. We have already noted that in evolution the molecules responsible for the transmission of genetic information are distinct from those responsible for function. While genetic information is found in DNA and functional activity is generally found in proteins, RNA has both. Furthermore, RNA can be replicated – making it an ideal molecule to evolve. However, there are far fewer RNA catalysts than catalytic proteins, and applications involving proteins currently dominate the field of directed evolution. A review of RNA-based evolution is available[5] as is a review describing evolution of RNA catalysts.[6]

As noted above, the two entities, gene and protein, need to be connected if we are to evolve a protein over a number of generations and express the mutant proteins at the conclusion of the experiments. The most commonly used procedure is to place the gene of interest in a plasmid and transform it into bacteria – this restricts the size of the mutant library and introduces problems in selection (or screening) of mutant proteins – typically library sizes are well below 10^9, a relatively small number compared with the number of possible mutations that could be made to small proteins. At any one position, any one of 20 different amino acids could be introduced in theory, so the numbers get very large very quickly when mutations are planned at multiple sites. Despite its shortcomings, this is the most widely used method for evolving enzymes and this review will concentrate on it. However, other methods are available for evolving proteins and a brief mention of a few will be made here.

In phage display systems, the gene of interest can be placed in a phage genome so that the protein is expressed on the exposed surface of the phage envelope. This allows the protein activity to be monitored and the genetic material to be subsequently recovered. This technology was developed in the mid-1980s and has

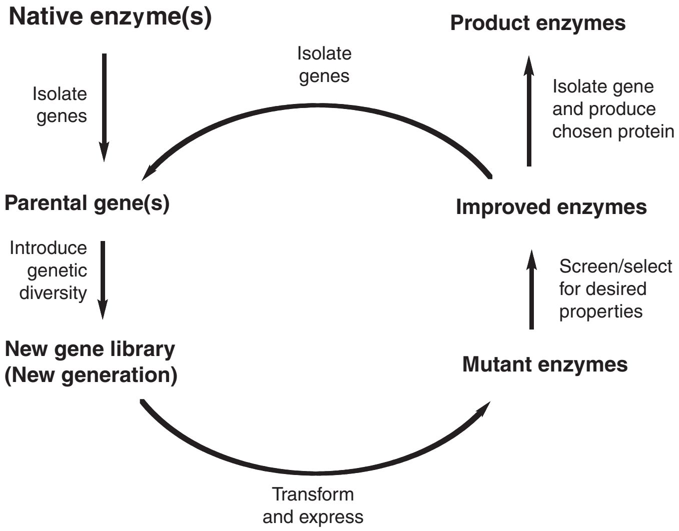

Native enzyme(s)
Isolate genes
Parental gene(s)
Introduce genetic diversity
New gene library (New generation)
Transform and express
Mutant enzymes
Screen/select for desired properties
Improved enzymes
Isolate genes
Isolate gene and produce chosen protein
Product enzymes

Figure 2 The general process by which *in vitro* evolution works.

been used mainly for evolving binding proteins. A recent review of the field covers developments in the method and recent results.[7] In other methods, the protein to be evolved can be coupled to its mRNA from which the DNA sequence can be later recovered. The linkage can be either chemical as in mRNA display or via the ribosome as in ribosome display. In both cases, there is no requirement to deal with bacteria: protein–RNA complexes are formed *in vitro* so that extremely large libraries can be generated. The applications have involved the evolution of peptides and binding proteins – large libraries present practical problems for screening enzymes. Reviews that cover these two techniques are available,[8,9] as is a paper that describes a particularly impressive piece of work.[10] These methods are potentially quite powerful, but they are not straightforward to implement. Furthermore, screening large libraries for enzyme activities presents practical problems, and so most studies of this type have focused on binding proteins.

9.20.1.2 Directed Molecular Evolution Is Evolving Rapidly

The idea of evolving enzymes has been around for some time. So, why has progress in the field been so rapid in recent years? The answer undoubtedly is due to a number of factors, particularly the automation of repetitive procedures along with the development of the polymerase chain reaction (PCR) and DNA shuffling. These techniques are relatively easy to use and are capable of generating significant genetic diversity. These techniques are described below and for comparison, the interested reader is directed to a review of prior technology that was published in 1985.[11] Progress in directed molecular evolution has also been motivated by the desire to turn enzymes into practical tools – for this reason, practical consequences usually dominate the discussions of enzyme evolution. However, it is becoming increasingly apparent that the technique is capable of answering questions about how enzymes evolve and how they function. These subjects are of more than academic interest. For example, enzymes are frequently the targets of drugs – antibiotics are frequently enzyme inhibitors that selectively target bacterial enzymes. It is well known that antibiotic resistance is a growing problem and one that may require a new approach. Directed evolution can be used to predict antibiotic resistance. It can also contribute to an understanding of how resistance arises. Enzymes that are not inhibited by antibiotics could be used as a starting point for the design of a new generation of drugs.

In terms of understanding how enzymes might function or can be altered, directed evolution does not give the same sense of satisfaction that one obtains from the successful redesign of an enzyme active site. To rationally alter the active site of an enzyme, one requires a detailed knowledge of the structure and a very thorough understanding of how the enzyme functions – neither is required for directed evolution. However, the evolutionary approach does have its compensations – one can acquire multiple and independent solutions to a problem. One could also obtain a number of enzymes that cover a range of activities – it may be possible to use this library of enzymes to come to grips with some of the more subtle aspects of enzyme mechanism. It should also be noted that directed evolution is not without its intellectual challenges. Working out a convenient way to screen a large library of mutant proteins is frequently quite difficult. In addition, interpreting the results may be difficult. Although the intent of a screen or selection may be clear, the response of nature may not be particularly clear. It is true that you get what you select for, but this can often be achieved in a number of ways. For instance, the easiest way to increase the activity of an enzyme in a cell may be to simply increase the concentration of the enzyme in the cell and to leave its catalytic properties unaltered.

9.20.1.3 Application of Directed Molecular Evolution to Enzymes

Most of the recent literature devoted to directed evolution deals with efforts to improve enzymes – the focus of the present work. Even before the advent of directed evolution, enzymes were recognized as truly remarkable catalysts and the advances made in molecular biology have merely accelerated their use in many branches of science including chemistry, medicine, and biotechnology. There are too many industries that use enzymes to name them all – however, special note should be made of the pharmaceutical and chemical industries, which make extensive use of enzymes in the synthesis of intermediates.[12] For example, tissue-type plasminogen activator (tPA) has been used in the treatment of stroke and heart attack.[13] For organic synthesis, numerous studies have been reported in which chemists have used enzymes to alter the stereospecificity of a reaction.[14] In addition to synthesizing chemicals, enzymes are also used in breaking down and thereby detoxifying chemicals –

for example, an enzyme capable of degrading organophosphate (OP) pesticides has already found commercial application.[15]

In instances where enzymes have been found to perform a particularly desirable function, there is some merit in investigating whether or not better enzymes could be found or perhaps evolved. For example, it has been proposed that a bacterial cocaine hydrolase be used to treat patients suffering from a cocaine overdose.[16] Unfortunately, a bacterial enzyme could elicit an immune response that could limit its use – evolving a human enzyme to degrade cocaine might produce a treatment even though the catalytic properties of the bacterial enzyme might (or might not) be better.[17] Clearly, other enzyme properties deserve some consideration and these will be introduced and discussed below.

9.20.1.3.1 Kinetic parameters of enzymes – more than just rate constants

Most enzymes have evolved over extremely long periods of time and one might expect that their properties would have been optimized long ago – at least for the organism that produced them. An enzyme that catalyzes a single reaction in a multistep pathway will evolve so that the output of the pathway is optimal – this may not require a particularly high catalytic turnover rate. For example, the enzyme dihydrofolate reductase (DHFR) catalyzes a reaction that produces tetrahydrofolate (THF) – an important intermediate that is used for a variety of essential processes. The enzyme is found in most organisms and the bacterial enzyme is the target of the antibiotic trimethoprim (TMP). Evolving *E. coli* DHFR to be resistant to TMP resulted in a number of variants that had increased rate constants (k_{cat}) compared with the wild-type enzyme.[18] It appears that the k_{cat} for DHFR has evolved only to the point that it caters for the requirements of the pathway. However, the K_m for the evolved enzymes was remarkably close to that observed for the wild-type enzyme. This was a little surprising since the substrate and TMP bear considerable similarity and it was expected that mutations that decreased the enzyme's affinity for TMP would also decrease its affinity for the substrate causing an increase in K_m (an apparent dissociation constant). In this case, the K_m for DHFR is important for the organism's survival and only mutations that did not affect K_m were selected. An inspection of the Michaelis–Menten curve shown in Figure 3 should convince the reader that if the substrate concentration is at or below the K_m of an enzyme, then increasing K_m will result in a dramatic reduction in the rate at which substrate is processed. The K_m for DHFR is consistent with the observation that enzymes tend to evolve such that their K_m is around the level of substrate in the cell. This observation has important consequences for enzyme engineering studies. In bioremediation studies, enzymes may be required to catalyze the degradation of toxic pollutants that occur at low levels in the environment – OP pesticides, for example. In this case, the primary concern should be to find an enzyme with a low K_m and not necessarily one that has a high k_{cat}.

9.20.1.3.2 Enzyme stability and the ability to express enzymes at high levels are important

The catalytic properties of an enzyme are not the only properties that deserve consideration when looking at the suitability of an enzyme for a particular task. If an enzyme is required in trace amounts within the cell, then it may not need to be expressed at high levels or be particularly stable or soluble. However, if large quantities

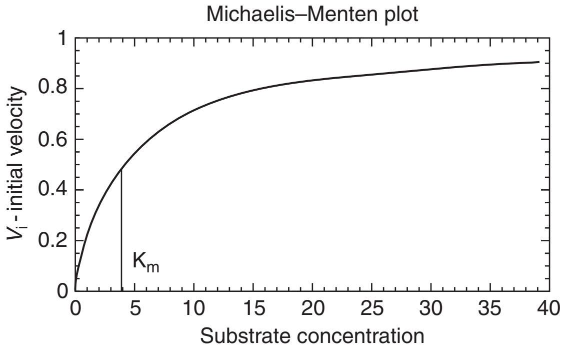


Figure 3 Michaelis–Menten plot.

are needed for industrial or scientific needs, then the expression level needs to be turned up and stability and solubility become important issues. In addition, enzymes often work only over a narrow temperature range – usually close to the temperature of the organism's environment. Enzymes that work at ambient temperatures are frequently unstable at higher temperatures and those that have evolved in thermostable organisms usually have low activities at ambient temperatures. Most enzymes have also evolved to function in an aqueous environment at physiological pH so that they may be unstable or lose activity when placed in the organic solvents with which many chemists prefer to work. However, the effects of organic solvents are not entirely negative – indeed, there are a number of applications for which organic solvents are the preferred option and for which enzymes can be adapted.[19,20] Clearly, there are a number of properties that could usefully be optimized by directed evolution. These properties should also be considered when interpreting the results of evolutionary experiments. For example, looking for increased enzymatic activity in bacteria may result in increased expression or solubility or stability of the enzyme being evolved and a change in the catalytic properties of the enzyme will not be apparent by simply looking at the level of enzyme activity in crude extracts. If the object of experiments is to improve the catalytic properties of an enzyme, then measurements should be undertaken with purified enzyme.

9.20.1.3.3 Substrate specificity – facts and fiction

Substrate specificity is another important enzyme property. Rather general statements concerning enzyme specificity are made in biochemistry texts. One gets the impression that all enzymes act on a single substrate to the exclusion of all others. The restriction endonucleases have exquisite specificity and fall into this first category. However, not all enzymes act only on a single substrate – substrate specificity is quite variable. Enzymes have usually evolved to catalyze one reaction, or a particular class of reaction, and the level of specificity will depend on the function of the particular enzyme. In some cases, enzymes will act on a broad range of substrates, just as nature intended. Examples of the latter type of enzymes are the cytochromes P-450 and the glutathione-*S*-transferases (GSTs). The P-450s and the GSTs are families of enzymes that are found within the smooth endoplasmic reticulum and cytosol, respectively, of most cells. Some forms are involved in detoxification and their targets include the products of oxidative stress, carcinogens, drugs, and other xenobiotics. The function of these enzymes requires broad substrate specificity – a trait that makes them, and particularly the P-450s, suitable for a wide range of practical applications. P-450s have many potential uses, including the breakdown of recalcitrant toxins and the biocatalytic generation of drug metabolites for the pharmaceutical industry, but they are also highly relevant to natural product research owing to their catalysis of hydroxylation and/or epoxidation of numerous classes of compounds, from alkanes to heterocycles.

In addition to their physiological function, enzymes may 'moonlight' or display 'substrate promiscuity', that is, they may have a second activity that is unconnected to their primary activity.[21] This may be either fortuitous or reflect an evolutionary origin. These secondary activities may be important for understanding how enzymes evolve. They may also be useful starting points for directed evolution studies.[22] For example, bacteria have evolved enzymes to degrade OP pesticides in a relatively short time. They appear to have done this by using a moonlighting enzyme that is normally used by the cell to break down lactones.[23] A review giving examples of other promiscuous enzymes is available.[24] The existence of promiscuous enzymes and enzymes with broad substrate specificity suggests that active sites possess a degree of flexibility and that it is this flexibility that facilitates evolution. Enzymes are malleable not immutable – they possess a degree of plasticity. The rapidity with which new or modified enzymatic activities can appear suggests that enzymes can evolve quickly even in a natural environment. For example, enzymes are frequently the targets of antibiotics. In some cases, resistant forms of these enzymes appear within months. Enzymes can also be found to degrade chemicals that have only been found in the environment for a relatively short time. Clearly, substrate specificity is an important issue for directed evolution. Although the field has great potential and is growing rapidly, it is still in its infancy and the ability to generate a new enzyme to catalyze an arbitrary reaction is some way off. The ability to evolve an enzyme with a high level of a particular activity may depend on finding an enzyme with a low level of activity or an enzyme that acts upon a related substrate. The process of finding the appropriate enzyme with which to initiate directed evolution is usually left to microbiologists, but in the future it may become a branch of bioinformatics or computational chemistry.

9.20.1.3.4 Enzyme regulation may restrict possible applications

It should be noted that even if an enzyme has a required activity, there still may be problems in exercising this activity – the enzyme may be highly regulated so that high substrate concentration or activators are required for a significant level of activity to be obtained. For example, the enzyme pyruvate decarboxylase (PDC) is used by yeast to convert pyruvate into acetaldehyde. This reaction is required only under some conditions, such as when oxygen levels are low and the concentration of pyruvate builds up. PDC activity is low at low concentrations of pyruvate, but is switched on (or activated) when pyruvate reaches a critical level. While this regulation might suit the needs of yeast, it hampers the use of the enzyme for industrial purposes. By selecting for increased activity at low substrate concentration, enzyme regulation is altered – a highly cooperative enzyme can be converted to an essentially Michaelis–Menten enzyme.[25]

9.20.1.4 Structural Biology Has Revealed a Great Deal about How Enzymes Evolve

Before we proceed with a description of tools and results of laboratory evolution, it might be worth considering what we know of how enzymes evolve in nature. This question has a long history – Horowitz[26] put forward a theory in 1945 that had at its heart the idea that it was the ability to bind a substrate that drove evolution. In other words, enzymes with the ability to bind a substrate can evolve the necessary machinery for catalysis. This idea was put forward long before the structures of enzymes were known and in an attempt to explain the formation of metabolic pathways. It is a theory that is not without success and examples of binding-driven evolution can be found in a recent review.[27] However, it would appear that 'chemistry-driven' evolution provides a better explanation for how enzymes usually evolve.[28] According to this theory, it is the placement of catalytic residues in the active site of an enzyme that is crucial for enzyme evolution – this placement ensures that the reactions catalyzed by an enzyme and its immediate ancestor are similar although the substrates may differ. Consider, for example, the α/β-hydrolase fold enzymes.[29] These enzymes differ significantly in size, but all share a common structural core in which the secondary structure elements are joined in the same manner – their cores are said to have conserved topology. Most of these enzymes catalyze hydrolytic reactions with mechanisms that are very similar (although not identical in all cases) to that of the serine proteases. That is, they all have a triad of hydrogen-bonded residues that consist of a nucleophile, a histidine, and an acid residue – a catalytic triad. The location of these catalytic residues is conserved in the fold members – the nucleophile is found on a conserved sharp bend (an elbow) at the end of strand 5. Furthermore, there are elements of the structure that seem well suited for catalysis – the nucleophilic elbow allows close approach of the substrate to the catalytic residues and forms part of the oxyanion hole that stabilizes the transition state formed during catalysis. However, the substrate-binding regions of the hydrolase fold that are responsible for binding substrate vary greatly among the members of the fold. In some cases, a short piece of peptide is sufficient to bind the substrate, whereas in other proteins, a substantial domain is devoted to this purpose. In very simple terms, nature appears to have taken a catalytic core and grafted on different substrate-binding domains to produce a variety of different hydrolytic enzymes. This is not something that would be easy to achieve in an *in vitro* experiment, but it does suggest a basis for enzyme promiscuity and what is likely to happen when enzymes evolve. The reactions catalyzed by a promiscuous enzyme are likely to share a common mechanism and changes observed in evolution experiments are likely to alter the substrate-binding residues rather than the catalytic residues.

We now give a description of some of the techniques that can be used to generate genetic diversity, that is, 'library generation', followed by a review of techniques for selection or screening of such libraries. We conclude with a review of the recent literature that illustrates the technology and the results that can be achieved with it.

9.20.2 Library Construction Techniques and Consequences

The idea and practice of evolving proteins has a long history. The overall process of inducing mutations and then selection for desirable traits has always been present, but the tools to do this work have improved immensely in recent years. Automation has made selection processes much easier, but it is really the advances made in library generation, using PCR, that have revolutionized the field. Prior to PCR, mutations could be made using chemical methods and mutator strains of *E. coli*, but these processes were difficult to control and their effects were sometimes difficult to reproduce. PCR allowed the mutations to be concentrated on a specific gene or region of the gene.

9.20.2.1 Random Techniques

9.20.2.1.1 Error-prone PCR

PCR was originally designed to amplify lengths of DNA, but has been adapted to carry out most of the steps involved in library generation for directed molecular evolution.[30] There are a number of reviews available that explain the general principles of PCR in some detail.[31,32] For our purposes, it is important to realize that PCR is based on two facts. The first is that by increasing the temperature, the hydrogen-bonded strands of DNA can be separated. Second, DNA is replicated using a DNA polymerase. Initially, the Klenow fragment was used,[33] but it was soon realized that a polymerase from a thermophile was more efficient since it could withstand the repeated cycles of high temperature required for denaturation of the DNA after each round of replication.[34] Short oligonucleotides (primers) that are present in excess can bind (hybridize) selectively to one of the two strands of the target DNA in regions to which they are complementary. Polymerases bind to the primer termini and extend them according to the sequence specified by the template strand, as shown in **Figure 4**. The primers mark the extent of DNA to be replicated. They can also be used to add restriction sites to the ends of the amplified gene. These restriction sites are useful in allowing the amplified gene to be inserted into a plasmid. PCR is an iterative technique that starts by separation of the strands of DNA by increasing the temperature of the DNA solution. The solution is then cooled to a temperature that allows the primers to anneal selectively to the complementary regions of DNA. The temperature is then increased to allow the polymerase to replicate the DNA template – as noted above, the polymerases used in PCR are from thermophilic organisms so that they are stable at elevated temperatures and have optimal activity at temperatures higher than ambient. The PCR cycle is repeated numerous times and the yield of amplified DNA will increase by a factor of 2 raised to the power of

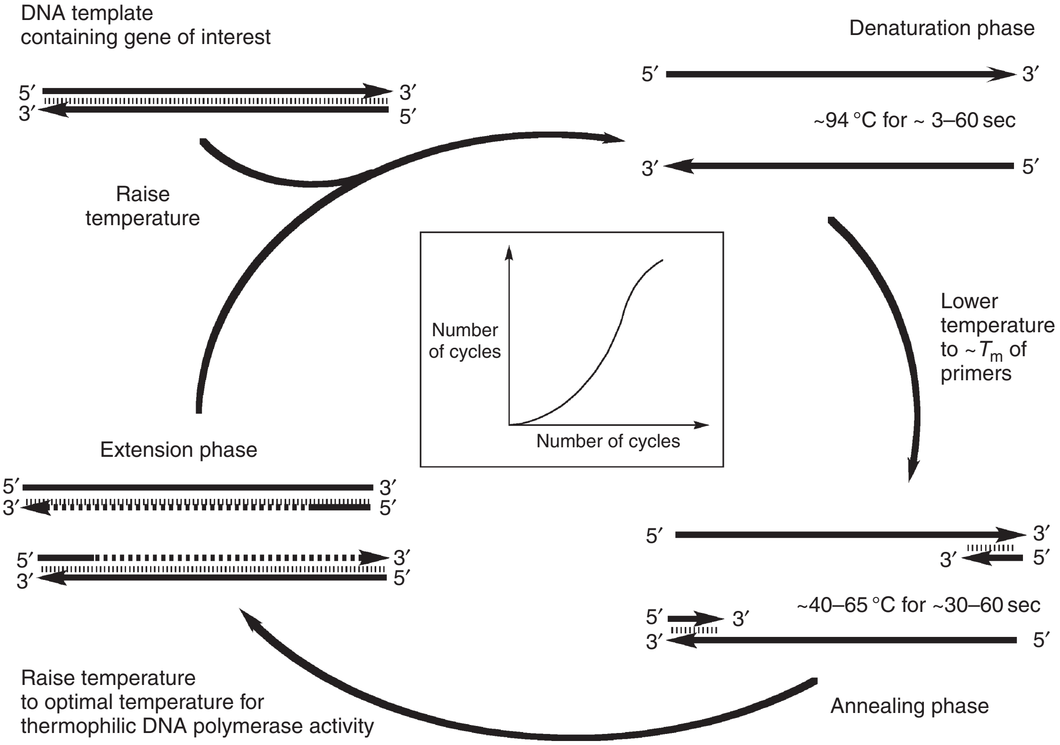


Figure 4 A schematic diagram illustrating the stages in a typical PCR cycle.

N, where *N* is the number of cycles. The choice of temperatures and timing is critical as is the choice of primers and their concentration. For example, if sufficient time is not allowed for replication, then partially completed DNA fragments will be present and could act as primers in subsequent cycles – this phenomenon is used to advantage in the staggered extension PCR (StEP) technique to be described later.

In addition to their polymerase activity, most DNA polymerases have a 3′ to 5′ exonuclease activity that has an editing function. This activity ensures that DNA is replicated with high fidelity and is usually found on a separate domain. In error-prone PCR (epPCR), a thermostable polymerase that lacks the domain responsible for editing is used. Enzymes like *Taq* polymerase (from *Thermus aquaticus*) incorporate bases incorrectly at a frequency of about 0.1×10^{-4} to about 2×10^{-4} depending on the reaction conditions.[35] Over the course of a PCR experiment, the polymerase could make over 30 passes of the DNA so that the cumulative error rate is about 1×10^{-3} per nucleotide. This is not a sufficiently high error rate to generate diversity, especially in stretches of DNA shorter than 1 kb. There are a number of protocols that have been developed to increase the error rate of *Taq* polymerase. These include increasing the concentration of $MgCl_2$,[36] the addition of Mn^{2+},[37] and the addition of triphosphate nucleoside analogues.[38] It should be clear that PCR could be used to construct a library with a specific average number of mutations per gene by tweaking the reaction conditions. For any given experiment, one should, of course, sequence a randomly selected number of genes to ensure that this is indeed the case.

There are a number of issues that should be considered when constructing a mutant library. First, what is to be achieved? Experience gained in a number of laboratories suggests that making useful changes to the properties of a protein can be achieved with a relatively small number of mutations – numbers of between 1 and 10 are commonly reported. This is usually achieved over a number of generations with modest mutation rates. Second, how is the library to be screened – or rather what is the size of the library that can be conveniently screened? The answer to this question will determine the level of genetic diversity that is to be put into the library. Unfortunately, if mutations are introduced randomly, then the frequency of beneficial mutations is generally much lower than that of detrimental mutations. In most cases, a mix of beneficial and detrimental mutations will occur and the latter will mask the effects of the former. The probability of finding an accumulation of beneficial mutations will be low and it is unlikely that they will be identified if only a small sample of all possible mutated genes is screened. There is little point in generating a huge library with a great deal of diversity if only a small fraction of it can be screened. The probability of finding a mutated gene with an accumulation of positive mutations is close to that of finding the proverbial needle in the haystack.

Not all nucleotide changes will manifest themselves as changes in the protein they encode. If the mutation rate is to be kept low so that there are only a modest number of mutations per gene, then the probability that these mutations will occur at adjacent nucleotides is small. This means that there is likely to be only one change per triplet encoding any amino acid. As a consequence, not all nucleotide changes will result in a change at the level of amino acids. For example, a change in a particular nucleotide may give rise to the same amino acid – a silent change – or it may cause the insertion of a premature stop codon in the gene. Furthermore, different mutations may give rise to the same amino acid change. On average, a single nucleotide change could result in a change of an amino acid to one of 5.7 other types – not all of the 19 other amino acids are accessible.[39] The observed changes in amino acid types tend to be conservative; that is, amino acids tend to be replaced by others that are chemically similar. This may not be harmful in terms of evolving proteins – the amino acid code may have evolved to minimize the effects of mutations. The protein engineer cannot expect to see all possible amino acid types inserted at a particular location of a protein and the observed mutations may not be optimal. For this reason, random mutagenesis is often followed by saturation mutagenesis. This is a form of site-specific mutagenesis in which a particular amino acid is replaced with all other amino acid types. Alternatively, one could mutate the amino acids in a short peptide centered on a particular position.

At this point, it may be useful to get a rough idea of the relationship between mutation rate and the size of the library needed to contain all possible mutations. Consider a small protein that contains 100 amino acids. If we assume that there are 20 possible amino acids at each position, then there are 10^{130} possible peptides – a number too large to even contemplate making or screening. Suppose we take one of these proteins – how many sequences are there that differ by one amino acid? If we assume that the mutation rate is low and that there are only (on average) 5.7 amino acids that are accessible at any given position, then we need to make a library that codes for only 570 proteins – not a large number. If random mutagenesis was used, then to generate such a library would require a larger number of mutant genes – at most the number would be around 900 so that each nucleotide is altered to the other possibilities. If we want all peptides that have changes at two locations, then we want a library that codes for 1.6×10^5 proteins – this would require a gene library of around 4×10^5. For changes at three locations, the library would have to accommodate 3.0×10^7 proteins and would require a gene library of around 1.2×10^8 genes. Although these numbers are within the realm of experimental possibility, the trend is disconcerting and this is for a relatively small protein. In these simple calculations, we have assumed an average mutation rate. In fact there will be a distribution in the number of mutations per gene that can be described by the Poisson distribution.[40] If the average mutation rate is 3 changes per gene, then the probability that a particular gene has either 2 or 3 mutations is 0.22 so that to generate a library containing all genes with 3 mutations will require around 5 times the number calculated above, that is around 10^9. Although these numbers are very approximate, they give some idea of the library sizes required to evolve proteins. If the change in function we hope to generate requires (say) 10 mutations, then multiple rounds of evolution are required. It should be noted that these calculations assume that the genes remain intact – that the mutated genes produced in one cycle form the starting point for generating the library to be screened in the next cycle. This process is analogous to asexual replication: there is no genetic recombination, probably the most important factor used to generate genetic diversity in nature.

9.20.2.1.2 DNA shuffling

In 1994, Stemmer[41,42] revolutionized directed molecular evolution by introducing DNA shuffling – iterative *in vitro* recombination of a set of parental genes. The process is shown schematically in Figure 5. It requires a pool of homologous genes that could be produced by epPCR. Alternatively, they could be genes in which a variety of site-specific mutations have been made or genes encoding a family of homologous proteins. The latter case is known as family shuffling and is capable of introducing considerable genetic diversity into the gene pool.[43] In shuffling, the DNA is first broken into relatively small fragments using a nuclease, DNase 1 for example. These fragments are then isolated and reassembled into a full-length gene using repeated cycles of 'primerless PCR' – in fact, the fragments prime each other. In practice, shuffling often introduces new mutations into gene – a useful feature in evolutionary experiments. Typically, considerable quantities of template DNA are required for shuffling to work. In addition, reassembly will work only if there is considerable homology between the

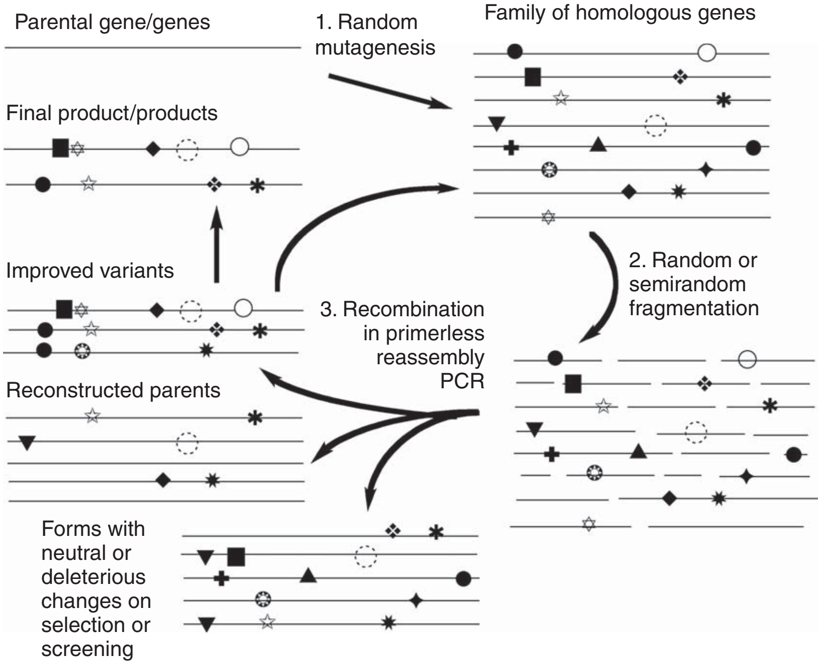

Parental gene/genes
1. Random mutagenesis
Family of homologous genes
Final product/products
Improved variants
2. Random or semirandom fragmentation
3. Recombination in primerless reassembly PCR
Reconstructed parents
Forms with neutral or deleterious changes on selection or screening

Figure 5 A schematic diagram illustrating the steps involved in DNA shuffling.

fragments – family shuffling, using classical shuffling methods, will usually work only if the initial genes are at least 70% homologous.

DNA shuffling reassembles genes from fragments – it can take blocks from different sources and join them together to form new genes. Ideally, the small fragments isolate the differences in the parental genes so that they can be fused together in a novel fashion – as in recombination. The importance of the shuffling process is that it allows favorable mutations to come together in a single gene. In the parental genes, these favorable mutations might have occurred with neutral or unfavorable mutations that masked their desirable effects. Alternatively, the effects of a single mutation may have been too small to detect, but may be readily detected when combined with other favorable mutations. Mutations that do not confer an advantage on the gene are quickly lost, provided they can be isolated on a separate DNA fragment. Neutral mutations that occur near to a favorable mutation may be carried through the evolutionary process. It should be noted that in the process of combining fragments to form a new gene from a number of parental genes that each have a few different mutations, then the most likely outcome would be the native gene – one of the possibilities shown in Figure 5. Some thought should be given to the choice of which genes should be used in shuffling experiments. The application of heavy selection pressures will eliminate all but the most fit of progeny – some would argue that these would serve as the best parents for subsequent generations. However, there is some evidence that this may be a counter-productive approach in the initial rounds of evolution. Tawfik and coworkers use the term 'neutral drift' to describe the process in which high mutational rates are used along with the application of mild selection conditions to give a large pool of mutants that serve as the starting point for the following round of shuffling. More stringent selection conditions are imposed on later generation of recombinants.[44]

Since its introduction, there have been a number of advances made in the way shuffling is accomplished. Computational methods can be used to model[45] and improve[46] shuffling. More importantly, the process can now be carried out in a single PCR step in the method known as StEP.[47] In this process, short annealing/extension times are used so that there is insufficient time for the complete gene to be replicated. The growing PCR fragments reanneal at random with templates and the extension continues. Full-length genes may take a number of cycles to complete, but in the process template switching effectively shuffles genes, as shown in Figure 6. StEP is much easier to perform than classical shuffling and it avoids the need for DNase 1 digestion so that less DNA template is required.

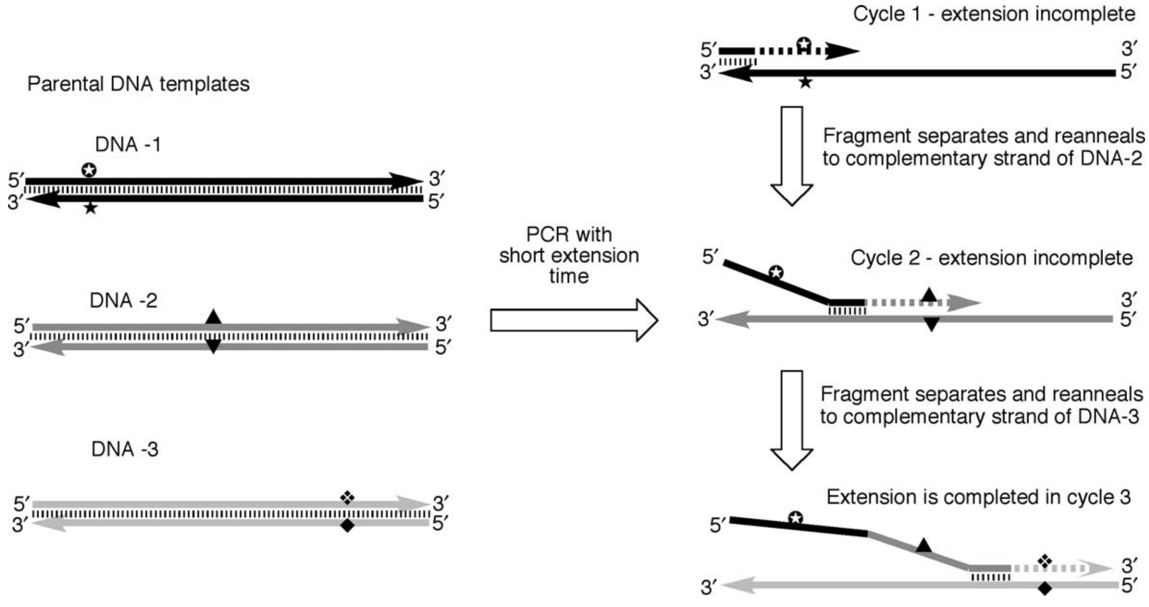

Parental DNA templates
DNA -1
5′
3′
3′
5′
DNA -2
5′
3′
3′
5′
DNA -3
5′
3′
3′
5′
PCR with short extension time
Cycle 1 - extension incomplete
5′
3′
3′
5′
Fragment separates and reanneals to complementary strand of DNA-2
Cycle 2 - extension incomplete
5′
3′
3′
5′
Fragment separates and reanneals to complementary strand of DNA-3
Extension is completed in cycle 3
5′
3′
3′
5′

Figure 6 A schematic diagram showing how recombination is achieved using staggered extension PCR (StEP).

9.20.2.1.3 There are many other ways to shuffle genes

One of the shortcomings of classical shuffling is that recombination tends to occur when fragments have considerable sequence identity and are sufficiently long to anneal.[48] This means that mutations may not be isolated into separate blocks so that detrimental mutations cannot be separated from those that confer an advantage. There are a number of methods that have been devised to increase the frequency of recombination. Random chimeragenesis on transient templates (RACHITT) reassembles trimmed parental genes on a single-stranded DNA (ssDNA) template and the genes are reconstructed with very high rates of recombination.[49] However, a good quality ssDNA template is frequently difficult to synthesize and so this method may not be widely useful. Another method, synthetic shuffling, uses partially degenerate overlapping oligonucleotides of about 60 bases in length to reconstruct genes that exhibit differences from the parental genes. For example, the analysis of 15 subtilisin genes was used to design a library of oligonucleotides that could be shuffled to produce a library that was screened to give a number of useful proteins that were encoded by recombinant genes.[50] The assembly of designed oligonucleotides (ADO) method was similar in terms of concept, but was implemented quite differently. Again, the method requires careful analysis of parental genes and design of oligonucleotides, but the results are impressive.[51] There are also techniques that allow recombination of parental genes that do not have high levels of sequence similarity. Incremental truncation for the creation of hybrid enzymes (ITCHY) is one such technique.[52] In essence, it produces a library that consists of genes formed by the fusion of fragments that code for the N- and C-termini of two parental genes. This is achieved by first cloning the two genes in tandem in an expression vector. This construct has three restriction sites that are located at the ends of the tandem repeat and in the middle. The restriction sites allow the linearized repeat to be isolated in good yield so that exonuclease III can be used to digest the two ends. The resulting fragments are then blunt ended and the two partially digested genes are then separated by digestion at the remaining unique site. The resulting fragments are blunt ended and ligated together to produce genes that are similar (though not necessarily identical) in length to the parents. Truncated and extended genes may also be produced – these are likely to be nonfunctional. Variation in the length of the recombined genes does generate products that cannot be obtained using classical shuffling techniques. The resulting library can be subjected to additional rounds of shuffling, the combined approach being known as SCRATCHY.[53] The ITCHY process is not without its shortcomings; it is lengthy and the nuclease digestion step is difficult to control. An alternative, procedure was developed with the idea of using nucleotide triphosphate analogues (such as α-phosphothionate) to protect the DNA from excessive nuclease digestion.[54]

Still other methods have been developed to allow distantly related sequences to be recombined. These processes go by a variety of names and involve some knowledge of structure that allows fragment boundaries to be chosen. Although the process by which fragments are recombined is random, the choice of boundaries is not, so that the processes could be described as semirational. The sequence-independent site-directed chimeragenesis (SISDC) method is a general method that allows targeted fragments to be recombined in a desired order[55] – a computer program (SCHEMA) has also been written to aid in the choice of these sites.[56] In terms of concept, this is similar to the structure-based combinatorial protein engineering (SCOPE) method that was used to shuffle the genes for two distantly related members of the X-family of rat DNA polymerase β.[57] In this study, the proteins had similar folds and subdomains were identified using sequence homology and the known three-dimensional structure of rat DNA polymerase β. This allowed new and improved polymerases to be created from the fragments of parent proteins.

There are also a variety of new methods of directed molecular evolution that relate more closely to biology. Exon shuffling is one such example.[58] The idea here is to isolate protein domains on gene fragments – exons – and to shuffle these using the mechanics associated with intron processing. Other workers are devising methods that utilize *in vitro* recombination as a means of shuffling genes. These processes have been carried out in *E. coli*[59] and yeast – the method is known as combinatorial libraries enhanced by recombination in yeast (CLERY).[60]

9.20.2.2 Directed Techniques That Focus on a Particular Section of the Gene

It has been noted above that the use of epPCR does not allow all possible mutations to occur. One may observe a mutation produced by epPCR, but is it the best possible mutation? Random mutagenesis may identify a site, where improvements can be made, but it may not provide the optimal replacement, since, as we have noted, error rates have to be kept at a modest level and not all possible mutations are accessible. Furthermore, there is strong evidence that, for some kind of improvement, a targeted approach may be superior to introducing mutations at random.[61] This situation could arise when the structure of an enzyme is known and the researcher has a good idea of the residues that confer substrate specificity. In this case, randomization of particular active site residues may be a far better approach than screening mutants that occur at random over the entire length of the protein. Clearly, this approach is at variance with the more general view of how evolution could or should occur – it assumes that human knowledge may be employed to accelerate the rate of evolution. Whether this type of intervention helps or hinders evolution remains to be seen – the technology to test it is available. The method to replace a single or a small number of amino acids with all possible other amino acids is referred to as site-saturation mutagenesis. These experiments can be done by cassette methods, as was carried out by Lim and Sauer[62] in their work with λ repressor. Alternatively, oligonucleotide-directed techniques can be applied – such a technique was successfully used to alter active site residues of a cyclodextrin glucanotransferase in order to change it into an α-amylase.[63] Much work in this area has been performed by Reetz and colleagues, who have developed the iterative saturation mutagenesis approach[64] and extended it to develop the combinatorial active site saturation mutagenesis test (CAST) approach, in which the residues of the active site are randomized and recombined.[65]

9.20.3 Selection and Screening Techniques

Identifying mutants with enhanced activity among the members of a large library – ‘interrogating a mutant library’ or ‘screening or selecting’ – is usually the most difficult step in directed evolution. DNA can be manipulated with relative ease and large libraries of mutant genes can be generated using procedures that are not dependent on the DNA sequence. However, the properties of proteins, both physical and catalytic, can differ dramatically from one another and there is no universal protocol for the identification of mutant proteins with enhanced activity or physical properties.

There are two approaches to examining a mutant library – selection and screening. The term ‘selection’ has traditionally implied methods that are based on enzymatic functions that confer growth or survival advantages to the host organism. By definition these are *in vivo* techniques – the organism survives or grows rapidly only if it possesses an enzyme with appropriate activity. Antibiotic selection is a good example. One ligates a mutant library into a plasmid that is transformed into a cell line and then selects for mutants that survive at levels of antibiotic that kill the cells harboring wild-type enzyme. When testing the effectiveness of DNA shuffling, Stemmer selected for mutant variants of β-lactamase that would enable the host to survive in the presence of elevated levels of the antibiotic cofotaxime.[42] Antibiotic selection is not restricted to evolving proteins that confer antibiotic resistance as it can be coupled to other properties, such as protein solubility,[66] as described in the next section. Apart from antibiotic resistance, selection can be based on the ability of an enzyme to produce an essential nutrient for the host organism. For example, *E. coli* can be engineered so that it can utilize the phosphorus in simple OP pesticides such as paraoxon as the sole form of phosphorus in its growth media. This allows variants of an enzyme capable of degrading OP pesticides to be selected on the basis of their ability to support the growth of the host.[67] There are still other ways that selection methods can be established. If essential genes can be inactivated in a host, then selection procedures can be created. For example, a strain in which the gene for glyceraldehyde-3-phosphate dehydrogenase (GAPDH encoded by gapA) has been inactivated has been used to select for variants of Rubisco (ribulose-1,5-P_2-carboxylase/oxygenase) that enable the normally essential GAPDH to be bypassed.[68] Selection techniques are generally applied to mutants grown on solid media (agar plates) as shown schematically in **Figure 7**. These techniques are typically used with libraries of less than 10^5 variants but larger library sizes can be dealt with. They are relatively easy to use, but they may not give a good quantitative response and a secondary screen is often required to get a better sense of the idea of the level of activity of mutant enzymes. Another problem is that nature may find an alternative means by which to overcome the deficiency that does not necessarily involve the protein of interest.

Screening involves direct measurement of mutant proteins by a high-throughput assay – commonly carried out using agar plates. The enzyme assay is carried out inside the cell – the reaction is followed by visual

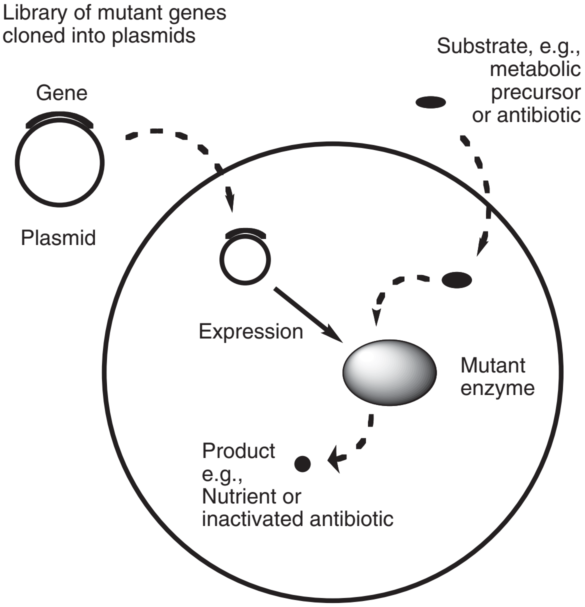

Library of mutant genes cloned into plasmids
Gene
Plasmid
Substrate, e.g., metabolic precursor or antibiotic
Expression
Mutant enzyme
Product e.g., Nutrient or inactivated antibiotic

Figure 7 The selection process. The large circle represents a cell that survives if a mutated gene produces an enzyme capable of producing either a nutrient or an antibiotic.

inspection of the plate. This may be done by a variety of ways. The most common method is probably the formation of a colored or fluorescent product or the disappearance of a colored or fluorescent substrate. For example, an OP pesticide that produces fluorescent products has been used to screen for enhanced activity of OP-degrading enzymes.[69] Color change can also be generated by a shift of pH and the presence of a suitable indicator. This approach has been used to screen for glycosynthase activity.[70] Many substrates form an insoluble precipitate when incorporated into agar so that enzymes capable of converting these materials to a soluble product will generate a 'halo' around the host cells. This technique was used to screen for enhanced activity in mutant OP-degrading enzymes with enhanced activity toward the pesticide chlorpyrifos.[71] These screening assays are usually easy to use but often do not give a quantitative answer – a secondary screen may be needed to discriminate if a large number of positives are obtained. Typically, libraries of less than 10^5 variants are monitored with these techniques.

Microtiter plates are frequently used to screen for enzyme activity. These plates usually contain 96 wells and are analyzed using spectrophotometers or fluorometers. In most applications, a single transformant is grown in each well, the plate is replicated, and the cells of one copied plate are lysed and used to measure the relevant activity. Useful clones can then be obtained from the original plate. Screening is usually done with a UV/visible spectrophotometer or a fluorometer that is built into the plate scanner. Many of the assays described in the previous paragraph can be used with these instruments. However, plate readers are much more versatile instruments with which to monitor enzyme activities compared to doing enzyme assays on agar plates. For example, reactions involving redox cofactors, such as NAD^+/NADH, can be monitored with these instruments. In addition, reactions can also be followed by coupling them to redox reactions. For example, by coupling PDC activity to that of alcohol dehydrogenase, 10^4 mutants could be screened relatively quickly.[72] Product formation can also be monitored using antibodies as in an enzyme-linked immunosorbent assay (ELISA).[73] Microtiter plates are frequently used to carry out a primary screen of a library of modest size (10^4), but this number can be increased if more than one clone is grown per well. Adding more than one clone per well has a number of advantages – libraries can be grown in liquid culture, the cell density determined, and then simply diluted into the microtiter plates for assaying. The technique will work if the activity of a mutant protein stands out against the background signal that would be expected to be higher than in the case of a single clone per well. This approach has been examined from a theoretical perspective and tested.[74] It has also been applied to the problem of evolving PDC.

Apart from the screening and selection techniques described above, there are a number of emerging techniques that may find general application in the future. Only a brief mention will be made of some of these – a more detailed account can be found in the literature.[75] Many of the problems that are encountered in the application of traditional screening and selection methods are due to the inability of substrates to diffuse into the cell. Lysing cells in a microtiter plate is one way of overcoming this problem, but accessible library size is compromised in the process. Another way around this problem is to attach the protein to be evolved to the external membrane of the cell. The mechanics of getting the molecule to the cell surface are understood for most proteins, but they do not always deliver large quantities of the protein and in some cases they do not work at all. However, cell surface display techniques may be the technique of choice when dealing with problematic substrates. The technology has been successfully applied to the generation of specific endopeptidase variants that cut at nonnative sites.[76] In these experiments, the reaction products adhered to the cell surface and gave rise to a fluorescence signal that could be detected with a fluorescence-activated cell sorting (FACS) system, an instrument capable of sorting $>10^4$ events per second. Indeed many of the new techniques rely on the production of fluorescent products that can be detected using FACS. These instruments can rapidly screen very large libraries and have great potential in facilitating enzyme evolution studies. One particularly elegant application of FACS involves allowing the substrate to diffuse into a cell and then trapping the fluorescent product inside the cell – the cell is essentially used as a 'microreactor'. In these experiments, a large library of mutants (10^7) was screened for the ability to sialylate a fluorescently labeled acceptor sugar.[77]

In these two applications of FACS, the cell provided a physical linkage between the enzyme, its genetic material, and the product of catalysis. There are other ways to maintain this linkage. One solution to this problem is to encapsulate cells and substrate in an oil droplet (water in oil) that can be sorted using a FACS machine. This approach has been used with some success in evolving the PON1 protein.[78] More ambitious methods may evolve from the idea of using oil droplets as a vehicle for holding the various components of evolutionary experiments together. *In vitro* compartmentalization (IVC) does away with the cell and encapsulates the genetic material, the necessary translational factors, and the substrate into a single oil drop that can be sorted by a FACS machine.[79] FACS is not the only means by which cells can be sorted. Microfluidic devices are also being built to rapidly sort cells contained within oil droplets.[80] This technology offers the ability to process enormous libraries very quickly.

9.20.4 Applications of Directed Molecular Evolution

Having provided an introduction to the techniques that are commonly used today in directed molecular evolution experiments, it is now appropriate to introduce a number of examples to illustrate its enormous utility. This is not intended to be a comprehensive review of this field, within which many hundreds of reports have been published over the past decade. Rather, it is hoped that this will serve to highlight some instances where the technique has been applied to achieve specific improvements, with a particular focus on the use of directed molecular evolution to improve enzymes involved in industrial applications and natural product synthesis.

9.20.4.1 Improving an Existing Activity

The most obvious use of directed molecular evolution involves its application to improve an existing activity. Often enzymes are isolated and cloned because they catalyze a reaction that is important in industry. Any improvement thus dramatically decreases the cost of the process. For example, if an enzyme's activity is increased 10-fold, 90% less enzyme will be required.

Perhaps the clearest example of this is the seminal publication in this field, the directed evolution of a β-lactamase for increased resistance to the antibiotic cefotaxime by Stemmer.[41] The TEM-1 β-lactamase gene was heterologously expressed in *E. coli*, providing resistance up to $0.02\ \mu g\ ml^{-1}$. Over five generations of DNA shuffling,[42] during which the selection pressure was stepped up by increasing the concentration of cefotaxime, an improved variant was obtained that conferred resistance at concentrations up to $640\ \mu g\ ml^{-1}$, a 32 000-fold improvement. Sequence analysis showed that this remarkable improvement was a result of only six amino acid mutations of the wild-type enzyme.

A second example of the use of directed evolution to improve a native enzymatic activity is the bacterial phosphotriesterases (PTEs) from *Pseudomonas diminuta.* These enzymes are of interest because of their extraordinarily high catalytic rates for the turnover of OP pesticides, such as parathion, that approach the diffusion limit.[81] By generating a very large library (3.4×10^7) of variant genes and screening through IVC using water-in-oil emulsions (described above), Griffiths and Tawfik[82] were able to increase the k_{cat} of the enzyme to approximately 8000 per second, one of the highest turnover rates ever reported.

9.20.4.2 Increasing Substrate Range

One of the most valuable traits of enzymes is their specificity; unfortunately, for industrial and biotechnological applications, this can be a major stumbling block. Enzymes have typically evolved over thousands of years to specifically catalyze a reaction necessary for the organism to survive or at least outcompete other organisms. The huge number of enzymes reflects the vast array of chemical reactions that occur *in vivo.* The anthropogenic selection pressure that is now being placed on enzymes means that they are required to function with new chemicals, often derived from very different ecological niches in the case of natural products, or only recently synthesized in the case of drugs or pesticides. Thus, there is great interest in expanding the substrate range of enzymes. Examples where the substrate ranges of the bacterial PTE, glycosyltransferases (GTs), aldolases, and cytochromes P-450 have been increased through directed molecular evolution are discussed below.

The rapid hydrolysis of paraoxon by the bacterial PTE has already been discussed. However, the PTE catalyzes the hydrolysis of other pesticides with significantly lower efficiency. Using PTE expressed on the surface of *E. coli* cells, Cho and coworkers improved its catalytic ability to hydrolyze the closely related OP pesticides chlorpyrifos by 725-fold.[71] This was accomplished starting from a previously evolved PTE variant,[83] followed by three rounds of directed evolution. Similarly, Ollis and coworkers increased the activity of PTE for dimethyl-substituted OPs significantly,[69] raising them to the level of another naturally occurring PTE, including an ~70-fold increase in the turnover of dimethyl demeton. Interestingly, most of this effect was found to originate from a single H254R mutation within the active site that led to a new hydrogen bond to the substrate (**Figure 8**).

Another area of research that has utilized directed evolution to satisfy the demands for an expanded substrate range is the use of GTs in natural product glycosylation. For a comprehensive discussion, the reader is directed to a recent review of the field.[84] Briefly, the D- or L-sugar substituents of many natural products often have significant effects on their biological activity.[85] The importance of natural products to drug development and the importance of glycosylation to the function of many natural products have prompted the development

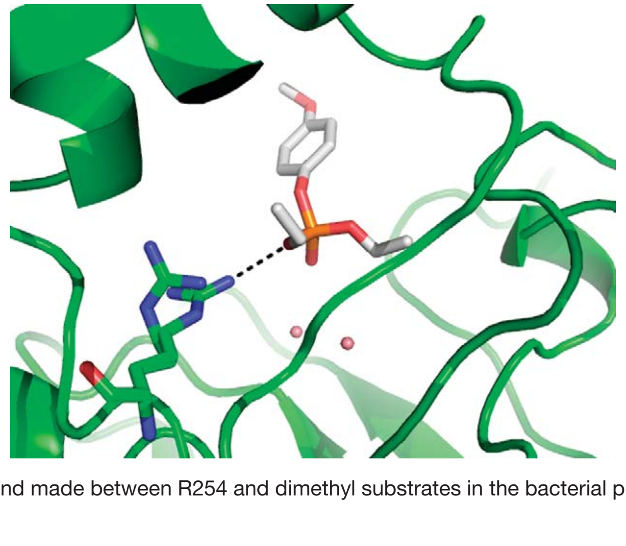


Figure 8 The hydrogen bond made between R254 and dimethyl substrates in the bacterial phosphotriesterase (PTE).

of tools by which the glycosylation of natural products can be facilitated. Enzymatic glycosylation is an attractive option, as it can allow *in vitro* glycorandomization and biosynthesis and has been used in the synthesis of improved antibiotics.[86] However, GTs are notoriously specific; in an analysis of the substrate range of the GT NovM using four aglycone analogues and over 40 nucleotide sugars, only three new 'unnatural' natural products were synthesized.[87] Recent work from the Thornson laboratory has used directed evolution to increase the substrate range of the oleandomycin GT dramatically, with the rate of glucoside formation with UDP-Glc and a number of compounds increasing between 2-fold and more than 180-fold, including a 5-fold increase with the antibiotic novobiocin.[88] Using the knowledge gained from this experiment regarding the location of the beneficial mutations, site-directed saturation mutagenesis was then performed to enhance the GT's activity with novobiocin.[89] An improvement resulted on the order of 300-fold compared with the wild-type enzyme and is a good example of the power of combining random and directed techniques.

Another important example of the use of directed evolution in increasing enzymatic substrate range is the aldol reaction, which has recently been reviewed.[90] Aldolases catalyze the asymmetric aldol reaction, a key reaction in natural product synthesis, owing to the formation of a new C–C bond and the introduction of chirality into molecules.[91] The native *N*-acetylneuraminic acid lyase (NAL) catalyzes the reversible aldol condensation of *N*-acetylmannosamine with pyruvate, to yield *N*-acetylneuraminic acid. Compounds related to *N*-acetylneuraminic acid, such as 6-dipropylcarboxamides, have attracted considerable interest owing to their efficacy as inhibitors of influenza A sialidase[92] and are difficult to synthesize chemically. Chemoenzymatic synthesis of such compounds is an attractive alternative, yet the activity of wild-type NAL toward 6-dipropylcarboxamides is relatively low. Using structure-based saturation mutagenesis, Berry and coworkers were able to achieve a dramatic (690-fold) shift in the substrate specificity of the enzyme from sialic acid to the dipropylamide analogue and a 49-fold increase in its activity toward the dipropylamide analogue used in the screening assay.[93] This example again demonstrates the power of using limited saturation libraries when the locations of probable enzyme/substrate contacts can be predicted with confidence.

The hydroxylation of small alkanes, such as propane and ethane, is particularly attractive to the chemical industry. However, monooxygenation of carbon centers is a reaction that is difficult to achieve chemically as it requires extreme temperatures. An attractive alternative is the use of cytochrome P-450 enzymes (P-450s) that can catalyze such reactions with a high degree of regio- and stereoselectivity. Thus, there has been much work directed toward evolving P-450s with altered substrate specificities that can catalyze hydroxylation of C1–C10 alkanes.

Most attention has been focused on evolving CYP102A1 (known historically as P-450 BM3). Farinas and coworkers[94] used a *p*-nitrophenol ether-linked model substrate (*p*-nitrophenoxy octane (8-pNPane)) to screen approximately 2000 mutants generated by epPCR for variants with enhanced activity toward octane, a poor substrate of the wild-type CYP102A1. Improvements of around fivefold based on an NADPH consumption assay were found in two successive rounds of evolution. This work was subsequently extended using StEP recombination with screening against 8-pNPane.[95] A CYP102A1 mutant (139-3) was identified that contained 11 amino acid substitutions and showed enhanced NADPH consumption with shorter chain alkanes (rates of up to $\sim 3900\,\text{min}^{-1}$ on hexane) and lauric and palmitic acids, two typical substrates for wild-type CYP102A1 (increased twofold up to $\sim 5160\,\text{min}^{-1}$). Mutant 139-3 hydroxylated propane, butane, pentane, hexane, cyclohexane, and octane (on the ω-1 to ω-3 positions on the linear substrates)[95] as well as monooxygenating benzene (to phenol), styrene (epoxide), cyclohexene (epoxide and 2-cyclohexene-1-ol), hexene (1-hexene-3-ol), and propylene (oxide) to the metabolites indicated.[96] K_d values for octane, hexane, and laurate were 10–25-fold lower than that for the binding of laurate to the wild-type enzyme. Interestingly, mutations found in 139-3 appeared to be almost entirely different from those proposed through rational redesign,[97,98] highlighting the power of random techniques.

Using an alterative approach, involving saturation mutagenesis of active site residues and recombination, Meinhold *et al.*[99] extended the lower limit of CYP102A1 substrate size to include ethane by targeting each of 11 active site residues within 5 Å of the substrate. The best mutant (53-5H) catalyzed ethane oxidation to ethanol, albeit at low rates. This study and similar work from Wong and coworkers, who achieved similar results with CYP101,[100] were particularly significant in that they demonstrated that the high bond energy of the C—H bond in ethane can be overcome by a P-450 biocatalyst.

9.20.4.3 Improving a ‘Promiscuous’ Activity

Increasing the substrate range of an enzyme could be thought of as improving a promiscuous activity. However, here we will use a more strict definition in which a ‘promiscuous activity’ involves catalysis of a reaction with a different class of substrate. Many enzymes are promiscuous in the sense that they can catalyze other reactions. This is not surprising considering that the enormous variety of enzymes that exist utilize only a small number of active site chemistries and structural scaffolds. Thus, an almost identical enzyme could catalyze lactone hydrolysis or phosphotriester hydrolysis. Because these activities are often very weak to begin with, directed molecular evolution experiments to improve these activities often result in remarkable improvements.

An excellent example of this is seen in the work performed with the mammalian serum paraoxonase. Although this enzyme was originally isolated because it could degrade the OP pesticide paraoxon, it has recently been shown that its most probable physiological function is as a lactonase.[101] Using directed evolution and a variety of selection strategies, Tawfik and coworkers were able to improve a number of low-level promiscuous activities, including thiolactonase (80-fold), esterase (31-fold), and PTE (155-fold), by significant amounts.[102,103] These improvements built upon the inherent hydrolytic function of the active site, which functions by polarizing a substrate and allowing nucleophilic attack by a water molecule.[104] This can occur in largely the same way with lactones, esters, and phosphotriesters: what differs, and what can be easily adjusted, is how suited the active site is to their binding. Using the technique of neutral drift introduced above, this work has been extended, demonstrating that libraries that number in the hundreds can produce many variants with greatly enhanced promiscuous activities without the requirement for the loss of the original activity.[103,105]

In enzymes that are naturally relatively promiscuous, there is some degree of overlap between studies showing increases in substrate range and those demonstrating enhanced substrate promiscuity, the distinction often being the sensitivity detection limit of the product being sought.

Wild-type CYP102A1 has been shown to metabolize certain drugs and related molecules, including chlorzoxazone, aniline, *p*-nitrophenol, propranolol, and nifedipine, typical substrates of human ‘drug-metabolizing’ P-450s, but with millimolar K_m values, significant uncoupling rates, and low turnover rates (in the same range as observed with human P-450s).[106] The SCHEMA algorithm was used to design a library of 6561 chimeras between the heme domains of CYP102A1, CYP102A2, and CYP102A3.[107] The final library was engineered using the SISDC method[55] to contain recombinations at seven sites, chosen so as to minimize disruption of important structural contacts. Crossovers were positioned within the elements of secondary structure, but blocks of fragments lined up with major structural elements or mobile regions of the CYP102A1 structure. Landwehr *et al.*[108] characterized 14-folded chimeras from the library and the three parents as both the isolated heme domains and with each of the three possible reductase domains against 11 substrates, including the drugs tolbutamide, chlorzoxazone, zoxazolamine, and propranolol. Against each substrate, a chimera was always the best enzyme. Five chimeras from the CYP102A1–A3 SCHEMA-derived library demonstrated novel activity toward verapamil metabolism and the two most active of these also metabolized astemizole, neither drug being a substrate of any of the parental forms. K-cluster analysis showed that mutants could be grouped on the basis of similarities in substrate specificity and that groupings of structurally similar substrates could be associated with these, giving hope that surrogate substrates might be useful for probing libraries for activity toward desired compounds. These studies provided the basis for the commercialization of this library by Codexis as a tool for drug development.

9.20.4.4 Improving Heterologous Enzyme Expression

For most screening procedures and bioindustrial enzymatic applications, it is essential that adequate amounts of enzyme can be produced through heterologous expression in *E. coli* or similar strains of bacteria. Unfortunately, difference between eukaryotic and prokaryotic protein folding means that enzymes that express well in eukaryotes may not express well in prokaryotes, because of different folding chaperones, and so on. Thus, there is often a pressing need to improve enzyme solubility and expression before enzymes can be used.

One excellent example is the human paraoxonase already described above. Before the work toward enhancing the promiscuous activities of the enzyme could be performed, the enzyme was required to be expressed in *E. coli*. This was achieved using family shuffling, as described above, and an activity screen,[102] and this led not only to its recombinant application, but also to the determination of its crystal structure.[104]

A second example is the use of fusion proteins to provide a selection method by which the amount of soluble protein can be increased. The first work in this area was carried out by Waldo and colleagues and involved fusing a green fluorescent protein (GFP) to the protein of interest.[109,110] Since formation of the GFP chromophore depends on correct folding of the fusion partner, it in effect serves as a reporter of the amount of folded protein that is produced. Numerous examples of the use of this technology exist, and the reader is directed to a recent review.[109] A similar approach has been applied by Ollis and coworkers, in which solubility was directly linked to survival through the creation of a fusion protein between the gene of interest and the DHFR gene from *E. coli*.[66] As DHFR can provide resistance against the antibiotic TMP, increased levels of overexpression of the protein of interest resulted in increased levels of DHFR production and antibiotic resistance. As an example of the application of this approach, five mutations were found to convert a highly insoluble *Vibrio fischeri* protein into a very soluble form (**Figure 9**).

9.20.4.5 Improving Coupling with Auxiliary Systems

A critical issue in the use of P-450s as biocatalysts is the cost of the cofactors NADPH and NADH (to a lesser extent) required to support P-450-mediated activity. The first approach used to address this issue has been to exploit the peroxide shunt, that is, the ability of P-450s to use peroxides as surrogate oxygen donors, thereby bypassing the need for supplying electrons from a pyridine cofactor, by evolution of the P-450 to function more effectively as a peroxygenase.

Joo *et al.*[111] used epPCR to generate ~20 000 mutants of CYP101, which were then screened in bacteria coexpressing horseradish peroxidase for H_2O_2-supported activity toward naphthalene. Roughly a sixth of the clones showed enhanced activity over wild type. The best mutants showed 5–11-fold increases for naphthalene over the wild type and more modest increases in activity toward 3-phenylpropionate. Mutations at one site were proposed to protect from H_2O_2-mediated inactivation but it was harder to ascribe roles to other residues and it was unclear whether mutations affected activity toward naphthalene or the ability to use H_2O_2. A second-generation SteP library identified mutants with 19- and 25-fold further increased activity.[111]

Efforts to bypass the need for pyridine nucleotide cofactors to supply electrons for the P-450 catalytic cycle have also involved the use of mediators. Nazor and Schwaneberg[112] evolved the CYP102A1 F87A variant to enhance electron transfer using Co(III) sepulchrate and zinc dust as cofactors. Saturation mutagenesis was performed on residues around the putative substrate access channel[113] followed by two rounds of epPCR

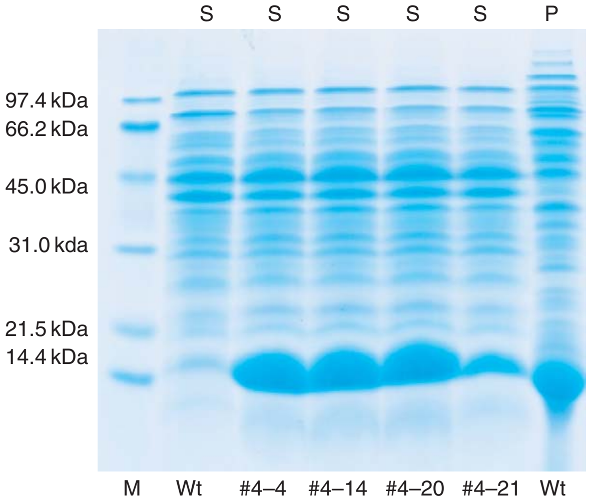


Figure 9 SDS-polyacrylamide gel showing the soluble (s) mutant proteins and the soluble and insoluble (p) wild-type protein.

throughout the whole CYP102A1 sequence to derive mutant M5. Using NADPH to support activity, K_m and k_{cat} values of mutants evolved for activity using the mediator system were inferior to the parental variant. However, with the electron donor system, K_m values were uniformly lower and k_{cat} values were elevated ~2.5-fold compared to the starting forms. Catalytic efficiencies were enhanced with the Zn/Co(III) sepulchrate system over NADPH for all mutants. It is unclear how mutations in the substrate access channel can promote exploitation of the mediator system.[112] One particular mutation, found to be critical in enhancing Zn/Co(III) sepulchrate-supported activity, was proposed to alter the enzyme structure in such a way as to enhance the access of critical residues proposed to be involved in an electron relay to Zn/Co(III) sepulchrate in the external medium.[113]

9.20.4.6 Improving Thermostability

Enzymes have (generally) evolved to function in neutral, aqueous environments at moderate temperatures (20–37 °C). However, they are often required to function in very different environments such as at high temperature. Listed below are several examples where an enzyme has been evolved to withstand high temperature.

One of the first examples of improving an enzyme's resistance to thermal denaturation was demonstrated by Arnold and coworkers in 1998.[114] A 4-nitrobenzyl esterase from *Bacillus subtilis* was evolved over six generations of random mutagenesis, recombination, and screening at high temperatures to increase its melting temperature by 14 °C. Importantly, this was accomplished without sacrificing its activity at mesophilic temperatures. The final variant contained only seven amino acid mutations out of 490, which were located in a flexible C-terminal domain of the enzyme. There was, however, no obvious trend in the location of the mutations, highlighting the importance of the random approach.

A more recent example of enhanced thermostability is the extraordinary improvement in thermostability of a lipase by Reetz *et al*.,[115] using the technique of iterative saturation mutagenesis.[64]. In this work, the 10 residues with the highest average *B*-values were chosen as sites for randomization. They were initially randomized individually, before being recombined in a sequential process. The final variant retained 50% of its activity after 1 h at 93 °C, compared with only 48 °C for the wild-type enzyme. This again demonstrates that with prior knowledge of the crystal structure, an advantage of iterative saturation mutagenesis over random mutagenesis-based methods such as DNA shuffling is that greater sequence space can be accessed.

Salazar *et al.*[116] used a CYP102A1 mutant designated 21B3 to evolve a more thermostable variant by epPCR, random mutagenesis, and DNA shuffling. Screening was for peroxygenase activity toward 12-*p*-nitrophenoxydodecanoic acid (12-pNCA) after heat treatment. Mutant 5H6 had a T_{50} (temperature at which 50% of the activity remains) of 61 °C compared to that of its cognate heme domain of 43 °C, which was in turn higher than both the 21B3 and F87A heme domains. Mutant 5H6 showed only around 50% of the peroxygenase activity of 21B3 but 10-fold that of F87A heme domain. Four of the introduced mutations were close to mutations that improved peroxygenase activity, suggesting that these compensate for temperature-destabilizing mutations accumulated in the search for increased peroxygenase activity.[116] Bloom *et al.*[107] then used the two CYP102A1 variant heme domains (21B3 and 5H6), which had similar activities but divergent stabilities toward temperature and urea, to assess whether stability affected the robustness of proteins to evolution. Of a total of 464 residues, 8 differed between the mutants. Two epPCR libraries were generated from the respective parental mutants with the same percentage and statistical distribution of mutations. The progeny in the library created from 5H6, however, was roughly twice as likely to fold correctly as indicated by characteristic spectra (61 vs. 33% for 21B3). Almost all folded mutants from both libraries retained activity toward 12-pNCA, but 13 mutants were found in the 5H6-derived library to have enhanced activity toward novel substrates as compared to only 4 in the 21B3-derived library. Whereas some of the 5H6-derived improved mutants had been destabilized to temperature compared with the parent (range 48–65 °C), all the 21B3-derived mutants were roughly as unstable as the parent (T_{50} values of 44–49 °C). Beneficial but destabilizing mutations identified in the 5H6-derived mutants were introduced into 21B3 but were found to cause irreversible loss of enzyme activity. Although the numbers of mutants found are too small to show a statistically significant enhancement of functional diversity in libraries from more stable versus less stable parents, the results suggest that stability enhancement facilitates molecular evolution.[107]

Arnold and colleagues[65,117] have also assessed the thermostability of mutants in the SCHEMA-derived library constructed from P-450s 102A1, 102A2, and 102A3, parental forms with T_{50} values of 54.9, 43.6, and 49.1 °C, respectively. A sample of 184 mutants showed values between 39 and 64 °C. Linear regression showed a correlation with fragment block composition, that is, within this set of chimeras, thermostability could be predicted based on sequence. Both the linear regression and a prediction based on a training set were sufficiently accurate to point to sequencing and other errors, for example, the expression of an incomplete or mutated sequence, which, when corrected, reconstituted proteins with the predicted stability. The most stable mutant had a half-life at 57 °C that was 108-fold longer than the most stable parent. The most stable chimeras also tended to be efficiently expressed, even in the absence of overt induction of the heterologous promoter.

9.20.4.7 Improving Activity in Organic Solvents

Much like temperature, many industrial applications of enzymes require activity in the presence of organic solvents. Interestingly, Burton *et al.*[118] have pointed out that thermostability and tolerance of organic solvents are characteristics that are often linked; sequence and structural modifications that improve activity to one denaturing influence often also improve tolerance to others. Unfortunately, enzymes have generally evolved to function in aqueous media; solvents are believed to denature and inactivate enzymes by exclusion of water from critical structural or functional components. Thus, many enzymes cannot tolerate common bioprocess conditions and techniques such as directed evolution must be used to improve their stability.[119]

One of the first such examples was again from the Arnold laboratory.[120] In this study, several rounds of random mutagenesis were performed to improve the activity of subtilisin in the presence of the organic solvent dimethylformamide (DMF) by 256-fold over the wild-type enzyme and this was later followed by a second publication describing further directed evolution and a 471-fold improvement over the wild-type enzyme. These studies started with the best rationally engineered enzyme,[121] the product of several years of work and containing three amino acid substitutions, and in a few generations exceeded its activity in DMF by a further sixfold as a result of 13 amino acid substitutions, many of which were located in regions that would not have been predicted to be important based on the prior understanding of structure–function relationships in the enzyme.

Many of the substrates of P-450s are organic compounds that have limited solubility in aqueous solution; there is accordingly a genuine need to improve cytochrome P-450 activity in organic solvents. The substrates of industrial relevance are hydrophobic to a greater or lesser degree. In addition, for maximal efficiency, substrates should be delivered at high concentrations and products build up to high concentrations, which presents problems for the solubility of many relevant chemicals in aqueous systems. Thus, organic solvents and other means of delivery and recovery have been explored, requiring investigation and/or optimization of the performance of the P-450 biocatalyst in the presence of solvents and other systems.

The solvent tolerance of CYP102A1 was improved by Wong *et al.*,[122] who used epPCR and saturation mutagenesis at several sites to evolve the wild-type enzyme and the F87A mutant for enhanced activity in THF and dimethyl sulfoxide (DMSO). Expression was performed using a temperature-controlled promoter that ensured that selected mutants also displayed a concomitant degree of thermal stability (42 °C). The first round of mutagenesis revealed two improved variants. Saturation mutagenesis at three positions, followed by random mutagenesis and reversion of A87 to F87 yielded a variant with approximately 6- and 3.4-fold improved activity in 25% v/v DMSO and 2% v/v THF, respectively. All mutants isolated appeared to have improved expression compared to the respective parental forms. Improved mutants also showed enhanced activity in the presence of a range of other solvents. Some mutations were found in the linker region between the heme and reductase domains, suggesting that the relative orientation of the two domains was important for solvent tolerance.

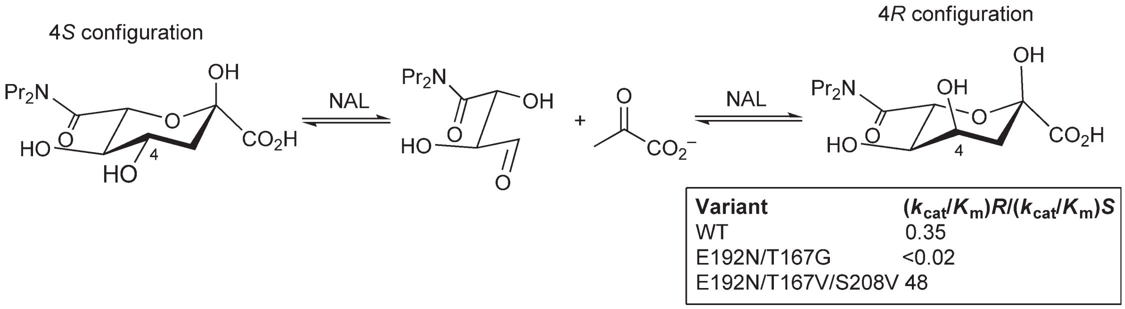
9.20.4.8 Improving Enzymatic Synthesis of Enantiopure Starting Materials

Stereochemistry is of utmost importance to natural product synthesis, as different enantiomers of the same natural product can often produce very different physiological effects. Several methods can be used to generate enantiopure starting materials or to introduce chirality along the synthetic route, including asymmetric catalysis and kinetic resolution. One of the most effective methods is the use of enzymes. Unfortunately, the odds of simply finding an enzyme with the desired enantioselectivity, specificity, and efficiency are extremely low and enzymes must therefore be engineered to better satisfy the requirements of the process.

The aldol reaction is a very important reaction in natural product synthesis because it can introduce chirality into the molecule.[91] The NAL has received much attention owing to its potential to participate in the biosynthesis of potentially antiviral sialic acid-related compounds.[92] However, as is often the case, although the wild-type enzyme had much potential, further efforts were required to tailor it to industrial needs. One shortcoming of the native enzyme is its poor facial selectivity, in which the cleavage of the 4*S*-configured product was only threefold favored over the 4*R*-configured product from the reversible condensation between pyruvate and an aldehyde ((2*R*,3*S*)-2,3-dihydroxy-4-oxo-*N*,*N*-dipropylbutyramide). To overcome this limitation, Berry and coworkers have extended their previous work in expanding the substrate range.[93] Starting from an E192N single mutant, which has improved activity toward analogues of *N*-acetylneuraminic acid in which the glycerol moiety is replaced by a dialkylaminocarbonyl group, they utilized directed molecular evolution to generate two complementary variants (E192N/T167G and E192N/T167V/S208V) that were 50-fold more selective toward the respective 4*S*- and 4*R*-configured condensation products (Williams 2006) (**Figure 10**).

A second example of the use of directed molecular evolution for natural product synthesis is the use of lipases by Reetz and colleagues. This work is based on the kinetic hydrolytic resolution of racemic mixtures, in which one enantiomer is preferentially hydrolyzed and the chiral product is thus enriched. Utilizing both random mutagenesis and directed techniques such as CAST,[64] they have improved the stereoselectivity of a lipase from *Pseudomonas aeruginosa* (PAL) on a number of occasions with different substrates. One of the first examples utilized the model substrate 2-methyldecanoic acid *p*-nitrophenyl ester, for which the wild-type enzyme has an enantioselectivity of $E = 1.1$. As a consequence of five mutations accumulated through random mutagenesis, followed by saturation mutagenesis, the enantioselectivity was increased to $E = 25.8$.[123] More recent work has involved applying the CAST methodology described above to PAL in order to generate a highly enantioselective ($E = 111$) enzyme for the kinetic resolution, in which one enantiomer is preferentially hydrolyzed and the chiral product is thus enriched, of an axially chiral allene, *p*-nitrophenyl 4-cyclohexyl-2-methylbuta-2,3-dienoate. Interestingly, the large improvement in enantioselectivity was found to result from a single L162P amino acid substitution in the active site.[65]

Finally, the bacterial PTE mentioned above has also been exhaustively studied with regard to its enantioselectivity. Initial studies used the known crystal structure of PTE to identify the substrate-binding pocket. This was then rationally evolved for enhancement and relaxation of the stereospecificity.[97] Most recently, a combinatorial library has been screened for the resolution of chiral phosphate, phosphonate, and phosphinate esters.[124] This work identified two variants with markedly different preferences for *S*p- and *R*p-enantiomers of 4-acetylphenyl methyl phenyl phosphate. One variant preferentially catalyzed hydrolysis of the *S*p-enantiomer by a factor of 3.7×10^5, while the other preferentially catalyzed hydrolysis of the *R*p-enantiomer by a factor of 9.7×10^2 – an enantioselective discrimination of 3.6×10^8.



Variant	$(k_{cat}/K_m)R/(k_{cat}/K_m)S$
WT	0.35
E192N/T167G	<0.02
E192N/T167V/S208V	48

Figure 10 The condensation reaction between pyruvate and (2*R*,3*S*)-2,3-dihydroxy-4-oxo-*N*,*N*-dipropylbutyramide catalyzed by NAL. The large changes in the facial selectivity, in either direction, as a result of mutations identified during directed evolution are listed in the table (inset).

9.20.5 Conclusions

Since its slow beginnings, and its renaissance after the introduction of DNA shuffling in 1994, the technique of directed molecular evolution has steadily increased in both popularity and number of applications. Many inventive molecular techniques and the incorporation of directed, or rational, approaches in order to focus on the search of sequence space have led to several major technical advances. Directed molecular evolution is now widely used in many industries, not least of which is the pharmaceutical industry, in which the use of evolved enzymes to generate enantiopure starting materials or to contribute to chemoenzymatic syntheses is becoming increasingly popular. The technical approaches and examples of the application should serve to provide the reader with an appreciation of both the potential and the limitations of utilizing this technique to improve the already significant catalytic benefit that enzymes can provide in a wide range of applications.

Abbreviations

8-pNPane	*p*-nitrophenoxy octane
12-pNCA	*p*-nitrophenoxydodecanoic acid
ADO	Assembly of designed oligonucleotides
CAST	combinatorial active-site saturation mutagenesis test
DHFR	dihydrofolate reductase
DMF	dimethylformamide
DMSO	dimethyl sulfoxide
epPCR	error-prone PCR
FACS	fluorescence-activated cell sorting
GAPDH	glyceraldehyde-3-phosphate dehydrogenase
GFP	green fluorescent protein
ITCHY	incremental truncation for the creation of hybrid enzymes
IVC	*in vitro* compartmentalization
P450 or CYP	cytochrome P-450
PCR	polymerase chain reaction
PDC	pyruvate decarboxylase
PTE	phosphotriesterase
RACHITT	random chimeragenesis on transient templates
SCRATCHY	combination of ITCHY and DNA shuffling
SISDC	sequence-independent site-directed chimeragenesis
ssDNA	single-stranded DNA
StEP	staggered extension process
THF	tetrahydrofuran
Thio-ITCHY	ITCHY with phosphothionate dTNPs
TMP	trimethoprim

References

1. B. I. Dahiyat; S. L. Mayo, *Science* **1997**, *278* (5335), 82–87.
2. B. Kuhlman; G. Dantas; G. C. Ireton; G. Varani; B. L. Stoddard; D. Baker, *Science* **2003**, *302* (5649), 1364–1368.
3. L. Regan, *Curr. Opin. Struct. Biol.* **1999**, *9* (4), 494–499.
4. D. Rothlisberger; O. Khersonsky; A. M. Wollacott; L. Jiang; J. DeChancie; J. Betker; J. L. Gallaher; E. A. Althoff; A. Zanghellini; O. Dym; S. Albeck; K. N. Houk; D. S. Tawfik; D. Baker, *Nature* **2008**, *453* (7192), 190–195.
5. G. F. Joyce, *Nature* **2002**, *420* (6913), 278–279.
6. A. Jaschke; B. Seelig, *Curr. Opin. Chem. Biol.* **2000**, *4* (3), 257–262.
7. M. Paschke, *Appl. Microbiol. Biotechnol.* **2006**, *70* (1), 2–11.
8. D. Lipovsek; A. Pluckthun, *J. Immunol. Methods* **2004**, *290* (1–2), 51–67.
9. X. Yan; Z. Xu, *Drug Discov. Today* **2006**, *11* (19–20), 911–916.

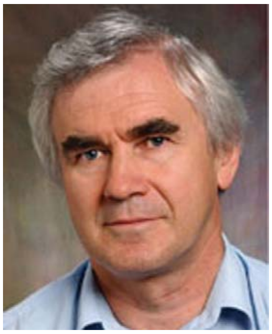
10. A. D. Keefe; J. W. Szostak, *Nature* **2001**, *410* (6829), 715–718.
11. D. Botstein; D. Shortle, *Science* **1985**, *229* (4719), 1193–1201.
12. A. Zaks, *Curr. Opin. Chem. Biol.* **2001**, *5* (2), 130–136.
13. G. W. Albers; V. E. Bates; W. M. Clark; R. Bell; P. Verro; S. A. Hamilton, *JAMA* **2000**, *283* (9), 1145–1150.
14. N. J. Turner, *Curr. Opin. Chem. Biol.* **2004**, *8* (2), 114–119.
15. T. D. Sutherland; I. Horne; K. M. Weir; C. W. Coppin; M. R. Williams; M. Selleck; R. J. Russell; J. G. Oakeshott, *Clin. Exp. Pharmacol. Physiol.* **2004**, *31* (11), 817–821.
16. N. A. Larsen; J. M. Turner; J. Stevens; S. J. Rosser; A. Basran; R. A. Lerner; N. C. Bruce; I. A. Wilson, *Nat. Struct. Biol.* **2002**, *9* (1), 17–21.
17. H. Sun; M. L. Shen; Y. P. Pang; O. Lockridge; S. Brimijoin, *J. Pharmacol. Exp. Ther.* **2002**, *302* (2), 710–716.
18. M. Watson; J. W. Liu; D. Ollis, *FEBS J.* **2007**, *274* (10), 2661–2671.
19. A. M. Klibanov, *Nature* **2001**, *409* (6817), 241–246.
20. A. L. Serdakowski; J. S. Dordick, *Trends Biotechnol.* **2008**, *26* (1), 48–54.
21. S. D. Copley, *Curr. Opin. Chem. Biol.* **2003**, *7* (2), 265–272.
22. R. J. Kazlauskas, *Curr. Opin. Chem. Biol.* **2005**, *9* (2), 195–201.
23. L. Afriat; C. Roodveldt; G. Manco; D. S. Tawfik, *Biochemistry* **2006**, *45* (46), 13677–13686.
24. O. Khersonsky; C. Roodveldt; D. S. Tawfik, *Curr. Opin. Chem. Biol.* **2006**, *10* (5), 498–508.
25. B. J. Stevenson; J. W. Liu; D. L. Ollis, *Biochemistry* **2008**, *47* (9), 3013–3025.
26. N. H. Horowitz, *Proc. Natl. Acad. Sci. U.S.A.* **1945**, *31* (6), 153–157.
27. J. A. Gerlt; P. C. Babbitt, *Annu. Rev. Biochem.* **2001**, *70*, 209–246.
28. G. A. Petsko; G. L. Kenyon; J. A. Gerlt; D. Ringe; J. W. Kozarich, *Trends Biochem. Sci.* **1993**, *18* (10), 372–376.
29. D. L. Ollis; E. Cheah; M. Cygler; B. Dijkstra; F. Frolow; S. M. Franken; M. Harel; S. J. Remington; I. Silman; J. Schrag; J. L. Sussman; K. H. G. Vershueren; A. Goldman, *Protein Eng.* **1992**, *5* (3), 197–211.
30. C. Neylon, *Nucleic Acids Res.* **2004**, *32* (4), 1448–1459.
31. N. Arnheim; H. Erlich, *Annu. Rev. Biochem.* **1992**, *61*, 131–156.
32. H. A. Erlich; D. Gelfand; J. J. Sninsky, *Science* **1991**, *252* (5013), 1643–1651.
33. R. K. Saiki; S. Scharf; F. Faloona; K. B. Mullis; G. T. Horn; H. A. Erlich; N. Arnheim, *Science* **1985**, *230* (4732), 1350–1354.
34. R. K. Saiki; D. H. Gelfand; S. Stoffel; S. J. Scharf; R. Higuchi; G. T. Horn; K. B. Mullis; H. A. Erlich, *Science* **1988**, *239* (4839), 487–491.
35. R. C. Cadwell; G. F. Joyce, *PCR Methods Appl.* **1994**, *3* (6), S136–S140.
36. Y. H. Zhou; X. P. Zhang; R. H. Ebright, *Nucleic Acids Res.* **1991**, *19* (21), 6052.
37. R. C. Cadwell; G. F. Joyce, *PCR Methods Appl.* **1992**, *2* (1), 28–33.
38. M. Zaccolo; D. M. Williams; D. M. Brown; E. Gherardi, *J. Mol. Biol.* **1996**, *255* (4), 589–603.
39. K. Miyazaki; F. H. Arnold, *J. Mol. Evol.* **1999**, *49* (6), 716–720.
40. S. Shafikhani; R. A. Siegel; E. Ferrari; V. Schellenberger, *Biotechniques* **1997**, *23* (2), 304–310.
41. W. P. Stemmer, *Nature* **1994**, *370* (6488), 389–391.
42. W. P. Stemmer, *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91* (22), 10747–10751.
43. M. Kikuchi; K. Ohnishi; S. Harayama, *Gene* **1999**, *236* (1), 159–167.
44. S. Bershtein; D. S. Tawfik, *Curr. Opin. Chem. Biol.* **2008**, *12* (2), 151–158.
45. N. Maheshri; D. V. Schaffer, *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100* (6), 3071–3076.
46. C. A. Voigt; S. L. Mayo; F. H. Arnold; Z. G. Wang, *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98* (7), 3778–3783.
47. H. Zhao; L. Giver; Z. Shao; J. A. Affholter; F. H. Arnold, *Nat. Biotechnol.* **1998**, *16* (3), 258–261.
48. J. M. Joern; P. Meinhold; F. H. Arnold, *J. Mol. Biol.* **2002**, *316* (3), 643–656.
49. W. M. Coco; W. E. Levinson; M. J. Crist; H. J. Hektor; A. Darzins; P. T. Pienkos; C. H. Squires; D. J. Monticello, *Nat. Biotechnol.* **2001**, *19* (4), 354–359.
50. J. E. Ness; S. Kim; A. Gottman; R. Pak; A. Krebber; T. V. Borchert; S. Govindarajan; E. C. Mundorff; J. Minshull, *Nat. Biotechnol.* **2002**, *20* (12), 1251–1255.
51. D. Zha; A. Eipper; M. T. Reetz, *ChemBioChem* **2003**, *4* (1), 34–39.
52. M. Ostermeier; J. H. Shim; S. J. Benkovic, *Nat. Biotechnol.* **1999**, *17* (12), 1205–1209.
53. S. Lutz; M. Ostermeier, *Methods Mol. Biol.* **2003**, *231*, 143–151.
54. S. Lutz; M. Ostermeier; S. J. Benkovic, *Nucleic Acids Res.* **2001**, *29* (4), E16.
55. K. Hiraga; F. H. Arnold, *J. Mol. Biol.* **2003**, *330* (2), 287–296.
56. M. M. Meyer; J. J. Silberg; C. A. Voigt; J. B. Endelman; S. L. Mayo; Z. G. Wang; F. H. Arnold, *Protein Sci.* **2003**, *12* (8), 1686–1693.
57. P. E. O'Maille; M. Bakhtina; M. D. Tsai, *J. Mol. Biol.* **2002**, *321* (4), 677–691.
58. J. A. Kolkman; W. P. Stemmer, *Nat. Biotechnol.* **2001**, *19* (5), 423–428.
59. S. Xu; J. Ju; H. Misono; K. Ohnishi, *Gene* **2006**, *368*, 126–137.
60. V. Abecassis; D. Pompon; G. Truan, *Methods Mol. Biol.* **2003**, *231*, 165–173.
61. K. L. Morley; R. J. Kazlauskas, *Trends Biotechnol.* **2005**, *23* (5), 231–237.
62. W. A. Lim; R. T. Sauer, *Nature* **1989**, *339* (6219), 31–36.
63. R. M. Kelly; H. Leemhuis; L. Dijkhuizen, *Biochemistry* **2007**, *46* (39), 11216–11222.
64. M. T. Reetz; J. D. Carballeira, *Nat. Protoc.* **2007**, *2* (4), 891–903.
65. J. D. Carballeira; P. Krumlinde; M. Bocola; A. Vogel; M. T. Reetz; J. E. Backvall, *Chem. Commun. (Camb.)* **2007**, *21* (19), 1913–1915.
66. J. W. Liu; Y. Boucher; H. W. Stokes; D. L. Ollis, *Protein Expr. Purif.* **2006**, *47* (1), 258–263.
67. S. Y. McLoughlin; C. Jackson; J. W. Liu; D. L. Ollis, *Appl. Environ. Microbiol.* **2004**, *70* (1), 404–412.
68. O. Mueller-Cajar; S. M. Whitney, *Biochem. J.* **2008**, *414* (2), 205–214.
69. H. Yang; P. D. Carr; S. Y. McLoughlin; J. W. Liu; I. Horne; X. Qiu; C. M. Jeffries; R. J. Russell; J. G. Oakeshott; D. L. Ollis, *Protein Eng.* **2003**, *16* (2), 135–145.

70. A. Ben-David; G. Shoham; Y. Shoham, *Chem. Biol.* **2008**, *15* (6), 546–551.
71. C. M. Cho; A. Mulchandani; W. Chen, *Appl. Environ. Microbiol.* **2004**, *70* (8), 4681–4685.
72. B. J. Stevenson; J. W. Liu; D. L. Ollis, *Biochemistry* **2008**, *47* (9), 3013–3025.
73. H. Leemhuis; K. P. Nightingale; F. Hollfelder, *FEBS J.* **2008**, *275* (22), 5635–5647.
74. K. M. Polizzi; M. Parikh; C. U. Spencer; I. Matsumura; J. H. Lee; M. J. Realff; A. S. Bommarius, *Biotechnol. Prog.* **2006**, *22* (4), 961–967.
75. H. Leemhuis; R. M. Kelly; L. Dijkhuizen, *IUBMB Life* **2009**, *61* (3), 222–228.
76. N. Varadarajan; S. Rodriguez; B. Y. Hwang; G. Georgiou; B. L. Iverson, *Nat. Chem. Biol.* **2008**, *4* (5), 290–294.
77. A. Aharoni; K. Thieme; C. P. Chiu; S. Buchini; L. L. Lairson; H. Chen; N. C. Strynadka; W. W. Wakarchuk; S. G. Withers, *Nat. Methods* **2006**, *3* (8), 609–614.
78. A. Aharoni; G. Amitai; K. Bernath; S. Magdassi; D. S. Tawfik, *Chem. Biol.* **2005**, *12* (12), 1281–1289.
79. V. Taly; B. T. Kelly; A. D. Griffiths, *ChemBioChem* **2007**, *8* (3), 263–272.
80. A. Huebner; L. F. Olguin; D. Bratton; G. Whyte; W. T. Huck; de A. J. Mello; J. B. Edel; C. Abell; F. Hollfelder, *Anal. Chem.* **2008**, *80* (10), 3890–3896.
81. S. R. Caldwell; J. R. Newcomb; K. A. Schlecht; F. M. Raushel, *Biochemistry* **1991**, *30* (30), 7438–7444.
82. A. D. Griffiths; D. S. Tawfik, *EMBO J.* **2003**, *22* (1), 24–35.
83. C. M. Cho; A. Mulchandani; W. Chen, *Appl. Environ. Microbiol.* **2002**, *68* (4), 2026–2030.
84. G. J. Williams; R. W. Gantt; J. S. Thorson, *Curr. Opin. Chem. Biol.* **2008**, *12* (5), 556–564.
85. A. C. Weymouth-Wilson, *Nat. Prod. Rep.* **1997**, *14* (2), 99–110.
86. X. Fu; C. Albermann; J. Jiang; J. Liao; C. Zhang; J. S. Thorson, *Nat. Biotechnol.* **2003**, *21* (12), 1467–1469.
87. C. Albermann; A. Soriano; J. Jiang; H. Vollmer; J. B. Biggins; W. A. Barton; J. Lesniak; D. B. Nikolov; J. S. Thorson, *Org. Lett.* **2003**, *5* (6), 933–936.
88. G. J. Williams; C. Zhang; J. S. Thorson, *Nat. Chem. Biol.* **2007**, *3* (10), 657–662.
89. G. J. Williams; R. D. Goff; C. Zhang; J. S. Thorson, *Chem. Biol.* **2008**, *15* (4), 393–401.
90. A. Bolt; A. Berry; A. Nelson, *Arch. Biochem. Biophys.* **2008**, *474* (2), 318–330.
91. R. Mahrwald, *Chem. Rev.* **1999**, *99* (5), 1095–1120.
92. N. R. Taylor; A. Cleasby; O. Singh; T. Skarzynski; A. J. Wonacott; P. W. Smith; S. L. Sollis; P. D. Howes; P. C. Cherry; R. Bethell; P. Colman; J. Varghese, *J. Med. Chem.* **1998**, *41* (6), 798–807.
93. G. J. Williams; T. Woodhall; A. Nelson; A. Berry, *Protein Eng. Des. Sel.* **2005**, *18* (5), 239–246.
94. E. T. Farinas; U. Schwane; A. Glieder; F. H. Arnold, *Adv. Synth. Catal.* **2001**, *343*, 601–606.
95. A. Glieder; E. T. Farinas; F. H. Arnold, *Nat. Biotechnol.* **2002**, *20* (11), 1135–1139.
96. E. T. Farinas; M. Alcalde; F. H. Arnold, *Tetrahedron* **2004**, *60* (3), 525–528.
97. M. Chen-Goodspeed; M. A. Sogorb; F. Wu; F. M. Raushel, *Biochemistry* **2001**, *40* (5), 1332–1339.
98. O. Lentz; Q.-S. Li; U. Schwaneberg; S. Lutz-Wahl; P. Fisher; R. D. Schmid, *J. Mol. Catal. B Enzym.* **2001**, *15*, 123–133.
99. P. Meinhold; M. W. Peters; M. M. Chen; K. Takahashi; F. H. Arnold, *ChemBioChem* **2005**, *6* (10), 1765–1768.
100. F. Xu; S. G. Bell; J. Lednik; A. Insley; Z. Rao; L. L. Wong, *Angew. Chem. Int. Ed. Engl.* **2005**, *44* (26), 4029–4032.
101. O. Khersonsky; D. S. Tawfik, *Biochemistry* **2005**, *44* (16), 6371–6382.
102. A. Aharoni; L. Gaidukov; S. Yagur; L. Toker; I. Silman; D. S. Tawfik, *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101* (2), 482–487.
103. G. Amitai; R. D. Gupta; D. S. Tawfik, *HFSP J.* **2007**, *1* (1), 67–78.
104. M. Harel; A. Aharoni; L. Gaidukov; B. Brumshtein; O. Khersonsky; R. Meged; H. Dvir; R. B. Ravelli; A. McCarthy; L. Toker; I. Silman; J. L. Sussman; D. S. Tawfik, *Nat. Struct. Mol. Biol.* **2004**, *11* (5), 412–419.
105. R. D. Gupta; D. S. Tawfik, *Nat. Methods* **2008**, *5* (11), 939–942.
106. G. Di Nardo; A. Fantuzzi; A. Sideri; P. Panicco; C. Sassone; C. Giunta; G. Gilardi, *J. Biol. Inorg. Chem.* **2007**, *12* (3), 313–323.
107. J. D. Bloom; S. T. Labthavikul; C. R. Otey; F. H. Arnold, *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103* (15), 5869–5874.
108. M. Landwehr; M. Carbone; C. R. Otey; Y. Li; F. H. Arnold, *Chem. Biol.* **2007**, *14* (3), 269–278.
109. G. S. Waldo, *Methods Mol. Biol.* **2003**, *230*, 343–359.
110. G. S. Waldo; B. M. Standish; J. Berendzen; T. C. Terwilliger, *Nat. Biotechnol.* **1999**, *17* (7), 691–695.
111. H. Joo; Z. Lin; F. H. Arnold, *Nature* **1999**, *399* (6737), 670–673.
112. J. Nazor; U. Schwaneberg, *ChemBioChem* **2006**, *7* (4), 638–644.
113. J. Nazor; S. Dannenmann; R. O. Adjei; Y. B. Fordjour; I. T. Ghampson; M. Blanusa; D. Roccatano; U. Schwaneberg, *Protein Eng. Des. Sel.* **2008**, *21* (1), 29–35.
114. L. Giver; A. Gershenson; P. O. Freskgard; F. H. Arnold, *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95* (22), 12809–12813.
115. M. T. Reetz; J. D. Carballeira; A. Vogel, *Angew. Chem. Int. Ed. Engl.* **2006**, *45* (46), 7745–7751.
116. O. Salazar; P. C. Cirino; F. H. Arnold, *ChemBioChem* **2003**, *4* (9), 891–893.
117. C. R. Otey; M. Landwehr; J. B. Endelman; K. Hiraga; J. D. Bloom; F. H. Arnold, *PLoS Biol.* **2006**, *4* (5), e112.
118. S. G. Burton; D. A. Cowan; J. M. Woodley, *Nat. Biotechnol.* **2002**, *20* (1), 37–45.
119. S. Luetz; L. Giver; J. Lalonde, *Biotechnol. Bioeng.* **2008**, *101* (4), 647–653.
120. K. Chen; F. H. Arnold, *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90* (12), 5618–5622.
121. K. Q. Chen; F. H. Arnold, *Biotechnology (NY)* **1991**, *9* (11), 1073–1077.
122. T. S. Wong; F. H. Arnold; U. Schwaneberg, *Biotechnol. Bioeng.* **2004**, *85* (3), 351–358.
123. K. Liebeton; A. Zonta; K. Schimossek; M. Nardini; D. Lang; B. W. Dijkstra; M. T. Reetz; K. E. Jaeger, *Chem. Biol.* **2000**, *7* (9), 709–718.
124. C. Nowlan; Y. Li; J. C. Hermann; T. Evans; J. Carpenter; E. Ghanem; B. K. Shoichet; F. M. Raushel, *J. Am. Chem. Soc.* **2006**, *128* (49), 15892–15902.

Biographical Sketches

Colin J. Jackson was born in Hamilton (New Zealand) and received his B.Sc. (Hons) in 2002 from the University of Otago. In 2006 he completed his Ph.D. under the supervision of Professor David Ollis at the Australian National University. He subsequently completed a short postdoctoral fellowship under Dr. John Oakeshott at the Commonwealth Scientific and Industrial Research Organization (CSIRO, Australia) before being appointed as a research scientist. He is currently on leave to take up a Marie Curie Research Fellowship at the Institut de Biologie Structurale in Grenoble, France. He is broadly interested in the structural and chemical determinants of enzymatic catalysis and enzyme engineering.

Elizabeth M. J. Gillam studied biochemistry at the University of Queensland and graduated with a B.Sc. (Hons) in 1986. In 1990 she received a D.Phil (Pharmacology) from the Oxford University before undertaking postdoctoral research with Fred Guengerich in the Department of Biochemistry and Center in Molecular Toxicology, Vanderbilt University, USA. She returned to the University of Queensland to take up an academic position in the Department of Physiology and Pharmacology (later the School of Biomedical Sciences) in 1993 and was made a Professor in Biochemistry in the School of Chemistry and Molecular Biosciences in 2009. Her research interests involve cytochrome P-450 enzymes, their engineering and application in drug discovery and biotechnology, and their roles in drug metabolism and toxicology.



David L. Ollis completed a B.Sc. (Hons) at the University of New South Wales before moving across town to complete a Ph.D. in the Department of Chemistry in Sydney University. He then worked in Tom Steitz's laboratory in Yale University before taking a faculty position in the Department of Biochemistry, Molecular Biology, and Cell Biology in Northwestern University, Chicago. He returned to Australia in 1992 to take up a position within the Research School of Chemistry of the Australian National University. His research interests include the application of directed evolution and crystallography in protein structure–function studies.

9.21 Single Molecule Fluorescence Methods in Enzymology

Peng Chen and Nesha May Andoy, Cornell University, Ithaca, NY, USA

9.21.1 Introduction

The rapid advances in single-molecule methods have enabled many innovative studies in the life sciences. These single-molecule measurements have unraveled the detailed workings of many macromolecular machineries,[1] that would otherwise be hidden in measuring the average behaviors of a population of molecules. Breakthrough discoveries continue to emerge in areas such as molecular motors,[2–6] protein–DNA interactions,[7–9] RNA activities,[10–14] protein folding and dynamics,[15–19] enzymology,[20–25] and gene expression.[26–28]

The power of the single-molecule approach stems from its many distinctive features. First, it removes population averaging so that heterogeneous behaviors of biomolecules can be revealed and subpopulations analyzed. This is particularly important for biological system since heterogeneity easily arises in biomacromolecules, for example, proteins in different conformational states. Second, it removes the need for synchronization of molecular actions for studying time-dependent processes, as it monitors one molecule at a time. This feature also allows us to visualize the actions of individual biomolecules in real time, which is particularly useful in capturing reactive intermediates and elucidating the mechanisms of biochemical reactions.

Experimental methods to investigate single biomolecules fall into three categories: optical (e.g., fluorescence and nonlinear optical microscopy), mechanical (e.g., optical tweezers, magnetic tweezers, scanning probe microscopy, and microfluidics), and electrical measurements (e.g., patch clamp and nanopores). All these methods allow real-time observation of dynamic processes of individual biomolecules. Among these methods, single-molecule fluorescence techniques are perhaps one of the most popular because of their straightforward instrumentation and easy operation.[4,29–32] One can monitor the fluorescence intensity, spectrum, polarization, or lifetime of a biomolecule to investigate its molecular properties; among these, following fluorescence intensities is the most straightforward. Introducing exogenous fluorescent labels is a general strategy when the target biomolecules are not naturally fluorescent.

To achieve single-molecule fluorescence detection, there are three common experimental practices. First, experiments are done at low concentrations (10^{-9}–10^{-12} mol l^{-1}) to spatially separate molecules, so each of them can be studied without interference from surrounding molecules. Second, fluorescence signal detection is confined to a small volume ($<10^{-15}$ l) to minimize background noises for single-molecule sensitivity. Third, biomolecules are often immobilized, so a single molecule can be studied over time.

Two experimental setups are widely used for single-molecule fluorescence detection: total internal reflection (TIR) fluorescence microscopy and confocal fluorescence microscopy. The evanescent field from TIR confines the laser excitation to a thin layer (~50–300 nm), whereas the confocal scheme focuses on the laser beam to a diffraction limited volume and uses a pinhole to confine the signal detection around the focus ($\sim 300 \times 300 \times 600$ nm^3). The TIR fluorescence microscopy typically uses electron-multiplying cameras as detectors and can image hundreds of molecules simultaneously; the time resolution is about milliseconds limited by the camera speed, although submillisecond imaging is possible with the state-of-the-art hardware and exceptionally bright probes.[33] The confocal microscopy uses point detectors, such as single-photon avalanche photodiodes, and examines one molecule at a time; the time resolution can be up to microseconds for following dynamic processes of biomolecules. For both TIR and confocal microscopy, multiple detection channels, such as different colors and polarizations, can be readily implemented. The laboratory manual edited by Selvin and Ha[34] provides a comprehensive account of the technical and application aspects of single-molecule techniques. Many reviews on the principles and applications of these single-molecule methods are also available.[1,4,5,13,18,28–32,35–53]

The scope of this chapter is limited to single-molecule studies of enzymology, in particular using single-molecule fluorescence techniques. A continuing challenge in structure–function studies of enzymes is to understand the contribution of conformational dynamics to enzyme function. Single-molecule enzymology, with its ability to monitor enzyme reactions in real time at the single-molecule level, can gain insight into how slow conformational dynamics of enzymes are coupled to catalysis.[12,20–25,54–62]

This chapter is organized as follows: We divide the single-molecule enzymology studies according to their approaches; for each approach, we focus on the principle, features and generality, and experimental challenges, and use examples in the recent literature for illustrations. For different perspectives on single-molecule enzymology studies, we refer the readers to previous reviews,[47–50,62] in particular the extensive review by Hammes and coworkers.[48] We also want to point out that some content of this chapter overlaps with a recent review article by the authors on single-molecule studies of biological inorganic systems.[50]

9.21.2 Fluorescent Active Site

9.21.2.1 Principle

Many enzymes use organic cofactors, such as flavin and porphyrin, at the active sites for catalysis. Some of these organic cofactors, especially flavin, have intrinsic fluorescence, and can be imaged readily at the single-molecule level. If the fluorescence of these cofactors is coupled with the state of the active site in the catalytic cycle, monitoring the fluorescence of the active site can directly probe the catalysis. The classic example of this approach is the study of cholesterol oxidase (COx) by Xie and coworkers.[20] Gafni, Palfey, Steel, and coworkers further applied this approach to study dihydroorotate dehydrogenase[63,64] and *p*-hydroxybenzoate hydroxylase.[24] Here, we will use the COx study to exemplify the approach.

9.21.2.2 Example: Catalytic Dynamics of Cholesterol Oxidase

COx catalyzes the oxidation of cholesterol by oxygen. The active site of the enzyme contains a flavin adenine dinucleotide (FAD), which is naturally fluorescent in its oxidized form (FAD) but not in its reduced form ($FADH_2$). During the catalytic cycle, the FAD at the active site is reduced to $FADH_2$ by the substrate cholesterol, which is then oxidized back to FAD by O_2 (**Figure 1(a)**). Therefore, each catalytic turnover of a COx is accompanied by a cycle of the active site between the fluorescent oxidized FAD form and the nonfluorescent reduced $FADH_2$ form. Monitoring the fluorescence of the FAD can thus directly probe the catalysis of a single COx molecule in real time.

Xie and coworkers utilized the intrinsic fluorescence properties of the FAD active site of COx to study its catalytic dynamics.[20] Using confocal fluorescence microscopy, they recorded real-time fluorescence intensity trajectories of single COx molecules trapped in agarose gels that also contain the substrate cholesterol and oxygen. The fluorescence trajectory of a single COx molecule shows digital on–off events where each on–off event represents a single catalytic turnover between the FAD and $FADH_2$ forms of the enzyme (**Figure 1(b)**). The fluorescence on-times in the trajectory are the stochastic waiting times for the reduction reaction of Enzyme-flavin adenine dinucleotide (E-FAD) and the off-times are those for the oxidation reaction of E-$FADH_2$. Both waiting times can be statistically analyzed to extract out the kinetics of the associated reactions. The distribution of on-times in **Figure 1(c)** exemplifies the kinetic information contained in the single-turnover waiting times: the delayed maximum of the on-time distribution reflects the presence of a kinetic intermediate during the reduction of FAD by cholesterol, which is an enzyme–substrate complex as in the classic Michaelis–Menten mechanism.

The real-time single-turnover trajectories also enabled Xie and coworkers to analyze the time-dependent activity of each enzyme molecule. They found that individual COx molecules show temporal activity fluctuations (i.e., dynamic disorder in activity), attributable to the slow conformational dynamics of the enzyme. The timescale of the activity fluctuation is the timescale of the conformational dynamics that are longer than the catalytic turnovers and can be obtained from the autocorrelation function of the waiting times (**Figure 1(d)**), which shows an exponential decay behavior versus the index of turnovers (m) and whose decay constant is the fluctuation timescale. This conformational dynamics-coupled enzyme catalysis is fundamental to enzyme catalysis and extremely challenging to study with traditional methods measuring the average behaviors of a population of molecules.

9.21.2.3 Features and Generality

Many enzymes use the flavin cofactor at the active site; the fluorescent active site approach can be applied to study these enzymes at the single-molecule level. Other naturally fluorescent enzymes, like those that contain NAD cofactors, can in principle be studied, although the bluer fluorescence of NAD poses a technical challenge for single-molecule fluorescence detection. As the approach uses the natural fluorescence of the enzyme, no labeling with fluorescent probes is needed, offering no or minimum perturbation on the enzyme structure and function.

9.21.2.4 Challenges

Common to most fluorescence-based single-molecule methods, photobleaching limits the observation time window using the fluorescent active site approach. For COx, its FAD cofactor showed a significantly better stability than common dye molecules, possibly due to the protection by the protein.[20] As this approach relies on the natural fluorescence of the enzyme, the application is largely limited to flavin enzymes. And, more often than not, the natural fluorescence of the FAD cofactor inside a protein is intrinsically quenched by nearby redox-active residues (e.g., tyrosine, tryptophan) through photo-induced electron transfer (see Section 9.21.7). This natural fluorescence quenching can sometimes be circumvented by mutating the residues that cause quenching, although caution needs to be exercised that the mutation will not perturb the enzyme function significantly.

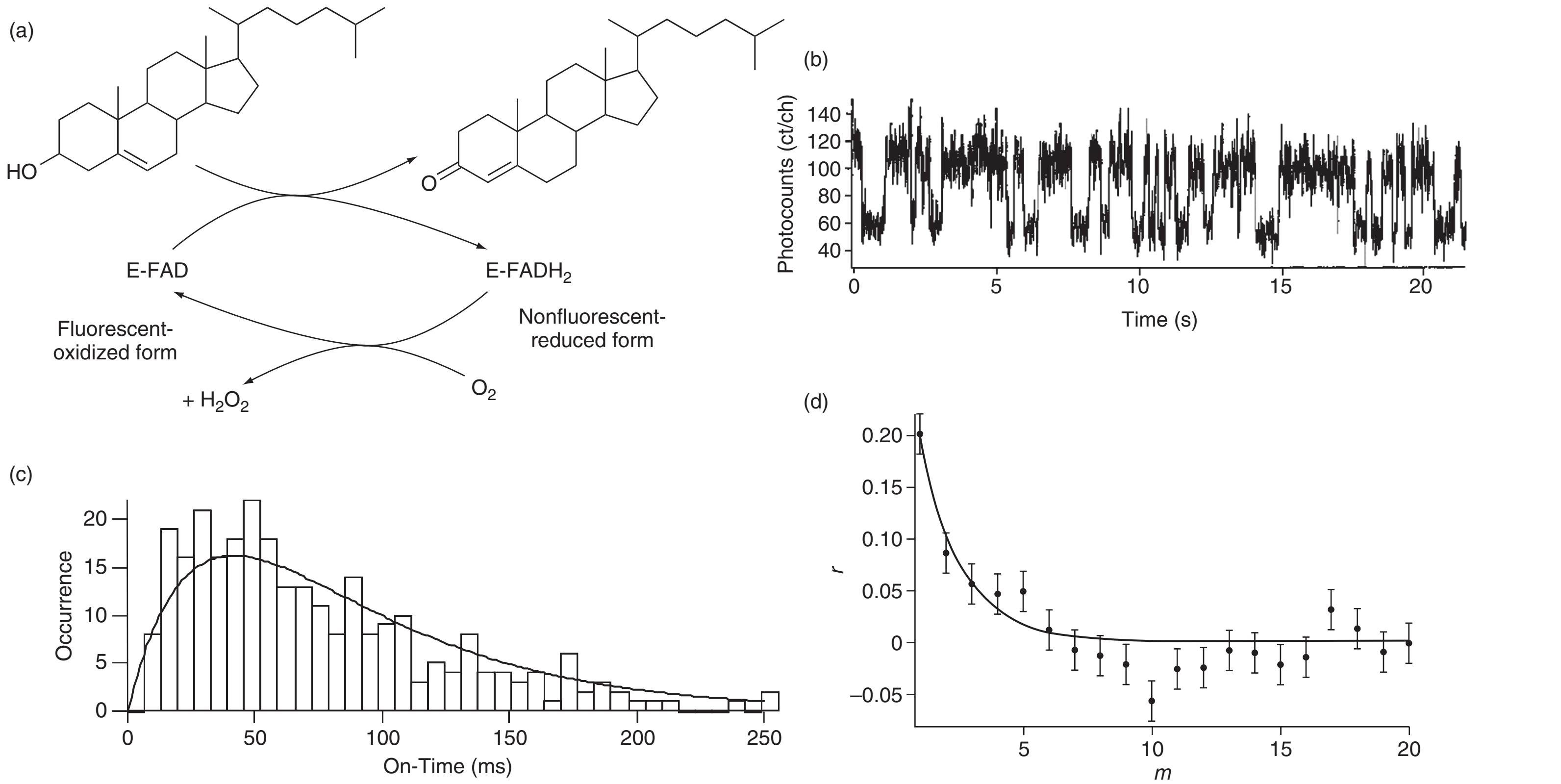


Figure 1 (a) Catalytic cycle of COx. E: enzyme. (b) Fluorescence intensity trajectory of a single COx molecule during catalysis. (c) Distribution of the on-times (bars) derived from the fluorescence intensity trajectory of a single COx molecule at $2\,\text{mmol}\,\text{l}^{-1}$ cholesterol. (d) The autocorrelation function of on-times $r(m) = \langle \Delta t(0)\Delta t(m)\rangle / \langle \Delta t2 \rangle$; m is the index of turnovers. Reproduced from H. P. Lu; L. Y. Xun; X. S. Xie, *Science* **1998**, *282*, 1877–1882. Reprinted with permission from AAAS.

9.21.3 Fluorogenic Reaction

9.21.3.1 Principle

The chemical transformation catalyzed by an enzyme also provides an opportunity to study catalysis at the single-molecule level. If this chemical transformation generates a fluorescent product (i.e., a fluorogenic reaction), monitoring the fluorescence of the product can directly probe the catalysis. This fluorogenic reaction approach is arguably the most popular in studying enzymes at the single-molecule level. Two experimental designs have been reported in applying this approach to assay single enzyme activity: one to detect the generation of every product molecule in real time by a single enzyme at single-turnover resolution,[21–23,54–56,65–67] and the other to monitor the accumulation of all the product molecules over time produced by a single enzyme encapsulated in a confined volume.[68–75]

Using the real-time single-turnover detection design, Rigler and coworkers first examined the activity of horseradish peroxidase at the single-molecule level,[22,54] which was further studied by Engelkamp and coworkers.[66] Velonia *et al.*,[23] Flomenbom *et al.*,[55] and Carette *et al.*[67] used this approach to study lipase catalysis. Xie *et al.*[21] and Moerner *et al.*[56] studied β-galactosidase, a favorite system in single-molecule enzymology studies. Hofkens and coworkers also used this to study chymotrypsin.[65] On the contrary, using encapsulation in confined volume, Yeung *et al.*[68] studied lactate dehydrogenase and Dovichi *et al.*[69] studied alkaline phosphatase by trapping individual enzymes in capillary tubes. Noji and coworkers[70] and Collier and coworkers[75] used microchambers formed in polydimethylsiloxane (PDMS) stamps, and Walt and coworkers[71,72] used microchambers formed at the tips of optical fibers to trap individual β-galactosidase molecules for activity assays. Liposomes[74] and water-in-oil emulsions[73] were also used to encapsulate individual alkaline phosphatase[74] and chymotrypsin[73] enzymes for activity assays with fluorogenic reactions.

Here, the studies by Xie and coworkers[21] and Noji and coworkers[70] using the fluorogenic reaction approach are used to illustrate these two experimental designs, both of which studied β-galactosidase as a model enzyme.

9.21.3.2 Example 1: β-Galactosidase with Single-Turnover Detection

β-Galactosidase (β-gal) catalyzes the hydrolysis of lactose. A nonfluorescent substrate resorufin-β-D-galactopyranoside (RGP) can be hydrolyzed by β-gal to generate the fluorescent resorufin (R) as one of the reaction products (**Figure 2(a)**). Using this fluorogenic reaction, Xie and coworkers studied the enzyme kinetics of individual β-gal molecules in real time at single-turnover resolution.[21] To follow the catalysis of a single β-gal molecule, they immobilized the enzyme onto polystyrene beads, which are in turn immobilized on a glass surface, and monitored the fluorescence signal from a single β-gal molecule in real time using confocal fluorescence microscopy (**Figure 2(a)**). The micron-sized beads are visually identifiable under an optical microscope, helping to locate individual β-gal molecules. The fluorescence intensity trajectory from a single β-gal molecule shows stochastic fluorescence bursts in the presence of the substrate RGP; each burst represents a single turnover of catalytic product formation (**Figure 2(b)**). The intervals between any two neighboring fluorescence bursts are the microscopic waiting times (τ) for individual product formation events, and these waiting times are the most important observables in the trajectory and contain most information about the catalytic dynamics.

Since each catalytic reaction generates a new product molecule, the fluorogenic reaction approach is not limited by photobleaching and long turnover trajectories can be obtained. Counting the burst frequency in sequential time segments of a single trajectory directly determines the time evolution of the activity of a single β-gal molecule (**Figure 2(c)**). The activity of a single β-gal molecule shows strikingly temporal fluctuations that occur over a wide range of timescales. This temporal activity fluctuation is especially significant at high substrate concentrations where the catalytic conversion reaction is rate-limiting in the overall turnover cycle (**Figure 2(c)**, blue trajectory), indicating the large fluctuation of the catalytic rate constant of the enzyme. This dynamic fluctuation of enzyme activity, that is, dynamic disorder, is attributable to the interconversion dynamics of different enzyme conformers that have different activities. Xie and coworkers further showed that the classic Michaelis–Menten still holds for fluctuating single enzymes, as shown by the Lineweaver–Burke plot of $\langle \tau \rangle$ versus $1/[S]$, where $\langle \tau \rangle$ is the time average of single-turnover waiting times (**Figure 2(d)**).

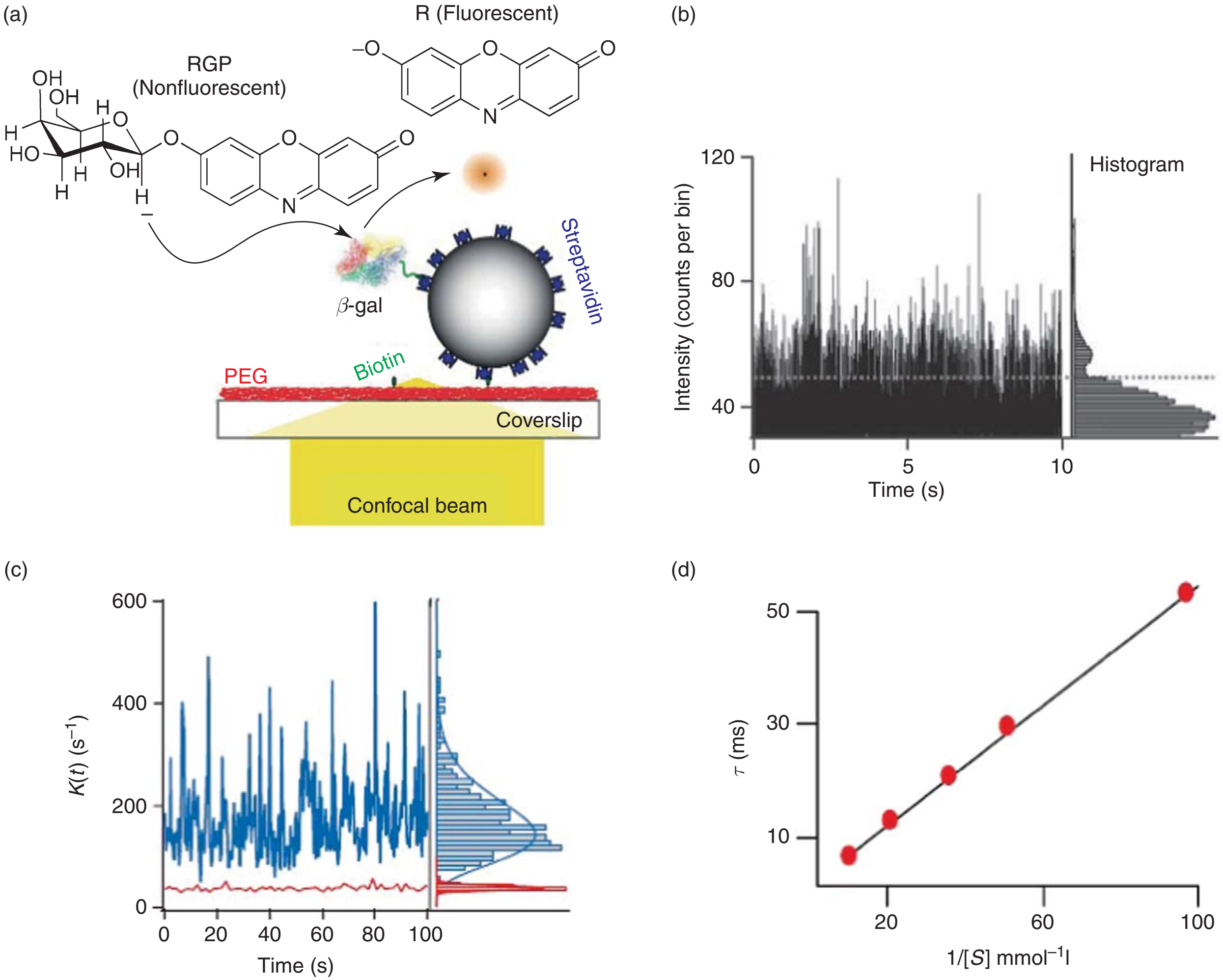


Figure 2 (a) Schematic of enzyme immobilization and fluorescence detection. A single β-gal molecule is linked to a streptavidin-coated polystyrene bead, which is immobilized on a biotin-coated glass surface. RGP: resorufin-β-D-galactopyranoside; R: resorufin. (b) Fluorescence intensity trajectory of a single β-gal molecule at 100 μmol l^{-1} RGP. Dashed line is the threshold separating the signal from the background noise. On the right is the fluorescence intensity histogram. (c) Two trajectories of turnover rates $k(t)$ of two β-gal enzymes at 20 μmol l^{-1} (red) and 100 μmol l^{-1} (blue) RGP concentration. On the right are the corresponding histograms of $k(t)$. (d) Single-molecule Lineweaver–Burke plot, $\langle\tau\rangle$ versus 1/[*S*]. The line is a fit with the classic Michaelis–Menten equation. Reprinted by permission from Macmillan Publishers Ltd: B. P. English; W. Min; A. M. van Oijen; K. T. Lee; G. Luo; Y. Sun; B. J. Cherayil; S. C. Kou; X. S. Xie, *Nat. Chem. Biol.* **2006**, *2*, 87–94, copyright (2006).

9.21.3.3 Example 2: β-Galactosidase with Microchamber Encapsulation

Although the real-time single-turnover detection is powerful with fluorogenic reactions, a few challenges still exist under certain circumstances. First, for enzymes having slow overall turnover rates but fast product dissociation, one needs to maintain a 'short' time resolution in fluorescence detection to catch every product molecule before it dissociates from the enzyme and at the same time needs to monitor the same enzyme for a 'long' time to observe statistically significant number of turnover events. This broad time range poses a technical challenge in practice, because maintaining high time resolution for a long period will generate trajectories with a huge number of data points that often push the capacity of computers. Second, surface immobilization sometimes affects the enzyme activity significantly. Third, the nonfluorescent substrates for fluorogenic reactions still have some fluorescence (although weak), and the experiments are often limited to the relatively low concentration range of substrates to achieve single-molecule sensitivity of the product molecule.

Encapsulation of single enzyme molecule in confined volumes offers a way to overcome these challenges. By trapping a single enzyme molecule in a confined volume with substrates inside, the product will be confined in the same volume and its accumulation over time can be detected without maintaining a high time resolution to catch the product. Immobilization is also unnecessary due to confinement. Moreover, as one can monitor the accumulation of many product molecules produced by a single enzyme in the same confined volume, high substrate concentration conditions can also be studied without worrying about suppressing background for single-molecule detection.

This confined volume strategy is well demonstrated by the microchamber encapsulation work by Noji and coworkers.[70] Using soft-lithography, they fabricated arrays of femtoliter microchambers on a PDMS stamp (**Figure 3(a)**). They then encapsulated enzyme–substrate solutions in these microchambers, and kept the enzyme concentration low (picomolar–nanomolar) to ensure statistically there would be one or two enzyme molecules in each microchamber. They studied β-gal as a model system and used fluorescein-di-β-galactopyranoside as the fluorogenic substrate that will generate fluorescein as a product. Imaged by a wide-field fluorescence microscope that can observe many microchambers simultaneously, the enzyme activity is manifested by the steady increase of fluorescence intensity inside individual microchambers, reporting the product accumulation because of enzyme catalysis (**Figure 3(b)**). The characteristic of single-enzyme catalysis is the discrete sloping of the fluorescence intensity versus time trajectories; the discrete sloping here results from the discrete number of enzyme molecules in each microchamber (i.e., zero, one, or two enzymes) (**Figure 3(b)**). From these slopes, the rate of product accumulation can be derived for the catalysis in each microchamber. The histogram of the rates from many microchambers shows multiple peaks, corresponding to microchambers that contain zero, one, two, and three enzyme molecules (the distribution here follows the Poisson distribution).

9.21.3.4 Features and Generality

As the fluorescent product is continuously generated during catalysis, the biggest advantage of the fluorogenic reaction approach is that there is no photobleaching limit, which is common to many other fluorescence-based single-molecule techniques. The nonfluorescent (or weakly fluorescent) nature of the substrate also enables study over a wide range of substrate concentrations; even for high substrate concentrations where the increased background prevents single-molecule product detection, the confined-volume strategy still allows study of single enzyme catalysis. As no fluorescent labeling of the enzyme is involved, there is no perturbation on the enzyme function.

The fluorogenic reaction approach is also widely applicable so long as suitable fluorogenic reactions are available. With modern synthetic methods, suitable fluorogenic substrates can probably always be designed and synthesized for a particular enzymatic reaction; examples are already available in the literature.[76–80]

9.21.3.5 Challenges

To study a particular enzyme using the fluorogenic reaction approach, the first challenge is to find a fluorogenic substrate. As discussed above, with modern synthetic chemistry, the availability of suitable substrates is only limited by researchers' creativity in designing molecules.

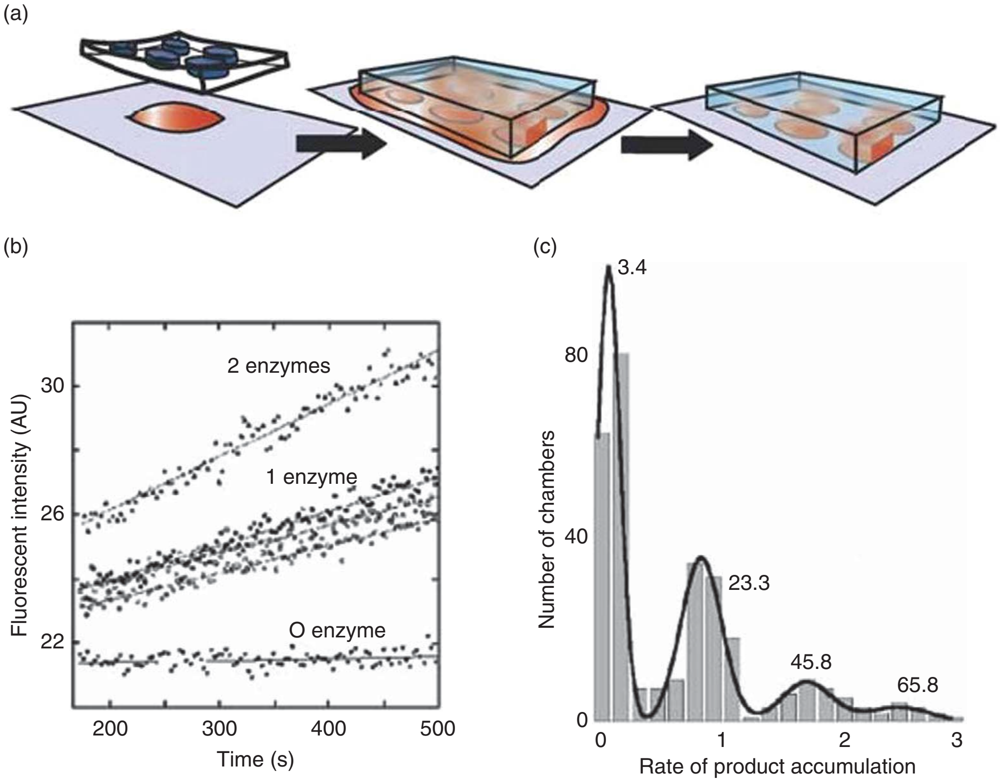


Figure 3 (a) Schematic of solution encapsulation by PDMS stamps with embedded femtoliter microchambers. (b) Exemplary fluorescence intensity versus time trajectories from individual microchambers. The time resolution here is much longer than the turnover time. The different slopes reflect the different number of enzyme molecules in each microchamber. (c) Histogram of product accumulation rates from many microchambers. From low to high values, the multiple peaks correspond to microchambers containing zero, one, two, and three enzyme molecules, respectively. Reprinted by permission from Macmillan Publishers Ltd: Y. Rondelez; G. Tresset; K. V. Tabata; H. Arata; H. Fujita; S. Takeuchi; H. Noji, *Nat. Biotechnol.* **2005**, *23*, 361–365, copyright (2005).

For the real-time single-turnover detection of fluorogenic reactions, catalytic reactions with fast turnover rates (or fast product dissociation) pose a time resolution challenge, as single-molecule fluorescence detection often requires detection of hundreds of photons to obtain statistically significant information. For these fast enzymes, one can vary the substrates, use enzyme variants from different organisms, or use mutants to slow the catalysis down.

For the encapsulation in confined volume strategy, the observation time is limited by the total amount of substrate molecules available in the enclosed chambers, and the catalysis ends when all substrate molecules are consumed. Therefore, it is desirable to have free exchange of reaction solutions. For the PDMS-based microchambers, one way is to couple the microchambers to microfluidic channels, which allow solution exchange and are controllable through pressure valves, as demonstrated by Collier and coworkers.[75]

9.21.4 Fluorescent Substrate

9.21.4.1 Principle

Besides detecting the fluorescence of the catalytic product, another approach to study enzymes at the single-molecule level is to detect the fluorescence of the substrate. Here, the substrate, either naturally fluorescent or fluorescently labeled, is monitored at the single-molecule level to directly follow its association with the

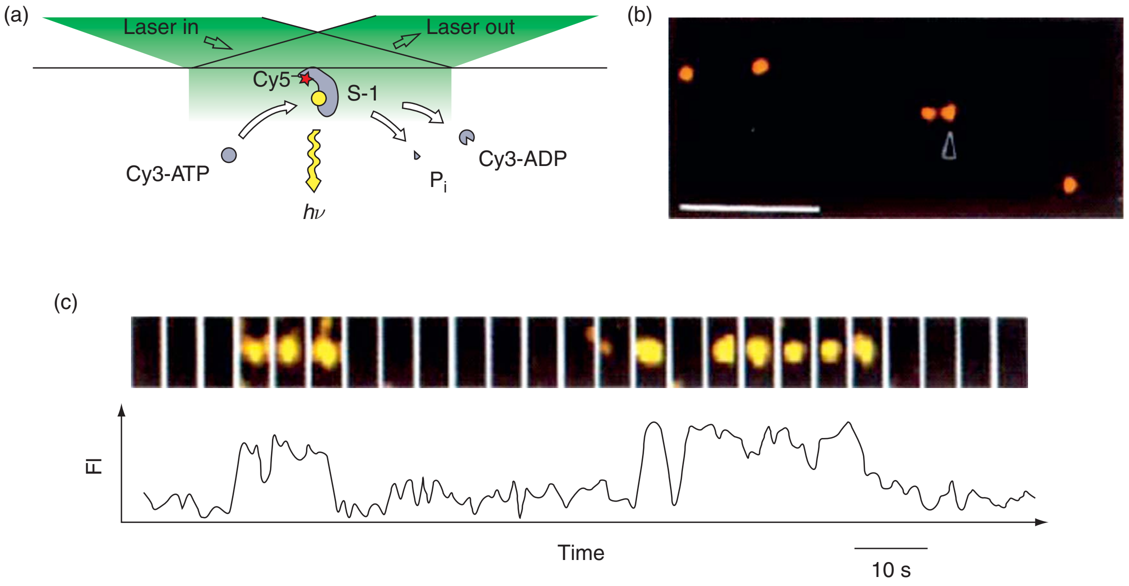


Figure 4 (a) Schematic of experimental design of detecting single-head myosin catalysis using a fluorescent substrate. The laser excitation is in TIR geometry. The Cy5-label on the enzyme helps to identify the location of individual enzymes. (b) Fluorescence image of Cy5-labeled myosin immobilized on the surface under solution. Cy5 is excited at 632 nm. The image is artificially colored red. (c) ATP turnovers by a single myosin molecule. The upper panel shows typical fluorescence images of Cy3-nucleotide (ATP or ADP) coming in and out of focus by associating and dissociating with a myosin molecule marked by an arrowhead in (b). Cy3-fluorescence is excited at 532 nm. The image is artificially colored yellow. The lower panel shows the time trajectory of the corresponding fluorescence intensity. Figure (a) reprinted with permission from Macmillan Publishers Ltd: T. Funatsu; Y. Harada; M. Tokunaga; K. Saito; T. Yanagida, *Nature* **1995**, *374*, 555–559, copyright (1995).

enzyme and the subsequent dissociation. Yanagida and coworkers used this approach to detect the ATP hydrolysis activity of single myosin molecules,[61,81] which is described below.

Myosin is a motor protein and hydrolyzes ATP to power its mechanical movement along actin filaments. The ATP hydrolysis comes from the ATPase activity of the myosin molecule. Yanagida and coworkers used the fluorescent substrate approach to visualize its ATPase activity at the single-molecule level.[61,81] They synthesized a Cy3-dye-labeled ATP derivative, Cy3-ATP, as a fluorescent substrate for myosin. To monitor the ATP hydrolysis turnovers by single myosin molecules, they immobilized myosin molecules on a glass surface (Figure 4(a)). The myosin was labeled with a Cy5-dye to facilitate its identification. They first obtained the Cy5-fluorescence image to locate individual myosin molecules immobilized on the glass surface using TIR fluorescence microscopy and 632-nm excitation (Figure 4(a,b)). The Cy3-ATP substrate was supplied in the solution. They then switched to a 532-nm laser to directly excite the Cy3-fluorescence to probe the ATP hydrolysis by single myosin molecules. Each association event of a Cy3-ATP molecule to the enzyme is reported by a sudden appearance of the Cy3-fluorescence signal at the position of a myosin molecule; the subsequent dissociation of the hydrolyzed product Cy3-ADP or unhydrolyzed substrate is reported by the sudden disappearance of the Cy3-fluorescence (Figure 4(c)). By analyzing the distribution of the fluorescence residence times, that is, the on-times in the Cy3-fluorescence intensity trajectories (Figure 4(c)), they were able to determine the dissociation rate that agrees with the ATP turnover rate of Cy3-ATP by myosin.

9.21.4.2 Features and Generality

The fluorescent substrate approach is in principle generalizable, if suitable fluorescent substrates are available. As the fluorescence signal is associated with the substrate, substrate binding to the enzyme is directly observable, and so is the lifetime of enzyme–substrate complex (possibly including the lifetime of the enzyme–product complex if the product also carries the same fluorescence). The direct observation of substrate actions is advantageous compared to the fluorogenic reaction approach, which cannot observe substrate binding to the enzyme. This approach is also not limited by photobleaching, because the fluorescent substrate can be supplied continuously with solution flow and typical photobleaching lifetimes of fluorescent probes are longer than the substrate residence time on the enzyme.

9.21.4.3 Challenges

A big limitation of the fluorescent substrate approach is the concentration limit – because the substrate is fluorescent, the experiments have to be done at low substrate concentrations (10^{-12}–10^{-9} mol l^{-1}) to suppress the fluorescence background for single-molecule detection. (The TIR scheme in Yanagida and coworker's study[61,81] helps suppress the background owing to the small excitation volume.) At these low concentrations, enzyme turnover rate is often diffusion limited. Moreover, because the fluorescence signal is associated with the substrate, whether the chemical transformation of catalysis has indeed happened is unclear, and the disappearance of the signal could simply result from substrate dissociation, not dissociation of the catalytic product. Supplementary experiments are needed here to confirm that catalysis does happen at the single-molecule imaging conditions.

9.21.5 Fluorescence Resonance Energy Transfer (FRET)

9.21.5.1 Principle

FRET is widely applied in single-molecule fluorescence studies of biomolecules.[31,32,82] Governed by the Förster mechanism,[83] the spectral overlap between the fluorescence spectrum of a donor molecule and the absorption spectrum of an acceptor can result in efficient energy transfer from the donor to the acceptor. The energy transfer efficiency is dependent on the donor to acceptor distance, r, within the nanometer range (energy transfer efficiency $= 1/[1+(r/r_0)^6]$, where r_0 is the Förster radius of the donor–acceptor pair, a constant typically of a few nanometers[83,84]). The most general experimental design of applying FRET is to label biomolecules with a fluorescent donor–acceptor pair and then monitor the fluorescence intensities of both the donor and the acceptor at the single-molecule level. For studying enzymes, if an enzyme reaction involves nanometer-scale distance changes between any parts of the enzyme complex, the enzyme reaction can be studied by the single-molecule FRET method by labeling the enzyme (or the enzyme complex) with a pair of FRET probes at appropriate locations.

Weiss and coworkers pioneered in using single-molecule FRET to study the conformational dynamics and reaction mechanism of staphylococcal nuclease.[85] Lu and coworkers used it to study the conformational dynamics of T4 lysozyme during catalysis.[58] Hammes, Benkovic, and coworkers studied dihydrofolate reductase,[25] the enzymes involved in T4 primosome[86] and replisome.[87,88] Yang and coworkers studied adenylate kinase.[89] Here, we use the T4 lysozyme study to illustrate the approach.[58]

9.21.5.2 Example: Conformational Dynamics-Coupled Catalysis of T4 Lysozyme

T4 lysozyme catalyzes the hydrolysis of polysaccharide chains of *E. coli* cell wall matrices. This enzyme has two domains connected by an α-helix (**Figure 5(a)**), and the active site sits between the two domains. During catalysis, the two domains undergo hinge-bending motions that are coupled with substrate binding. Lu and coworkers used single-molecule FRET to study this conformational dynamics of T4 lysozyme under hydrolysis reactions.[58] They attached a FRET donor–acceptor pair, tetramethylrhodamine (TMR) and Texas Red, to Cys54 and Cys97 of the enzyme. By exploiting the reactivity difference of maleimide and iodoacetamide with the sterically constrained Cys54, they were able to label these two cysteines with TMR and Texas Red site specifically. At these two labeling positions, the TMR–Texas Red interdistance directly reports the hinge-bending motions of the two domains of the lysozyme. With a Förster radius of ~50 Å for this FRET pair and an average Cys54–Cys97 distance of ~36 Å, the energy transfer from TMR to Texas Red is sensitive to their interdistance changes from the enzyme conformational dynamics.

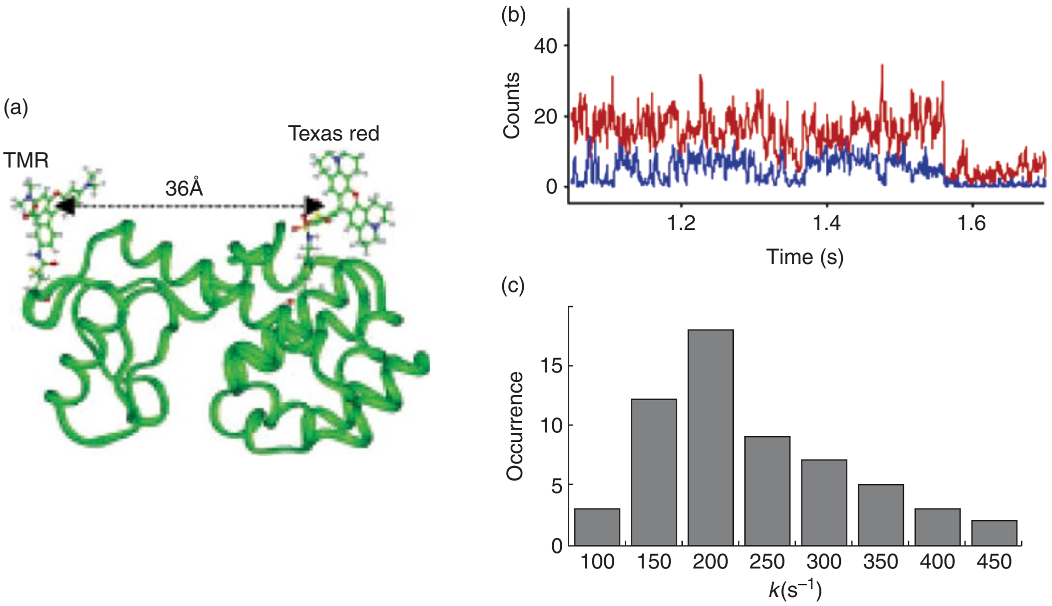


Figure 5 (a) The structure of T4 lysozyme with the two dye labels schematically shown. (b) Fluorescence intensity trajectories of the TMR donor (blue) and the Texas Red acceptor (red) of a single T4 lysozyme in the presence of *E. coli* B cell wall. (c) Distribution of the decay rate constants (*k*) of the donor intensity autocorrelation functions. Reproduced with permission from Y. Chen; D. Hu; E. R. Vorpagel; H. P. Lu, *J. Phys. Chem. B.* **2003**, *107*, 7947–7956. Copyright (2003) American Chemical Society.

To follow the conformational dynamics of T4 lysozyme, Lu and coworkers used confocal fluorescence microscopy to monitor the real-time energy transfer between the two probes on a single enzyme molecule by monitoring the fluorescence intensities of both probes.[58] The donor–acceptor fluorescence intensities of a single lysozyme show large anticorrelated fluctuations in the presence of the substrate (i.e., *E. coli* cell wall), directly reporting the enzyme conformational dynamics (**Figure 5(b)**). By analyzing the autocorrelations of the fluorescence intensity trajectories, they determined the rate constant of the conformational dynamics for individual enzymes. They observed large static heterogeneity in the distribution of the rate constant (**Figure 5(c)**), indicating the widely varying dynamic properties of individual enzyme molecules. The large static heterogeneity was attributed to enzyme searching for reactive sites on the substrate.

9.21.5.3 Features and Generality

FRET is a general approach for studying enzymes. So long as there are structural changes within the enzyme or the enzyme complex during catalysis, FRET can be used to monitor the conformational dynamics and thus the enzyme reaction. A wide variety of fluorescent probes are available.[84] With different combinations of these probes, FRET pairs of different Förster radii can be designed to cover a large range of distances.

9.21.5.4 Challenges

As FRET methods require a pair of fluorescent probes, to label a protein with two probes site-specifically is a general challenge. Orthogonal labeling chemistry is needed to place the donor and the acceptor at the specific locations. Luckily, for many occasions, one merely needs to label the enzymes with two labels at two locations, obtaining a mixture of donor–donor-, acceptor–acceptor-, donor–acceptor-, and acceptor–donor-labeled molecules. Under single-molecule imaging conditions, the donor–donor- and acceptor–acceptor-labeled molecules are easily distinguishable and discarded; and the donor–acceptor and acceptor–donor-labeled ones are often effectively the same for FRET measurements.

Based on the Förster mechanism and the available probes suitable for simultaneous single-molecule two-color detection, the sensitivity of FRET is limited to nanometer-scale conformational changes. Angstrom-scale conformational changes, which are also common for biomacromolecules, are beyond the sensitivity of FRET. For detecting angstrom-scale dynamics, the fluorescence quenching via electron transfer provides an alternative (see Section 9.21.7). Engineering biomolecules to amplify small-scale structural changes is a viable strategy to enable single-molecule FRET measurements, such as our work of using engineered DNA Holliday junctions to study protein–DNA interactions that involve merely small structural changes. Photobleaching also limits the observation time of FRET measurements. Nonetheless, with a good oxygen scavenging system, a single fluorescent probe molecule can last for up to a few minutes before being photobleached;[90] complication here is that the O_2 scavenging system must not interfere with the enzyme reaction.

9.21.6 Fluorescence Quenching via Energy Transfer

9.21.6.1 Principle

Instead of using a pair of donor–acceptor fluorescent probes, a variant of FRET, fluorescence quenching via energy transfer, uses only one fluorescent probe as a donor. The acceptor is nonfluorescent and acts as a quencher – it has strong absorption bands that overlap spectrally with the fluorescence of the donor. This energy-transfer-caused quenching changes the donor fluorescence, from which the chemical state of the acceptor (i.e., quencher) or the distance between the donor and the quencher can be deduced. Hammes and coworkers used this fluorescence quenching approach to examine the substrate binding and catalysis of dihydrofolate reductase.[59,60] Visser and coworkers used this approach to probe the catalysis of *p*-hydroxybenzoate hydroxylase.[91] Also, Canters and coworkers used this to follow the reaction of nitrite reductase (NiR).[92] Here, we use the NiR study to exemplify the approach.

9.21.6.2 Example: Catalysis of Nitrite Reductase

NiR catalyzes the reduction of nitrite (NO_2^-) to nitric oxide (NO). The enzyme is a homotrimer (**Figure 6(a)**); each monomer contains two mononuclear copper sites: one so-called type-1 copper site that acts as an electron transfer center and the other so-called type-2 copper site where the catalysis occurs. These two copper sites have distinct absorption properties. The type-1 site belongs to the large family of blue copper sites in biology and has a cysteine ligand. At the oxidized form (Cu^{2+}), this cysteine ligand gives rise to an intense sulfur-to-Cu^{2+} charge transfer absorption near 600 nm ($\varepsilon > 4000\,\mathrm{mol^{-1}\,l\,cm^{-1}}$); at the reduced form ($Cu^{1+}$), this site has no absorption in the visible region because of the d^{10} electron configuration of Cu^{1+} ions. On the contrary, the type-2 copper site has histidine and water-based ligands and shows little or no absorption in the visible region regardless of its oxidation state.

Canters and coworkers utilized the different spectral features between the oxidized and reduced forms of the type-1 copper site to follow NiR catalysis at the single-molecule level.[92] They labeled one monomer of NiR at the N-terminus with the fluorescent probe ATTO-655, whose fluorescent spectrum overlaps with the absorption band of the oxidized type-1 copper site. Because of this spectral overlap, the fluorescence intensity of ATTO-655 is quenched by energy transfer to the type-1 copper site when this copper site is at the oxidized form. If the type-1 copper site is at the reduced form, no quenching via energy transfer occurs and the fluorescence intensity of ATTO-655 is high. During the catalysis of NiR, its type-1 copper site cycles through oxidized and reduced forms repetitively, and the ATTO-655 consequently undergoes quenching and no-quenching, resulting in its temporal fluorescence intensity fluctuations that are coupled with the catalysis (**Figure 6(a)**). Based on this scheme, Canters and coworkers used confocal fluorescence microscopy to monitor the real-time catalysis of surface-immobilized NiR molecules at the single-molecule level. They observed large fluorescence intensity fluctuations that report the turnover dynamics of individual NiR molecules (**Figure 6(b)**). By analyzing the fluctuation behaviors of individual trajectories, they revealed a distribution of electron transfer rates between the type-1 and type-2 copper centers during the catalytic cycle, which is related to the disorder of the catalytic site of the enzyme.

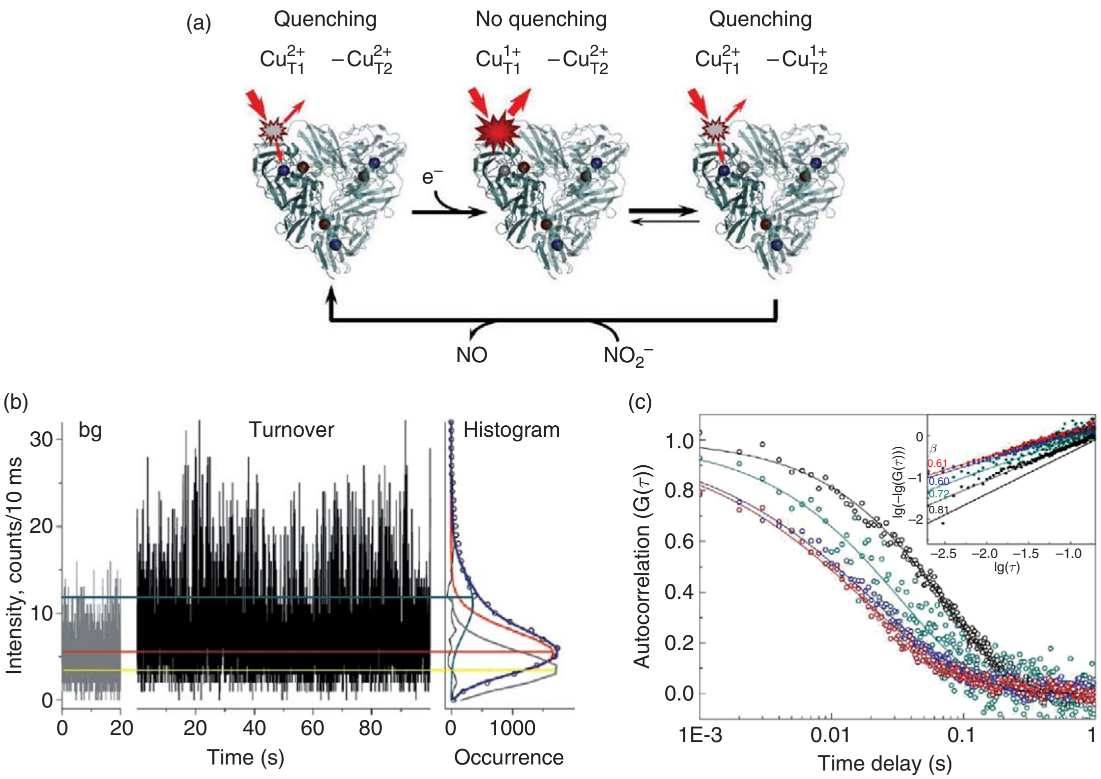


Figure 6 (a) Schematic of sequence of events during turnover of a labeled NiR, showing the fluorescence quenching via energy transfer that reports the turnover. Left: the resting enzyme, both type-1 Cu (Cu_{T1}, blue) and type-2 Cu (Cu_{T2}, red) are at the 2+ oxidized state Cu^{2+}_{T1}–Cu^{2+}_{T2}, and fluorescence of the label is low due to fluorescence quenching via energy transfer to the oxidized Cu^{2+}_{T1}. Center: upon reduction of Cu_{T1} by an external reductant, the label fluorescence increases, due to the disappearance of energy transfer from the label to the reduced Cu^{1+}_{T1} (gray). Right: An electron is transferred to from Cu_{T1} to Cu_{T2}, which becomes reduced (gray); the label fluorescence is again quenched by the oxidized Cu^{2+}_{T1}. (b) Fluorescence intensity trajectory of a single labeled NiR enzyme undergoing catalysis ('turnover'), in comparison with that in the absence of catalysis ('bg'). The corresponding fluorescence intensity histogram is shown on the right. (c) Fluorescence intensity autocorrelation functions of individual NiR enzymes at 5 μmol l^{-1} (black), 50 μmol l^{-1} (green), 500 μmol l^{-1} (blue), and 5 mmol l^{-1} (red) $NaNO_2$. Normalized data (circles) were fitted with stretched exponential functions (lines), $G(t) = \exp(-(t/\tau 0)\beta)$. Inset: same curves plotted as log(−log(G(t))) versus log(*t*). Reproduced with permission from S. Kuznetsova; G. Zauner; T. Aartsma; H. Engelkamp; N. Hatzakis; A. E. Rowan; R. J. M. Nolte; P. C. M. Christianen; G. W. Canters, *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 3250–3255; Copyright (2008) National Academy of Sciences, U.S.A.

9.21.6.3 Features and Generality

The fluorescence quenching via energy transfer approach could be widely applicable for studying many other enzymes. In principle, any enzyme that has active sites with intense absorption properties can be targeted using this approach. In particular, for metalloenzymes, which use transition metals at the active site (e.g., NiR), the metal-based catalysis often involves species that have intense ligand-to-metal charge transfer absorptions.[93] These strong chromophoric species can be exploited as quenching centers for single-molecule fluorescence detection.

The use of external fluorescent probes is also general. Site-specific labeling of proteins is readily achievable with many accessible labeling schemes including site-directed mutagenesis, green fluorescent protein (GFP) fusion, and unnatural amino acids.[84,94] Many fluorescent probes suitable for single-molecule detections are also available covering a wide spectral range.[84] Besides targeting the enzyme active site as a quencher, such as in the study of NiR[92] and *p*-hydroxybenzoate hydroxylase,[91] chromophoric substrates can also act as a quencher for this scheme, as demonstrated in the study of dihydrofolate reductase.[59,60]

9.21.6.4 Challenges

Common to most fluorescence-based single-molecule methods, photobleaching limits the observation time window using the fluorescence quenching via energy transfer approach. In addition, this approach only obtains the fluorescence intensity from one probe; thus, fluorescence intensity fluctuations due to probe photophysics, such as fluorescence blinking, can complicate the results and data analyses. Triplet quenchers such as Trolox®[95] can reduce fluorescence blinking. Careful control experiments are in any case necessary.

9.21.7 Fluorescence Quenching via Electron Transfer

9.21.7.1 Principle

A molecule at excited states often has different redox potentials from that at its ground state, and photo-induced reduction (or oxidation) of a molecule via electron transfer with another reductant (or oxidant) frequently occurs. For a fluorescent molecule, photo-induced electron transfer at an excited state can significantly shorten its fluorescence lifetime and quench its fluorescence intensity. This fluorescence quenching via electron transfer can be utilized to study enzymes on a single-molecule basis. Xie and coworkers[96,97] have used this single-molecule approach to probe the conformational dynamics of a flavin reductase (Fre),[96] as well as the conformation dynamics of an antibody–antigen complex.[97] Sauer and coworkers also have incorporated this strategy into molecular beacons to detect nucleic acids.[98,99] Here, we use the Fre study to exemplify the approach.

9.21.7.2 Example: Conformational Dynamics of Flavin Reductase

Fre binds a naturally fluorescent FAD cofactor tightly (**Figure 7(a)**). Upon photoexcitation, the protein-bound FAD becomes reduced by a nearby tyrosine residue (Tyr35) forming a transient charge-transferred state (**Figure 7(b)**).[96] The separated charges then quickly recombine, and the Fre–FAD complex returns to the ground state. The photo-induced electron transfer from Tyr35 to FAD significantly shortens the fluorescence lifetime, γ^{-1}, of FAD ($\gamma^{-1} = 1/(k_r + k_{nr} + k_{ET}) \approx 1/k_{ET}$, when $k_{ET} >> k_r$ and k_{nr}; k_r, the radiative decay rate; k_{nr}, the nonradiative decay rate; k_{ET}, the electron transfer rate), which leads to quenching of the FAD fluorescence. The electron transfer rate, k_{ET}, is exponentially dependent on and thus highly sensitive to the distance, r, between the electron donor and the acceptor ($k_{ET} = k_0 \exp(-\beta r)$, k_0 is a constant, $\beta \approx 1.4\,\text{Å}^{-1}$ for electron transfer in proteins[100]). Fluctuations of r due to conformational dynamics of the protein will thus cause fluctuations of k_{ET} and of γ^{-1}. Consequently, monitoring the time-dependent changes of γ^{-1} of the FAD cofactor can probe the FAD–Tyr35 distance fluctuations and the conformation dynamics of Fre on a single-molecule basis.

Xie and coworkers measured the real-time fluctuations of γ^{-1} of single Fre–FAD complexes using a time-stamped photon-by-photon detection technique, which registers the arrival time of each emitted photon and the delay time between each detected fluorescence photon and the corresponding excitation laser pulse.[96] They found that γ^{-1} of FAD in a Fre–FAD complex fluctuates over time (**Figure 7(c)**). Autocorrelation analysis indicates that the fluctuation of the fluorescence lifetime of a single Fre–FAD complex occurs at a broad range of timescales, from hundreds of microseconds to tens of seconds (**Figure 7(d)**). The broad time range of fluctuation suggests the existence of multiple interconverting conformers of a Fre–FAD complex on a rugged energy landscape, and the interconversions can be described by an anomalous diffusion model (**Figure 7(d)**). The existence of these interconverting conformers also relates to the fluctuating catalytic reactivity of the flavin enzyme COx, discussed earlier in this review.[20]

9.21.7.3 Features and Generality

Because the electron transfer rate decays rapidly over a few angstroms, the fluorescence quenching via electron transfer approach is sensitive to subtle distance changes on the angstrom scale. This distance range is complementary to that of the widely used FRET-based techniques, such as FRET and fluorescence quenching

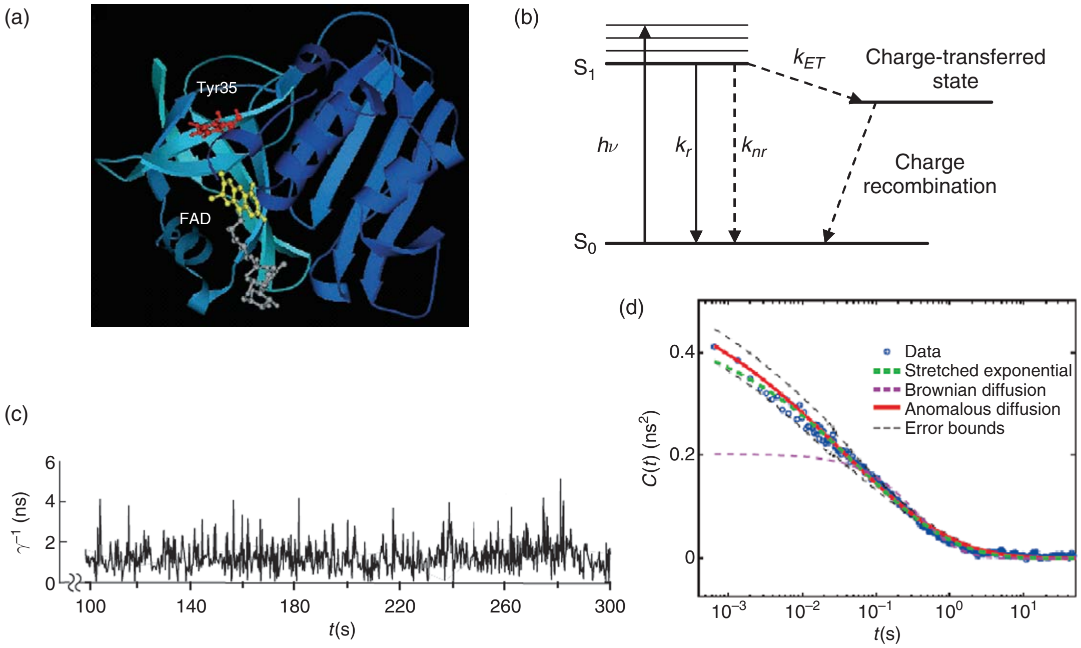


Figure 7 (a) Structure of the Fre–FAD complex. The FAD and tyrosine-35 are highlighted. (b) Energy diagram and transition schemes for the fluorescence quenching via photo-induced electron transfer process. k_r: radiative decay rate; k_{nr}, all other nonradiative decay rate; k_{ET}, electron transfer rate. (c) Trajectory of the fluorescence lifetime of a Fre–FAD complex, indicating the fluctuations of the fluorescence lifetime. (d) Autocorrelation function of the fluorescence lifetime fluctuations of a single Fre–FAD complex and fits with different models. Figures (b–d) reproduced from H. Yang; G. Luo; P. Karnchanaphanurach; T.-M. Louie; I. Rech; S. Cova; L. Xun; X. S. Xie, *Science* **2003**, *302*, 262–266. Reprinted with permission from AAAS.

via energy transfer discussed in Sections 9.21.5 and 9.21.6, which are effective in studying nanometer-scale distance changes. Many structural changes in biology, for example, the geometry reorganizations of the metal active sites in redox-active enzymes, are in the angstrom range.[93] The fluorescence quenching via electron transfer approach offers an opportunity to study these types of structural dynamics at the single-molecule level.

The fluorescence quenching via electron transfer needs two redox-active centers: one as a quencher and the other as a probe that is also fluorescent. For the quencher, there are organic redox-active groups (e.g., tyrosine, tryptophan, and guanine) as well as metal active sites in biology that have rich redox properties and cover a broad range of redox potentials.[93] For the probe, in addition to the naturally fluorescent cofactors, such as FAD and flavin mononucleotide (FMN), one could label proteins externally with fluorescent probes that can undergo photo-induced electron transfer. With careful experimental design, the fluorescence quenching via electron transfer approach could have applications in many systems.

9.21.7.4 Challenges

The fluorescence quenching via electron transfer approach is also limited by photobleaching and complicated by the intrinsic photophysics of the fluorescent probe, as mentioned previously. Competition with fluorescence quenching via energy transfer can also be a problem. For example, the oxidized form of flavin, the oxidized blue copper centers, and cytochromes, all have strong absorption features and can act as a quencher via energy transfer. Selecting fluorescent probes that have fluorescence emission well separated spectrally from the absorption spectra of the quencher can circumvent this problem. Because the photo-induced electron transfer quenches the probe's fluorescence, the low detectable photon number and the shortened fluorescence lifetime of the probe are also challenges. High quantum yield (>30%) and fast response (<100 ps) detectors are needed to detect the weak fluorescence and resolve the short photon delay time in the photon-by-photon approach.[96] To use an externally introduced fluorescent probe to measure angstrom-scale distance changes, the attachment of the probe is critical. The probe needs to be anchored to the protein rigidly so its conformational flexibility around the attachment point will not overwhelm the measurements of the angstrom-scale protein conformational dynamics. Although there are many challenges in using the fluorescence quenching via electron transfer approach, we think that the rich redox properties in redox-active enzymes make it attractive to explore this approach for more applications and discoveries.

9.21.8 Summary

With the continuing development and application of single-molecule methods, more single-molecule enzyme studies are to follow. Here, we have reviewed single-molecule fluorescence methods in studying enzymes, focusing on the principles of their approaches with discussions on their features and challenges. The scope and views here are certainly limited, and many more innovative approaches and applications are expected to emerge in the coming years, for example, using semiconductor quantum dots as fluorescent probes[101,102] and monitoring enzymatic processes in live cells.[26–28] Moreover, the single-molecule studies of catalysis need not be limited to enzymes; catalyses of solid microcrystals by De Vos and coworkers[103] and of nanoparticles[104–107] and carbon nanotubes[108] by our group are new evolving directions.

Acknowledgments

We thank Cornell University, Camille and Henry Dreyfus Foundation, National Science Foundation, Wilson Disease Association, American Chemical Society Petroleum Research Foundation, Cornel Center for Materials Research, and National Institute of Health for funding our research.

Abbreviations

FRET	fluorescence resonance energy transfer
PDMS	polydimethylsiloxane
TIR	total internal reflection

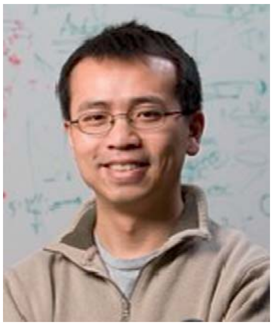
References

1. W. E. Moerner, *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 12596–12602.
2. A. Ishijima; T. Yanagida, *Trends Biochem. Sci.* **2001**, *26*, 438–444.
3. J. N. Forkey; M. E. Quinlan; Y. E. Goldman, *Prog. Biophys. Mol. Biol.* **2000**, *74*, 1–35.
4. E. Toprak; P. R. Selvin, *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 349–369.
5. W. J. Greenleaf; M. T. Woodside; S. M. Block, *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 171–190.
6. H. Park; E. Toprak; P. R. Selvin; *Q. Rev. Biophys.* **2007**, *40*, 87–111.
7. T. Ha, *Biochemistry* **2004**, *43*, 4055–4063.
8. I. Amitani; R. J. Baskin; S. C. Kowalczykowski, *Mol. Cell* **2006**, *23*, 143–148.
9. T. T. Perkins; R. V. Dalal; P. G. Mitsis; S. M. Block, *Science* **2003**, *301*, 1914–1918.
10. X. Zhuang, *Annu. Rev. Biophys. Biomol. Struct.* **2005**, *34*, 399–414.
11. J. Liphardt; B. Onoa; S. B. Smith; I. Tinoco, Jr; C. Bustamante, *Science* **2001**, *292*, 733–737.
12. D. Rueda; G. Bokinsky; M. M. Rhodes; M. J. Rust; X. Zhuang; N. G. Walter, *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 10066–10071.
13. N. G. Walter, *Mol. Cell* **2007**, *28*, 923–929.
14. M. A. Ditzler; E. A. Alemán; D. Rueda; N. G. Walter, *Biopolymers* **2007**, *87*, 302–316.
15. E. Rhoades; E. Gussakovsky; G. Haran, *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 3197–3202.
16. A. A. Deniz; T. A. Laurence; G. S. Beligere; M. Dahan; A. B. Martin; D. S. Chemla; P. E. Dawson; P. G. Schultz; S. Weiss, *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 5179–5184.
17. E. A. Lipman; B. Schuler; O. Bakajin; W. A. Eaton, *Science* **2003**, *301*, 1233–1235.
18. B. Schuler; W. A. Eaton, *Curr. Opin. Struct. Biol.* **2008**, *18*, 16–26.
19. J. J. Benitez; A. M. Keller; P. Ochieng; L. A. Yatsunyk; D. L. Huffman; A. C. Rosenzweig; P. Chen, *J. Am. Chem. Soc.* **2008**, *130*, 2446–2447.

20. H. P. Lu; L. Y. Xun; X. S. Xie, *Science* **1998**, *282*, 1877–1882.
21. B. P. English; W. Min; A. M. van Oijen; K. T. Lee; G. Luo; Y. Sun; B. J. Cherayil; S. C. Kou; X. S. Xie, *Nat. Chem. Biol.* **2006**, *3*, 87–94.
22. L. Edman; R. Rigler, *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 8266–8271.
23. K. Velonia; O. Flomenbom; D. Loos; S. Masuo; M. Cotlet; Y. Engelborghs; J. Hofkens; A. E. Rowan; J. Klafter; R. J. M. Nolte; F. C. de Schryver, *Angew. Chem. Int. Ed.* **2005**, *44*, 560–564.
24. J. R. Brender; J. Dertouzos; D. P. Ballou; V. Massey; B. A. Palfey; B. Entsch; D. G. Steel; A. Gafni, *J. Am. Chem. Soc.* **2005**, *127*, 18171–18178.
25. N. M. Antikainen; R. D. Smiley; S. J. Benkovic; G. G. Hammes, *Biochemistry* **2005**, *44*, 16835–16843.
26. J. Yu; J. Xiao; X. Ren; K. Lao; X. S. Xie, *Science* **2006**, *311*, 1600–1603.
27. L. Cai; N. Friedman; X. S. Xie, *Nature* **2006**, *440*, 358–362.
28. X. S. Xie; P. J. Choi; G.-W. Li; N. K. Lee; G. Lia, *Annu. Rev. Biophys.* **2008**, *37*, 417–444.
29. W. E. Moerner; D. P. Fromm, *Rev. Sci. Instrum.* **2003**, *74*, 3597–3619.
30. P. V. Cornish; T. Ha, *ACS Chem. Biol.* **2007**, *2*, 53–61.
31. X. Michalet; S. Weiss; M. Jaeger, *Chem. Rev.* **2006**, *106*, 1785–1813.
32. R. Roy; S. Hohng; T. Ha; *Nat. Methods* **2008**, *5*, 507–516.
33. X. Nan; P. A. Sims; P. Chen; X. S. Xie, *J. Phys. Chem. B* **2005**, *109*, 24220–24224.
34. P. R. Selvin; T. Ha; Eds, *Single Molecule Techniques: A Laboratory Manual*; Harbor Laboratory Press: Cold Spring, 2008.
35. X. S. Xie; J. K. Trautman, *Annu. Rev. Phys. Chem.* **1998**, *49*, 441–480.
36. G. Bokinsky; X. Zhuang, *Acc. Chem. Res.* **2005**, *38*, 566–573.
37. P. F. Barbara; A. J. Gesquiere; S.-J. Park; Y. J. Lee, *Acc. Chem. Res.* **2005**, *38*, 602–610.
38. T. Basche; W. E. Moerner; M. Orrit; U. P. Wild, *Single-Molecule Optical Detection, Imaging and Spectroscopy*; VCH Verlagsgesellschaft mbH: Weinheim, 1997.
39. C. Bustamante; J. C. Macosko; G. J. L. Wuite, *Nat. Rev. Mol. Cell Biol.* **2000**, *1*, 130–136.
40. A. M. van Oijen, *Biopolymers* **2007**, *85*, 144–153.
41. G. Charvin; T. R. Strick; D. Bensimon; V. Croquette, *Annu. Rev. Biophys. Biomol. Struct.* **2005**, *34*, 201–219.
42. S. A. Rosenberg; M. E. Quinlan; J. N. Forkey; Y. E. Goldman, *Acc. Chem. Res.* **2005**, *38*, 583–593.
43. P. Tinnefeld; M. Sauer, *Angew. Chem. Int. Ed.* **2005**, *44*, 2642–2671.
44. C. Dekker, *Nat. Nanotechnol.* **2007**, *2*, 209–215.
45. B. Sakmann; E. Neher; Eds., *Single-Channel Recording*; 2nd ed.; Plenum Press: New York 1995.
46. C. Joo; H. Balci; Y. Ishitsuka; C. Buranachai; T. Ha, *Annu. Rev. Biochem.* **2008**, *77*, 51–76.
47. X. S. Xie, *Single Mol.* **2001**, *2*, 229–236.
48. R. D. Smiley; G. G. Hammes, *Chem. Rev.* **2006**, *106*, 3080–3094.
49. C. R. Bagshaw; P. B. Conibear, *Single Mol.* **2000**, *4*, 271–277.
50. P. Chen; N. M. Andoy, *Inorg. Chim. Acta* **2008**, *361*, 809–819.
51. K. C. Neuman; A. Nagy, *Nat. Methods* **2008**, *5*, 491–505.
52. N. G. Walter; C. Y. Huang; A. J. Manzo; M. A. Sobhy, *Nat. Methods* **2008**, *5*, 475–489.
53. M. Vrljic; S. Nishimura; W. Moerner, *Methods Mol. Biol.* **2007**, *398*, 193–219.
54. L. Edman; Z. FiSldes-Papp; S. Wennmalm; R. Rigler, *Chem. Phys.* **1999**, *247*, 11–22.
55. O. Flomenbom; K. Velonia; D. Loos; S. Masuo; M. Cotlet; Y. Engelborghs; J. Hofkens; A. E. Rowan; R. J. M. Nolte; M. van der Auweraer; F. C. de Schryver; J. Klafter, *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 2368–2372.
56. M. Paige; D. P. Fromm; W. E. Moerner, *Proc. Soc. Photo-Opt. Instrum. Engr.* **2002**, *4634*, 92–103.
57. T. Ha; A. Y. Ting; J. Liang; W. B. Caldwell; A. A. Deniz; D. S. Chemla; P. G. Schultz; S. Weiss, *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 893–898.
58. Y. Chen; D. Hu; E. R. Vorpagel; H. P. Lu, *J. Phys. Chem. B.* **2003**, *107*, 7947–7956.
59. P. T. R. Rajagopalan; Z. Zhang; L. McCourt; M. Dwyer; S. J. Benkovic; G. G. Hammes, *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 13481–13486.
60. Z. Zhang; P. T. R. Rajagopalan; T. Selzer; S. J. Benkovic; G. G. Hammes, *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 2764–2769.
61. T. Funatsu; Y. Harada; M. Tokunaga; K. Saito; T. Yanagida, *Nature* **1995**, *374*, 555–559.
62. H. Engelkamp; N. S. Hatzakis; J. Hofkens; F. C. de Schryver; R. J. M. Nolte; A. E. Rowan, *Chem. Commun.* **2005**, 935–940.
63. J. Shi; J. Dertouzos; A. Gafni; D. G. Steel; B. A. Palfey, *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 5775–5780.
64. J. Shi; B. A. Palfey; J. Dertouzos; K. F. Jensen; A. Gafni; D. Steel, *J. Am. Chem. Soc.* **2004**, *126*, 6914–6922.
65. G. D. Cremer; M. B. J. Roeffaers; M. Baruah; M. Sliwa; B. F. Sels; J. Hofkens; D. E. D. Vos, *J. Am. Chem. Soc.* **2007**, *129*, 15458–15459.
66. M. Comellas-Aragones; H. Engelkamp; V. I. Claessen; N. A. J. M. Sommerdijk; A. E. Rowan; P. C. M. Christianen; J. C. Maan; B. J. M. Verduin; J. J. L. M. Cornelissen; R. J. M. Nolte, *Nat. Nanotechnol.* **2007**, *2*, 635–639.
67. N. Carette; H. Engelkamp; E. Akpa; S. J. Pierre; N. R. Cameron; P. C. M. Christianen; J. C. Maan; J. C. Thies; R. Weberskirch; A. E. Rowan; R. J. M. Nolte; T. Michon; J. C. M. van Hest, *Nat. Nanotechnol.* **2007**, *2*, 226–229.
68. Q. Xue; E. S. Yeung, *Nature* **1995**, *373*, 681–683.
69. D. B. Craig; E. A. Arriaga; J. C. Y. Wong; H. Lu; N. J. Dovichi, *J. Am. Chem. Soc.* **1996**, *118*, 5245–5253.
70. Y. Rondelez; G. Tresset; K. V. Tabata; H. Arata; H. Fujita; S. Takeuchi; H. Noji, *Nat. Biotechnol.* **2005**, *23*, 361–365.
71. D. M. Rissin; H. H. Gorris; D. R. Walt, *J. Am. Chem. Soc.* **2008**, *130*, 5349–5353.
72. H. H. Gorris; D. M. Rissin; D. R. Walt, *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 17680–17685.
73. A. I. Lee; J. P. Brody, *Biophys. J.* **2005**, *88*, 4303–4311.
74. T.-M. Hsin; E. S. Yeung, *Angew. Chem. Int. Ed.* **2007**, *46*, DOI: 10.1002/anie.200702348.
75. S.-Y. Jung; Y. Liu; C. P. Collier, *Langmuir* **2008**, *24*, 4439–4442.
76. M. Halim; M. S. Tremblay; S. Jockusch; N. J. Turro; D. Sames, *J. Am. Chem. Soc.* **2007**, *129*, 7704–7705.
77. J. Steffen; Q. Zheng; G. S. He; H. E. Pudavar; D. J. Yee; V. Balsanek; M. Halim; D. Sames; P. N. Prasad; N. J. Turro, *J. Phys. Chem. C* **2007**, *111*, 8872–8877.

78. D. J. Yee; V. Balsanek; D. R. Bauman; T. M. Penning; D. Sames, *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 13304–13309.
79. M. K. Froemming; D. Sames, *J. Am. Chem. Soc.* **2007**, *129*, 14518–14522.
80. D. J. Yee; V. Balsanek; D. Sames, *J. Am. Chem. Soc.* **2004**, *126*, 2282–2283.
81. M. Tokunaga; K. Kitamura; K. Saito; H. A. Iwane; T. Yanagida, *Biochem. Biophys. Res. Commun.* **1997**, *235*, 47–53.
82. S. Myong; B. C. Stevens; T. Ha, *Structure* **2006**, *14*, 633–643.
83. B. W. van der Meer; G. Coker, III; S.-Y. S. Chen, *Resonance Energy Transfer: Theory and Data*; VCH Publishers, Inc.: New York, 1994.
84. R. P. Haugland, *The Handbook: A Guide to Fluorescent Probes and Labeling Technologies*; Invitrogen Corp.: California, U.S.A., 2005.
85. T. J. Ha; A. Y. Ting; J. Liang; W. B. Caldwell; A. A. Deniz; D. S. Chemla; P. G. Schultz; S. Weiss, *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 893–898.
86. Z. Zhang; M. M. Spiering; M. A. Trakselis; F. T. Ishmael; J. Xi; S. J. Benkovic; G. G. Hammes, *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 3254–3259.
87. J. Xi; Z. Zhuang; Z. Zhang; T. Selzer; M. M. Spiering; G. G. Hammes; S. J. Benkovic, *Biochemistry* **2005**, *44*, 2305–2318.
88. J. Xi; Z. Zhang; Z. Zhuang; J. Yang; M. M. Spiering; G. G. Hammes; S. J. Benkovic, *Biochemistry* **2005**, *44*, 7747–7756.
89. J. A. Hanson; K. Duderstadt; L. P. Watkins; S. Bhattacharyya; J. Brokaw; J.-W. Chu; H. Yang, *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 18055–18060.
90. T. Ha, *Methods* **2001**, *25*, 78–86.
91. A. H. Westphal; A. Matorin; M. A. Hink; J. W. Borst; W. J. H. van Berkel; A. J. W. G. Visser, *J. Biol. Chem.* **2006**, *281*, 11074–11081.
92. S. Kuznetsova; G. Zauner; T. Aartsma; H. Engelkamp; N. Hatzakis; A. E. Rowan; R. J. M. Nolte; P. C. M. Christianen; G. W. Canters, *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 3250–3255.
93. R. H. Holm; P. Kennepohl; E. I. Solomon, *Chem. Rev.* **1996**, *96*, 2239–2314.
94. B. N. G. Giepmans; S. R. Adams; M. H. Ellisman; R. Y. Tsien, *Science* **2006**, *312*, 217–224.
95. I. Rasnik; S. A. McKinney; T. Ha, *Nat. Methods* **2006**, *3*, 891–893.
96. H. Yang; G. Luo; P. Karnchanaphanurach; T.-M. Louie; I. Rech; S. Cova; L. Xun; X. S. Xie, *Science* **2003**, *302*, 262–266.
97. W. Min; G. Luo; B. J. Cherayil; S. C. Kou; X. S. Xie, *Phys. Rev. Lett.* **2005**, *94,* 198302.
98. J.-P. Knemeyer; N. Marme; M. Sauer, *Anal. Chem.* **2000**, *72*, 3717–3724.
99. O. Piestert; H. Barsch; V. Buschmann; T. Heinlein; J.-P. Knemeyer; K. D. Weston; M. Sauer, *Nano Lett.* **2003**, *3*, 979–982.
100. H. B. Gray; J. R. Winkler, *Annu. Rev. Biochem.* **1996**, *65*, 537–561.
101. A. P. Alivisatos, *Science* **1996**, *271*, 933–937.
102. X. Michalet; F. F. Pinaud; L. A. Bentolila; J. M. Tsay; S. Doose; J. J. Li; G. Sundaresan; A. M. Wu; S. S. Gambhir; S. Weiss, *Science* **2005**, *307*, 538–544.
103. M. B. Roeffaers; B. F. Sels; H. Uji-i; F. C. De Schryver; P. A. Jacobss; D. E. De Vos; J. Hofkens, *Nature* **2006**, *439*, 572–575.
104. W. Xu; J. S. Kong; Y.-T. E. Yeh; P. Chen, *Nature Mater.* **2008**, *7*, 992–996.
105. W. Xu; J. S. Kong; P. Chen, *J. Phys. Chem. C* **2009**, *113*, 2393–2404.
106. W. Xu; J. S. Kong; P. Chen, *Phys. Chem. Chem. Phys.* **2009**, *11*, 2767–2778.
107. P. Chen; W. Xu; X. Zhou; D. Panda; A. Kalininskiy, *Chem. Phys. Lett.* **2009**, *470*, 151–157.
108. W. Xu; H. Shen; Y. J. Kim; X. Zhou; G. Liu; J. Park; P. Chen, *Nano Lett.* **2009**, doi:10.1021/nl900988f.

Biographical Sketches



Peng Chen received his B.S. from Nanjing University, China in 1997. After spending a year at the University of California at San Diego with Professor Yitzhak Tor, he moved to Stanford University and did his Ph.D. with Professor Edward Solomon in bioinorganic and physical inorganic chemistry. In 2004, he joined Professor Sunney Xie's group at Harvard University for postdoctoral research in single-molecule biophysics. He started his assistant professorship at Cornell University in 2005. His current research interests focus on the single-molecule imaging of bioinorganic chemistry and nanoscale catalysis. He has received a Camille and Henry Dreyfus New Faculty award, an NSF Career award, and an Alfred P. Sloan Research Fellowship.

Nesha May Andoy obtained her B.S. in chemistry at the University of the Philippines in 2001. She is currently a graduate student at Cornell University in the Department of Chemistry and Chemical Biology working in Professor Peng Chen's group. Her research covers single-molecule studies of metalloregulator–DNA interactions, bioinorganic enzymology, and nanoscale catalysis.