Epistemology of the Cell

IEEE Press
445 Hoes Lane
Piscataway, NJ 08854

Epistemology of the Cell

A Systems Perspective on Biological Knowledge

Edward R. Dougherty

Michael L. Bittner

IEEE Press Series in Biomedical Engineering
Metin Akay, *Series Editor*

IEEE PRESS

WILEY

A John Wiley & Sons, Inc., Publication

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Dougherty, Edward R.
Epistemology of the cell : a systems perspective on biological knowledge / Edward R. Dougherty and Michael L. Bittner.
p. ; cm. – (IEEE-Press series on biomedical engineering)
Includes bibliographical references.
ISBN 978-1-118-02779-0 (cloth)
1. Biology–Philosophy. 2. Research. 3. Stochastic processes. I. Bittner, Michael L. (Michael Louis) II. Institute of Electrical and Electronics Engineers. III. Title. IV. Series: IEEE Press series in biomedical engineering.
[DNLM: 1. Biology–methods. 2. Research. 3. Biological Phenomena. 4. Knowledge. 5. Models, Theoretical. 6. Stochastic Processes. QH 315]
QH331.D68 2011
570.1–dc22
2011002186

Printed in Singapore

oBook ISBN: 978-1-118-10486-6
ePDF ISBN: 978-1-118-10485-9
ePub ISBN: 978-1-118-10487-3

10 9 8 7 6 5 4 3 2 1

To Professor Michael Kosok
Without those long nights many years ago
he spent with E. R. D. drinking German beer
and discussing Hume, Kant, Hegel,
Russell, Einstein, et al., this book would
never have been written.

Contents

Preface

The driving force behind this book is a comment by Albert Einstein that heads the first chapter: “Science without epistemology is—insofar as it is thinkable at all—primitive and muddled.” Einstein said these words in 1949 on the heels of a revolution in epistemology driven by the monumental advance in physics during the first half of the twentieth century, where the unintelligibility of physics in terms of everyday physical categories had become clear to everyone. It would, however, be a mistake to place the aim of Einstein’s quote on the twentieth century. Indeed, the dénouement of intelligibility began with Galileo and Newton, when knowledge of the physical world was freed from the requirement of a causal description, the Aristotelian ground of science, and became associated with mathematical equations serving as quantitative descriptions of the effects exerted by hypothesized relationships, such as gravitation. The potency of these descriptions derived from the ability to check the accuracy of predictions based upon them through experimentation and the ability to derive new relations from them whose accuracy of prediction could also be experimentally established. The first half of the twentieth century brought the full flowering of the mathematical–experimental duality underlying the epistemology of science along with the full appreciation that the scientific *truth*, or validity, of a mathematical model lay with the concordance of model-based predictions with experimental observations.

Given the advances in experimental technology and computational power, along with the enormous importance of biology to life, it is not surprising that many are predicting that the twenty-first century will belong to biology; however, according to Einstein, this will only happen if the proper epistemological ground is set. It is our contention that this has not been done; on the contrary, there is a lingering, and sometimes strident, call for biology to maintain a pre-Galilean stance. This manifests itself in various ways: a lack of demarcation between biological science, metaphysics, and everyday categories of understanding; a desire for intelligible explanation rather than a strict scientific epistemology based on a mathematical–experimental duality;

"empty talk," to use Einstein's phrase, instead of rigorous mathematical models; little or no attention to proper statistical validation; and a forlorn hope for simple description of extremely difficult and complex phenomena. It is our contention, present throughout the book, that the successful pursuit of biological science requires a proper attention to scientific epistemology.

The epistemology of biology is governed, both experimentally and mathematically, by the nature of the subject matter, that is, the phenomena being studied. Since the cell is the basic unit of life, its epistemology must drive biological epistemology in general, up to and including the organism level. The following 1935 quotation from Conrad Waddington applies to the organism level but it arrives there from its starting point in the cell, the details of which were unknown at the time: "To say that an animal is an organism means in fact two things: firstly, that it is a system made up of separate parts, and secondly, that in order to describe fully how any one part works one has to refer either to the whole system or to the other parts." Biology is not physics and it is not chemistry. The study of cells at the biological level is not about chemical relations; rather, it is first and foremost about the regulatory apparatus that governs the integration of cellular activities in such a way as to form a living system. As Waddington put it, "If there is a 'secret life,' it is here we must look for it."

Taking a historical perspective, Waddington's systems view of biological science was timely because it was in the 1930s when Norbert Wiener and others were formulating the basis of systems theory in the framework of the newly developing theory of random processes. In 1948, Wiener noted "the essential unity of the set of problems centering about communication, control, and statistical mechanics, whether in the machine or in living tissue." This unity results from the demands placed on a self-organizing dynamical system, whether it be natural or man-made.

The fruitful pursuit of biological knowledge requires one to take Einstein's admonition as a practical demand for scientific research, to recognize Waddington's characterization of the subject matter of biology, and to embrace Wiener's conception of the form of biological knowledge in response to its subject matter. It is from this vantage point that we consider the epistemology of the cell.

Edward R. Dougherty
Michael L. Bittner

Acknowledgments

We thank Mohammadmahdi Yousefi for preparing the figures for the book. We offer our appreciation to the people whose work contributed to examples in the book. These include Xiaoning Qian, Jianping Hua, Chao Sima, Ulisses Braga-Neto, David Martins, Blaise Hanczar, Mohammadmahdi Yousefi, Ranadip Pal, Aniruddha Datta, Amin Zollanvari, and Marcel Brun. We acknowledge the many discussions on epistemology with Ulisses Braga-Neto and Ilya Shmulevich that helped to fine-tune our thinking. Most of all, we acknowledge the continuing support of Jeffrey Trent, under whose leadership we have been given the freedom, opportunity, and encouragement to think about how biology fits into the framework of science and how that framework, when properly employed, can lead to the advancement of medicine.

This book incorporates certain work that has appeared in previous journal publications, and we are appreciative of the fact that the publishers have given us permission to use certain passages from those sources. We list them here. (1) Dougherty, E. R., and U. Braga-Neto, "Epistemology of Computational Biology: Mathematical Models and Experimental Prediction as the Basis of Their Validity," *Journal of Biological Systems*, 14(1), 65–90, 2006, with permission of World Scientific Publishing Co. Pte. Ltd.; (2) Dougherty, E. R., "Validation of Inference Procedures for Gene Regulatory Networks," *Current Genomics*, 8(6), 351–359, 2007, with permission of Bentham Science Publishers Ltd.; (3) Dougherty, E. R., "On the Epistemological Crisis in Genomics," *Current Genomics*, 9(2), 69–79, 2008, with permission of Bentham Science Publishers Ltd.; (4) Dougherty, E. R., "Translational Science: Epistemology and the Investigative Process," *Current Genomics*, 10(2), 102–109, 2009, with permission of Bentham Science

Publishers Ltd.; (5) Dougherty, E. R., and M. L. Bittner, “Causality, Randomness, Intelligibility, and the Epistemology of the Cell,” *Current Genomics*, 11(4), 221–237, 2010, with permission of Bentham Science Publishers Ltd.

E. R. D
M. L. B

CHAPTER 1

Science and Knowledge

Science without epistemology is—insofar as it is thinkable at all—primitive and muddled.

—*Albert Einstein*

Science is a product of the human mind—and a very recent product at that. One can reasonably argue that science began in ancient Greece, certainly to the extent that it was given formalization by Aristotle; however, science, in the sense that it has become understood over the last four centuries, was unknown to the ancient civilizations. There is a point in time that will concern us throughout this book. That point is Galileo. There is a pre-Galilean and a post-Galilean mind. The pre-Galilean mind is "everyday" as it looks out upon the physical world. It can be mathematical, as in the case of Archimedes; it can be philosophical, as in the case of Plato; it can be empirical, as in the case of Aristotle. Nonetheless, the categories with which it organizes its perceptions and the manner in which those perceptions arise through sensation take the data of the world as it presents itself to everyday understanding. The mind remains within the bounds of a naïve realism, one that sees a rock as a hard, extended body, or gravity as a cause. Words like "body" and "cause" arise from a noncritical view of knowledge, one that takes the physical world as it appears (the solidity of rock) and identifies physical relations with explanations created to make sense of the world (cause and effect). The great epistemological

Epistemology of the Cell: A Perspective on Biological Knowledge, First Edition.
Edward R. Dougherty, Michael L. Bittner.

achievement of science has been to overcome the lure of appearance and explanation. This is not to say that appearance and explanation do not have their place in human knowledge; it is that they do not have a place in scientific knowledge.

Implicit in the latter statement is that we can characterize a specific kind of knowledge to be called “scientific.” This characterization lies outside of science, from whose perspective it is a priori. This is not to say that it is prior to experience, rather, that it is prior to the organization of experience within scientific categories. Scientific epistemology must address the representation of knowledge and its truth. These are not unrelated and their characterization is at the core of scientific knowledge; indeed, the modern understanding of representation and truth with regard to sensibility differentiates the pre-Galilean and post-Galilean minds. Wilhelm Windelband defines epistemology in the following way: “The problems, finally, which arise from the questions concerning the range and limit of man’s knowing faculty and its relation to the reality to be known form the subject-matter of epistemology or theory of knowledge” (Windelband, 1958). Taking the word “range” to refer to the kind, or nature, of the knowledge under consideration, the nature of scientific knowledge is determined by its manner of representation and its criteria for truth; its limitations are determined by the limits of its form of representation and the degree to which its criteria of truth can be applied, and its relation to reality is determined by the manner in which its representation is connected to physical phenomena and the relation between scientific truth and physical phenomena.

These are tough issues. If they were easy, then the Greco-Roman world would have answered them with its store of philosophical, mathematical, and empirical skills. One is not born with a post-Galilean perspective. It requires wrenching the everyday mind out of its natural condition and into one at odds with the natural one. Ordinary intuition and the beliefs of the tribe must be stripped from the adolescent mind by a rigorous education that tears one loose from the safe moorings of everyday common sense. The desire for full understanding and certainty must be relegated to the immaturity of youth and the mature scientist must live with stringent limitations and radical uncertainty.

Burning questions that are most natural for the scientist to ask, simply because, from the perspective of the human person, science seems to point that way, must be left untouched by science because they transcend the limitations of scientific representation and truth. The following words of Immanuel Kant open the preface of the first edition of the *Critique of Pure Reason*:

> Human reason, in one sphere of its cognition, is called upon to consider questions, which it cannot decline, as they are presented by its own nature, but which it cannot answer, as they transcend every faculty of the mind. It falls into this difficulty without any fault of its own. It begins with principles, which cannot be dispensed with in the field of experience, and the truth and sufficiency of which are, at the same time, insured by experience. With these principles it rises, in obedience to the laws of its own nature, to ever higher and more remote conditions. But it quickly discovers that, in this way, its labors must remain ever incomplete, because new questions never cease to present themselves; and thus it finds itself compelled to have recourse to principles which transcend the region of experience, while they are regarded by common sense without distrust. It thus falls into confusion and contradictions, from which it conjectures the presence of latent errors, which, however, it is unable to discover, because the principles it employs, transcending the limits of experience, cannot be tested by that criterion. The arena of these endless contests is called Metaphysic. (Kant, 1952)

Keep these words in mind as you read this book. These are hard words, perhaps too hard for those who want science to satisfy their spiritual needs. These words are a warning to those who desire to understand "remote conditions" with principles "regarded by common sense without distrust." They tell of a region outside the domain of science, one in which legitimate scientific inquiry cannot venture. Kant aims to demarcate the proper domain of science, one that cannot "transcend the region of experience."

As the scientist, if remaining within the confines of science, must forego the answers to questions toward which science points, the scientist must forego the natural human desire for certainty. This occurs at two levels. First, there is uncertainty in the sense that a scientific law will give exact knowledge of events in the future, given the present state of nature. There are no such laws. As stated by Hans Reichenbach in *The Rise of Scientific Philosophy*,

> Gone is the idea of the scientist who knows the absolute truth. The happenings of Nature are like rolling dice rather than like revolving stars; they are controlled by probability laws, not by causality, and the scientist

> resembles a gambler more than a prophet. He can tell you only his best posits (predictive statements)—he never knows beforehand whether they will come true. (Reichenbach, 1971)

Second, there is uncertainty in the sense that one is assured that a scientific law will stand the test of time and remain unchanged in perpetuity. There is no such assurance. Not only do scientific laws lack certainty in what they say about the future, the laws themselves stand open to refutation with new observations. The Newtonian laws give way to relativity theory and the fundamental dogma of molecular biology—DNA to RNA to protein—gives way to posttranscriptional regulation. Better technology leads to an ever-expanding scope of observations, going ever further outside the range of everyday human sensibility. Among scientific theories there is a kind of survival of the fittest. Philipp Frank states, "Experience is responsible for the natural selection that determines which system is the fittest for survival and which has to be dropped" (Frank, 1961). Karl Popper agrees, "The scientific method's aim is not to save the lives of untenable systems but, on the contrary, to select the one which is by comparison the fittest, by exposing them all to the fiercest struggle for survival" (Popper, 1959).

Science is a paradoxical enterprise. It cannot answer questions whose answers we long for. It cannot tell us with certainty what events will occur no matter how accurate our measurements. It lacks objectivity, in the sense that there is knowledge independent of the mind in which that knowledge is manifested. In these ways, science resembles faith. The following words of Soren Kierkegaard could easily be applied to science by simply changing the word "faith" to the word "science": "If I wish to preserve myself in faith I must constantly be intent upon holding fast the objective uncertainty, so as to remain out upon the deep, over seventy thousand fathoms of water, still preserving my faith" (Kierkegaard, 1941). Of course, the epistemology of the scientist is very different from that of Kierkegaard's knight of faith; indeed, they are so different that no stress arises from being both a scientist and a knight of faith. Where they agree, however, is in acting in the face of uncertainty.

Unless the scientist is going to stand on the sidelines, like Hamlet in a lab coat, he or she must act on posits. The bridge must be built and the patient treated, and both of these actions can best be accomplished as translations of science into action, and translation is performed best when the proper epistemological outlook is maintained. Reichenbach writes his own soliloquy for Hamlet. It concludes with the following words (where one should interpret the word "logic" as "science"): "There I am, the eternal Hamlet. What does it help me to ask the logician, if all he tells me is to make posits? His advice confirms my doubt rather than giving me the courage I need for my action. Logic is not made for me. One has to have ever more courage than Hamlet to be always guided by logic" (Reichenbach, 1971). Perhaps the scientist is not out with Kierkegaard's knight of faith over seventy thousand fathoms—perhaps only over twenty thousand fathoms.

Those with open eyes cannot return to the childhood of science, back before Galileo into a world where science merged with metaphysics and both were presumed to hold the promise of objective truth, independent of the epistemological constraints of the human mind. Today, the limitations and uncertainty of science are more limiting and uncertain than in Kant's day. Nevertheless, science and its manifestations in technology now play a major role in human destiny and a cavalier attitude toward its epistemology is dangerous. The situation is most problematic in biology, which is the science of life. In biology, the answers we desire are among the deepest and actions derived from biology, particularly in medicine, are among the most consequential. In biology, perhaps more than anywhere else in science, one might be tempted to weaken the epistemological demands in order to satisfy a desire for metaphysical knowledge or arrive at a new treatment for cancer.

Many researchers appear to believe that epistemological issues are too arcane and not relevant to their interest in advancing science. On the contrary, epistemology is primary. How can one intentionally and efficiently advance science without knowing the nature of science? Inattention to epistemology results in research that appears scientific but fails to have depth, or even worse, is scientifically unsound. Albert Einstein writes, "The reciprocal relationship of epistemology and science is of a noteworthy kind. They are dependent upon each other. Epistemology without contact with science becomes an empty scheme. Science without epistemology is—insofar as it is thinkable at all—primitive and muddled" (Einstein, 1949). Only through deep reflection on epistemology can one come to grasp what it means to possess scientific knowledge of Nature and therefore be in a position to effectively seek such knowledge. Significant effort must be spent escaping the naïve realism of everyday consciousness so that the deep relationships within Nature shine forth. Einstein penned the following words in a letter:

> I fully agree with you about the significance and educational value of methodology as well as history and philosophy of science. So many people today—and even professional scientists—seem to me like somebody who has seen thousands of trees but has never seen a forest. A knowledge of the historic and philosophical background gives that kind of independence from prejudices of his generation from which most scientists are suffering. This independence created by philosophical insight is – in my opinion – the mark of distinction between a mere artisan or specialist and a real seeker after truth. (Einstein, 1944a)

In accepting the proposition that "science without epistemology is—insofar as it is thinkable at all—primitive and muddled," a biologist must possess both a general perspective on scientific epistemology and an appreciation of the specific attributes of biological knowledge that specialize it within the general framework. As energy and matter lie at the basis of physical knowledge, the cell, as the basic unit of life, lies at the basis of biological knowledge. A properly biological epistemology must recognize that the cell is more than a product of many physical/chemical reactions, that these interactions must be regulated by the cellular components toward specific goal states that may be aimed at producing a cell locked into a particular differentiated state or shifting the cell to another differentiated state. This means viewing the cell as a system that fulfills the functions necessary for survival, such as regulation of protein production, communication among components, information integration, response to external signals, self-organization in response to internal changes or external stimuli, and reproduction.

Biology is the study of organisms, physical systems capable of retaining and utilizing information to execute processes that utilize available energy to organize matter for facilitation of their own persistence and reproduction. Reproduction can involve the passing on of slightly varied copies of the information as well as the combining of information from individuals possessing somewhat different sets of information. This constantly generated variance in organisms over time can produce different levels of fitness of the offspring organisms for particular environments. If the differences are sufficiently large to be a selective advantage, such variances can spread throughout the population of such organisms.

The appearance of a theory of evolution in combination with a growing appreciation of the magnitude of the changes in extant organisms over time, through studies of the fossil record, served to focus attention on biology as a very long-running, continuous process. This in turn raised questions about how the information in organisms is coded and used to produce the range of capabilities within and across organisms. The clear adaptation of organisms to their environments raised questions about how the processes operating in biological systems are controlled to allow highly variable responses appropriate to both rapid and long-term changes in the environment. The persistence of organisms focused attention on how the extraordinarily complex biological processes required for an organism's survival could be made sufficiently robust to account for long life spans.

As in physics, the first types of relationships characterized in biology were ones where the process involved relied on simple, linear relationships. Study of the metabolic products common in organisms was an early and fruitful branch of chemical research and, by the beginning of the 1900s, clear patterns of Mendelian heredity could be seen for diseases such as alcaptonuria, where the enzyme that catabolizes homogentisic acid is inactive and persons with the disease produce black urine, a result of the oxidation of excreted homogentisic acid. The general method of associating mutations of specific enzymes with failures to metabolize a particular substrate, biochemical genetics, was extremely successful in producing a clear understanding of the stepwise enzymatic manipulation of small molecules involved in anabolism, catabolism, and energy production.

The prodigious success of biochemical genetics along with its very intuitive, easily understood methodology has deeply influenced how biologists think about and approach the study of biological processes. Metabolic processing relies on chains of enzymatic transformation of small molecules, where each step in the process is obligatory and each is typically carried out by a single catalytic entity. The processing is extremely efficient, with very little redundancy of activity and only modest branching and merging of the process chains. Much of the regulation of metabolic pathways is carried out at the level of the individual steps through feedback based on product levels, where either the amounts of enzyme made or the activity level of the enzyme are adjusted depending on changes in the concentration of its metabolic product. In analogy to physics, anywhere in biology that simple linear relationships are an appropriate approximation of the key interactions responsible for a particular phenomenon, model building is rapid and produces useful predictions that enable control. Unfortunately, biologists now face the barrier that confronted physics at the end of the nineteenth century: Most processes have sufficiently many conditioning influences that simple linear relationships cannot produce useful predictive models for them.

Human beings seem to possess a deeply rooted desire to characterize Nature in terms of simple relationships whose effects are easily intuited once a descriptive model is constructed. In physics, this desire can be best illustrated in the study of gravity. Newton's model is very straightforward and can be thought of in a way that appeals to our own commonsense description of the world, namely, things fall as though attracted in some way to the Earth. However, by 1900, there was sufficient experimental data to show that predictions made concerning Mercury's orbital precession using this simple distance-attenuated, attractive force model were sufficiently inaccurate to warrant reconsidering the model. Most effort was aimed at trying to find an explanation of the discrepancy that would fit within the Newtonian model. In the end, only a very different model, Einstein's, was able to produce a better fit and make new predictions beyond the reach of the old model, such as gravitational lensing, and thus become the model of choice. In getting to this model, all of the comfortable assumptions about invariance in time and space had to be rejected and, as in quantum-mechanical physics, much of our universe became alien and nonintuitive. Recently, as physicists have been acquiring more and better data on larger scale objects, such as galaxies, discrepancies from the Einstein model have arisen and are provoking the same desires to find an accommodating adjustment or produce a theory with a substantial difference. In the current phase, there are competing hypotheses, such as one that would remove the discrepancies by assuming that 94% of the universe is composed of "dark" matter that can only be sensed indirectly and one that would alter the way in which gravitational force changes over distance (Moffat, 2008). Whatever change eventually occurs will further distance us from our intuitive understandings.

Acceptance of the need for models that will not provide biologists with simple, intuitive, and easily intelligible pictures of what they are studying will no doubt prove as difficult as it has been in physics. Even now, where it seems clear that the models physicists can produce will always be provisional, with new and different forms of data showing that the existing model only provides usefully accurate predictions for some restricted range of situations where further possible complications are not in play, there remains a desire among some for a grand unifying theory that will provide a simple and intuitive model of everything. Like physicists, biologists must confront the issue of intelligibility. In particular, no notion played a more central role in the ancient and medieval concept of the knowledge of Nature than did causality, and that is where we begin.

CHAPTER 2

Causality and the Three Pillars of Aristotelian Science

> The law of causality, I believe, like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm.
>
> —*Bertrand Russell*

Biological science is threatened from two opposite sides, the rational and the empirical. The rationalists are unhappy with the empirical requirements of science and the empiricists are unhappy with the rational requirements of science. Neither are happy with the limitations and uncertainty engendered by the mathematical–experimental duality that underlies modern science. Owing to the long history of entanglement between science and metaphysics, it is not surprising that the rationalist threat has a much longer history and is still vibrant today. The radical empiricist agenda is more recent and has gained enormous momentum in the last half century.

In this chapter, we will begin at the beginning, which for Western philosophy means Plato and a strong rejection of the empirical as a basis for knowledge. The critical statement in this regard is given in

Epistemology of the Cell: A Perspective on Biological Knowledge, First Edition.
Edward R. Dougherty, Michael L. Bittner.

Plato’s allegory of the cave in *The Republic*. The following words of Socrates to Glaucon form the heart of the matter:

> And now, let me show in a figure how far our nature is enlightened or unenlightened:—Behold! human beings living in an underground den, which has a mouth open towards the light and reaching all along the den; here they have been from their childhood, and have their legs and necks chained so that they cannot move, and can only see in front of themselves, being prevented by the chains from turning round their heads. Above and behind them a fire is blazing at a distance, and between the fire and the prisoners there is a raised way; and you will see, if you look, a low wall built along the way, like the screen which marionette players have in front of them, over which they show the puppets . . . And do you see men passing along the wall carrying all sorts of vessels, statues and figures of animals made of wood and stone and various materials, which appear over the wall? Some of them are talking, others silent. . . . [The prisoners] are like ourselves and see only their own shadows, or the shadows of one another, which the fire throws on the opposite wall of the cave? (Plato, 1952)

We are the prisoners who are condemned by the human condition to see only shadows, the ephemeral shadows of sensibility that are thin reflections of a deeper reality, one that is permanent and, unlike the shadow world, not always passing away. True knowledge is knowledge of the *forms* that constitute that deeper reality and these can only be reached by reason. Empirical knowledge is shadow knowledge and leaves us in perpetual darkness. Mathematics, which to the ancient Greek mind meant geometry, is unchanging and independent of the senses. As a mathematical entity, a triangle is a form that has permanence and mathematical knowledge of triangles is true knowledge, whereas any physical instance of a triangle is only a crude shadow of a mathematical triangle and knowledge of physical triangles is a vulgar kind of knowledge. Like mathematics, metaphysical knowledge is not transient and concerns the truly real, not shadows. It is not surprising then that Plato took so little interest in natural science. Nonetheless, his placing the physical far below the metaphysical has had great impact for over 2000 years. The metaphysician is enlightened; the physical scientist is not. In deprecating the natural sciences in favor of metaphysics, Plato has had the effect of encouraging scientists to weaken the scientific enterprise by infusing it with metaphysical speculation.

Aristotle diverged from Plato in the sense that he paid serious attention to the physical world, in particular, biology. He made many observations and made serious efforts to record and explain them. Nevertheless, whereas Plato's view of science had a general negative effect on the progress of science by disparaging the empirical ground of knowledge, Aristotle's negative influence was much more specific and perhaps more detrimental. Of course, the benefit of his emphasis on observation cannot be overstated; however, when it comes to the authenticity of knowledge, he is no less metaphysical than Plato. In Book III of the *Physics*, Aristotle writes, "Knowledge is the object of our inquiry, and men do not think they know a thing till they have grasped the 'why' of it (which is to grasp its primary cause)" (Aristotle, 1952). The shadows in Plato's cave point to a deeper reality beyond the shadows. Insisting upon an answer as to why points to a deeper reality (cause) beyond the phenomena (shadows).

Whereas Plato left the deeper reality to the abstract, mystical world of forms, and therefore had little impact on actual scientific inquiry, Aristotle related the "why" to the phenomena via the concept of causality, thereby having a huge impact on the future development of science. As described by Aristotle, causality has to do with providing categories of explanation. Knowledge is explanation surrounding the question of why and based on four causes, which, according to Aristotle, "perhaps exhausts the number of ways in which the term 'cause' is used."

Let us describe the four causes of Aristotle as defined in the *Physics* (Aristotle, 1952). A material cause is "that out of which a thing comes to be and persists." It is "the bronze of the statue, the silver of the bowl, and the genera of which the bronze and the silver are species." A formal cause is "the form or the archetype, i.e. the statement of the essence, and its genera, . . . and the parts in the definition." An efficient cause is "the primary source of the change or coming to rest; e.g. the man who gave advice is a cause, the father is the cause of the child, and generally what makes of what is made and what causes change of what is changed." A final cause is "the end, or that for the sake of which a thing is done, e. g. health is the cause of walking about. . . . The same is true also of all the intermediate steps which are brought about through the action of something else as means toward the end." The same analysis is provided by Aristotle in the *Metaphysics*.

It is clear what Aristotle means by a material cause but this does not agree with the modern use of the word "cause." It is also clear that he uses the terminology “final cause” to refer to purpose, or design, and this too is at odds with current commonplace usage. The meaning of a formal cause is more obscure, referring to essence, for instance, what makes the statue a statue. Again, this lacks connection with current usage. On the surface, an efficient cause seems more in accord with our ordinary understanding of causality. But what does it mean to be “the primary source of the change or coming to rest?” Perhaps if one thinks of a moving billiard ball hitting another billiard ball at rest, then a casual observer might say in the vernacular that the moving billiard ball is the “cause” of the motion of the previously stationary billiard ball. But this everyday appeal to causality lacks any quantitative description. The latter would involve velocity, impact angle, elasticity, friction, air resistance, and so on. Note that we have avoided trying to define “causality” in its current usage, instead allowing the reader to simply recognize the obvious difference or agreement with Aristotle’s usage. Not only are we not interested in parsing Aristotle’s usage, since our ultimate interest is in modern science, but, as will become apparent, defining causality in any meaningful sense is problematic.

Our concern with the epistemology characterized by Aristotle’s conception of causal knowledge is the orientation toward the science of Nature engendered by it and the resulting impact on the future development of scientific epistemology. In this regard, three points are fundamental to Aristotle’s epistemology:

1. To know is to explain;
2. Explanation must involve a causal relation; and
3. There is no demarcation between physics and metaphysics, so that the same causal categories are stated in both the *Physics* and the *Metaphysics*.

Much of the history of scientific epistemology has been about demolishing these three pillars of Aristotelian epistemology and overcoming their retarding effect on the development of science.

Since clarity regarding the proper domains of science and metaphysics plays a key role throughout this book, having discussed Aristotle’s four causes, let us provide a working definition of metaphysics. When characterizing the philosophical problems concerning “our knowledge of the actual world,” Windelband forms a partition:

> The general questions which concern the actual taken as a whole are distinguished from those which deal with single provisions of the actual. The former, viz. the highest principles for explaining the universe, and the general view of the universe based on these principles, form the problems of *metaphysics* . . . The special provisions of the actual are Nature and History. (Windelband, 1958)

Natural science comes under the province of Nature.

The grand issues that concern explaining the universe as a whole comprise the problems of metaphysics. Metaphysical explanations go beyond explanations of individual conditions (provisions) within the world to a universality encompassing all individual conditions, not simply as a collection of conditions, but integrated within the context of the whole. Metaphysics does not concern this or that scientific principle but rather the deeper reality governing scientific principles in general. For instance, as a metaphysical category, final causality does not refer to a specific purpose but rather to the teleological principle itself, that actions within the world have purpose.

At the beginning of the seventeenth century, Francis Bacon agrees with Aristotle that causality is the ground of knowledge; however, Bacon separates Aristotle's four causes as to whether they apply to physics or metaphysics: material and efficient causes to physics, formal and final causes to metaphysics. But Bacon does not make a demarcation between science and metaphysics. While he sees no place for final causes in science, his preference for authentic scientific understanding lies with formal causes. In the *Novum Organum* he writes,

> It is a correct position that "true knowledge is knowledge by causes." And causes again are not improperly distributed into four kinds: the material, the formal, the efficient, and the final. . . . The efficient and the material (as they are investigated and received, that is, as remote causes, without reference to the latent process leading to the form) are but slight and superficial, and contribute little, if anything, to true and active science. . . . For though in nature nothing really exists besides individual bodies, performing pure individual acts according to a fixed law, yet in philosophy this very law, and the investigation, discovery, and explanation of it, is the foundation as well of knowledge as of operation. And it is this law with its clauses that I mean when I speak of *forms*. . . . Now if a man's knowledge be confined to the efficient and material causes (which are unstable causes, and merely vehicles, or causes which convey the

> form in certain cases) he may arrive at new discoveries in reference to substances in some degree similar to one another, and selected beforehand; but he does not touch the deeper boundaries of things. But whosoever is acquainted with forms embraces the unity of nature in substances the most unlike, and is able therefore to detect and bring to light things never yet done. (Bacon, 1952)

Bacon separates himself from Plato by noting that forms do not give existence and only individual bodies exist in Nature. These bodies act according to a fixed law and "investigation, discovery, and explanation of it" is the foundation of knowledge. This law, which by Bacon is called a "form," is not within Nature; rather, it is metaphysical and governs Nature. It is in the domain of metaphysics where "true and active science" resides. Knowing the material out of which something comes to be or the source of change for a body's change of motion is "superficial" in comparison with knowledge of form. Efficient and material causes do not touch "the deeper boundaries of things."

Bacon distinguishes physics and metaphysics, and science intersects both, with the more important aspect of science, that being formal cause, lying within metaphysics. While the language of Bacon might be muddled, one should not overlook the advance in scientific perspective. Bacon drops final cause and regards efficient and material causes as superficial. Suppose we go a bit further than he and drop all reference to efficient and material causes. Then we are left with only what he calls a formal cause. Let us examine this formal "cause." First, it is not within Nature. Second, it represents "true science." Third, it corresponds to a law governing natural behavior. Fourth, it allows the scientist "to detect and bring to light things never yet done." Thus, if we drop the word "cause," drop the appeal to explanation, and drop the characterization of a natural law as being metaphysical, then it would be seen that Bacon has at least one foot in modernity. We are not saying that Bacon dropped Aristotle's efficient and material cause, nor that he disagreed with Aristotle regarding explanation, nor that by law he meant anything beyond simple cause and effect, nor that he put aside metaphysics, but we are saying that one can see the outlines of modern science forming in his mind.

Bacon desires a method to ascertain scientific knowledge based on experiment, not the abstract reasoning common in the medieval period, and he recognizes that more is required than Aristotle's anecdotal observations. Given that true knowledge rests upon causality, then the form of knowledge and its acquirement should conform to the causal relation. Thus, causality becomes inextricably linked to induction: When we observe that event *B* follows whenever event *A* is observed, then a cause-and-effect relation is in some (unspecified) sense "logically" induced between *A* and *B*. For Bacon, this relation is a formal cause and goes beyond the list of observations to a deeper knowledge of reality. For Bacon, scientific knowledge is causal knowledge and this knowledge is reached by the "logical" process of induction upon observing one event, the effect, repeatedly following the other, the cause, without exception. Think of a billiard ball *A* repeatedly sent into billiard ball *B*. Each time, ball *B* begins to move when hit by ball *A*, the latter being the efficient cause. For Bacon, a deeper relation, one possessing true scientific knowledge, is induced in the relation that any moving body *A* hitting a stationary body *B* will always result in the stationary body moving. This more general relation about bodies in general would constitute a formal cause. It is metaphysical and it is induced from repeated observations.

Bacon recognizes that haphazard observation will not yield the kind of structured observations that lead to the discovery of inductive relationships. Perhaps his salient contribution is recognizing the need for experiments. He writes,

> There remains simple experience which, if taken as it comes, is called accident; if sought for, experiment. But this kind of experience is no better than a broom without its band, as the saying is—a mere groping, as of men in the dark, that feel all round them for the chance of finding their way, when they had much better wait for daylight, or light a candle, and then go. But the true method of experience, on the contrary, first lights the candle, and then by means of the candle shows the way; commencing as it does with experience duly ordered and digested, not bungling or erratic, and from it educing axioms, and from established axioms again new experiments. (Bacon, 1952)

Because causality lies at the basis of knowledge, Bacon formulates experimental design with the hope of revealing sequences of events from which to induce causal relations. The fact that he ties causality to induction means that, in some sense, he has dropped efficient and final causes from his understanding of science and transformed a formal cause into the metaphysical counterpart of a collection of efficient causes.

Modern science, in particular, breaking the dependency of science on causality, arrives with Galileo, in great part because he recognizes that science should concern itself with quantifiable relations among phenomena. Galileo does not deny causality; rather, he sets the issue aside and gets on with pragmatic description. In *Dialogues Concerning Two New Sciences*, Galileo puts these words into the mouth of Salviati:

> The present does not seem to me to be an opportune time to enter into the investigation of the cause of the acceleration of natural motion, concerning which various philosophers have produced various opinions, some of them reducing this to approach to the center; others to the presence of successively less parts of the medium [remaining] to be divided; and others to a certain extrusion by the surrounding medium which, in rejoining itself behind the moveable, goes pressing and continually pushing it out. Such fantasies, and others like them, would have to be examined and resolved, with little gain. For the present, it suffices our Author that we understand him to want us to investigate and demonstrate some attributes of a motion so accelerated (whatever be the cause of its acceleration) that the momenta of its speed go increasing, after its departure from rest, in that simple ratio with which the continuation of time increases, which is the same as to say that in equal times, equal additions of speed are made. (Galileo, 1954)

In the terminology of phenomenology, Galileo *brackets* causality, ignores it, and gets on with the business of obtaining relations between phenomena. There would be “little gain” in examining the kind of “fantasies” put forth by philosophers to explain acceleration in terms of causality. It is more beneficial to “investigate and demonstrate some attributes of motion.” Although Galileo does not deny causality, he rejects it as a requirement for knowledge. Aristotle’s grip is broken.

In general, Galileo is dissatisfied with words. These constitute ersatz knowledge, the result being both an illusion of knowledge and an impediment to actual knowledge owing to satisfaction with empty phrases. In *Dialogue Concerning the Two Chief World Systems*, when the Aristotelian Simplicio comments that everyone knows that bodies fall on account of gravity, Salviati responds,

> You are wrong, Simplicio; you should say that everyone knows that it is called “gravity.” But I am not asking you for the name, but the essence of the thing. Of this you know not a bit more than you know the essence of the mover of the stars in gyration. We don’t really understand what principle or what power it is that moves a stone downwards, any more

> than we understand what moves it upwards after it has left the projector, or what moves the moon round. (Galileo, 2001)

Observation shows that bodies fall, and perhaps something called causality is operating here, but to simply say that there is a cause and to name it provides no knowledge. A name tells us nothing about the object being named or even if such an object exists. Moreover, understanding the power that moves a stone downwards is not a prerequisite for providing a quantitative relation between the stone and the earth. In general, cogitating on words can lead one away from the phenomena rather than toward a characterization of their attributes.

While Galileo may have initiated modern science, the epistemology of scientific knowledge takes shape with Isaac Newton: its mathematical structure, its relational nature, its predictive connection with phenomena, and its idealization, in the sense that relations between phenomena are characterized under the assumption of unrealistic conditions with the recognition that in actuality such conditions will have some effect, for instance, Galileo's assumption of a frictionless plane. Consider gravity. Newton formulates a mathematical law of gravitation that relates the distance, mass, and acceleration. The gravitational law is mathematical, relational, idealized insofar as, when put into practice, it ignores confounding effects such as air resistance, and it can be related to phenomena via experiment. The gravitational law mathematically characterizes a relation in such a way that the relation can be used to make predictions, thereby providing a means for validation and application. The mathematical structure represents a precise, intersubjective, and operational form of knowledge.

The gravitational law contains no reference to some physical process behind the relations, in particular, there is no mention of a cause of acceleration. Regarding causality, Bertrand Russell states,

> In the motions of mutually gravitating bodies, there is nothing that can be called a cause, and nothing that can be called an effect; there is merely a formula. Certain differential equations can be found, which hold at every instant for every particle of the system, and which, given the configuration and velocities at one instant, or the configurations at two instants, render the configuration at any other earlier or later instant theoretically calculable. . . . But there is nothing that could be properly called "cause" and nothing that could be properly called "effect" in such a system. (Russell, 1913)

Like Galileo, Newton is not denying causality; he is bracketing it. Like Galileo, he is breaking with Aristotle and Bacon in formulating knowledge that does not depend on causality.

Near the beginning of *The Principia: Mathematical Principles of Natural Philosophy*, Newton makes his intent clear when he writes, "For I here design only to give a mathematical notion of these forces, without considering their physical causes and seats" (Newton, 1952). The following words, written near the end of *The Principia*, are striking for their time:

> Hitherto I have not been able to discover the cause of those properties of gravity from the phenomena, and I frame no hypothesis; for whatever is not deduced from the phenomena is to be called an hypothesis; and hypotheses, whether metaphysical or physical, whether of occult qualities or mechanical, have no place in experimental philosophy. In this philosophy particular propositions are inferred from the phenomena, and afterward rendered general by deduction. Thus it was the impenetrability, the mobility, and the impulsive forces of bodies, and the laws of motion and of gravitation were discovered. And to us it is enough that gravity does really exist, and acts according to the laws which we have explained, and abundantly serves to account for all the motions of the celestial bodies, and of our sea. (Newton, 1952)

Newton has not discovered the cause from the phenomena and, until this is done, cause has no place in experimental philosophy (science). Later when speaking of gravity, he adds,

> But our purpose is only to trace out the quantity and properties of this force from the phenomena, and to apply what we discover in some simple cases as principles, by which, in a mathematical way, we may estimate the effects thereof in more involved cases: for it would be endless and impossible to bring every particular to direct and immediate observation. We said, in a mathematical way, to avoid all questions about the nature or quality of this force. (Newton, 1952)

The knowledge of which Newton speaks is mathematical, but it is not mathematics devoid of relation to human experience. It is empirically grounded.

From an epistemological perspective, there are two critical points. First, Newton is "to avoid all questions about the nature of gravity." As he said earlier, "it is enough that gravity does really exist." Something exists, but as Galileo had said, we know nothing of its substance.

Second, the mathematical system is not meant to include all factors, but is of sufficient predictive power that it can "estimate" effects in a more general setting. Owing to the predictive nature of the mathematical system, it can be empirically tested independently of the reasoning leading to it.

Galileo and Newton do not deny causality as a category of knowledge, but they widen the scope of knowledge to include mathematical systems that relate phenomena, while bracketing "questions about the nature" of the phenomena. The physical substance behind the mathematical relations is bracketed so that physical knowledge is constituted by mathematical knowledge, with the proviso that the mathematical knowledge be explicitly related to observations. Although both Galileo and Newton held on to the notion of causality, thereby not breaking completely free of Aristotle's influence, they brought about a radical epistemological transformation by describing relations among phenomena with mathematical formulas, absent a causal explanation, the kind of explanation that had ultimately led to "fantasies," to use Galileo's terminology.

When Galileo and Newton bracket causality, they not only begin a search for noncausal knowledge, thereby going beyond Aristotle, they also permit themselves the luxury of not coming to grips with the meaning of causality. In particular, if we focus on Bacon's perspective, which is essentially a metaphysical formalization of Aristotle's efficient cause, then there is a temporal aspect to causality in that the cause occurs prior to the event and this temporality plays a key role in the inductive method, as understood by Bacon. David Hume raises a crucial epistemological question for science and metaphysics: Are a cause and its effect merely related via temporal priority, with the cause prior to the effect, or is there more than temporal contiguity? To wit, is there something that touches "the deeper boundaries of things," as Bacon would have it? Is there a necessary connection between the cause and the effect? Hume argues that in using the phrase "cause and effect," we mean the latter. In *An Enquiry Concerning Human Understanding*, he writes:

> When one particular species of events has always, in all instances, been conjoined with another, we make no longer any scruple of foretelling one upon the appearance of the other, and of employing that reasoning, which alone can assure us of any matter of fact or existence. We then call one object, Cause; and the other, Effect. We suppose that there is some

> connexion between them; some power in the one, by which it infallibly produces the other, and operates with the greatest certainty and strongest necessity. (Hume, 1952)

But do repeated conjoined observations warrant the supposition of a necessary connection? Is there a ground in reason or a physical ground for judging there to be a necessary connection? Hume states emphatically that there is no such ground. Belief in causality rests not on reason, but on habit. In one of the key passages in scientific epistemology, he writes,

> But there is nothing in a number of instances, different from every single instance, which is supposed to be exactly similar; except only, that after a repetition of similar instances, the mind is carried by habit, upon the appearance of one event, to expect its usual attendant, and to believe that it will exist. This connexion, therefore, which we *feel* in the mind, this customary transition of the imagination from one object to its usual attendant, is the sentiment or impression from which we form the idea of power or necessary connexion. Nothing farther is in the case. Contemplate the subject on all sides; you will never find any other origin of that idea. This is the sole difference between one instance, from which we can never receive the idea of connexion, and a number of similar instances, by which it is suggested. The first time a man saw the communication of motion by impulse, as by the shock of two billiard balls, he could not pronounce that the one event was *connected*: but only that it was *conjoined* with the other. After he has observed several instances of this nature, he then pronounces them to be *connected.* What alteration has happened to give rise to this new idea of *connexion*? Nothing but that he now *feels* these events to be *connected* in his imagination, and can readily foretell the existence of one from the appearance of the other. When we say, therefore, that one object is connected with another, we mean only that they have acquired a connexion in our thought. (Hume, 1952)

In *A Treatise of Human Nature*, Hume states,

> [The] supposition that the future resembles the past is not founded on arguments of any kind, but is derived entirely from habit, by which we are determined to expect for the future the same train of objects to which we have been accustomed. . . . All our reasonings concerning causes and effects are derived from nothing but custom and belief is more properly an act of the sensitive than of the cogitative part of our nature. (Hume, 1951)

The sticking point is necessity. In the *Treatise*, Hume writes, "From the mere repetition of any past impression, even to infinity, there never will arise any new original idea, such as that of a necessary connexion; and the number of impressions has in this case no more effect than if we confined ourselves to one only" (Hume, 1951). Repetition may lead to increased expectation, but not necessity—and certainly not to some deeper relationship. Induction does not depend upon causality; in fact, it is the opposite. Belief in causality is itself an unwarranted leap from repeated observations.

If, as Aristotle and Bacon believe, scientific knowledge is knowledge of causes, and if causality rests on habit and custom, then the ground of scientific knowledge is brought into question. If, as Hume argues, the concept of a necessary connection between phenomena is subjective, then does not this entail the subjectivity of scientific knowledge? Hume did not miss this point. Regarding his conclusion that the connection between cause and effect is arrived at by habit and exists only in human thought, in the *Enquiry*, he writes,

> And what stronger instance can be produced of the surprising ignorance and weakness of the understanding than the present? For surely, if there be any relation among objects which it imports to us to know perfectly, it is that of cause and effect. On this are founded all our reasonings concerning matter of fact or existence. By means of it alone we attain any assurance concerning objects which are removed from the present testimony of our memory and senses. The only immediate utility of all sciences is to teach us, how to control and regulate future events by their causes. Our thoughts and enquiries are, therefore, every moment, employed about this relation: Yet so imperfect are the ideas which we form concerning it, that it is impossible to give any just definition of cause, except what is drawn from something extraneous and foreign to it. Similar objects are always conjoined with similar. Of this we have experience. (Hume, 1952)

In these few words, Hume rattles the foundations of scientific knowledge. If all reasoning concerning matter of fact or existence is founded on causality and the utility of all sciences is to control Nature through the regulation of events via their causes, and if causality is simply a product of habit, then scientific understanding rests on habit, or custom, not on objective physical relations, in which case it is indeed very weak.

All reasoning concerning matter of fact is not founded on causality and Hume should have been aware of this. While he may have shown there to be nothing of consequence in the brackets that Galileo and Newton put aside, his skeptical assault does nothing to undercut the mathematical–experimental structure of modern science as conceived by its founders. Their scientific theories do not rest upon causality. Nevertheless, in showing that the brackets contain a ghost—at least insofar as causality represents some intrinsic physical reality apprehensible to the understanding—Hume deals a severe blow to our conception of our place in the universe. Einstein writes, "Man has an intense desire for assured knowledge. That is why Hume's clear message seems crushing: the sensory raw material, the only source of our knowledge, through habit may lead us to belief and expectation but not to the knowledge and still less to the understanding of lawful relations" (Einstein, 1944b). Hume forever buried the Aristotelian concept of science. Moreover, he fundamentally went beyond Galileo and Newton. The latter knew that the mathematical theories of science are idealized and can only be used to "estimate" behavior. When Hume wrote, "the mind is carried by habit, upon the appearance of one event, to expect its usual attendant," he made the monumental shift from causality to expectation, thereby recognizing that scientific statements are inherently probabilistic; indeed, in the *Treatise*, the section dealing with the fundamental issues surrounded causality is entitled, "Of the Probability of Causes."

Modernity fully arrives with Hume (and not just in science). He does not bracket causality as a scientific category; he dismisses it as a scientific category by showing that it has no grounding in reason or in Nature, at least insofar as is empirically discernable. Necessary connections are subjective impressions, not objective relations. Observations lead to expectation, a probabilistic category, not to certainty. Scientific certitude is a fiction, a product of a leap of thought. Hume wrote the *Treatise* in 1739. Two centuries later, Erwin Schrödinger writes, "It can never be decided experimentally whether causality in Nature is 'true' or 'untrue.' The relation of cause and effect, as Hume pointed out long ago, is not something that we find in Nature but is rather a characteristic of the way in which we regard Nature" (Schrödinger, 1957).

Pierre-Simon Laplace, who is among the founders of probability theory and a great physicist, recognizes the uncertainty in making predictions but attributes this uncertainty to ignorance. In *A Philosophical Essay on Probabilities*, he writes,

> The curve described by a single molecule in air or vapor is regulated in a manner just as certain as the planetary orbits; the only difference between them is that which comes from our ignorance. Probability is relative, in part to this ignorance, in part to our knowledge. We know that of three or greater number of events a single one ought to occur; but nothing induces us to believe that one of them will occur rather than the others. In this state of indecision it is impossible for us to announce their occurrence with certainty. (Laplace, 1953)

While he recognizes the need for a probabilistic approach to Nature and is writing more than a half century after Hume's *Treatise*, Laplace holds on to causality as existing in Nature. In the following famous passage, he advocates a complete determinism:

> We ought then to regard the present state of the universe as the effect of its anterior state and the cause of the one which is to follow. Given for one instant an intelligence which could comprehend all the forces by which nature is animated and the respective situation of the beings who compose it – an intelligence sufficiently vast to submit this data to analysis – it would embrace in the same formula the movements of the greatest bodies of the universe and those of the lightest atom; for it, nothing would be uncertain and the future, as the past, would be present in its eyes. (Laplace, 1953)

By conditioning full deterministic knowledge on a "sufficiently vast" intelligence, Laplace does not claim that human beings can achieve a completely deterministic theory of Nature; nevertheless, he postulates determinism in Nature based on causality in Nature. This causality is not simply temporal contiguity. His words, "the present state of the universe as the effect of its anterior state and the cause of the one which is to follow," clearly suggest that there is more to cause and effect than anterior to posterior. The advent of probability theory does not bring about the demise of causality in science; rather, it is Hume's recognition that belief in causality derives from custom and "is more properly an act of the sensitive than of the cogitative part of our nature."

Laplace prefaces his determinism with causality but this need not have been the case. He could have hypothesized a superior intelligence that knew all the laws of mathematical physics and could at one instant make all the necessary measurements, and that Nature is completely described by these laws. Still, this hypothesis is not rooted in human knowledge of Nature. Indeed, it is a vacuous hypothesis from which virtually any desirable conclusion could follow.

Immanuel Kant agrees with Hume that the principle of causality is not a product of reason. In the *Prolegomena to Any Future Metaphysics*, he writes, "[Hume] maintained justly that we can in nowise discern through reason the possibility of causation, namely, the reference of the existence of one thing to the existence of another, which is necessitated by the former" (Kant, 1891). However, whereas for Hume, habit underlies belief in causality, for Kant, causality is a category of understanding. It is a form imposed on phenomena by the nature of the human mind. The mind imposes forms on the data of sensation, and scientific knowledge is limited by these forms. The way things appear, such as being spatially coordinated and connected by causality, are due to subjective a priori conditions for knowledge. One cannot know things apart from the manner in which they conform to these a priori mental forms. Of these categories of understanding, of which causality is one, Kant writes in the *Critique of Pure Reason*:

> Now the question is whether there do not exist, *a priori* in the mind, conceptions of understanding also, as conditions under which alone something, if not intuited, is yet thought as object. If this question be answered in the affirmative, it follows that all empirical cognition of objects is necessarily conformable to such conceptions, since, if they are not presupposed, it is impossible that anything can be an object of experience. Now all experience contains, besides the intuition of the senses through which an object is given, a conception also of an object that is given in intuition. Accordingly, conceptions of objects in general must lie as *a priori* conditions at the foundation of all empirical cognition; and consequently, the objective validity of the categories, as *a priori* conceptions, will rest upon this, that experience (as far as regards the form of thought) is possible only by their means. For in that case they apply necessarily and *a priori* to objects of experience, because only through them can an object of experience be thought. (Kant, 1855)

The last line of this quotation is the crux: Only through the categories can an object of experience be thought. The mind, in its very structure, imposes causality on our experiences as a prior condition for thinking about the experiences.

Kant's argument imposes causality upon the phenomena we experience but not on the *things-in-themselves* that underlie the phenomena, the *noumena*, as he calls them. We cannot experience the things-in-themselves because they lie outside our sense experience. Kant asserts the existence of things-in-themselves, which for a strict empiricist like

Hume cannot be asserted. Kant does not ascribe causality to the things-in-themselves, only to the phenomena we experience, and that because our minds impose causality on the phenomena as a condition of thinking about them. Whereas Galileo and Newton bracket causality, Kant moves it from Nature to the mind.

Relative to Hume, Kant takes three steps forward and one step back. As for the latter, he returns to causality by making causality, not expectation, a category of understanding. Among Kant's categories, causality is a category of relation, in this case, between cause and effect. Surely the mind relates events. But if there is contiguity between a prior event *A* and a posterior event *B*, then why insist that the mind imposes the category of causality as the relation between them? If causality is more than mere temporal contiguity, then the category seems to say that the mind imposes the belief that there is some occult connection, precisely the notion that Newton bracketed and Hume rejects as having no logical or empirical foundation. Hume has already seen that the functional category of understanding is expectation. Observation of event *A* leads one to expect event *B*. Hume sees correctly that expectation is a probabilistic concept and considers it subjective. Laplace thinks expectation is a consequence of ignorance. Kant, Galileo, Newton, and Laplace, all among the greatest geniuses in human history, remain tied to the Aristotelian epistemology. There is simply no empirical or logical reason to raise the idea of causality. If experience shows that event *A* tends to precede event *B*, or even if in our experience event *A* has always preceded event *B*, then why go beyond saying that upon observation of event *A* we expect to observe event *B*? Hume recognizes that there is no empirical or logical reason for introducing a category beyond expectation. What he fails to see, and what would await the twentieth century to be understood, is the manner in which expectation would be incorporated into a rigorous mathematical theory of probability and how scientific knowledge would be constituted in a probabilistic framework.

Kant moves the situation forward in three regards. First, he insists that mind imposes human categories on the way in which Nature is humanly understood. He agrees with Hume that causality cannot be grounded in Nature, but argues that it is more than habit because, in conjunction with other categories of understanding, it is imposed upon experience. He writes, "I proceeded to the deduction of these conceptions, which I was now assured could not, as Hume had pretended, be derived from experience but must have originated in the pure understanding" (Kant, 1891). One need not agree with Kant that the categories lie in the domain of metaphysics, in the sense that they "determine the whole range of the pure Reason, in its limits as well as in its content, completely according to universal principles." Yet, the point remains that human experience does not arrive *qua* experience; rather, as human experience it arrives via the structure of the human mind. The mind imposes connectivity upon events. For Hume, there is no connecting mind. For him, experience is a succession of atomistic sense impressions disconnected from each other. We ourselves do not escape his critique. The mind is nothing but a bundle of perceptions. In the *Treatise*, Hume writes,

> The mind is a kind of theatre, where several perceptions successively make their appearance; pass, repass, glide away, and mingle in an infinite variety of postures and situations. There is properly no *simplicity* in it at one time, nor *identity* in different [times], whatever natural propension we may have to imagine that simplicity and identity. The comparison of the theatre must not mislead us. They are the successive perceptions only that constitute the mind. (Hume, 1951)

For Hume, there is no mind to organize successive perceptions into a coherent whole because the perceptions, themselves, "constitute the mind." Kant puts mind, as an organizing and connecting entity, prior to experience.

Einstein recognizes Kant's error in postulating Euclidean geometry and causality as a priori categories of understanding and he does not believe that specific categories are intrinsic; nevertheless, he contends that thinking requires the use of concepts not dependent on sensory experience. He writes,

> Then Kant took the stage with an idea which, though certainly untenable in the form in which he put it, signified a step towards the solution of Hume's dilemma: whatever in knowledge is of empirical origin is never certain (Hume). If, therefore, we have definitely assured knowledge, it must be grounded in reason itself. This is held to be the case, for example, in the propositions of geometry and in the principle of causality. These and certain other types of knowledge are, so to speak, a part of the implements of thinking and therefore do not previously have to be gained from sense data (i. e., they are *a priori* knowledge). Today everyone knows, of course, that the mentioned concepts contain nothing of the certainty, of the inherent necessity, which Kant had attributed to them. The following,

> however, appears to me to be correct in Kant's statement of the problem: in thinking we use, with a certain "right," concepts to which there is no access from the materials of sensory experience, if the situation is viewed from the logical point of view. (Einstein, 1944b)
>
> . . . The theoretical attitude here advocated is distinct from that of Kant only by the fact that we do not conceive of the "categories" as unalterable (conditioned by the nature of the understanding) but as (in the logical sense) free conventions. They appear to be *a priori* only insofar as thinking without the positing of categories and of concepts in general would be as impossible as is breathing in a vacuum. (Einstein, 1949)

Einstein does not assume that the categories appear as intrinsic to the understanding but that they are "free conventions"; nevertheless, he does assert the necessity of categories for thinking. Whereas for Hume the mind is nothing but a succession of perceptions, for Einstein there is a mind that thinks in the framework of categories.

In arguing that, even if causality is the underlying principle upon which science is based, its application lies at the level of the phenomena, Kant is making a second, fundamental point: whatever ultimately lies behind the phenomena is outside the domain of science. A strict empiricist like Hume dogmatically asserts that one cannot speak of anything lying behind the phenomena. Kant argues otherwise and, in doing so, is more in line with Newton, who believes that gravity exists, although he can say nothing about it except what is revealed by the mathematical formulas expressing phenomenal relations. Insofar as science is concerned, Galileo, Newton, and Kant bracket physical substance, but among the three, Kant does not bracket causality. He places it in a different place—in the mind, but not as Hume would have it, as habit, but as a prior condition for experience.

This does not end the story. Kant, as an epistemologist and moral philosopher, tries to accomplish two goals at once. First, he wants to establish causality as a basis for science, and as a proper subject for metaphysics. This is accomplished by making causality a category of understanding because, for Kant, the categories are proper subjects for metaphysics. Second, he wants to preserve human freedom, or else morality disappears. His solution is a Cartesian duality. As a subject of science, human action is viewed in the light of cause and effect, so that the necessary condition for moral action, freedom, does not exist. However, causality and its consequent elimination of moral action only apply to the phenomenal world because that is the world experienced through the categories of understanding. Causality does not apply to the noumenal world, and freedom resides therein. Putting metaphysics aside, the key point for science is that science studies phenomena and whatever categories our understanding imposes on phenomena apply to phenomena. As a subject of science, human behavior is determined, but only as a subject of science. While Kant no doubt had the intent of establishing the possibility of metaphysics, this does not mean that he did not hit upon a third key point: Science is a product of the human mind and, because science is limited by its epistemology, the mind is only bound to the conclusions of science when it operates within the categories of the understanding, which themselves are limited to phenomenal experience and therefore are not operative outside the domain of that experience.

Whereas Kant sees Hume's arguments concerning the lack of empirical ground for causality as definitive, the empiricist John Stuart Mill wishes to empirically ground science in the aftermath of Hume, which, for him, means grounding induction and, in turn, causality. In *A System of Logic, Ratiocinative and Inductive*, he writes, "At the root of the whole theory of induction is the notion of physical cause. To certain phenomena, certain phenomena always do, and, as we believe, always will, succeed. The invariable antecedent is termed the 'cause,' the invariable consequent, the 'effect'"(Mill, 2002). There are four salient points regarding Mill's view:

1. No necessary connection is implied by causality;
2. The effect must be the "invariably and unconditionally consequent" of the cause;
3. Causality makes no reference to what is behind the phenomena; and
4. Causality is "coextensive with human experience."

In one of those instances where a philosopher neatly sums up his view, Mill writes,

> The notion of causation is deemed by the schools of metaphysics most in vogue at the present moment to imply a mysterious and most powerful tie, such as cannot, or at least does not, exist between any physical fact and that other physical fact on which it is invariably consequent, and which is popularly termed its cause: and thence is deduced the supposed necessity of ascending higher, into the essences and inherent constitution

> of things, to find the true cause, the cause which is not only followed by, but actually produces, the effect. No such necessity exists for the purposes of the present inquiry, nor will any such doctrine be found in the following pages. The only notion of a cause which the theory of induction requires is such a notion as can be gained from experience. The Law of Causation, the recognition of which is the main pillar of inductive science, is but the familiar truth that invariability of succession is found by observation to obtain between every fact in nature and some other fact which has preceded it, independently of all considerations respecting the ultimate mode of production of phenomena, and of every other question regarding the nature of "Things in themselves." (Mill, 2002)

In one sense, Mill escapes Hume's criticism by abandoning any notion of necessary connection and making induction purely sequential, but he falls completely flat by missing Hume's critical scientific point regarding the impossibility of arriving at the unconditional invariability of succession by any finite number of observations.

Mill recognizes that causality cannot be as simple as that of a single event being the sole cause of an effect. Regarding the complexity of causation, he states, "But the real cause is the whole of the antecedents, the whole of the contingencies of every description, which being realized, the consequent invariably follows. Yet even invariable sequence is not synonymous with causation. The sequence, besides being invariable, must be unconditional" (Mill, 2002). Clearly, "the whole of the antecedents, the whole of the contingencies of every description" has no bounds and may very well be the entire universe, which would reduce the entire notion of cause and effect to a statement about universal determinism. This would be a restatement of Laplacian determinism absent any individual causal relations within the universe. It is therefore not surprising that Mill adopts an essentially Laplacian position, except that unlike Laplace, who appeals to a "sufficiently vast" intelligence, Mill remains within the realm of human experience. He writes,

> The state of the whole universe at any instant, we believe to be the consequence of its state at the previous instant; insomuch that one who knew all the agents which exist as the present moment, their locations in space, and all of their properties, in other words, the laws of their agency, could predict the whole subsequent history of the universe, at least unless some new volition of a power capable of controlling the universe should supervene. (Mill, 2002)

If causality depends on knowing all the antecedents composing a cause, then surely it is not coextensive with human experience. On the other hand, expectation is very much coextensive with human experience.

Mill recognizes that, when applying induction in the course of scientific discovery, haphazard observation will not do. On finding causal relations, he writes,

> In the analysis of sequences into conditional and unconditional, the first operation is to ascertain and distinguish antecedents and consequents. The next step is to trace the connexion between antecedents and consequents, and this we can do only by a consideration of some of the antecedents or consequents under other conditions; we must either find an instance in nature suited to our purposes, or by an artificial arrangement of circumstances make one. When we make an artificial arrangement, we are said to experiment; and experimentation has great advantages over observation in that it often enables us to obtain innumerable combinations of circumstances which are not to be found in nature. (Mill, 2002)

But instead of the Galilean–Newtonian recognition that experimental constraint leads to relations that "estimate" relations among naturally occurring phenomena, Mill wants to use experiment to obtain "innumerable combinations of circumstances," a goal that on its face is impossible.

In trying to circumvent Hume's attack on causality on strictly empiricist grounds, Mill returns to a pre-Galilean world in the sense that, although necessary connection is abjured, causality remains a requirement for knowledge. Hume's analysis regarding uncertainty and the impossibility of concluding a necessary connection, one that is unconditional and invariable, is impenetrable because the certainty of formal logic does not apply to human interaction with Nature. Expectation, not causality, is coextensive with human experience. Indeed, it may be coextensive with animal experience, at least with those possessing greater degrees of intelligence, for instance, dogs. After surprisingly little training, one can point in a direction and Maggie will run that way, expecting a thrown ball to land in her pathway, and, never looking back, grab the ball in her mouth off the bounce. Should we presume from repeated observations of this behavior that Maggie has come upon the notion of an efficient cause, a group of antecedents which when occurring somehow in the proper conjunction result in the ball invariably falling in her path? Mill's problem is that he wants to bring metaphysics in through the backdoor. Aristotle was correct in placing the four forms of causality in the *Metaphysics*, but not correct in placing them in the *Physics*. Mill's hope of grounding causality in invariable and unconditional empirical sequences had already been doomed by Hume. Whereas Kant had recognized Hume's achievement, Mill did not.

In his famous essay, *On the Notion of Cause*, Russell demonstrates the impossibility of giving precise meaning to several different attempts to define "cause." He settles on the previously cited definition of Mill as perhaps the best attempt at a viable definition: "The Law of Causation, the recognition of which is the main pillar of inductive science, is but the familiar truth that invariability of succession is found by observation to obtain between every fact in nature and some other fact which has preceded it." But this attempt fails owing to the impossibility of supplying it with a suitable notion of event and the "insuperable difficulties," which Russell carefully articulates, of trying to define the timing between a cause and an effect. Recognizing that Mill's reasoning regarding induction and causality are based on the appearance of uniformities in nature, Russell addresses the issue:

> It must, of course, be admitted that many fairly dependable regularities of sequence occur in daily life. It is these regularities that have suggested the supposed law of causality; where they are found to fail, it is thought that a better formulation could have been found which would have never failed. I am far from denying that there may be such sequences which in fact never do fail. It may be that there will never be an exception to the rule that when a stone of more than a certain mass, moving with more than a certain velocity, comes in contact with a pane of glass of less than a certain thickness, the glass breaks . . . What I deny is that science assumes the existence of invariable uniformities of sequence of this kind, or that it aims at discovering them. All such uniformities, as we saw, depend upon a certain vagueness in the definition of the "events." That bodies fall is a vague qualitative statement; science wishes to know how fast they fall. This depends upon the shape of the bodies and the density of the air. It is true that there is more nearly uniformity when they fall in a vacuum; so far as Galileo could observe, the uniformity is then complete. But later it appeared that even there the latitude made a difference, and the altitude. Theoretically, the position of the sun and moon must make a difference. In short, every advance in a science takes us farther away from the crude uniformities which are first observed, into greater

> differentiation of antecedent and consequent, and into a continually wider circle of antecedents recognized as relevant. The principle 'same cause, same effect,' which philosophers imagine to be vital to science, is therefore utterly otiose. As soon as the antecedents have been given sufficiently fully to enable the consequent to be calculated with some exactitude, the antecedents have become so complicated that it is very unlikely they will ever recur. Hence, if this were the principle involved, science would remain utterly sterile. (Russell, 1913)

Russell neatly sums up his view of causality: "The law of causality, I believe, like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm" (Russell, 1913).

No doubt Hume is disturbing to those who desire certitude. To the extent that science must be grounded on certainty, or unconditional and invariable sequences, his analysis is devastating. In *The Rise of Scientific Philosophy*, Hans Reichenbach writes, "Empiricism broke down under Hume's criticism of induction, because it had not freed itself from a fundamental rationalist postulate, the postulate that all knowledge must be demonstrable as true. For this conception the inductive method is unjustifiable, since there is no proof that it will lead to true conclusions" (Reichenbach, 1971). The point is that science does not depend on unconditional sequences, does not base its formulations on a notion of "logical" induction, and does not does not have a notion of certainty. One need not turn to physics to see this; it is readily recognized in biology, where the subject matter begins with the cell, whose behavior is conceptualized as a random dynamical process. This does not mean that science is ungrounded, only that it must be grounded in probability theory and statistical inference, not in deterministic logic and induction.

CHAPTER 3

Scientific Knowledge

A concept without a percept is empty; a percept without a concept is blind.

—*Immanuel Kant*

The ancients, in particular, Aristotle, recognized that observation was necessary to gain knowledge of the physical world. Reason applied to observations, not reason alone, yields pragmatic knowledge of Nature. This is emphasized by the second-century Greek physician Galen in his treatise, *On the Natural Faculties*, when, in regard to the effects of a certain drug, he refutes the rationalism of Asclepiades when he writes, “This is so obvious that even those who make experience alone their starting point are aware of it . . . In this, then, they show good sense; whereas Asclepiades goes far astray in bidding us distrust our senses where obvious facts plainly overturn his hypotheses” (Galen, 1952). For the ancients, the philosophy of Nature might have dealt with principles of unity, ideal forms, and final causes, but natural science was observation followed by rational analysis. This was especially so during the Roman period, as evidenced by their remarkable engineering achievements.

What the ancients lacked is the idea of a controlled scientific experiment. Nor was this idea familiar to Ptolemy. The modern experimental method, as promulgated by Bacon and put into practice by Galileo, does not rely on accidental observations but instead constrains

Epistemology of the Cell: A Perspective on Biological Knowledge, First Edition.
Edward R. Dougherty, Michael L. Bittner.

its focus to the phenomena of interest in order to mitigate to the extent possible the effects of confounding variables. For modern science, reason does not enter the picture following observations; rather, it first provides a protocol for the observations so their analysis will characterize relations of interest and not be confounded by a multitude of secondary variables. For modern science, reason steps outside of Nature and constrains the manner in which Nature presents herself for analysis. While such constraint causes inexactitude relative to the knowledge of all variables and their interactions, Nature's complexity precludes such full knowledge anyway. For modern science, reason brings focus to the scientific enterprise.

Everything begins with the notion of a designed experiment—that is, methodological as opposed to unplanned observation. Rather than being a passive observer of Nature, the scientist structures the manner in which Nature is to be observed. The monumental importance of this change is reflected by the inclusion of the following statement concerning the early modern scientists, in particular, Galileo and Torricelli, by Immanuel Kant in the preface of the second edition of the *Critique of Pure Reason*:

> They learned that reason only perceives that which it produces after its own design; that it must not be content to follow, as it were, in the leading-strings of nature, but must proceed in advance with principles of judgment according to unvarying laws, and compel nature to reply to its questions. For accidental observations, made according to no preconceived plan, cannot be united under a necessary law. But it is this that reason seeks for and requires. It is only the principles of reason which can give to concordant phenomena the validity of laws, and it is only when experiment is directed by these rational principles that it can have any real utility. Reason must approach nature with the view, indeed, of receiving information from it, not, however, in the character of a pupil, who listens to all that his master chooses to tell him, but in that of a judge, who compels the witnesses to reply to those questions which he himself thinks fit to propose. To this single idea must the revolution be ascribed, by which, after groping in the dark for so many centuries, natural science was at length conducted into the path of certain progress. (Kant, 1952)

Kant, after surveying a century and a half of the scientific landscape, echoes the words of Bacon, who prior to the breakthroughs of Galileo had referred to accidental experience as "a mere groping, as of men in the dark, that feel all round them for the chance of finding their way." With a controlled experiment, reason devises a design to probe Nature in accordance with a conceptual model that has been partially formed by reason itself. Nature is not viewed as an unlimited store of empirical information to be randomly gathered up as if the scientist were a squirrel groping about an infinite field in search of nuts; instead, reason formulates a constrained space of relations and an experimental design to further constrain the space toward a conceptual system that better expresses the phenomenal relations of interest.

The product of an experiment is a set of measurements that form the data of sensibility, the empirical (as opposed to a rational) basis for knowledge. In themselves, measurements do not constitute scientific knowledge. They must be integrated into a conceptual system. Scientific knowledge is constituted via synthesis of the observed measurements. These are related to variables and relations among the variables. The change brought about by the "new science" of the seventeenth century is based on the integration of two fundamental principles: (1) the design of experiments under constrained circumstances to extract specifically desired information; and (2) the mathematical formulation of knowledge. The two principles arise from the two sides of the scientific problem, the source of knowledge and the representation of knowledge in the knower.

No doubt there was a strong tendency in ancient Greece toward discovering truth in mathematics without concern for applications. Plato, for instance, saw the pure forms of mathematics being real, not the shadows on the wall of the cave but the real forms giving rise to the shadows. Yet, one need only think of Archimedes' mathematical analyses of fluidics and mechanics to see that some ancients recognized an important role for mathematics in understanding the physical world. But just as the ancient and medieval worlds had observations without an experimental framework, they had mathematics without a clearly defined relationship to the observations. Bacon tried to address both of these deficiencies via designed experiments aimed at induction of causal relationships. He had the right inclination but the wrong concept—causality. Galileo and Newton had the right concept—mathematical relations.

Scientific knowledge necessarily takes the form of mathematics for four reasons: (1) scientific knowledge is based on quantitative measurements, be they logical or numeric; (2) scientific knowledge concerns relations and mathematics provides the formal structure for relations;

(3) the validity of a scientific theory depends on predictions and this requires a quantitative structure from which to generate predictions and a theory of probability in which the goodness of predictions can be quantified; and (4) mathematics provides a formal language sufficiently simple so that both the constituting theory and the experimental protocols for prediction are intersubjective, once the underlying mathematical representation of the theory is agreed upon.

There is much more to a model than the defining relations, that is, the general principles of the model. A great power of the scientific epistemology lies in the deducibility of logically necessary relations from the defining relations—the *hypothetico-deductive method*. This deduction can reveal critical relations not at once apparent in the defining relations. A full mathematical model consists of the defining relations and all relations logically deduced from these. The knowledge constituted by the derived relations is implicit in the defining structure but only becomes apparent when derived explicitly. Often, the most striking aspects of a scientific theory are represented by derived relations—for instance, the consequences of Newton's gravitational law and of Maxwell's equations. Key applications are typically the result of consequences of the basic model.

A mathematical model alone does not constitute a scientific theory. The model must be related to phenomena: a model's formal structure must be related to the empirical ground of science. Verification of a system requires that the symbols be tied to observations by some semantic rules that relate not necessarily to the general principles of the mathematical model themselves but to conclusions drawn from the principles. In other words, the theory is tested by checking measurable consequences of the theory. The conceptual system must be related to the experimental methodology. Phillipp Frank summarizes the situation both historically and epistemologically:

> Reichenbach had explicitly pointed out that what is needed is a bridge between the symbolic system of axioms and the protocols of the laboratory. But the nature of this bridge had been only vaguely described. Bridgman was the first who said precisely that these *relations of coordination* consist in the description of physical operations. He called them, therefore, *operational definitions*. (Frank, 1961)

The operational definitions are an intrinsic part of a scientific theory, for without them there would be no connection between the mathematics and observation, between the conceptual system and the experiments. The conceptual system must have consequences that can be checked via their relation to sensory observations. The mathematical equations may relate abstract symbols, but there must be a well-defined procedure for relating the consequences of the equations to quantifiable observations, such as the compression of a spring, the level of mercury in a thermometer, or the quantity of a particular mRNA in a cell. In sum, a scientific theory consists of two parts: (1) a mathematical model composed of symbols (variables and relations between the variables); and (2) a set of operational definitions that relate the symbols to data.

In Kantian terminology, the mathematical model *constitutes* the object of our knowledge, but the validity of that knowledge must be assessed via experiment. Thus, the experiment and the mathematical model form two inseparable requirements for scientific knowledge. Either without the other cannot yield scientific knowledge. Kant famously puts the duality this way, "A concept without a percept is empty; a percept without a concept is blind." The key issue is the manner and the extent to which the model must be related to experimental outcomes.

Recalling Hume and taking expectation as the ground of scientific knowledge leads to the crux of scientific epistemology: Predictive relations characterize model validity and are necessary for the existence of scientific knowledge. Scientific truth is pragmatic truth and this truth is contained in the predictive capacity of a scientific theory. Scientific knowledge is about the future. Past observations may lead one to construct a theory but the theory must predict the future. Riechenbach writes, "A mere report of relations observed in the past cannot be called knowledge. If knowledge is to reveal objective relations of physical objects, it must include reliable predictions. A radical empiricism, therefore, denies the possibility of knowledge" (Reichenbach, 1971).

Prediction is not certitude. Instead of causality, science involves conditional distributions that describe the probability of a *target* random variable Y given the values of a set of *predictor* random variables, $X_1, X_2, \ldots, X_m$. In particular, given the predictor random variables, the best prediction (relative to mean square error) for the value of Y is its conditional expectation. Causality is replaced by conditioning. Statements concerning conditional prediction can be validated via experimentation. The meaning of a statement can be rigorously defined within the framework of probability theory and its relation to measurable phenomena can be mathematically characterized within the theory of statistics. If the predictor variables are temporally antecedent to the variable to be predicted, then we have forward prediction. The terms "cause" and "effect" never appear.

There must be a predictive framework for validation because the scientific truth, or validity, of the model depends on the accuracy of predictions arising from the model. A model's formal structure must lead to experimental predictions in the sense that there are relations between model variables and observable phenomena such that experimental observations are in accord with the predicted values of corresponding variables. The model is connected to the experimental methodology by the operational definitions and these serve to transform mathematical predictions within the model to empirical predictions within the experimental methodology.

The general epistemological perspective seems clear, but its application to particular settings is not specified. Where is the model to come from and how does one characterize model validity relative to a measurement process? Einstein states,

> In order that thinking might not degenerate into "metaphysics," or into empty talk, it is only necessary that enough propositions of the conceptual system be firmly enough connected with sensory experiences and that the conceptual system, in view of its task of ordering and surveying sense experience, should show as much unity and parsimony as possible. Beyond that, however, the system is (as regards logic) a free play with symbols according to (logically) arbitrarily given rules of the game. (Einstein, 1944b)

According to Einstein, the model (conceptual system) is a creation of the "imagination." The manner of this creation is not part of the scientific theory. The classical manner is that the scientist combines an appreciation of the problem with reflections upon relevant phenomena and, based upon mathematical knowledge, creates a model. As Einstein states, this creation is free except that it must conform to the rules of the mathematical game.

Einstein's prescription does not lead to a unique, absolute truth because validation is a process and the "truth" of the theory is relative to that process. At issue is what is meant by "enough propositions" being "firmly enough connected with sensory experiences." The model must be connected to observations but the specification of this connection in a given circumstance is left open. This specification constitutes an epistemological issue that must be addressed in mathematical (including logical) statements. Absent such a specification, a purported scientific theory is meaningless. Reichenbach states, "The reference to verifiability is a necessary constituent of the theory of meaning. A sentence the truth of which cannot be determined from possible observations is meaningless" (Reichenbach, 1971). Because a model consists of mathematical relations and system variables must be checked against quantitative experimental observations, there is no nonmathematical way to describe the requirements and protocols to assess model validity. Hence, mathematics is essential to the structure of the model and its verification.

Suppose a geneticist recognizes phenotypic effects from blocking the promoter region of a gene to prevent transcription or from using RNAi to suppress signaling. The geneticist might then propose a mathematical model of the form $(g \rightarrow 0) \Rightarrow (p1 \rightarrow p2)$, where $g \rightarrow 0$ means that the protein product of gene g never reaches its target, $p1 \rightarrow p2$ means phenotype $p1$ is transformed to phenotype $p2$, and $\Rightarrow$ is probabilistically interpreted as prediction. The model is validated by an experiment designed to reflect conditions under which the model is hypothesized. If the geneticist were to make observations without specifying a precise mathematical model (including a probability distribution to characterize the probabilistic aspects of the model) and a protocol for predictive validation, then there would be no scientific knowledge.

A scientific theory is incomplete without the formal specification of achievable measurements that can be compared with predictions derived from the conceptual theory and the manner in which the measurements are to be compared with the conceptual system, in particular, the choice of validity criteria and the mathematical properties of those criteria as applied in different circumstances. The validity of a theory is relative to this specification, but what is not at issue is the necessity of a set of relations tying the conceptual system to operational measurements. It makes no sense to argue about the validity of a scientific theory without specifying the validation protocol. A scientific theory is intersubjective, but the epistemological criteria underlying a particular validation are open to debate. Once the validation requirements are specified, the mathematical model (conceptual system) is valid relative to the validation criteria and to the degree that the requirements are satisfied, that is, to the degree that predictions demanded by the validation protocol and resulting from the mathematical model agree with experimental observations.

One might question the decisive role of prediction by asking whether it not be called "science" if one simply categorizes observations based on some defined set of criteria, such as with taxonomy. Certainly such categories represent a form of knowledge and their assembly, which can require great effort and ingenuity, are part of the scientific enterprise, but they no not constitute scientific knowledge unless they are utilized within some predictive framework. Reichenbach states, "Scientific philosophy has constructed a *functional* conception of knowledge, which regards knowledge as an instrument of prediction and for which sense observation is the only admissible criterion of nonempty truth" (Reichenbach, 1971). Scientific knowledge is worldly knowledge in the sense that it points into the future by making predictions about events that have yet to take place. Richard Feynman firmly asserts, "Knowledge is of no real value if all you can tell me is what happened yesterday" (Feynman, 1998). Scientific knowledge is contingent, always awaiting the possibility of its invalidation. Its truth or falsity lies in the verity of its predictions and, since these predictions depend upon the outcomes of experiments, ultimately the validity of scientific knowledge is relative to the methodology of verification. William James writes, "Truth happens to an idea. It becomes true, is made true by events. Its verity is in fact an event, a process, the process namely of its verifying itself, its verification. Its validity is the process of its validation" (James, 1963). This is a long way from Plato's cave. The prisoners in the cave see only shadows but reason can reach deeper to the true forms casting the shadows. These exist in some timeless place where there is no idea of process. It is also a long way from Aristotle's three pillars: causality, explanation, and metaphysics. For Aristotle, reason could explain the observations by placing them within some rational structure, in his case it being a causal structure, intrinsic to the whole of reality, it thereby being metaphysical. In both cases, for Plato and Aristotle, truth is metaphysical, it being a property of an idea that, while it might be only partially revealed in observations, is intrinsic to the idea. For science, the truth of an idea depends on the process of validating its truth. Since many processes might be used there are many truths. Change this process and the truth may change.

Aristotle provides four causes as the basis for explanation, in particular, explanation of the physical world. Irrespective of the continuing appeal to causality, explanation remains ubiquitous and is perhaps the greatest impediment to meaningful scientific inquiry. In *Webster's Unabridged Dictionary*, the first usage for the word "explain" is "to make plain, clear, or intelligible"; and intelligible means understandable or comprehensible, in particular, in philosophy, understandable by the intellect (*Webster's New Twentieth Century Dictionary*, 1978). Aristotle's four causes represent categories of intelligibility whose explanatory usage makes the world understandable. Humans observe the world around them and try to understand it. Aristotle sees regularity in change, objects being shaped out of material, ideas of form guiding material changes, and purpose to change. Two millennia before the critical philosophy of Kant, he naturally formalizes these everyday observations into categories of understanding. They make the world intelligible by explaining it in terms of categories grasped by the intellect. They satisfy the desire to give order to the physical world and comprehend the "why" of that order. Using the four causes, Nature is grasped directly. She becomes accessible to the human intellect. The result is a rationalist approach to Nature: reason working a posteriori on observations, or perhaps in the absence of observations, to construct a mental picture of the world. For physical science, this would mean a picture in terms of overtly physical categories corresponding to physical substance, such as particles, gravity, and force.

When Newton writes, "And to us it is enough that gravity does really exist," he is bracketing causality along with whatever "physical" substance is represented by the phenomena observed. What perhaps Newton did not realize is that this bracketing would become permanent in the sense that today there is no explanation of gravitation as a physical substance; indeed, one is hard pressed to say what is meant by a "physical substance." What is certain, however, is that the Newtonian gravitational law and the more modern theories in terms of the curvature of space are mathematically clear and make excellent predictions.

When discussing the enormity of the transformation wrought by Galileo and Newton, Morris Kline writes, "What science has done, then, is to sacrifice physical intelligibility for the sake of mathematical description and mathematical prediction" (Kline, 1985). Sacrificing physical intelligibility does not involve an abandonment of knowledge; on the contrary, it involves the recognition that everyday human categories concerning Nature—those that arise from the ordinary interaction with the physical world, such as pushing and pulling—are not suitable for describing phenomenal relations. Kline goes on to say,

> The insurgent seventeenth century found a qualitative world whose study was aided by mathematical abstractions. It bequeathed a mathematical, quantitative world that subsumed under its mathematical laws the concreteness of the physical world. In Newton's time and for two hundred years afterwards, physicists spoke of the action of gravity as "action at a distance," a meaningless phrase that was accepted as a substitute for explaining the physical mechanism, much as we speak of spirits or ghosts to explain unseen phenomena. (Kline, 1985)

A mathematical theory is intelligible because it is a product of the human intellect; the world about us is not the product of human intellect.

Consider the electromagnetic field theory that is responsible for so much technology in the modern world. The theory, rooted in James Clerk Maxwell's equations, is completely understood because it is a mathematical theory. Its applications depend on the behavior of detectors as predicted by the theory. But what is the nature of the physical substance behind these? The thoughts of Maxwell on this subject are given in his paper, *On Faraday's Lines of Force*, and should be read by every student of science:

> The first process therefore in the effectual study of the science, must be one of simplification and reduction of the results of previous investigation to a form in which the mind can grasp them. The results of this simplification may take the form of a purely mathematical formula or of a physical hypothesis. In the first case we entirely lose sight of the phenomena to be explained and though we may trace out the consequences of given laws, we can never obtain more extended views of the connexions of the subject. If, on the other hand, we adopt a physical hypothesis, we see the phenomena only through a medium, and are liable to that blindness to facts and rashness in assumption which a partial explanation encourages. We must therefore discover some method of investigation which allows the mind at every step to lay hold of a clear physical conception, without being committed to any theory founded on the physical science from which that conception is borrowed, so that it is neither drawn aside from

> the subject in pursuit of analytical subtleties, nor carried beyond the truth by a favorite hypothesis. (Maxwell, 1855)

Maxwell recognizes that the mathematical approach "may lose sight of the phenomena to be explained," but the danger of adopting a "physical hypothesis" is that it may lead to "blindness to facts and rashness." To wit, human intuition based upon everyday interaction with Nature can distort one's view by leading to a rationalist approach to science that results in bending the observations to fit a "favorite hypothesis." How polite!

After discussing how analogies with physically based models are often useful for the arrival at satisfactory theories, even when a model may relate to a different physical setting than the one being considered, Maxwell comments that he will analogize lines of force as "fine tubes of variable section carrying an incompressible fluid." After discussing the aim and methodology of the fluid analogy, he writes,

> I propose, then, first to describe a method by which the motion of such a fluid can be clearly conceived; secondly to trace the consequences of assuming certain conditions of motion, and to point out the application of the method to some of the less complicated phenomena of electricity, magnetism, and galvanism; and lastly to shew how by an extension of these methods, and the introduction of another idea due to Faraday, the laws of the attractions and inductive actions of magnets and currents may be clearly conceived, without making any assumptions as to the physical nature of electricity, or adding anything to that which has been already proved by experiment. By referring everything to the purely geometrical idea of the motion of an imaginary fluid, I hope to attain generality and precision, and to avoid the dangers arising from a premature theory professing to explain the cause of the phenomena. If the results of mere speculation which I have collected are found to be of any use to experimental philosophers, in arranging and interpreting their results, they will have served their purpose, and a mature theory, in which physical facts will be physically explained, will be formed by those who by interrogating Nature herself can obtain the only true solution of the questions which the mathematical theory suggests. (Maxwell, 1855)

Maxwell proceeds "without making any assumptions as to the physical nature of electricity." Nevertheless, he remains hesitant, adding that the mathematical theory is only suggestive of the "true solution." Looking back to Aristotle and the desire for physical intelligibility, he hopes for “a mature theory, in which physical facts will be physically explained.” Maxwell is not alone in this dissatisfaction. Kline writes,

> Despite the Herculean efforts to determine physically what an electric field and a magnetic field are, scientists are unsuccessful. . . . We do not have any physical account of the knowledge of the electromagnetic waves as waves. Only when we introduce conductors such as radio antennae in electromagnetic fields do we obtain any evidence that those fields exist. Yet we send radio waves bearing complex messages thousands of miles. Just what substance travels through space we do not know. (Kline, 1985)

As Newton brackets causality and the physical nature of gravity in favor of mathematical relations, Maxwell brackets the physical waves behind the field theory. The upshot of all this bracketing is that the subject of physics (as science) is embedded within mathematics. We accept that the physical world is not intelligible (in the standard Aristotelian sense, which is really the everyday sense), but we are not debilitated because the mathematical structures allow us to build devices that respond according to the equations and thereby produce pragmatic effects in the physical world.

Classically, the scientist worked with models whose basic terms referred to ideas whose origins lay in prescientific perceptual experience, terms such as “particle,” “wave,” and “force.” Moreover, the frames of experience, such as Euclidean three-dimensional space and linear time, had their origins in the commonplace perception of everyday phenomena. With the advent of quantum mechanics and general relativity, recognition of the prime importance of the mathematical apparatus with regard to representation and prediction increased, along with the recognition that any intuitive appreciation of this apparatus is secondary. In *The Mysterious Universe*, James Jeans writes,

> The final truth about phenomena resides in the mathematical description of it; so long as there is no imperfection in this, our knowledge is complete. We go beyond the mathematical formula at our own risk; we may find a [nonmathematical] model or picture that helps us to understand it, but we have no right to expect this, and our failure to find such a model or picture need not indicate that either our reasoning or our knowledge is at fault. (Jeans, 1930)

Nonmathematical reasoning may be useful for the scientist in exploratory thinking, but scientific knowledge is constituted in a mathematical model. One might use a metaphor of observers holding lights on approaching trains to make an intuitive point concerning relativity, but the scientific theory lies properly within the equations. Any attempt to force a nonmathematical understanding creates the risk of having a diminished (or erroneous) scientific theory because it substitutes readily understandable and often convincing descriptions in place of strict scientific knowledge, which must take a mathematical form.

We cannot expect to have scientific knowledge within the categories of everyday understanding because everyday understanding is inadequate for quantitative predictive models. This point is strongly emphasized by Feynman in the following statement made before beginning a series of lectures on quantum electrodynamics to an audience of nonspecialists:

> What I am going to tell you about is what we teach our physics students in the third or fourth year of graduate school—and you think I'm going to explain it to you so you can understand it? No, you're not going to be able to understand it . . . You see, my physics students don't understand it either. That is because I don't understand it. Nobody does . . . It is whether or not the theory gives predictions that agree with experiment. It is not a question of whether a theory is philosophically delightful, or easy to understand, or perfectly reasonable from the point of view of common sense. The theory of quantum electrodynamics describes Nature as absurd from the point of view of common sense. And it agrees fully with experiment. So I hope you can accept Nature as she is—absurd. (Feynman, 1985)

The absurdity is not intrinsic to Nature. Absurdity is a human category and the absurdity of Nature is relative to human intelligibility. The philosophical notion that the human mind has the capacity to understand Nature in everyday categories has gone by the wayside.

Science has not abandoned reason; rather, the role of reason has changed. Scientific knowledge is constituted in a most pure form of reason, mathematics, but the truth of that knowledge is not ascertained directly by reason, nor is that knowledge required to conform to ordinary categories of intelligibility. In one sense, reason loses its lofty position because it cannot remain independent in its judgments; they must be tied to phenomena in well-defined ways. To put the matter more forcefully, reason is no longer trusted. The Enlightenment, in the person of its two greatest philosophers, Hume and Kant, turns reason upon itself and exposes its limitations, at least in its pure form. When Maxwell speaks of discovering a method that allows the mind not to be "carried beyond the truth by a favorite hypothesis," he is warning of the danger of unchecked reason, a warning given more forcefully by Hume, who, in the *Treatise*, asserts, "Reason is, and ought only to be the slave of the passions, and can never pretend to any other office than to serve and obey them" (Hume, 1951). Whereas Maxwell is concerned about tilting one's reason in the direction of a favorite hypothesis owing to "that blindness to facts and rashness in assumption which a partial explanation encourages," Hume, with is usual flare for directness, states that reason is a servant of desire and therefore cannot be trusted as an arbiter of its own deliberations. One not only should be wary of blindness to the facts affecting explanations but also recognize that explanations may be constructed in such a way as to "serve and obey" the passions. Consider two protagonists who firmly believe in the products of their individual reason. Even if it were possible to decompose their arguments into their logical components, it may be next to impossible to find an error in either, the problem being that somewhere down deep they are arguing from competing premises. A critical aspect of scientific validity is that we need not consider their reasoning. We need only test their claims, which can be done because they must each provide operational definitions in conjunction with their models.

Perhaps modernity has to some extent deprived reason of its lofty position; however, it has also made reason more powerful in other ways. First, it has made an extraordinary leap away from the immediate perceptions that were previously the grist for its understanding of the natural order. This entails a huge leap in creativity. Einstein writes, "Experience, of course, remains the sole criterion for the serviceability of mathematical constructions for physics, but the truly creative principle resides in mathematics" (Einstein, 1933). The veracity of a scientific model lies in experience, but its conception arises from the imagination, an imagination freed from the fetters of Euclidean geometry, linear time, certainty, causality, and other relics of the past. Second, when confronting Nature, reason no longer is confined to groping through aimlessly collected data; instead, it views Nature though an experimental filter based upon its own needs. In this regard, William Barrett writes,

> Rationalism does not surrender itself here to the brute facts. Rather, it sets itself over the facts in their haphazard sequence; it takes the audacious steps of positing conditions contrary to fact, and it proceeds to measure the facts in the light of these contractual conditions. Reason becomes "legislative of experience" – this was the decisive point that Kant's genius perceived as the real revolution of the new science. (Barrett, 1979)

Third, science has abandoned the rational explanation of Nature and reason no longer is stuck looking backward in an attempt to explain the past; rather, its role is to foretell the future. Regarding the mathematical theory that constitutes scientific knowledge, Reichenbach states,

> If the abstract relations are general truths, they hold not only for the observations made, but also for observations not yet made; they include not only an account of past experiences, but also predictions of future experiences. That is the addition which reason makes to knowledge. Observation informs us about the past and the present, reason foretells the future. (Reichenbach, 1971)

To be able to predict the future puts great power into the hands of scientists, because it facilitates the predictable transformation of Nature resulting from human action in the world. According to Barrett, "The scientist constructs models, which are not found among the things given him in his experience, and proceeds to impose those models upon Nature" (Barrett, 1979). This is the key to powerful translational science, a subject on which we will subsequently have much to say.

An advantage of a causality-based epistemology or, more generally, a deterministic epistemology, is that, assuming sufficient knowledge, there is no uncertainty. This is Laplace's position, given a sufficiently intelligent being. In practice, measurements are not perfectly precise, so there is always uncertainty as to the value of any variable. This uncertainty is not part of a deterministic epistemology but rather pertains to the actualization of the epistemology in the measurement process. From a classical perspective, one might look forward to ever increasingly precise measurements without limit, so that, in principle, the measurement error could be negligible. This assumption vanishes with the quantum theory, where, in principle, there is a hard limit. According to the Heisenberg uncertainty principle, at any moment in time, the product of the uncertainty in position and momentum of a particle must exceed $h/2\pi$, where h is Planck's constant. The position and momentum can be measured separately without an intrinsic bound on accuracy, but not jointly. The uncertainty posited by Heisenberg is intrinsic to human interaction with Nature. Epistemologically, the state of Nature, as perceived by human beings, is not independent of human observation and therefore there is no "objective" scientific reality independent of human observation.

This conclusion does not compel one to relinquish a deterministic natural philosophy. Just as Hume recognized that causality is "a characteristic of the way in which we regard Nature," to use Schrödinger's phrase, so too is determinism. Recalling Windelband's definition of metaphysics, determinism represents a worldview and therefore a metaphysical, not a scientific, category. It can neither be proven nor disproven by empirical observations.

There are, however, fundamental constraints imposed on science by observational limitations. Since a model can only be verified to the extent that its symbols can be tied to observations in a predictive framework, the ability to design and perform suitable experiments, including the availability of technology to make the desired measurements, is mandatory. Limitations on experimentation can result in limitations on the complexity or content of a theory. To be validated, a theory must not exceed the experimentalist's ability to conceive and perform appropriate experiments. With the uncertainty theory, modern physics appears to have brought us beyond the situation of where the limitations on observation result only from insufficient experimental apparatus to the point where the limitations are unsurpassable in principle. In this vein, Schrödinger writes,

> It really is the ultimate purpose of all schemes and models to serve as scaffolding for any observations that are at all conceivable. . . . There does not seem to be much sense in inquiring about the real existence of something, if one is convinced that the effect through which the thing would manifest itself, in case it existed, is certainly not observable. (Schrödinger, 1957)

In other words, without observable effects due to an object, the object is not a suitable subject for scientific inquiry.

Yet we need not go to the uncertainty theory to appreciate Schrödinger's point. The inability to experience absolute simultaneity and other such absolutes played a key role in Einstein's approach to relativity theory. He writes,

> A further characterization of the theory of relativity is an epistemological point of view. In physics no concept is necessary or justifiable on an *a priori* basis. A concept acquires a right to existence solely through its obvious and unequivocal place in a chain of events relating to physical experiences. That is why the theory of relativity rejects concepts of absolute simultaneity, absolute speed, absolute acceleration, etc.; they can have no unequivocal link with experiences. Similarly, the notions of "plane," and "straight line," and the like, which form the basis of Euclidean geometry, had to be discarded. Every physical concept must be defined in such a way that it can be used to determine in principle whether or not it fits the concrete case. (Einstein, 1993)

A second imposition on scientific theory imposed by observational limitations concerns the kind of mathematical models to be employed in scientific theories. If there is intrinsic uncertainty in the measurements relating to a model, then a deterministic model is intrinsically limited in its ability to lead to accurate predictions because phenomenal predictions tied to the model via its operational definitions will be affected by the uncertainty and therefore validation is problematic. Consequently, probabilistic models, taking uncertainty into account, are preferable. Whereas imprecise measurements always affect model validation, the uncertainty principle makes this problem intrinsic. This does not mean that deterministic models are no longer useful. In the classical setting, when measurement error is very small in comparison with the values being measured, it can be ignored. This is also the situation in the macroscopic world when it comes to intrinsic measurement uncertainty because Planck's constant is very small and the uncertainty can be practically ignored.

Classical measurement uncertainty plays a significant role in biology. Consider the large amount of processing involved in gene expression measurements on microarrays or the even greater amount required for mass spectrometry-based proteomics. In these cases, the measurements are highly variable and dependent on the particular filtering methods employed in obtaining them. Even should these measurement processes be greatly improved over time, the uncertainty arising from model constraint will remain, this latter uncertainty being inherent in the model paradigm. Cell function involves the interaction of hundreds of thousands of genes and proteins, so that any functional model must greatly constrain its focus. Hence, the relations among the model variables are stochastic because they are affected by changes of latent variables outside the model.

Deterministic models may be suitable for phenomena not subject to consequential changes outside those internal to the system; however, they typically are unsatisfactory for modeling complex interactive physical systems subject to consequential external latent variables. In *Theory of Random Functions*, after accepting phenomenological interdependence as a fundamental law of dialectical materialism, Vladimir Pugachev explains the compatibility of that determinist metaphysical position with a stochastic scientific epistemology:

> By virtue of this [law], each observable phenomenon is causally related to innumerable other phenomena and its pattern of development depends on a multiplicity of factors . . . Only a limited number of these factors can be established and traced. For this reason, if we observe the same phenomenon many times, it is seen that besides its general properties, there are certain special features which are only typical of a particular observation. (Pugachev, 1965)

If we repeatedly observe a dynamical process and make measurements on some set of variables over time, we cannot expect the measurements to remain the same across the different trials because, even if we could somehow replicate the initial state of the variables for each trial, unless the process were completely isolated so that the variables being measured were unaffected by no others but themselves, its evolution will depend upon variables outside the set.

Like determinism, interpreted as a world view, randomness is a metaphysical category that can neither be proven nor disproven by empirical observations. The assumption of a stochastic model is a scientific decision, not a metaphysical perspective. Andrei Kolmogorov, discoverer of the modern measure-theoretic approach to probability theory, puts the modeling issue in the following way: “The possibility of using, in the treatment of a real process, schemes of well-determined or of only stochastically definite processes stands in no relation to the question whether the real process is itself determined or random” (Kolmogorov, 1931). The so-called “real process” is not a subject of scientific knowledge. Even if cell function were deterministic, this determinism would not likely be reflected in a practical gene network model because the genes in the model would undoubtedly be affected by events (latent variables), including genes, outside the model, thereby imparting a stochastic nature to the model. This would be the case even without considering experimental effects. This recognition is critical to the science of the cell and to translational science related to controlling cell behavior.

This brings us to the general topic of randomness in biology. All too often the word "random" is tossed about without heed to a rigorous definition. For instance, the time between two events, such as between completion of transcription to initiation of translation, is said to be random, or the occurrence of an event, such as a specific gene mutation, may be described as random. What is the meaning, if any, of these expressions?

Webster's Unabridged Dictionary defines the adjective "random" as, "without aim or purpose; haphazard" (*Webster's New Twentieth Century Dictionary*, 1978). Purpose is defined in two ways. The first is "that which a person sets before himself as an object to be reached or accomplished; aim; intention; design." Under this definition, a random occurrence is one not intended or designed. For the two aforementioned biological examples, this would say that there is no intended or designed time between the completion of transcription and the initiation of translation, and that mutations are neither intended nor designed. This is clearly not what is meant by "random" in science because science, as an empirically based discipline, is not concerned with intentions outside the phenomena. The second *Webster* definition of purpose is an "end in view; the object for which something exists or is done." Under this definition, a random occurrence is one without object. But this puts us back into the domain of causality because it says that events occur without final cause, and even Bacon places final causality within metaphysics.

Although we cannot appeal to *Webster's Dictionary* for a definition of randomness for science, the relation to final cause opens the door to an alternative view of randomness, that a random event is one not determined by causality. Here there is a mixture of two concepts, determinism and causality. Let us first dismiss causality because if a random event were defined as a noncausal event, then randomness would be defined as a negation of a nonscientific category and therefore would, itself, be nonscientific. Hence, we arrive at a random event being one that is not determined. But this throws us back upon determinism, absent a notion of causality, and randomness would be defined as nondeterminism, a metaphysical, not a scientific, category.

Randomness is indeed related to determinism, not in terms of phenomena but in how phenomena are modeled within science. Consider the time r between transcription and translation. For a specific instance, r represents a single measurement and takes the mathematical form of a real number. However, our interest is not with a single observation of time but rather the class of measurements. Thus, the time between transcription and translation varies depending on a host of conditions within the cell and the time is represented as a random variable, which we will denote by R. Here, the word "random" appears as part of the term "random variable," which has a precise mathematical definition. The nature of the random variable R is more subtle than that of a simple real number and it was not until the twentieth century that we had a suitable definition of a random variable, that being a measurable function from a probability space into the space of real numbers. This definition requires the definitions of a measurable function and a probability space, which in turn require the definition of a probability measure, the upshot being that it requires the development of the mathematical theory of measure to be able to give meaning to the word "random" in regard to its scientific usage.

To extend this simple case to one more representative of a biological system, consider a dynamical process, say the amount of the protein product, Wnt5a, corresponding to the gene WNT5A, in a cell. Dynamically, this measurement is represented as a variable of the form $x(t)$, where t denotes time and $x(t)$ is a numerical value whose units depend upon the measurement procedure. If we track this abundance for a single cell, we get a time function that is deterministic, the latter meaning, by definition, that there is a certain value at each time point. However, we are typically interested in the behavior of Wnt5a for an arbitrary cell and, then, the measurement is not deterministic, the abundance trajectory being different for different cells. In this case the measurement is represented as a time-dependent random variable, denoted $X(t)$, again the word "random" appearing as part of the mathematical term "random variable." The deterministic variable $x(t)$ takes values in some numerical space, such as the logical space $\{0, 1\}$, the integers, or the real line, depending on the quantization of the measurement procedure. The random variable $X(t)$ is a function from a probability space into a numerical space. Whereas $x(t)$ is referred to as a " time function," $X(t)$ is referred to as a "random time function," a "random time process," or a "stochastic process." In every instance the word "random" is used, it requires a definition in terms of the underlying mathematical spaces. None of this makes any suppositions concerning things-in-themselves. In the context of science, "random" is simply a word adopted by mathematics and defined therein within the framework of axiomatic probability theory.

To see what happens when one tries to use the word "random" loosely, consider the following statement by Francisco Ayala:

> Mutations are said to be accidental, undirected, random, or chance events. These terms are often used as synonyms, but there are at least three different senses in which they are predicated of the mutation process. First, mutations are accidental or chance events, in the sense that they are rare exceptions to the regularity of the process of DNA replication, which normally involves precise copying of the hereditary information, encoded in the nucleotide sequences. Second, mutations are accidental, random, or chance events also because there is no way of knowing whether a given gene or genome will mutate in a particular cell or in a particular generation. We cannot predict which individuals will have a new mutation and which ones will not, nor can we predict which gene will mutate in a given individual. This does not imply that no regularities exist in the mutation process; the regularities are associated with stochastic processes, to which probabilities can be assigned. There is a definite probability (although it may not have been ascertained) that a given gene will mutate in any given individual. Moreover, it is not true that a mutation is just as likely to occur as any other mutation. Third, mutations are accidental, undirected, random, or chance events in a sense that is very important for evolution; they are unoriented with respect to adaptation. (Ayala, 2008)

Note the terms Ayala is grouping with "random." Both "accidental" and "undirected" agree with *Webster's* because they relate to unintended events. So it seems that he is in agreement with *Webster's* and at the outset leaves the domain of science for psychology or metaphysics. Yet his first definition does not speak of intent; rather, he uses randomness to describe a process that exhibits rare exceptions to regularity, hence, seeming to imply that randomness applies to a noncausal process, in the sense of Mill, thereby making it, at best, a metaphysical category or, at worst, meaningless. Skipping momentarily to the third definition, this defines random as being noncausal in the sense of final cause, again metaphysical. The second definition is the only one suitable for science. Although he does not go into a careful mathematical characterization of mutations forming a stochastic process, one can be given. Ayala's contention that all three definitions are used in biology is where the problem lies. The first and third have nothing to do with science.

Many important problems of classical physics can be addressed with deterministic models and, to a great extent, conform to human understanding. This happens for two reasons. First, the ranges of the variables, distance, velocity, mass, and so on, are within the range of human experience. Second, very simple models, such as the Newtonian gravitational law, lead to good predictions under everyday circumstances because the effects of latent variables are fairly negligible. Neither of these conditions tend to hold in biology: The complexity of the operations within the cell is far outside the normal scope of human experience and the massive interrelationships among the cell components make it difficult to find simple models unless the focus is narrowed extensively, and even when this is possible it is likely that model stochasticity resulting from latent variables will be much too great to ignore. Consider, for example, Conrad Waddington's concluding remarks in his 1966 book, *Principles of Development and Differentiation*:

> In my opinion, at least, the three problems immediately in front of us are these: What is the nature of the change that renders a cell competent, so that it is ready to be switched into a particular developmental path? What is it that triggers off the switch and puts the cell into a state of determination, which is only with difficulty reversible, and can normally be transmitted through several cell generations? Finally, how are the activities of all the genes concerned in any developmental pathway tied together, so that they proceed in an integrated and orderly manner—or does this, perhaps, follow from the answers to the first two questions? (Waddington, 1966)

At biology's fundamental place within natural science, that being regulation, Waddington points to the requirement that biological knowledge depend on the theory of multivariate dynamical processes, which will of necessity be random processes owing to the massive complexity.

There is no hope that simple models and elementary deterministic mathematics can constitute biological knowledge. Biological knowledge is more removed from everyday understanding than is a great deal of physical knowledge and, concomitantly, its mathematical representation is more abstract.

Consider genomics, where, as suggested by Waddington, concern is with cellular control mechanisms based on the manner in which information stored in DNA is converted into molecular machines with various capabilities, including those required to carry out the copying of DNA and the transformation of its code into RNA and protein. Via interactions among the proteins present in the cell and interactions of protein complexes with the DNA, logical relations are produced that maintain highly varied patterns of gene expression among the differing cell types present in an organism. Cellular control, and its failure in disease, results from multivariate decision making and to the degree that human understanding of decision making is represented in logic, it is natural to employ logical models, perhaps multivalued, to constitute biological knowledge. Since the cell is an information processing system, knowledge representation and information theory are fundamental aspects of biological knowledge, as is the mathematics of control as it pertains to such a system.

Taking into account randomness, in the proper sense, and cell dynamics, gene regulatory modeling involves stochastic nonlinear dynamical systems. These may be continuous or discrete, and they can be synchronous or asynchronous. As in all modeling situations, the more detailed the model, the greater the computational complexity and the more difficult the inference from data. Given a network model, at least two basic issues arise: (1) the phenotypic issue—characterizing the steady-state behavior of the system, where it settles following transient behavior; and (2) the translational issue—determination of intervention strategies to favorably alter the steady-state behavior of the system. It is usually very difficult to characterize the steady-state distribution of the system in terms of system parameters. Even if this is done, can one really claim to have an understanding of the steady-state distribution in terms of sensory intuitions regarding the genes? Even under the coarsest quantization, a binary network, and only 20 genes, the transition probability matrix of a Markov chain regulatory model possesses dimensions 1,048,576 × 1,048,576 and this matrix determines a steady-state distribution with 1,048,576 states. The behavior of such a network is virtually unintelligible. One is often mystified at how tiny variations in the parameters dramatically alter steady-state behavior. Often, mathematical analysis in terms of low-order statistical characteristics of a dynamical process allows application of the system, albeit with varying degrees of performance loss, but even then intuition of properties entailed by the low-order analysis is rare except for very simple covariance structures, which themselves usually arise from simplifying the full covariance structure of the model. The effort to glean some physical, in this case, physical biological, intelligibility by simply looking at a graphical model composed of arrows between related genes, absent the dynamical structure of the network, is a striking example of ignoring Jeans's warning about moving away from the mathematics to mental pictures, often known as visualizations.

The dependency on mathematics and the lack of intuition are even more extreme when one wants to use the regulatory model to determine therapeutic policies (Shmulevich and Dougherty, 2010). Fundamental, and often difficult, mathematical analyses must be performed to arrive at control strategies, and these are especially involved if one wishes to achieve robust strategies not overly sensitive to inaccurate system identification or imperfect application of control, both of which are ubiquitous in complex settings. There is no hope of obtaining categorical understanding of a control policy's performance by considering the phenomena themselves. Moreover, again referring to Jeans, graphical visualizations convey no dynamical information and depending on such visualizations for therapeutic applications is reckless.

If human beings had sensory experience of traveling near the speed of light, then perhaps our ordinary understanding would grasp changing masses and clocks slowing or speeding up. If we had sensory experience at the quantum level, then perhaps we would display no surprise at the behavior of a photon in the famous double-slit experiment. Our difficulties of understanding arise because the categories of our ordinary understanding relate to our sensory experiences. These difficulties extend to biology. We have no sensory experience with networks of thousands of nonlinearly interacting nodes exhibiting feedback, distributed regulation, and massive redundancy. Recalling Feynman, Nature is absurd from the human perspective because we lack the categories of understanding with which to intuit it—be it physics or biology.

CHAPTER 4

Cells and Factories

> The processes which keep an animal alive have to be quite as highly organized as the operations in the most complicated mass-production factory.
>
> —*Conrad Waddington*

Biology concerns living organisms. These exist in the physical world. Therefore biology depends upon physics. Each cell consists of a host of molecules that form the building blocks of structures within the cell and are involved with interactions both interior and exterior to the cell. Therefore, biology depends upon chemistry. But the subject matter of biology is not that of physics or chemistry; otherwise, biology would be a branch of physics or chemistry. Biology concerns the operation of the cell in its pursuit of life, not the molecular infrastructure that forms the physiochemical underpinnings of life. The activity within a cell is much like that within a factory. In the latter, machines manufacture products, energy is consumed, information is stored, information is processed, decisions are made, and signals are sent to maintain proper factory organization and operation. All of these functions also take place within a cell and it is through analogy with a factory that we approach the epistemology of the cell.

The factory analogy has been used before. In his 1935 book, *How Animals Develop*, Waddington writes,

> The processes which keep an animal alive have to be quite as highly organized as the operations in the most complicated mass-production

Epistemology of the Cell: A Perspective on Biological Knowledge, First Edition.
Edward R. Dougherty, Michael L. Bittner.

> factory. If there is a 'secret life,' it is here we must look for it, among the causes which bring about the arrangement of innumerable separate processes into a single harmonious living organism.… To say that an animal is an organism means in fact two things: firstly, that it is a system made up of separate parts, and secondly, that in order to describe fully how any one part works one has to refer either to the whole system or to the other parts. (Waddington, 1935)

Whereas Waddington is referring to an animal as a system composed of organized macrostructures, our concern is the cell as a system of microstructures, a system whose constituent parts were beyond the experimental capability of Waddington's day. Nonetheless, the principle is the same. The task of biology is to look for the "secret life," as Waddington calls it. We can forgive Waddington for referring to the "causes" that bring about a harmonious living organism as he was writing before the advent of systems theory and concomitant with the development of quantum mechanics and stochastic processes. One need only change the word "causes" to "regulatory apparatus" and Waddington's statement becomes fully modern. The function of no constituent part in a factory, or in the cell, can be described fully without reference to the system.

The hardware units within a factory, whether mechanical, electrical, or chemical, do not constitute the factory. These require specialized knowledge to build and are necessary for the factory to function but, in and of themselves, they simply compose a collection. They become part of a factory when their functioning is organized and regulated according to a logical program that integrates and orders their activities in such a way as to produce the desired products and maintain their proper functioning within the overall operation of the factory. If we strip away all of the components—the robots, the computers, the communication devices, the relays, and so on—that is, the units within the factory that could be individually used for any number of purposes, what remains, and what constitutes the factory as an entity, is the regulatory logic that controls the dynamics of the factory.

The same can be said for a cell if we strip away that which is purely physical and chemical. While it is true that transcription factors are required to implement the regulatory cell logic, the chemical interactions involved in the functioning of the transcription factors are a subject for chemistry, in the same way that the electrical impulses that carry the instructions to robots in a factory are a subject for physics.

One can know all of these reactions but be no closer to understanding the livingness of the cell. One can list a multitude of the interactions between molecules within the cell, as one could write down the entire instruction set of a computer, but without the program that regulates the manner in which the instructions are used, there are only the symbols of codes, not functioning codes that convey information.

A first and necessary step for modeling cellular processes and their regulation is, therefore, to begin to consider what level of regulation would be a useful target of study. This consideration is a form of detail triage that must be applied as a consequence of the considerable complexity in some types of cellular regulation. At the simplest level of regulation, the core functions of metabolism deal with the most basic and ubiquitous functions required for the cell to be able to carry out any further function. As would be expected for functions that have been continuously selected for continuity of operation and maximal efficiency for as long as organisms have existed, their regulation is tuned to maximize the operational utility of individual steps. There are some cases where these processes may need a large regulatory adjustment, such as hypoxia, or scarcity of an exogenously supplied carbon source used to derive energy or construct macromolecules. Yet for the most part, adaptations of the processes to variation in source materials and requirements for energy on the input side and energy expenditure and molecular construction on the output side fall well within the capability of adjustments determined at the local level.

There are critical issues that a well-run factory must confront, each with an analogue within the cell. A factory, or computer system, must handle interrupts. A factory is not a closed system. It has inputs and outputs, and it also has unplanned emergencies, such as the failure of some component, the loss of its primary energy supply, or a change in the demand for its products. Interrupts do not occur with fixed regularity, and in this sense every input can be considered an interrupt because the timing of input arrivals cannot be synchronously regulated. In order that the factory not completely shut down for an extended time, which could result in economic failure, its operational program must have procedures for redirecting the activities within the factory to handle the interrupt until such time that the effects of the interrupt have passed and the factory can return to normal function, a sort of homeostasis. Consequently, a factory's activities must be modeled stochastically. Basically, while all nodes within the factory may appear to function with complete regularity under “normal” functioning, in fact, the factory is affected by many latent variables unaccounted for no matter how complex the model and these give the factory a stochastic character.

Cells have a similar approach to managing their responses to critical changes that could lead to death or substantial damage to the organism. These actions fall in the category of stress responses, system-wide alterations that deal with various environmental insults and cellular malfunctions. In these situations, many processes may need to be halted and many others instituted. A familiar example of this kind of regulation is the response to damage from ionizing radiation. At this level of regulation, the concept of cellular context becomes evident. An organism has many different kinds of cells and reacts to damage to them in quite different ways. Cells that produce the precursors of short-lived cells needing frequent replacement, such as blood cells or cells that line the gut, are typically much more likely to invoke a death response when their DNA is damaged than cell types like neurons, which are very slow to replicate. The same stimulus is interpreted in different ways in these cells even though the mechanism of recognizing the damage is the same. In these cases, the interpretation of the recognition of damage is conditioned by interactions with different genes present in the different cell types. In regulatory processes, where chains of signals are used to induce systemic changes in the functions a cell is currently performing, the presence or absence of particular gene products that mediate the turning on and off of the production or function of the gene products targeted by the regulatory action can be used to specify whether one or another particular bank of genes will be acted on and whether the action will be an induction or cessation of their action. This capacity to use a single detector of a particular environmental shift to specify differing particular responses for cell types posing distinct types or levels of risk to the organism in their reaction to a threatening environmental action or an internal malfunction is one of the ways in which cells have developed to provide the organism in which they reside with the optimal response to a particular type of damage.

It is common for a factory, or computer system, to function on a clock, meaning that the timing of activities is quantized so that all activities are begun and completed in discrete time intervals, 0, τ, 2τ, 3τ, ... This is accomplished via delays that hold operations so that new operations begin only at times that are multiples of the basic period, τ. For instance, think of a classical human assembly line, where each task is executed during a time interval $k\tau$ to $(k + 1)\tau$. No matter when within the interval the task is completed, the next task is not begun until the next time interval. Another example, and one highly relevant to cells, concerns processes that require multiple inputs, for instance, the assembly of a product that requires multiple preassembled components. When assembly is clock-regulated, product assembly begins at a clock tick, even if all components are in place. This kind of coordination by a clock results in *synchronous* operation. A more efficient way to operate is to do away with a clock and base all activity on readiness. Product assembly commences once all components have arrived, excess components being held in queues awaiting their part in assembly. In a computer system, this means that instructions are executed once all necessary inputs (data and logical inputs) are ready. This kind of execute-when-ready system results in *asynchronous* operation and, in the context of computation, is sometimes referred to as a *data flow* system. Asynchronous operation is more efficient but also more difficult to control, requiring more complex regulatory logic. Ignoring interrupts, a synchronous system can be modeled deterministically, but an asynchronous system is inherently stochastic because variability in individual operation times is not normalized by clock periods. Nevertheless, if the regulatory logic can handle an asynchronous system, then there is significant gain in efficiency.

A factory requires redundancy to keep operations running smoothly. In the worst-case scenario, the failure of a single operation can halt factory operations. These kinds of catastrophic failures are avoided by eliminating their points of occurrence or, perhaps more simply, by building in redundant operations, for instance, backup generators. Multiple subsystems can be employed with backup capabilities. Optimization of redundancy is nontrivial because too much redundancy renders the overall operation too costly. Redundancy is made more efficient by using subsystems that can perform multiple tasks and therefore be able to serve as backups when needed. If a unit fails, the regulatory logic needs to be able to reconfigure the operations automatically to maintain productivity, albeit perhaps at a reduced level if backup systems are less efficient (and therefore less costly). Fault tolerance is enhanced by the ability for autonomous correction of failures, for instance, by error correction code that checks for faults and corrects these faults when discovered. Regarding the development of control systems to regulate this kind of autonomous reconfiguration, Pugachev writes,

> It is also feasible to distinguish another category of control systems which are capable of analyzing their own operating conditions and using this information to produce an optimum performance. The simplest systems of this type, which incorporate elements for automatically adjusting particular parameters according to an analysis of input and output data, are called *self-adjusting* systems. Complex systems of this kind are capable of adapting themselves completely at each instant to the results of their analysis of external conditions and previous performance. These are said to be *self-organizing*. It is quite clear that no theory of error under average operating conditions is adequate for the design of self-adjusting and self-organizing systems. A special theory is required which will solve the complex problems involved in processing the input data and utilizing it to best advantage in any particular case. Both problems can be tackled by the modern theory of optimal systems. (Pugachev, 1965)

Self-organization is more than simple redundancy. It allows a system to reconfigure itself to achieve optimal (practically, close to optimal) performance under varying conditions.

A fundamental way to achieve redundancy, as well as efficiency, is through the use of parallelism. Parallel assembly of independent components is obviously beneficial, as is regulatory parallelism. If a sequence of signals must be sent to various points in the system to result in a final instruction, then fault tolerance is achieved by sending multiple signals through different paths. If one path is blocked, the signal will still arrive via another. In fact, the final instruction may be assembled from packets of code that have been sent through multiple channels, with the packets including instructions on how they should be assembled at the end point. This approach provides both redundancy and enhanced speed of operation in cases where one channel is slowed owing to too much traffic or technical problems. In such a system, any channel or processor may be carrying or implementing many tasks simultaneously.

Cells also use redundancy and parallelism to deal with damage and malfunction. Redundancy is commonly observed in a cell's response to ionizing radiation. One gene involved in a wide variety of stress related responses is tumor protein 53 (TP53). TP53 serves as a central hub in the network of stress response and it can activate an array of responses, yet it is not always required for the occurrence of such responses. Many stress response processes can be successfully mounted even in the absences of this protein, even though when TP53 is present it drives that particular process. In these cases, other proteins have been identified that are competent to drive the response in TP53's absence. Such redundancy offers a sound way for cells to minimize the risk of failure of a critical function.

To study regulation in metabolic processes, the appropriate experimental designs and analyses will necessarily differ from the designs and analyses used for examining regulation in stress response. In shifting from regulatory relationships that are simple, linear, context-independent, and not highly branched to those that are complex, nonlinear, context-dependent, redundantly represented, and both highly branched and interpenetrating, one must take into account the vastly increased number of ways the process of interest can be configured. Consider the consequences of carrying out a stress response study where the experimental plan is based on linear expectations, such as challenging a cell line with wild-type TP53 and a mutated derivative of that cell line not producing functional TP53, and then determining which genes are induced by radiation in the $TP53^{wt}$ line that were not induced in the $TP53^{mut}$ line. If one were to interpret these results as indicating that only those genes induced in the $TP53^{wt}$ line and not in the $TP53^{mut}$ line are normally dependent on $TP53^{wt}$, one would be substantially in error. The numerous $TP53^{wt}$ responses that can be induced independently of $TP53^{wt}$ through redundant mechanisms would be incorrectly considered to be not normally dependent on $TP53^{wt}$. Confronting this problem requires one to envision a different type of network architecture, where the possible antecedents of a step in a pathway are multiple and produce the same outcome. One way to do so is to formulate a test that asks what happens when the gene of interest is active. If the results in that case are in agreement with the results when it is inactive, it must be considered as a possible controlling gene for which there is a redundant controller. This is obviously is not definitive; however, it will identify realistic possibilities that would not even be considered by an analysis that makes linear assumptions.

Closely related to parallelism is locality: Operating decisions should, wherever possible, be made at the local level. This means that control is distributed throughout the factory. Hierarchical control suffers from at least three serious flaws. First, it is unable to efficiently respond to changed conditions at the local level. If a machine is beginning to operate unsatisfactorily, perhaps needing overhaul or replacement, this is seen immediately at the local level and, presuming the ability to rectify the situation exists at the local level, is most efficiently handled there, or as close to the operational location as possible to maintain overall functioning of the system. If a long chain of command is required to make the decision as to how to proceed, this takes time and can lead to a delay in decision making, and thus a consequent downtime. A second problem with hierarchical control is fragility. The longer the chain of command, the more likely it will be broken along the way and no decision be forthcoming—in an extreme case, the center of the hierarchical regulatory system might fail, thereby bringing the entire factory to a stop. Finally, a long hierarchical chain can result in the decision resting in the hands of a decision maker less qualified relative to the specific machine.

As stated in our discussion about metabolism, the functions most commonly shared and heavily engineered by selection have extremely local regulation of activity. In many cases, the enzymes that carry out a particular function are autoregulatory, having interactions with the metabolite that they produce that let them adjust their level of catalytic activity based on the local abundance of their product. The fineness of this control is sufficient to produce both high levels of rapid adaptability to fluctuations anywhere in the network of operations and a level of stability that centrally driven regulation cannot achieve.

When observing a factory involving many subsystems, the high dimensionality of the operation is typically apparent, but what might be overlooked by the casual observer is the multivariate character of the individual decisions or operations within the overall structure. Multiple inputs are often required before execution. It is not simply that multiple components are required to execute a specific assembly; more significantly, multiple signals are required for a regulatory decision. For instance, there may be numerous sensors detecting changes in performance at various points and a decision to check a unit or pull it out of service may depend on multiple sensory signals. So too might a decision to override the standard control within some part of the factory and change to some specialized logic, for instance, to deal with an interrupt. The incoming signals may be quantized to binary form, so that binary logic is used to evaluate the multivariate information and make a binary decision as to whether action is to be taken. In the case of response to a potentially catastrophic interrupt, information is directed into some control point whose default value is 0 but whose value changes to 1 to canalize part of the system into a reconfigured state of operational control until the threat has passed. A cursory view of control points within the factory might reveal a sequence of changes, thereby giving the impression that the behavior of such a pathway might reveal the regulatory logic; however, although each point in this pathway might influence its successor, it is likely that each point is influenced by multiple signals and that the pathway only represents a trace of this activity along a certain set of points, not a dynamical trajectory in the full state space. In this sense, such a pathway represents marginal knowledge of the operations.

In cells, not only are there multiple inputs involved in a decision, there also are alterations in the hardware components that interpret the inputs, making the responses context-sensitive. In these cases, in addition to having multiple controller genes direct the same operation, a controller gene will now provoke a particular response only part of the time. This occurs when a controlling gene is capable of acting to produce a certain set of regulatory results only when its actions are interpreted by a particular set of gene products that are variably present. If one examines a set of samples in which such variable interpretation is acting, using an analytical approach that measures correlation of gene transcription on the assumption that a gene exerting a controlling effect on a target gene should show strong correlation in expression activity, then the context-dependent controller will likely be overlooked. Its correlation with its contextually controlled targets is only evident when the controller is in the proper context. This is a widespread problem in gene control, since in the cell, every gene being expressed is regulated by other genes and frequently there are multiple regulatory conditions for a gene to be expressed, so that any set of samples is likely to have many genes being controlled by different genes in different samples, thereby making simple correlation a poor way to identify regulatory connections.

An important control mechanism that leads to this kind of confounding occurs with canalizing genes. Just as a factory has control points that canalize components of the system into reconfigured states based under certain critical conditions, Waddington recognized the existence of genes that canalize a biological system into constrained channels of behavior (Waddington, 1942). The constraints he had in mind were those that facilitate reliable development of an organism even when presented with large perturbations. A key characteristic of a canalizing gene is its ability to override other regulatory instructions. In terms of a model of transcriptional consequences, this kind of control leads to a significant number of genes at various transcript abundance levels, and under the transcriptional control of other genes, to move to a single state when the broadly controlling gene is active—thereby exercising overriding control. We are not referring to sequential canalization, whereby a specific action of the master enforces a cascade of actions among a single highly correlated cohort of genes important in a single process, but rather where a gene has such broad regulatory power and its action sweeps across a such a wide swath of processes that the full set of affected genes are not highly correlated under normal conditions.

Canalizing genes are frequently found in signaling pathways that deliver information from a variety of sources to the machinery that enacts central cellular functions such as cell cycle, survival, apoptosis and metabolism. DUSP1 antagonizes the activity of the p38 mitogen-activated kinase, MAPK1 (ERK), which is a central component of the pathway by which extracellular signal-regulated kinases send mitogenic signals (Chang and Marin, 2001); thus, MAPK1 is canalizing in its phosphorylated state and DUSP1 is canalizing when it dephosphorylates MAPK1. As is often the case in such key pathways, there are multiple opportunities for canalizing behavior along the signal transducing pathway. Some of the earliest observations of canalization along the mitogenic pathway involved the RAS gene family, members of which were found, in cancer, to have frequent mutations in their 12th codon that produce uncontrolled proliferation (Tabin and Weinberg, 1985). Another locus of signal integration by a canalizing gene is in the area of stresses to the genome by the TP53 gene. The list of cellular stress responses where this gene exerts strong control continues to expand into new regulatory territories (Gomez-Lazaro et al., 2004). While canalizing genes can be extremely potent, their potency is nonetheless circumscribed by other features of the regulatory apparatus operating in the particular cell where control is attempted. A clear and very instructive example of this is the capability of a translocation gene formed from the fusion of the BCR and ABL genes to induce a cancer phenotype, a canalizing result of extraordinary scope, but an inducement only possible in a very particular subset of precursor cells at a specific point in their differentiation into myeloid cells.

Once a factory exceeds a very small number of interconnected components, coordinating its operations goes beyond a commonsense, nonmathematical approach. Reminded of Galileo's disparagement of words as constituting knowledge in the case of gravity, the situation with a complex system is orders of magnitude more resistant to everyday language and intuition. There are two basic operational issues concerning a factory: characterization and control of its operations. First, we want to characterize the input and output of the factory; second, we want to organize the operations so as to achieve optimal (or at least satisfactory) performance. Both characterization and control require a suitable conceptualization of the factory. Such a conceptualization must be mathematical for two reasons. First, characterization and control involve relations among the components and mathematics provides a relational language, and second, mathematics provides a language in which complexity can be represented in such a way as to be amenable to analysis. Not only are complex systems beyond ordinary intelligibility and intuition—indeed, their performance is often highly counterintuitive—but they also typically cannot be fully represented mathematically because there are too many relations and, even should one achieve a very precise and highly involved mathematical description, it may well be intractable relative to solutions of the problems of interest, such as optimizing some set of relations within the system. Hence, rather than completely characterize system outputs in terms of system inputs, we satisfy ourselves with characterizing properties of the output in relation to certain properties of the input. In such typical situations, we try to select variables on the input side that have big impact on important variables on the output side.

Given that biology appears to be solving its control problems using the same sorts of approaches but doing so in a much more complex environment, it would be reasonable to assume that progress in biological science will require adopting the same stance as engineering has already found useful, and necessary, in dealing with complexity—with the constraints being even more demanding in biology owing to a much greater degree of complexity. Hence, research must focus on finding levels of operation of biological control where prediction based on some input variables produces useful levels of prediction on the output variables.

Just as a factory's constituent parts—electrical, mechanical, and chemical—are required for the factory to exist, the physical-chemical constituent parts of a cell are required for the cell to exist. Moreover, just as the constituent parts of a factory do not constitute the factory, but rather the regulatory (operational) logic of the factory defines the factory as an operational system whose purpose is to consume energy, maintain itself, and produce an output, the constituent parts of a cell do not constitute the cell, but rather the regulatory (operational) logic of the cell defines the cell as an operational system whose purpose is to consume energy, maintain itself, and propagate. For both factory and cell, the regulatory logic determines the relations between the physical structures within the system and between the system and its environment. By regulatory logic we do not simply refer to simple binary deterministic logic but to mathematical functions, possibly binary in nature, that provide operational control within the framework of random processes. The roles of regulatory logic in the factory (or complex machine) and the cell are congruent because the key to the characterization of this logic lies in communication (between components) and control (of components)—that is, in systems theory, which therefore determines the epistemology of the cell.

This basic insight was put forward more than 60 years ago, well before the discovery of the double helix, when, in the original 1948 edition of *Cybernetics: or Control and Communication in the Animal and Machine*, Norbert Wiener stated in regard to himself and Arturo Rosenblueth, his physiologist collaborator, "Thus, as far back as four years ago, the group of scientists about Dr. Rosenblueth and myself had already become aware of the essential unity of the set of problems centering about communication, control, and statistical mechanics, whether in the machine or in living tissue" (Wiener, 1948).

Biology studies relations between molecules (chemical structures), not the molecules or the forces between molecules. The recognition that biological knowledge concerns regulatory logic and the consequent intracell operational organization of molecular structures, as well as, by extension, intercell organization, entails the concomitant recognition that biological systems, in their extraordinary complexity, are beyond everyday intelligibility and intuition. Moreover, it facilitates answers to two fundamental epistemological questions: (1) What form does biological knowledge take? (2) How is biological knowledge validated? The question as to form relates to the type of mathematics involved in modeling the relations that characterize regulatory knowledge. This depends on the nature of the relations being considered;

however, the general mathematical framework will be formed within the theory of stochastic multivariate dynamical processes. Validation depends on the mathematical model constituting the biological knowledge and, since this knowledge concerns operational regulation, validation will involve operational predictions derived from the mathematical regulatory model.

At this point we recall Maxwell's desire for "a mature theory, in which physical facts will be physically explained." The high complexity and fluidity of interactions engendered by biological processes and regulation engenders an analogy, albeit one that is no doubt oversimplified. The changing states of the macromolecules we measure in the cell are like the measurements made by sprinkling iron filings on a surface exposed to an electromagnetic field in the sense that we see the induction of orderly behavior due to the operation of processes that we do not experience directly and we try to characterize those processes according to the simple measurements that we can make. We suffer the same limitations as the physicists: The manifestations of the action of the system we want to study reveal only a shadow of what is going on.

EXAMPLE: A CONTEXTUAL MODEL OF GENE REGULATION

To illustrate key epistemological points, we consider a regulatory model that incorporates latency, context-dependence, distributed regulation, multivariate gene interaction, and stochasticity (Dougherty et al., 2009). Because regulation is parallel and distributed, if one views the cascade of activities resulting from the action of a single regulatory gene, both the strength and specificity of subsequent activities in the cascade may be expected to diffuse through subsequent steps in the cascade. As the regulatory effects propagate, they are progressively modified or limited by interactions with other factors modulating gene transcription.

One can view genes at various positions in a regulatory cascade as being either masters or slaves, keeping in mind that this is a relative characterization and that in certain situations a gene might act as a master, while in others in might act as a slave. If the situation were one of strict, complete control of one gene by another at all times, then gene g being a master of gene g_1 in a binary *ON–OFF* genetic regulatory model would mean that $g \to ON$ implies $g_1 \to ON$, and that $g \to OFF$

Table 4.1
Truth Table Showing the Consequences of Regulatory Inputs from Genes g, g_0, and g_{00} on Genes g_1 and g_2

Contexts	g_0	g_{00}	g	g_1	g_2
C_1	1	1	1	1	1
	1	1	0	0	0
C_2	1	0	1	0	1
	1	0	0	1	0
C_3	0	1	1	1	0
	0	1	0	0	1
C_4	0	0	1	0	0
	0	0	0	1	1

implies $g_1 \rightarrow OFF$. This kind of strict control is not indicative of distributed regulation; indeed, in a distributed environment, $g \rightarrow ON$ would not necessarily imply $g_1 \rightarrow ON$, since g may only be able to set $g_1 \rightarrow ON$ in coordination with other genes.

To illustrate the resulting context-dependent behavior of a model system, suppose genes g_1 and g_2 are fully controlled by genes g, g_0, and g_{00} (which may be in turn regulated, or affected, by other genes in any number of cascades). Table 4.1 shows a possible regulatory structure for five genes and Fig. 4.1 shows a network diagram consistent with this structure. Genes g_0 and g_{00} are not part of the model; however, their states are physically codeterminative along with the model master g of the model slaves g_1 and g_2. The four possible combinations of the states of g_0 and g_{00} determine four possible contexts, C_1, C_2, C_3, and C_4, for the model. Given the context, the relationship of the state of g to that of its slaves is determinative; however, absent knowledge of the context, it is not. If $g_1 = 1$ and $g_2 = 1$ in context C_1, then $g = 1$; if $g_1 = 1$ and $g_2 = 1$ in context C_4, then $g = 0$. It cannot be that $g_1 = 1$ and $g_2 = 1$ in contexts C_2 and C_3.

Conceptually, the regulatory action within the model is viewed as a system with inputs corresponding to the regulating master genes for the slave genes; however, the system is not fully described by the input gene values alone, but by these inputs in conjunction with the context.

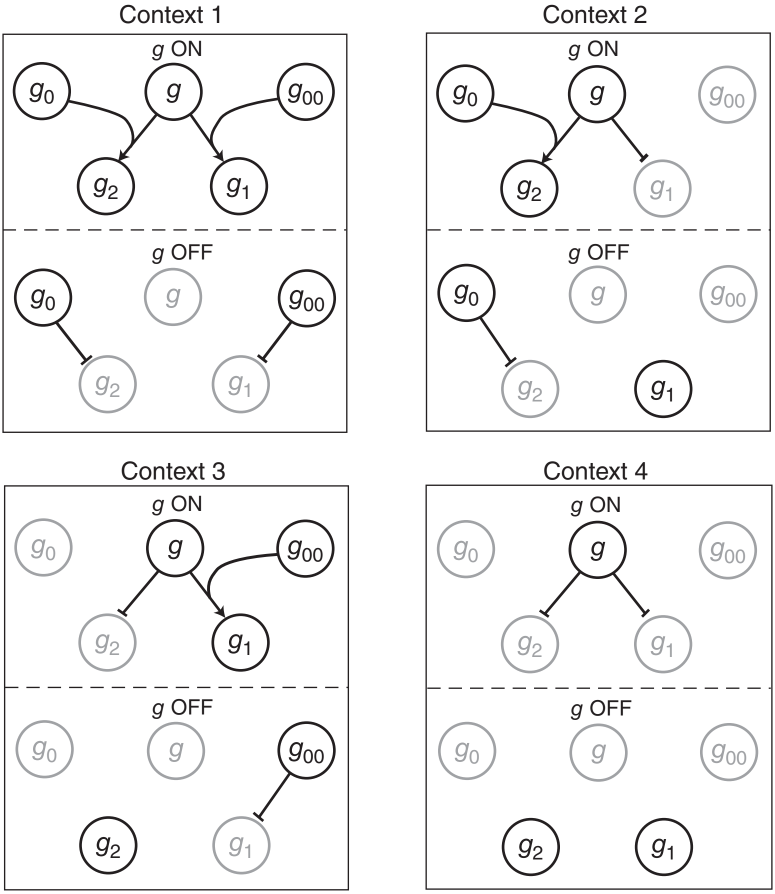

Figure 4.1 Regulatory inputs from genes g, g_0, and g_{00} on genes g_1 and g_2 for four contexts. [Dougherty, E. R., Brun, M., Trent, J. M., and M. L. Bittner, "A Conditioning-Based Model of Contextual Regulation," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6(2), 310–320, April, 2009. © 2009 IEEE].

Biologically, the context is determined by the manner in which the slaves are responding to latent genes external to the model network. Together, the latent genes act in a manner as to select a network (system) context. One can imagine a set of input lines entering the overall system, a family of subsystems (contexts) within the system, and the system output being a single line whose information is selected from among the subsystems. This would be the structure of a computer system whose output is determined by a multiplexer, with the multiplexer's decision being determined by a selection input to it. The model system behaves deterministically so long as it remains in a fixed context.

We now describe the master–slave model (Dougherty et al., 2009), restricting ourselves to a single master gene g and a corresponding set $S = \{g_1, g_2, \ldots, g_r\}$ of slaves (see Dougherty et al., 2009, for a more general formulation). The genes in S may be influenced by genes other than g. Let Y be the binary expression value for g and $\mathbf{X} = (X_1, X_2, \ldots, X_r)$ be the binary-valued expression vector for the slaves. Control by g is of the following form: if $Y = 1$, then all genes in S take on the value 1 with high probability. We let $p = P(Y = 1)$ be the probability that $g \rightarrow ON$.

If $Y = 1$, even though the master is *ON*, context-dependent regulation may affect the slaves. For any slave $g_k \in S$, the conditional probability of g_k being *ON* is given by

$$P(X_k = 1 \mid Y = 1) = 1 - \delta_k, \tag{4.1}$$

where the magnitude of δ_k depends on the extent to which the influence of the master on g_k is diminished by contextual effects. To illustrate the meaning of this conditional probability, consider Table 4.1. Partitioning the probability according to the contexts yields

$$P(X_1 = 1 \mid Y = 1) = \frac{\sum_{j=1}^{4} P(Y = X_1 = 1 \mid C_j)P(C_j)}{\sum_{j=1}^{4} P(Y = 1 \mid C_j)P(C_j)}, \tag{4.2}$$

where $P(C_j)$ is the probability of the context C_j. The size of δ_1 depends on the conditioning of the contexts and their probabilities. Suppose contexts C_2 and C_4 cannot occur, so that $P(C_2) = P(C_4) = 0$. Table 4.1 shows that

$$P(Y = X_1 = 1 \mid C_1) = P(Y = 1 \mid C_1) \tag{4.3}$$

and

$$P(Y = X_1 = 1 \mid C_3) = P(Y = 1 \mid C_3). \tag{4.4}$$

Thus, $P(X_1 = 1|Y = 1) = 1$ and $\delta_1 = 0$. Conditioning with the control that $X_1 = 1$ when $Y = 1$ occurs due to contexts C_2 and C_4, so that if they do not occur, there is no such conditioning. A similar analysis applies to $P(X_2 = 1|Y = 1)$, and in this case conditioning with the control that $X_2 = 1$ when $Y = 1$ occurs due to contexts C_3 and C_4. δ_k is called the *conditioning* parameter.

If $Y = 0$, then the probability that $X_k = 1$ depends on contextual effects when the master is not actively regulating the slaves. We let

$$P(X_k = 1 \mid Y = 0) = \eta_k. \tag{4.5}$$

From Table 4.1, partitioning the probability according to the contexts yields

$$P(X_1 = 1 \mid Y = 0) = \frac{\sum_{j=1}^{4} P(Y = 0, X_1 = 1 \mid C_j)P(C_j)}{\sum_{j=1}^{4} P(Y = 0 \mid C_j)P(C_j)}. \tag{4.6}$$

Again suppose contexts C_2 and C_4 cannot occur. From Table 4.1, we see that

$$P(Y = 0, X_1 = 1 \mid C_1) = P(Y = 0, X_1 = 1 \mid C_3) = 0, \tag{4.7}$$

so that $P(X_1 = 1|Y = 0) = 0$ and $\eta_1 = 0$. A positive value of η_1 means that it can be that $X_1 = 1$ absent the forcing control of the master when $Y = 1$. A similar analysis applies to $P(X_2 = 1|Y = 0)$. We refer to η_k as the *cross-talk* parameter because genes outside the model are turning the slaves on.

The model is determined by the two conditional probabilities defining the conditioning and cross-talk parameters. They characterize our understanding of regulation in the model. If there is very little conditioning and little cross talk, then η_k is substantially smaller than $1 - \delta_k$.

Cross talk poses implications for experimental design. Suppose one takes a large number of samples over unknown contexts. It may be that a master exhibits tight control (perhaps with no external conditioning) across all study samples for which the master is *ON*, but when that master is *OFF*, the behavior of the slaves is controlled by other genes. If under this other control the slaves are uniformly distributed *ON* and

OFF, we have the situation $\eta_k = 0.5$. If the probability mass of the contexts in which $g \to OFF$ greatly outweighs the mass of those contexts for which $g \to ON$, then the determinative effect of *g* on the slaves can be very low across the study samples. Essentially, the experimenter is blinded. Even worse, the experimenter can be severely fooled. If the slaves are mostly *OFF* outside the control of the master, so that the cross-talk parameter is very small, even if the master exhibits little control when it is *ON*, it might well show a stronger determinative effect than a master that exhibits tight control (when *ON*) but has slaves that respond significantly to experimental conditions outside the study examples for which the master is *ON*.

EXAMPLE: THE HYPOTHETICO-DEDUCTIVE METHOD IN CELL BIOLOGY—STRUCTURAL INTERVENTION

In this example, we provide an illustration of the deductive power of the scientific method in the context of cell biology. We will consider a network model for the mammalian cell cycle (Faure et al., 2006) and show how this basic network model (the hypothesis) leads via deduction to important knowledge for cell behavior and the development of therapeutic intervention, to the extent that the model is valid, an issue we do not consider here.

The proposed mammalian cell cycle model is a *Boolean network* model (Kauffman, 1993). In the Boolean model, gene expression is modeled as a binary value, 0 or 1, indicating down- or up-regulation, respectively, the binary value representing expression abundance being below or above some given threshold. The state of the network at any discrete time point, 0, 1, 2, . . . , is given by a vector of 0s and 1s, called the *gene activity profile* (*GAP*), which gives the state of each gene in the network. There are regulatory rules that give the value of each gene at time *t* as a logical function of some set of gene values at time $t - 1$. If there are *n* genes in the network, then there are 2^n states. As a Boolean network evolves through time, it eventually reaches some set of states through which it cycles endlessly, this set of states being known as an *attractor cycle* and each state in an attractor cycle being known as an *attractor state*. The attractor cycles of a Boolean network characterize its long-run behavior.

A *probabilistic Boolean network* (*PBN*) is a collection of Boolean networks in which the PBN is governed by the regulatory logic of one of the constituent networks, known as *contexts*, until a random switching variable calls for a change, in which case a context is randomly selected for governing the PBN until another switch is randomly called for (Shmulevich and Dougherty, 2010). The random switches represent changes in latent variables outside the model, just the kind of effect previously discussed in the contextual modeling of gene regulation. The PBN model is further randomized by allowing the possibility of random gene perturbations with some small probability. The general definition of a PBN does not restrict gene values to binary values, so that its contexts can be more general than Boolean networks; however, for simplicity we will assume binary context networks, that is, Boolean network contexts.

The PBN regulatory and probabilistic structures result in the model being a Markov chain in which the state transitions are governed by an $m2^n \times m2^n$ transition probability matrix, where m is the number of contexts. Specifically, if we label the $m2^n$ states, each of the form $(\kappa, \mathbf{x})$, where κ is a context and $\mathbf{x}$ a GAP, by $1, 2, \ldots, m2^n$, if the network is in state j at time $t - 1$, then there is a probability p_{jk} of transitioning to state k at time t and this probability can be derived from the defining network structure. Letting $N = m2^n$, the transition probability matrix is defined by

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1N} \\ p_{21} & p_{22} & \cdots & p_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ p_{N1} & p_{N2} & \cdots & p_{NN} \end{pmatrix}. \tag{4.8}$$

There exists a steady-state distribution giving the probabilities of being in each of the $m2^n$ states in the long run, meaning that there are probabilities $\pi(1), \pi(2), \ldots, \pi(m2^n)$ such that, no matter what state the network is currently in, the probability of being in state j after a very large number of transitions converges to $\pi(j)$. By definition, the attractors of a PBN consist of all attractors among its contexts. These play a major role in determining steady-state (long-run) behavior.

Returning to the mammalian cell cycle network, for a normal mammal, cell division coordinates with growth in a process tightly controlled via extracellular signals indicating whether a cell should divide or remain in a resting state. The positive signals (growth factors) instigate the activation of the key gene Cyclin D (CycD). Two other important genes are retinoblastoma (Rb) and p27. Rb is a tumor-suppressor gene expressed in the absence of the cyclins that inhibits Rb by phosphorylation. Gene p27 is also active in the absence of the cyclins. Whenever p27 is present, it blocks the action of CycE or CycA and Rb can also be expressed, even in the presence of CycE or CycA. Hence, it stops the cell cycle. In this wild-type cell cycle model, when p27 is active, the cell cycle can be stopped in cancerous situations.

Following a proposed mutation in Faryabi et al. (2008), assume p27 is mutated and always off. In this cancerous scenario, p27 can never be activated. This mutation introduces a situation where both CycD and Rb might be inactive. As a result, in this mutated phenotype, the cell cycles in the absence of any growth factor. We consider the logical states in which both Rb and CycD are down-regulated as undesirable states, when p27 is mutated. Table 4.2 summarizes the mutated Boolean functions derived from the functions given in Faryabi et al. (2008), where a PBN consisting of nine genes, CycD, Rb, E2F, CycE, CycA, Cdc20, Cdh1, UbcH10, and CycB, is constructed based on the regulatory logic of Table 4.2. The illustration of the relationship between these genes in the PBN is shown in Fig. 4.2. The above order of genes

Table 4.2
Logical Regulatory Functions for Mutated Boolean Cell Cycle Network

Order	Gene	Regulating Function
x_1	*CycD*	Extracellular signals
x_2	*Rb*	$\overline{CycD} \wedge \overline{CycE} \wedge \overline{CycA} \wedge \overline{CycB}$
x_3	*E2F*	$\overline{Rb} \wedge \overline{CycA} \wedge \overline{CycB}$
x_4	*CycE*	$E2F \wedge \overline{Rb}$
x_5	*CycA*	$(E2F \vee CycA) \wedge (\overline{Rb} \wedge \overline{Cdc20} \wedge (\overline{Cdh1 \wedge UbcH10}))$
x_6	*Cdc20*	$CycB$
x_7	*Cdh1*	$(\overline{CycA} \wedge \overline{CycB}) \vee Cdc20$
x_8	*UbcH10*	$\overline{Cdh1} \vee (Cdh1 \wedge UbcH10 \wedge (Cdc20 \vee CycA \vee CycB))$
x_9	*CycB*	$\overline{Cdc20} \wedge \overline{Cdh1}$

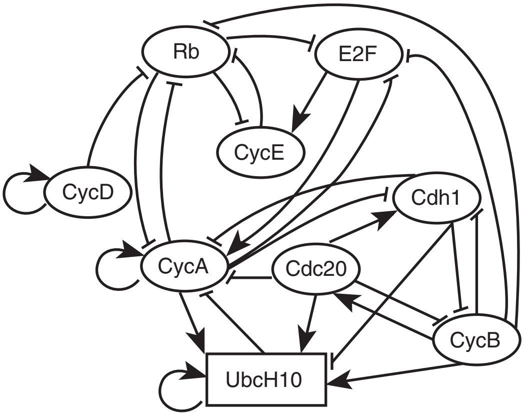

Figure 4.2 Logical regulatory graph for the mammalian cell cycle network: blunt arrows stand for inhibitory effects; normal arrows stand for activations. [Qian, X., and E. R. Dougherty, "Effect of Function Perturbation on the Steady-State Distribution of Genetic Regulatory Networks: Optimal Structural Intervention," *IEEE Transactions on Signal Processing*, 56(10), Part 1, 4966–4975, October, 2008. © 2009 IEEE].

is used in the binary representation of the logical states, with CycD as the most significant bit and CycB as the least significant bit. It is assumed that the extracellular signal to the cell cycle model is a latent variable. The growth factor is not part of the cell and its value is determined by the surrounding cells. The expression of CycD changes independently of the cell's content and reflects the state of the growth factor. Depending on the expression status of CycD, one of two context Boolean networks is obtained. The first context is determined from Table 4.2 when the value of CycD is equal to 0. Similarly, the second context is determined by setting the value of CycD to 1. In this cancerous scenario, a good therapeutic strategy would be to intervene in such a way as to avoid the states with simultaneously down-regulated CycD and Rb. Ignoring transient states and focusing on long-run behavior, this means intervening so as to reduce the steady-state probability mass of such states.

Suppose our intent is to intervene by making a one-bit perturbation to one of the logical regulatory functions, there being 18 such functions in all, nine for each context (note the use of the word "perturbation" in two senses, one representing a random change of gene value and the other an alteration to the regulatory logic). A one-bit perturbation alters the transition probability matrix because it alters gene regulation. To choose the optimal one-bit change, one can employ the mathematical theory developed in Qian and Dougherty (2008) that provides an analytic expression giving the new steady-state distribution after a one-bit intervention in terms of the original steady-state distribution and the original transition probability matrix. For each one-bit function perturbation, we use this theory to find the amount of beneficial shift in the steady-steady distribution, the beneficial shift being the gain in probability mass across all the desirable states. As found in Qian and Dougherty (2008), the two best one-bit perturbations are obtained with the regulatory functions for Rb and E2F in the second context, with these changes reducing the undesirable steady-state mass to 0.0346 and 0.0380, respectively. While this solves the mathematical problem in the abstract, as always one has to take other interests into account. For instance, we must choose a perturbation that can be physically implemented and we should choose one that alters the individual steady-state probabilities as little as possible while obtaining the most overall gain in the probability of desirable states. There are many possible criteria, the key being analysis within the framework of a basic mathematical model constituting biological knowledge (see Qian and Dougherty, 2008, for a more detailed discussion).

The intent in this example has been to demonstrate the hypothetico-deductive method in the framework mathematical model characterizing regulatory logic in a cell. The basic scientific model is a gene regulatory network, in this case, a PBN. From the model, one can deduce the steady-state distribution. Although we have not done so here, one can also quantitatively characterize long-run network sensitivity to alterations in the network (model) structure (Qian and Dougherty, 2009). Biological knowledge is represented by both the network model and its steady-state distribution. These represent biological knowledge just as Maxwell's equations and relations deduced therefrom represent physical knowledge. In the case of biology, the knowledge concerns regulation and system dynamics. Shifts in the steady-state distribution resulting from function perturbations also represent biological knowledge. Moreover, these can be used in a translational sense to arrive at therapeutic intervention strategies to alter the steady-state distribution so as to lower the probability of entering a cancerous state.

The model we have described incorporates only quantized gene expression and discrete time. It can be considered an approximation to a finer model, say, one involving genes and regulatory proteins (transcription factors), real values for the expression levels of genes and proteins, and continuous time rather than discrete time. In this case, one would have a differential equation model. However, to apply computational methods one may choose to discretize a continuous model to arrive at a matrix model, which would be an approximation of the continuous differential equation model (Ivanov and Dougherty, 2006). One could also reduce a gene–protein model to just a gene model (Goutsias and Kim, 2004).

EXAMPLE: INTRINSICALLY MULTIVARIATE PREDICTIVE (IMP) GENES

A key characteristic of a canalizing gene is its ability to override other regulatory instructions. In terms of a model of transcriptional consequences, this kind of control leads to a significant number of genes at various transcript abundance levels, and under the transcriptional control of other genes, to move to a single state when the broadly controlling gene is active—thereby exercising overriding control. This will result in a considerable change in the predictability of the controlling gene by those genes it controls. When not active, the controlling gene will not be predictable to any significant degree by its subject genes, either alone or in groups, since their behavior will be highly varied relative to the inactive controlling gene. When the controlling gene is active, its behavior may not be well predicted by any one of its targets, but can be very well predicted by groups of genes under its control. This property of being intrinsically multivariate predictive (IMP) can be characterized mathematically. In this example, we briefly describe the mathematical model and provide a biological example relating to the canalizing gene DUSP1. Although not necessary, for simplicity we assume that gene values are quantized to the binary values 0 and 1.

We utilize the *coefficient of determination* (*CoD*), which measures the degree to which the transcriptional levels of an observed gene set can be used to improve prediction of the transcriptional level of a target gene relative to the best possible prediction in the absence of observations. Formally, given two predictor random variables, X_1 and X_2, and a target random variable, Y, the CoD of the pair (X_1, X_2) with respect to Y is defined by

$$CoD_Y(X_1, X_2) = \frac{\varepsilon_Y - \varepsilon(X_1, X_2)}{\varepsilon_Y}, \tag{4.9}$$

where ε_Y is the error of the best predictor of Y in the absence of other observations and $\varepsilon_Y(X_1, X_2)$ is the error of the best predictor of Y based on observations X_1 and X_2. (Dougherty et al., 2000). The pair (X_1, X_2) is said to be *IMP* for Y with respect to λ and δ, for $0 \leq \lambda < \delta \leq 1$, if the maximum of $CoD_Y(X_1)$ and $CoD_Y(X_2)$ is less than or equal to λ and $CoD_Y(X_1, X_2) \geq \delta$ (Martins et al., 2008). The definitions of both the CoD and IMP extend directly to more than two predictor variables, but here we will restrict ourselves to two predictor variables. The mathematical properties of IMP predictors have been studied in regard to various issues such as predictive power, predictor logic, and the entropy among the predictors.

We now illustrate the manner in which intrinsically multivariate prediction relates to the canalizing gene DUSP1. DUSP1 can serve as an effective antagonist to a variety of processes stimulated by activated MAPK, including both proliferation and apoptosis (Smalley, 2003) and it has been recognized to be a functional antagonist to chronic replication driven by growth factors (Noguchi et al., 1993). In its role as a guard against the high levels of proliferation frequently associated with cancer in mature, differentiated tissue, DUSP1 acts by dephosphorylating a mitogen-activated kinase, MAPK1, a protein that serves as an integration point for a diverse set of cellular processes in addition to proliferation. A significant fraction of the downstream transcriptional consequences of the dephosphorylation of MAPK1 signaling derives from the inability of the dephosphorylated MAPK1 to activate transcription factors by phosphorylating them. Thus, a DUSP1-induced change in MAPK1 phosphorylation status is expected to have a very significant effect on the abundance of the transcripts of many genes.

The data set employed contains 31 melanoma samples with 587 gene expression measurements (Bittner et al., 2000). Gene expressions are binarized to indicate change or no change relative to a reference expression level for each gene individually. A change can be under- or over-expression. Both cases are labeled as 1, whereas no significant change from the reference is labeled as 0. Treating each gene as a target, and given λ and δ, the objective is to obtain a list of predictor pairs that are IMP with respect to λ and δ for the gene. For $\lambda = 0.2$ and both $\delta = 0.7$ and 0.8, DUSP1 has the largest number of IMP predictor pairs

Table 4.3
List of IMP Pairs for DUSP1 with $\lambda = 0.2$ and $\delta = 0.8$

X_1	X_2	*Logic*	CoD_1	CoD_2	CoD_{12}
RTN1	TEAD1	0001	0.0000	0.0000	1.0000
CHN1	TOP1	0001	0.0000	0.1429	0.8571
CASP3	STOM	0001	0.1429	0.0000	0.8571
EDG1	TEAD1	0001	0.0000	0.0000	0.8571
MMP3	TEAD1	0001	0.0000	0.0000	0.8571
TGFB1	FOS	0001	0.0000	0.0000	0.8571
UAP1	T0P1	0100	0.0000	0.1429	0.8571
TCF4	TEAD1	0001	0.0000	0.0000	0.8571
T0P1	SERPINE1	0010	0.1429	0.0000	0.8571
T0P1	TEAD1	0001	0.1429	0.0000	0.8571
T0P1	PLOD2	0010	0.1429	0.0000	0.8571
LAMA4	PCAF	0001	0.0000	0.0000	0.8571
SERPINE1	PSFL	1000	0.0000	0.0000	0.8571
IFIT1	TEAD1	0001	0.0000	0.0000	0.8571
NR4A3	FOS	0001	0.0000	0.0000	0.8571
CYP27A1	TEAD1	0001	0.0000	0.0000	0.8571
CYP27A1	ESTs	0010	0.0000	0.0000	0.8571
PCAF	FOS	0001	0.0000	0.0000	0.8571
ESTs	FOS	0001	0.0000	0.0000	0.8571

among all the genes, 176 for $\delta = 0.7$ and 19 for $\delta = 0.8$. The extent of DUSP1's control is exemplified by the fact that 21 different genes appear in the 19 IMP predictor pairs for $\lambda = 0.2$ and $\delta = 0.8$. These predictor pairs, along with the relevant CoDs, and the optimal predictor logics are shown in Table 4.3 (Martins et al., 2008). The logic codes refer to the following predictor logics: $Y = X_1 \wedge X_2$ (0001), $Y = X_1 \wedge \overline{X}_2$ (0010), and $Y = \overline{X_1 \wedge X_2}$ (1000). In the most extreme case, the CoDs for RTN1 and TEAD1 individually predicting DUSP1 are both 0 but the CoD for their joint prediction of DUSP1 is 1.

CHAPTER 5

Translational Science

> The intention and the result of a scientific inquiry is to obtain an understanding and a control of some part of the universe.
>
> —*Arturo Rosenblueth and Norbert Wiener*

More than providing predictions of the future, science provides the basis for controlling the future. Scientific knowledge is operational knowledge; indeed the operational definitions of a scientific theory provide the means of validation by relating the abstract symbols of the mathematical model to physical operations. Transforming the mathematical model therefore corresponds to transforming the physical world. In this sense, scientific knowledge is translated into action. *Translational science* transforms a mathematical model, whose purpose is to provide a predictive conceptualization of some portion of the physical world, into a model characterizing human intervention (action) in the physical world. Scientific knowledge is translated into practical knowledge by expanding a scientific system to include inputs that can be adjusted to affect the behavior of the system and outputs that can be used to monitor the effect of the external inputs and feed back information on how to adjust the inputs (Dougherty, 2009a).

The scientific enterprise is pragmatic, its conception of truth being based on predictions in the future, so that scientific knowledge is contingent, always open to refutation by new observations. Translational science goes further. It aims to characterize intentional intervention in

Epistemology of the Cell: A Perspective on Biological Knowledge, First Edition.
Edward R. Dougherty, Michael L. Bittner.

the physical world for the purpose of attaining a desired end. Since any physical action upon a physical system resulting from human action must be understood in terms of measurements relating to those physical actions, the translational scientific model is itself a scientific model. It is the purpose to which a model is put that makes it translational. Indeed, for the applied scientist, a model is ipso facto translational because the intent is to use it to accomplish some end. In this vein, Rosenblueth and Wiener write, "The intention and the result of a scientific inquiry is to obtain an understanding and a control of some part of the universe" (Rosenblueth and Wiener, 1945), where for Wiener "understanding" means a mathematical model. For them, science and translational science are inextricably linked, the ultimate purpose of acquiring scientific knowledge being to translate that knowledge into action.

The conceptualization of a transformation of a physical process takes the form of a mathematical operator on some mathematical system, which itself is a scientific model for the state of nature absent the transformation. There are two basic operator problems concerning systems. The first is *analysis*: Given a system and an operator, what can be said about the properties of the output system in terms of the properties of the input system? It might be mathematically difficult to characterize completely the output system given the complete input system or we may only know certain properties of the input system, so that the best we can hope for is to characterize related properties of the output system.

The second basic operator problem, the one most relevant to translational science, is *synthesis*: Given a system, we would like to design an operator to transform the system in some desirable manner. Whereas the purpose of science, absent translation, is to gain knowledge of the natural world, translational science is about changing it, and synthesis is the act of designing operations to make those changes. Synthesis represents the critical act for human intervention and forms the existential basis of engineering. One could proceed in a trial-and-error manner, trying one operation after another and observing the result. In this case, the operator is not constructed based on knowledge of the scientific system and synthesis is not part of translational science; rather, it is a form of groping in the dark, where one tries one operation after another in the hope of getting lucky, operator mining instead of data mining. Such groping in the dark does not preclude analysis, and therefore does not preclude translational scientific knowledge; however, the critical engineering aspect, that being operator creation for the purpose of transforming nature, is not translational in the scientific sense.

For synthesis to properly occur within translational science requires that synthesis begin with a mathematical theory constituting the relevant scientific knowledge and the theory be utilized to arrive at an optimal (or close to optimal) operator for accomplishing the desired transformation under the constraints imposed by the circumstances. In that sense, translational scientific synthesis, which is synonymous with modern engineering, begins with optimal time series filtering in the classic work of Kolmogorov (1941) and Wiener (1949)—although the latter was published in 1949, an unpublished version appeared in 1942. One begins with a scientific model and expands the model by adjoining operators with which to desirably alter the behavior of the original system. A criterion exists by which to judge the goodness of the response and the goal is to find an optimal way of manipulating the system. In the classic Wiener–Kolmogorov theory, the scientific model is a signal and the translational problem is to linearly operate on the signal so as to transform it to be more like some ideal (desired) signal. The synthesis problem is to find an optimal weighting function and the goodness criterion is the mean square difference between the ideal and filtered signals (for a detailed account of the translational nature of the Wiener–Kolmogorov theory, see Dougherty, 2009b).

Synthesis via mathematical optimization within the framework of a scientific model does not mean that one can obtain a corresponding physical transformation, but it does provide both a target for physical design and a benchmark for performance. Pugachev writes,

> The theory of optimal operators does not enable operators to be found directly which can be embodied forthwith in real constructions. It only enables those mathematical operations on input signals to be determined for which the theoretical limit of accuracy is achieved, for given probability characteristics of the mode of operation and noise, having regard to the nature of the problem and the intrinsic properties of the available data. Accordingly, the practical value of the theory of optimal operators consists mainly in the fact that it makes possible the determination of the theoretical optimum towards which the design engineer must strive in designing a real control system. (Pugachev, 1965)

Having conceptualized the translational problem and found an optimal solution within the mathematical formalization of the problem, the scientist and mathematical engineer can now turn to the technological design engineer to build a device that acts in the physical world in a manner corresponding to the optimal operator within the translational scientific model—or at least approximates to a satisfactory degree the action of the optimal operator.

Although there exists no standard protocol, synthesis in translational science generally involves four steps: (1) construct the mathematical model; (2) define the optimization problem; (3) solve the optimization problem; and (4) to the extent possible, physically implement the solution. One might argue that, unless a reliable model exists, costs determined, and an optimal operator implemented, posing and solving the optimization problem is of little benefit. On the contrary, the existence of a translational mathematical system can guide the scientist in building a model that can be fruitfully applied, the theoretical engineer in studying costs and benefits that accrue from certain kinds of actions, and the technological engineer in devising improved devices or treatments. In a properly functioning relationship, the scientist does not hand the engineer a set of data and ask the engineer to find something in it; instead, assuming there is a translational goal, then the overall enterprise should be guided by the goal and this goal should already have led to a carefully designed experiment aimed at elucidating relationships deemed useful for achieving the goal.

For translation, a critical issue is to form the conceptualization at the right level of abstraction. The model must be sufficiently complex to permit the translational problem to be formulated within it to a degree sufficient for the application at hand and it must be simple enough that the translational problem is not obscured by too much structure, the necessary parameters can be well enough estimated, and the optimization is mathematically and computationally tractable. The desire for simplicity drives much of the work of engineers: Reduce (compress) the model to achieve tractability while at the same time keeping sufficient information so that the resulting solution, while suboptimal from the perspective of the full model, is still acceptable. To achieve this end, it is important for the success of the translational enterprise that there be tight interaction between the scientist and engineer when it comes to model complexity.

There can be many ways of formulating a mathematical model constituting the same scientific knowledge, but the *right* representation (mathematical formulation) can be crucial both for recognizing how to pose the translational problem and for solving it. Many problems have stood unsolved for some time until someone discovers the appropriate transformation of the system that puts it in a form more suitable to solution (sometimes very easily once the system is transformed). In addition to facilitating a mathematical solution, the proper formulation may pave the way for successful implementation of the physical system.

Translational science is mathematical engineering, applied mathematics with a translational purpose. Whereas the pure mathematician is motivated by internal mathematical questions, the applied mathematician develops mathematics for science or engineering. Both can be excellent mathematicians; their domains are different, although there is no clear line of demarcation between them. The theoretical physicist and theoretical biologist are of necessity practitioners of applied mathematics. Their expertise is quite different from the experimental physicist or experimental biologist, who must design experiments to answer questions or validate hypotheses arising from theoretical speculation. All of these categories lie on a continuum; nevertheless, they must be recognized because each contributes to scientific knowledge in its own way.

Wiener recognized the difficulties that the mathematical requirement of science and translational science would present for medicine when, in 1948, he wrote,

> It is these boundary regions of science which offer the richest opportunities to the qualified investigator. They are at the same time the most refractory to the accepted techniques of mass attack and the division of labor. If the difficulty of a physiological problem is mathematical in essence, ten physiologists ignorant of mathematics will get precisely as far as one physiologist ignorant of mathematics. If a physiologist who knows no mathematics works together with a mathematician who knows no physiology, the one will be unable to state his problem in terms that the other can manipulate, and the second will be unable to put the answers in any form that the first can understand. Dr. Rosenblueth has always insisted that a proper exploration of these blank spaces on the map of science could only be made by a team of scientists, each a specialist in his own field but each possessing a thoroughly sound and trained acquaintance with the fields of his neighbors; all in the habit of working together, of knowing one another's intellectual customs, and of recognizing the significance of a colleague's new suggestion before it has taken on a full formal expression. The mathematician need not have the skill to conduct

> a physiological experiment, but he must have the skill to understand one, to criticize one, and to suggest one. The physiologist need not be able to prove a certain mathematical theorem, but he must be able to grasp its physiological significance and tell the mathematician for what he should look. (Wiener, 1948)

In the kind of collaboration envisioned by Wiener, the mathematician must be an expert mathematician and possess the ability to criticize and suggest appropriate experiments, and the experimentalist must be an expert scientist and be able to grasp the scientific significance of a mathematical proposition and guide the mathematician in the process of symbolic formalization. Wiener is not calling for so-called multidisciplinary individuals who are neither expert scientists nor expert mathematicians; rather, "each is a specialist in his own field." There is little benefit in a brilliant biologist surrounding himself with assistants possessing only superficial training in mathematics, nor for a first-rate applied mathematician to work with any but the best biologists. Rosenblueth was a brilliant physiologist and he chose to work with the greatest mathematical engineer of the twentieth century. Together they produced seminal work in systems biology (Wiener and Rosenblueth, 1946).

Wiener's conception of collaboration puts strong demands on the educational process. The mathematician cannot be a scientific dilettante carrying his toolbox from one discipline to another searching for some problem that directly fits some existing theory; rather, the mathematician should possess sufficient biological knowledge and experimental proficiency to recommend and evaluate experiments. The mathematician need not know the physical details of carrying out an experiment but must be able to reason from a mathematical model to an operational conception in the physical domain. In the other direction, the biologist need not know the formal mathematical structure behind a theory or possess the ability to prove theorems but should appreciate the manner in which relevant theorems manifest themselves in the physical world. Implicit is that the relevant theorems be understood. While a biologist need know neither the measure-theoretic underpinnings of probability theory nor proofs of theorems in the theory of random processes, because biological knowledge of the cell is constituted within the theory of stochastic dynamical systems, he or she should be familiar with basic definitions and theorems that relate to fundamental cellular processes—for instance, those pertaining to the covariance structure of stochastic processes and the steady-state distribution of a Markov chain. Indeed, how can a biologist appreciate the experimental or theoretical requirements relating to a Markovian regulatory model without knowing the basics? Wiener's investigative paradigm does not preclude the advancement of knowledge when the spheres of the mathematician and experimentalist are disjoint; rather, it proscribes the manner of efficient investigation and the path to deep discoveries. Basically, Wiener posits two requirements: (1) the mathematician and experimentalist are experts in their respective fields; and (2) the domain of intersection between their knowledge sets is sufficiently large to allow fruitful interaction. Education should aim to achieve these goals.

EXAMPLE: EXTERNAL CONTROL IN GENE REGULATORY NETWORKS

In Chapter 4, we considered translational synthesis in the form of structural intervention in a mammalian cell cycle network. The translational character of the intervention is reflected in four aspects of the intervention protocol:

1. The interaction among key genes in the cell cycle was modeled (via a probabilistic Boolean network);
2. An optimization criterion was posited (minimize the undesirable probability mass in the steady state);
3. A class of actions was specified (one-bit perturbations in the regulatory rules); and
4. Mathematical methods were used to find an optimal action (via Markov chain perturbation theory).

Whereas structural intervention involves a single change to the regulatory logic of a gene network, in the present example we will consider translational intervention in the form of external control over time. At each time instant whether and how an external action should be taken to best achieve a desired goal must be decided (Datta and Dougherty, 2007). External control takes advantage of the fact that the dynamic behavior of a probabilistic Boolean network can be modeled by a Markov chain, thereby making intervention in PBNs amenable to the theory of Markov decision processes. Control is generally based on flipping (or not flipping) the value of a control gene. The translational problem is to arrive at a series of intervention decisions whose objective is to decrease the long-run likelihood of states favorable to pathological cell functionality.

To accomplish this goal, the task of finding an effective intervention strategy has been formulated as a classical sequential decision-making optimization (Datta and Dougherty, 2007). Two kinds of costs contribute to measure the goodness of an intervention at any stage in the treatment process: (1) a cost that discriminates between the desirable and undesirable states of the system; and (2) a cost of intervention that quantifies the negative effect of an intervention, say, drug treatment or chemotherapy. These are combined into a total cost per stage and the objective of the decision maker is to minimize the accumulated cost associated with the progression of the network. To wit, given the state of the network, an effective intervention strategy identifies which action to take so as to minimize the overall cost. The devised intervention strategy can be used as a therapeutic strategy that alters the dynamics of aberrant cells to reduce the long-run likelihood of undesirable states favorable to the disease.

In this framework, the translational problem assumes the existence of an external regulator and a binary intervention input $u(t)$ at each stage t. The value, 0 or 1, of the intervention input $u(t)$ specifies the action on a control gene. Treatment alters the status of the control gene. If treatment is applied, $u(t) = 1$, then the state of the control gene is toggled; otherwise, the state of the control gene remains unchanged. Given the cost-per-stage function, the objective is to derive an optimal intervention strategy from among a class of allowable strategies. Essentially, a strategy is a function that, at each stage, takes as input the current state (and perhaps the history) of the network and outputs a decision: $u(t) = 0$ or $u(t) = 1$. The decision maker searches for an optimal strategy that minimizes the expected cost aggregated over the long-run progression of the regulatory network. These kinds of problems have a long history in control engineering (Bertsekas, 1995).

We illustrate the methodology in a case where the intervention objective is based on a study in which experimentally increasing the levels of the Wnt5a protein secreted by a melanoma cell line via genetic engineering methods directly altered the metastatic competence of that

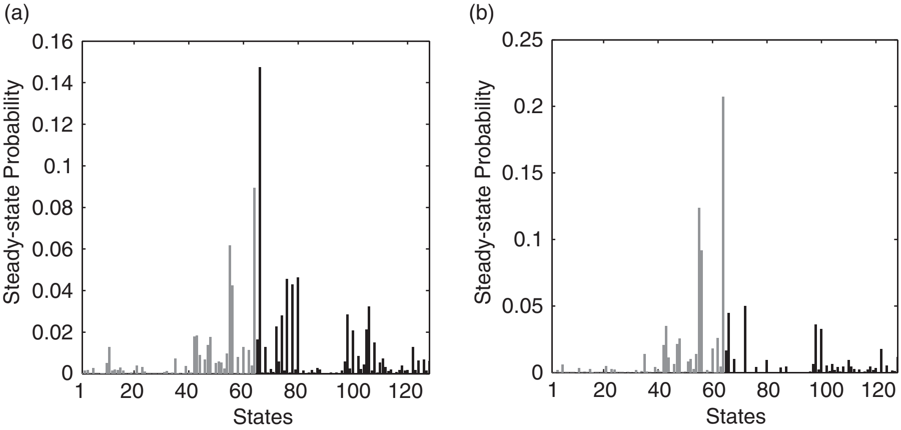

Figure 5.1 Steady-state distribution of PBN: (a) original network; (b) with intervention. Desirable states are gray and undesirable states are black. [Pal, R., Datta, A., and E. R. Dougherty, "Optimal Infinite Horizon Control for Probabilistic Boolean Networks," *IEEE Transactions on Signal Processing*, 54(6), Part 2, 2375–2387, June, 2006. © 2009 IEEE].

cell as measured by the standard in vitro assays for metastasis and in which an intervention that blocked the Wnt5a protein from activating its receptor, the use of an antibody that binds the Wnt5a protein, substantially reduced Wnt5a's ability to induce a metastatic phenotype (Weeraratna et al., 2002). These observations suggest a control strategy that reduces the WNT5A gene's action in affecting biological regulation, because disruption of this influence could reduce the chance of a melanoma metastasizing, a desirable outcome. In (Pal et al., 2006), a seven-gene PBN containing the genes WNT5A, pirin, S100P, RET1, MART1, HADHB, and STC2 was considered. Desirable states were those in which WNT5A was down-regulated, the control gene was pirin, a cost function was defined to reflect the goal and the cost of intervention, and the optimal long-run control strategy was derived by dynamic programming optimization techniques. Figure 5.1(a) shows the original steady-state distribution of the gene regulatory network absent control and Fig. 5.1(b) shows the steady-state distribution of the controlled network. We observe a marked decline in the undesirable probability mass (black) in the controlled network.

The classical intervention optimization just described has two drawbacks. First, it requires inference of the Markov chain associated with the PBN, and second, the computational complexity of the optimization algorithm increases exponentially with the number of genes in the model. Both the inferential and computational impediments to the design of control strategies can be eased by utilizing only partial information regarding the model. The resulting policy may be suboptimal from the perspective of the full model, but it may be all that is possible given the inference and computational requirements of the full model. One such suboptimal long-run policy is based on the mean first passage times of the network and utilizes the heuristic that it is preferable to reach desirable states as early as possible and preferable to leave undesirable states as early as possible (Vahedi et al., 2008). Another such policy checks to see if changing the transition probability matrix for a given control gene in a particular state beneficially alters the steady-state distribution by increasing the steady-state probabilities of desirable states and implements such changes over time (Qian et al., 2009). In the first case, the set of mean first passage times constitutes the partial knowledge required for the algorithm, and in the second, the steady-state distribution constitutes the required partial knowledge.

If we simply focus on the difficulty of inferring a full model, then there are two engineering approaches that can be employed: adaptive control (Kumar and Lin, 1982) and robust control (Nilim and El Ghaoui, 2005). In adaptive control, the model and the control strategy are simultaneously estimated online as the data sequence is input into the adaptive algorithm. Not only does this avoid the need for full model estimation, it also allows the controller to adapt to changes in underlying physical processes that would perturb the model. With robust control, it is assumed that the model is not known with certainty, so that the model is assumed to belong to an *uncertainty class*, and the control strategy is designed to take into account the performance across the entire uncertainty class. For instance, in the case of PBNs, robustness can be with respect to regulation or the effect of latent variables on the model (Pal et al., 2008). Adaptive and robust methods result in control strategies that are suboptimal relative to optimization for the full model, but they often provide good performance when certain knowledge of the full model is not available. Whereas the scientist might not be satisfied with such uncertainty with regard to the scientific model, the translational scientist has an application in mind and poses the problem in a way compatible with the state of partial knowledge.

If the goal is curing patients, then the experimentalist needs to be guided by the engineer to design the kind of experiments that best support the translational goal and the engineer needs to be guided by the clinician as to the kind of effects that the translational solution should provide.

CHAPTER 6

Stochastic Validation: Classifiers

> The elaborate mechanism built on the theory of infinitely large samples is not accurate enough for simple laboratory data. Only by systematically tackling small sample problems on their merits does it seem possible to apply accurate tests to practical data.
>
> —*Ronald Fisher*

Having discussed the role of validation in providing a notion of truth in scientific epistemology, in this chapter we will consider concrete examples of validation in stochastic modeling in order to illustrate various issues that arise. Because both model formation and model validation involve observations, we will see that the two are not unrelated and, in fact, can be strongly dependent. The discussion will focus on particular models because only in that way can the epistemological subtleties of stochastic modeling be made concrete. We will pay much attention to classifier models. There are three reasons for this: (1) they have been studied for many years; (2) they are relatively simple; and (3) although the mathematical issues are technically difficult, they can be described in relatively simple terms, so as to illustrate the epistemological issues. Owing to their fundamental role in biological knowledge, in the next chapter we will consider validation of regulatory network models. Consideration of two model families will illustrate

Epistemology of the Cell: A Perspective on Biological Knowledge, First Edition.
Edward R. Dougherty, Michael L. Bittner.

how validation criteria must be specifically tailored to a particular model family.

To begin, let us return to Newton, in particular, his second law of motion, which relates force and acceleration. Everyday human experience leads us to believe that there is some relation between pushing or pulling a body and the acceleration of the body. A billiard ball is stationary and remains so until hit by another billiard ball, the greater the impact, the greater the acceleration. To obtain scientific knowledge, one can design an experiment in which force and acceleration can be measured and obtain a relation between them by varying the amount of force. For instance, a block can be pulled across a frictionless plane with varying levels of force to produce varying levels of acceleration. It is observed that the force and acceleration are proportional, meaning that there is a relation $f = ma$, where f is the force, a the acceleration, and m the constant of proportionality, called the mass of the body. This equation represents Newton's second law. Force and acceleration are vectors and f is the resultant of all the forces acting on the body.

When the experiment is performed, the equation will not fit the data perfectly. One reason is measurement error. As experimental capability increases, measurement error will be reduced; nonetheless, it will never be completely eliminated. Moreover, even if there were perfect measurement accuracy, a second imperfection in the fit would remain, that being the effect of forces outside those being measured, for instance, friction from the "frictionless" plane, air resistance, and the gravitational effect of Saturn. Try as one might, there will be a host of latent factors that cannot be taken into account. Not only do these latent factors affect the data, they do so in a stochastic manner, such as the changing position of Saturn relative to the experimental setting. Thus, force and acceleration must be treated as random variables. In fact, the situation is more complicated than this because the supposed "constant" of proportionality is not constant, mass depending on velocity. Hence, the mass is changing during the duration of the experiment, so that the range of the experiment affects the distribution of the measurements. In actuality, even absent experimental error, one would not obtain a set of data points lying on the line $f = ma$, but a scattering of data pairs (f, a). Thus, a better scientific model than the equation $f = ma$ would be a probability distribution governing the random variable pair (F, A).

A scientific theory is neither true nor false in an absolute sense. Newton's second law is a good example. The "truthfulness" of the law depends on the accuracy of its predictions and predictive capability determines the validity of the law. Validity is conditional: The conditions under which the law is applied set the range of its predictive accuracy and therefore its degree of "truthfulness." If the body is small, then the gravitational effects of Saturn are negligible; at low velocities mass is virtually constant; and a carefully designed experimental apparatus can minimize effects such as friction. The model $f = ma$ possesses functional validity under certain conditions; that is, if the physical conditions are such that the effects of latent factors are negligible. Depending on one's desire to increase the scope of the relation between force, mass, and acceleration, the complexity of the model must be increased and a probabilistic framework employed to take into account, to the extent possible, the effects of factors outside the model. One might argue that, if one includes all forces acting on the body, so that there are no latent factors, then (assuming we ignore the change of mass) the model $f = ma$ would hold exactly, so that stochasticity is simply a result of ignorance or experimental incompleteness. But this argument would bring us back to Laplace's "sufficiently vast" intelligence. It is an empirically vacuous argument to suggest that all forces could be taken into account. One might make a metaphysical argument but this would be outside the domain of science.

All scientific models are idealizations and the scope of any scientific theory depends upon the intention of the scientist, who decides the portion of Nature to be probed. The wider the field of relations desired, the greater the experimental burden and the greater the complexity of the resulting mathematical model. In his seminal treatise on random geometrical measurements, Georges Matheron states,

> In general, the structure of an object is defined as the set of relationships existing between elements or parts of the object. In order to experimentally determine this structure, we must try, one after the other, each of the possible relationships and examine whether or not it is verified. Of course, the image constructed by such a process will depend to the greatest extent on the choice made for the system of relationships considered possible. Hence this choice plays *a priori* a constitutive role (in the Kantian meaning) and determines the relative worth of the concept of structure at which we will arrive. (Matheron, 1975)

Science is about modeling relations among phenomena and the set of relations in Nature is vast. Hence the scientist's choice of the universe of relations to be examined frames at the outset the very kind of mathematical structure that can potentially result because this choice determines the manner in which Nature is to be probed, that is, the aspects of Nature to be modeled. Ipso facto, the chosen model represents a choice of idealization on the part of the scientist. Since the model constitutes scientific knowledge relative to the phenomenal relations considered, the scientist's choice regarding the relations to be considered represents an a priori decision as to the nature and scope of that knowledge. This choice, if made in the framework of a sound conceptualization of the problem at hand, differentiates a prudently designed experiment from "groping in the dark." Science uses a long-barrel rifle, not a shotgun.

Scientific focus depends not only on mathematical and experimental considerations but also on the interest of the scientist. Consequently, subjectivity enters the scientific enterprise. A virtually unlimited number of experiments can be performed and those relatively few actually performed are somehow determined by psychological, cultural, and metaphysical considerations. Schrödinger writes, "A selection has been made on which the present structure of science is built. That selection must have been influenced by circumstances that are other than purely scientific" (Schrödinger, 1957). The selection is influenced by the interests and goals of the investigator. Schrödinger emphasizes the emotive drive in scientific practice and reinforces the inherent pragmatism of science when he writes, "The origin of science (is) without any doubt the very anthropomorphic necessity of man's struggle for life" (Schrödinger, 1957).

Suppose a biologist believes that a certain transcription factor is related to the expression of a certain gene. To model the relation, an experiment is performed with a particular cell line and measurements are randomly taken across different cells over time to produce a set of measurements, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, where x_i and y_i measure the mRNA abundances for the gene, g_1, governing the transcription factor and the gene, g_2, hypothesized to be under the regulatory control of the transcription factor, respectively. Notice that the model does not involve the transcription factor directly because the measurement process provides gene expression readings. If the abundances are proportional, then one would expect the correlation coefficient between them to be close to 1 and, if this is what the data show, then one might hypothesize a linear relation of the form $y = cx$. This is the situation in

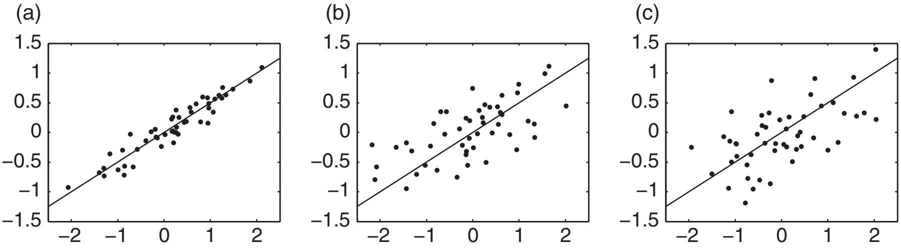

Figure 6.1 Scatter plots with different degrees of dispersion.

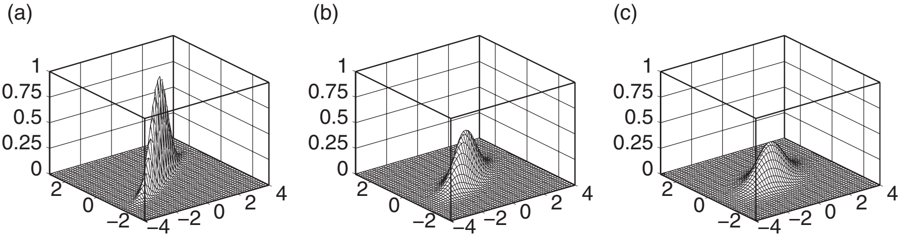

Figure 6.2 Distributions from which the data points of Fig. 6.1 have been randomly sampled.

Fig. 6.1(a), where the data points are close to the line $y = x/2$, and one might proceed under the assumption that the slight variation is mainly due to experimental error. In parts (b) and (c) of the figure, the data points are increasingly scattered and in these instances it would be imprudent of assume the variation is due only to experimental error; instead, one would better conclude that there is a probabilistic relation between the random variables X and Y, and that this relation should be modeled by a probability distribution. Model distributions corresponding to the parts of Fig. 6.1 are shown in Fig. 6.2. In each case, the surface is defined by a probability distribution function $F(x, y)$. The mathematical interpretation of $F(x, y)$ is that, given any region in the (x, y) plane, the probability of a random measurement pair (X, Y) falling in the region is given by the volume under the surface determined by the region. The validity of such a probability distribution depends on the accuracy of such probabilities in regard to future observations: Does the model $F(x, y)$ predict well the distribution of actual measurements?

The same line, $y = x/2$, appears in all parts of Figs. 6.1 and 6.2. If there were a proportional relation between the x and y measurements, then one could say that the line represents a deterministic model reflecting this relationship; however, such a statement makes no sense in a probabilistic model. To understand the meaning of the line, recall Hume's notion of expectation: If we observe event A, then we expect to observe event B, but there is no certainty. Also recall Newton's statement in the *Principia*: "Our purpose is . . . to apply what we discover in some simple cases as principles, by which, in a mathematical way, we may estimate the effects thereof in more involved cases." If there were a deterministic proportional relation between the expression, x, of gene g_1 and the expression, y, of gene g_2, then, given a measurement of the expression of g_1 we could directly compute the expression of g_2. However, in the probabilistic setting, given the expression of g_1, the expression of g_2 is a random variable possessing its own distribution, known as the *conditional distribution* of Y given x. The average value of Y given x is known as the *conditional expectation* of Y given x and is denoted by $E[Y|x]$. In Hume's terminology, this is the value of the expression of gene g_2 that we expect given the expression measurement x for gene g_1. In Newton's terminology, this is our estimate of the expression of g_2 given we have observed expression x for g_1. These terminologies refer to the uncertainty that must arise from focusing on a set of variables and ignoring others, but they refer to different aspects of it. Conditional expectation (Hume) refers to the average value of Y given the observation x, which is a statement about the model [the joint probability distribution, $F(x, y)$ for X and Y]. Estimation (Newton) refers to our predictions regarding future occurrences. Expectation and estimation are tied together because (with suitable mathematical definitions) the optimal estimate of future predictions is the conditional expectation.

For a specific point x, the conditional expectation $E[Y|x]$ is a number. If we let x vary then it becomes a function, which in our examples has been a straight line, but need not be. The conditional expectation function is a partial representation of the full probability distribution. The degree of partiality is characterized by the accuracy of predictions. If the conditional expectation is used as an estimate of future observations, then prediction accuracy for a specific observation x is reflected by tightness of the conditional distribution around the point $E[Y|x]$. This is measured by the *conditional variance*, which is simply the variance of the conditional distribution. From the perspective of the entire distribution of X and Y, the degree of partiality is characterized by the tightness of the full distribution $F(x, y)$ about the line $E[Y|x]$ determined by letting x vary. Referring to Fig. 6.2, the conditional variances increase from part (a) to part (b) to part (c), so that the partial description given by the conditional expectation line becomes poorer.

Continuing with the biological problem at hand, it is much more likely that the relation between gene expressions will be nonlinear in nature, for instance, due to thresholds governing activation and deactivation. Moreover, owing to the high degree of interaction in the regulatory system of the cell and the chemical basis of signal transmission, one is likely to observe a high degree of stochasticity owing to latent factors and variability in transmission timing. Thus, the data are likely to be scattered with no discernable linear relationship between the variables, thereby indicating a joint probability distribution for the gene expressions exhibiting substantial variance and a nonlinear conditional expectation. This is the situation in Fig. 6.3, where the data indicate three distinct regions, depending on the expression level x_1 of gene g_1: (1) a low, essentially constant, basal expression level, y_1, for gene g_2 for $x_1 \leq u_1$; (2) a somewhat linear increase in g_2 expression for increasing g_1 expression for $u_1 < x_1 < u_2$; and (3) a leveling off of g_2 expression

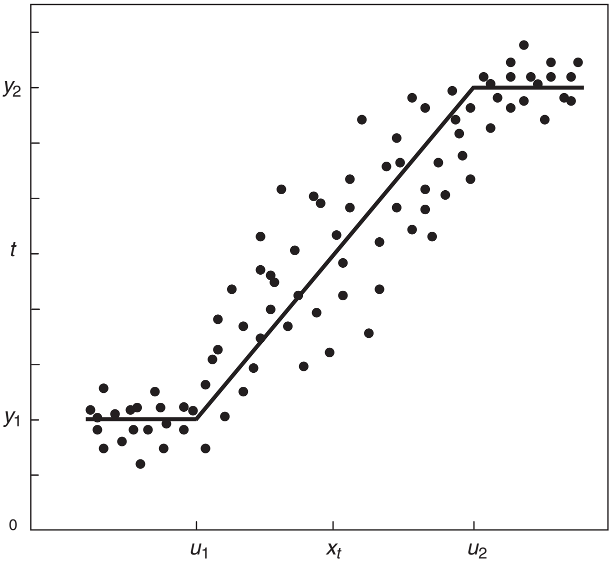

Figure 6.3 Scatter plot and conditional expectation.

at y_2 for $u_2 \leq x_1$, indicating that beyond u_2, greater g_1 expression has little or no effect. The solid line indicates the conditional expectation of a particular probability distribution (not shown) that provides a good model for the data. Note that, according to the data, a good model would have to exhibit small conditional variance in the regions of constant conditional expectation and greater conditional variance in the region of increasing conditional expectation.

Epistemologically, the conditional expectation can also serve as a mathematical model for the random variables X and Y, albeit one that contains less predictive power than the full joint distribution of X and Y. To see this, suppose we ask the following question: What is the probability that a randomly observed pair (X, Y) will fall in some region in the plane? This question is answerable from the joint distribution but it is not answerable from the conditional expectation function or from knowledge of all the conditional distributions. Now consider the following question: For a randomly observed pair (X, Y), if $X = x$, what is the probability that $y_1 \leq Y \leq y_2$? This question is answerable from the joint distribution and from the conditional distribution but it is not answerable from the conditional expectation. Finally, consider the question: For a randomly observed pair (X, Y), if $X = x$, what is the optimal prediction for Y? This question can be answered from the conditional expectation because the optimal prediction is the conditional expectation, as well as from the joint distribution and the conditional distributions because the conditional expectation can be derived from these. In sum, there is increasing predictive power as we move from the conditional expectation, to the conditional distribution, to the joint distribution.

The preceding propositions concerning increasing predictive power are based on the assumption that the full distribution is known, the conditional distributions are mathematically derived from the full joint distribution, and the conditional expectations are derived from the conditional distributions. Scientifically, however, one has a model whose accuracy depends on the accuracy of predictions made from the model and, in practice, certain model parameters are estimated from empirical data. While it may be desirable to model the joint distribution, it may require much less data to obtain an accurate estimate of the conditional expectation function, to the extent that predictions based on the directly estimated conditional expectation are better than those based on the conditional expectation derived from the estimate of the full joint distribution. Although in principle it may be better to have a more complex model that reduces to a simpler model by eliminating or fixing certain variables within the complex model, epistemological judgments concerning model superiority must ultimately rest with predictive capacity.

This issue is prominent in translational science, where often a scientist's conceptualization is a mean approximation, which is the conditional expectation function. Systems are then designed, not on a full probabilistic description of the variables, but on the conditional expectation, both in engineering and statistics, where the conditional expectation function is typically referred to as the regression function. To increase the predictive accuracy, and therefore the performance of any operational system derived from the model, it is common to replace the deterministic model $y = E[Y|x]$ by $y = E[Y|x] + N$, where N is a random variable called "noise." The "noise" term is put in to partially take into account the variability inherent in the joint distribution, which itself is difficult to model. The model $y = E[Y|x] + N$ is "between" the conditional expectation and the full joint distribution. Whether or not this additive noise model is beneficial depends on the degree to which it increases prediction accuracy relative to the deterministic model.

For the reasons mentioned at the beginning of the chapter, we will pay particular attention to classification. Consider the gene expression random variables corresponding to the data in Fig. 6.3 and suppose that, based on the value of X, we are interested in predicting when the regulated gene is expressing above some threshold; for instance, when $Y \geq t = (y_1 + y_2)/2$. A binary decision is to be made: Predict that the expression of the regulated gene will exceed t or not exceed t, depending on the observed value of X. If we label the regions $y \geq t$ and $y < t$ by 1 and 0, respectively, and denote the predicted label by $\psi(X)$, then, from the figure, it looks like a reasonable decision procedure should be of the form $\psi(X) = 1$ if $X \geq x_t$ and $\psi(X) = 0$ if $X < x_t$, where x_t is an appropriately chosen decision boundary on the x-axis. Stated in this manner, this is a binary classification problem. If the regulation were deterministic, then one would simply have to choose x_t to be the point on the x-axis corresponding to the point t on the y-axis and the decision would always be correct. In the current probabilistic setting, there is no value x_t that will give perfect classification accuracy. The misclassification error, which is the probability of an erroneous decision, is the sum of two probabilities: the probability that $Y = 1$ and $\psi(X) = 0$ plus the probability that $Y = 0$ and $\psi(X) = 1$. We would like a boundary point x_t that minimizes this error. If we knew the actual joint distribution of the random variables, then we could find a boundary value to minimize the error, but we do not. We have two choices, use the data to estimate the joint distribution and find the optimal boundary value for this estimated distribution as the boundary value or use the data to directly find a boundary value via some rule applied to the data. Unless there is a very large amount of data, the latter approach is typically better, assuming a good rule can be found, owing to the difficulty of obtaining a good estimate of the true joint distribution. This principle is often stated as classification is easier than distribution estimation. Having obtained a classification boundary, the true error is obtained using the true distribution, which we do not know. Thus, there needs to be a rule to estimate this error. When done, we will have a classifier defined by a boundary point and an estimated error. The epistemological question concerns the manner in which these constitute a mathematical model and the validity of that model. We shall now rigorously define this general classification problem and use it as a detailed case study of model validation.

The binary classification problem involves a random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ and a random variable Y, where Y is constrained to take on the values 0 and 1, called *labels*. $X_1, X_2, \ldots, X_n$ are called *features* and the joint distribution of $\mathbf{X}$ and Y, known as the *feature-label distribution*, takes the form $F(\mathbf{x}, y)$. This distribution provides the most complete description of the joint behavior of $\mathbf{X}$ and Y. Partial description is supplied by the conditional distributions of Y. Since Y is discrete, these are given by the probabilities of $Y = 0$ and $Y = 1$, given $\mathbf{X} = \mathbf{x}$. We denote these conditional probabilities by $P(Y = 0|\mathbf{x})$ and $P(Y = 1|\mathbf{x})$. In this setting, regression is replaced by classification, which is to provide the optimal prediction for the label given the value of the feature vector. A classifier ψ is a binary valued function of the form $\psi(\mathbf{x})$ that serves to predict the value of Y. The error of a classifier is the probability of misclassification. This error, denoted $\varepsilon[\psi]$, is given by a sum of two probabilities: the probability that $Y = 0$ and $\psi(\mathbf{X}) = 1$ plus the probability that $Y = 1$ and $\psi(\mathbf{X}) = 0$. An optimal classifier, known as a *Bayes classifier* for the feature-label distribution, is a classifier possessing minimum error among all classifiers. An optimal classifier is constructed by choosing, for each $\mathbf{x}$, the label having the greater conditional probability, namely, $\psi(\mathbf{x}) = 1$ if $P(Y = 1|\mathbf{x}) \geq P(Y = 0|\mathbf{x})$

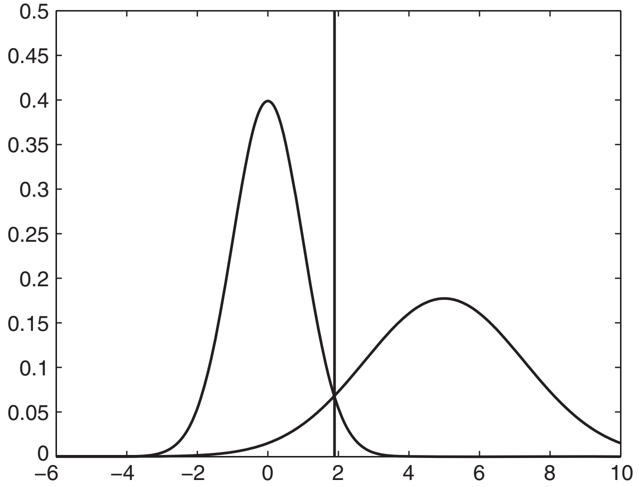

Figure 6.4 Two Gaussian class-conditional densities with the vertical line showing the Bayes classifier.

and $\psi(\mathbf{x}) = 0$ if $P(Y = 1|\mathbf{x}) < P(Y = 0|\mathbf{x})$. There may be more than one Bayes classifier but all possess minimum misclassification error, known as the *Bayes error*. In analogy to the conditional expectation, the feature-label distribution is the complete model and the classifier represents partial description.

Because the labels are binary, a classifier partitions the space of feature vectors into two decision regions: In one region the classifier labels all points 0 and in the other it labels them 1. The shape of the partition depends on the manner in which the labels are distributed over the feature space. These *class-conditional distributions* are defined by the distribution of **x** given label 0 and the distribution of **x** given label 1. Figure 6.4 shows two Gaussian class-conditional distributions for a one-dimensional classification problem. If the classes are equally likely, then the Bayes classifier is defined by value of x where the two Gaussian curves intersect.

From a scientific perspective, the preceding mathematical considerations are the consequence of a primary modeling decision: the relation between the features and the label is to be modeled by a probability distribution. The "real-world" situation is not known and must be approximated by a model created by reasoning concerning scientific theory, estimation from data, or a combination of both. The worth of any constructed feature-label distribution is judged by its predictive capacity, specifically, the agreement of probabilities deduced from the model and their frequency counterparts in experimental observations; nevertheless, a primary epistemological assumption has been made, that being the existence of a feature-label distribution governing the behavior of the features and label, and all further analysis is relative to it. We are not saying that this ideal distribution exists in some Platonic world of forms; rather, it is hypothesized in the mind of the scientist, even though the scientist does not know its specification.

If classification is our interest, then the goal is to obtain a classifier with minimal error among all classifiers and this goal must be expressed relative to the primary assumption. If the hypothesized ideal feature-label distribution were known, then a Bayes classifier and its error could be deduced from it. But it is not known. Therefore, there are two choices: (1) construct an approximating feature-label distribution and deduce a Bayes classifier and its error from this distribution; and (2) directly construct a classifier by reasoning concerning scientific theory, estimation from data, or a combination of both, and estimate its error by some procedure. Note the conundrum. Since one does not know the ideal distribution, how can the goodness of the approximation be judged in the first case and the goodness of the error estimate be judged in the second?

To concretize these remarks, we consider discrimination between phenotypes A_0 and A_1 based on the assumption that the different phenotypes result from production of a single protein λ controlled by transcription factors τ_1 and τ_2. Specifically, when τ_1 and τ_2 bind to the promoter region for gene g, the gene expresses, the corresponding mRNA is produced, and this translates into the production of protein λ, thereby resulting in phenotype A_1; on the other hand, in the absence of either τ_1 or τ_2 binding, there is no transcription and phenotype A_0 is manifested. Letting X_1 and X_2 denote the abundances of τ_1 and τ_2, respectively, and Y the label, 0 and 1 for phenotypes A_0 and A_1, respectively, the primary epistemological assumption is that the behavior of X_1, X_2, and Y is governed by a joint probability distribution.

Suppose it is believed that there exist expression levels κ_1 and κ_2 such that phenotype A_1 is manifested if $X_1 > \kappa_1$ and $X_2 > \kappa_2$, whereas A_0 is manifested if either $X_1 \leq \kappa_1$ or $X_2 \leq \kappa_2$. This assumption could be imposed on the hypothesized ideal feature-label distribution. If so, then we could make the following probability statement regarding the class-conditional distributions for the feature-label distribution:

$$P(X_1 > \kappa_1 \text{ and } X_2 > \kappa_2 \mid 1) = 1. \tag{6.1}$$

The assumption on the transcription factors can be used to construct a classifier. Define $\psi(X_1, X_2) = 1$ if $X_1 > \kappa_1$ and $X_2 > \kappa_2$, and $\psi(X_1, X_2) = 0$ if $X_1 \leq \kappa_1$ or $X_2 \leq \kappa_2$. If these conditions were to strictly hold, then the classifier would have zero error. Owing to concentration fluctuations, time delays, and the effects of other factors, one cannot expect to have such a simple physical situation. Consequently, Eq. 6.1 is not likely to hold and one cannot conclude that the just-defined classifier will have zero error. Moreover, absent specification of the ideal distribution, the error cannot be directly computed. We will return shortly to the issue of error estimation.

Suppose we do not know the thresholds κ_1 and κ_2, only that phenotype A_1 occurs if and only if the transcription factors are both sufficiently expressed. We could proceed by utilizing a procedure, called a *classification rule*, which upon being applied to sample data yields estimates $\hat{\kappa}_1$ and $\hat{\kappa}_2$ of κ_1 and κ_2, respectively, that can be used to construct a classifier. Going further, we might not have any biological knowledge giving us confidence that the classifier should be of the form $\psi(X_1, X_2) = 1$ if and only if $X_1 > \kappa_1$ and $X_2 > \kappa_2$. In this scenario, we need to use a classification rule that assumes some "reasonable" form for the classifier and then estimates the particulars of the classifier from sample data to form a designed classifier.

For illustration purposes, we consider a well-studied classification rule that constructs a hyperplane classification boundary, with values on one side of the boundary being classified into one class and values on the other side being classified into the other class. If the classes are equally likely, then the decision boundary is determined by a discriminant function

$$d_k(\mathbf{x}) = -(\mathbf{x} - \hat{\mathbf{u}}_k)^T \hat{\mathbf{K}}^{-1} (\mathbf{x} - \hat{\mathbf{u}}_k), \tag{6.2}$$

where $\hat{\mathbf{u}}_k$ the sample mean for class k computed from the data from class k ($k = 0, 1$), $\hat{\mathbf{K}}$ is the pooled sample covariance matrix computed from all the data, all operations are matrix-vector operations, and T denotes the transpose. An observed point $\mathbf{x}$ is classified into class 1 if $d_1(\mathbf{x}) > d_0(\mathbf{x})$ and into class 0 otherwise. The classification rule defined by the discriminant $d_k(\mathbf{x})$ is called *linear discriminant analysis* (*LDA*). If both class-condition distributions are Gaussian with common covariance matrix and we replace the sample means and sample covariance matrix in Eq. 6.2 by the actual means and actual common covariance matrix, then the discriminant would produce the optimal boundary between the two classes, this boundary determining the Bayes classifier. In practice, we do not know the actual class-conditional distributions, so we use the sample mean and sample covariance matrix estimates. Moreover, in practice, the assumptions regarding common covariance matrix and Gaussian class-conditional distributions are overly restrictive; nonetheless, the rule can work quite well so long as they are only mildly violated.

Once a classifier has been designed via a classification rule, it is necessary to estimate its error. Relative to the data, error estimation can be approached in two ways. One way is to split the data into independent *training* and *test* sets, with the classifier designed by the classification rule using the training data and the error estimated by the error rate on the independent test data. On account of the test data being held out from the classifier design, this method is known as *hold-out* error estimation. With hold-out, the error estimate is simply the proportion of incorrect classifications on the test data. Hold-out is a direct stochastic counterpart to the classical view of validity: A model exists (having been designed using training data) and it is tested by examining its accuracy on test data.

Given no limitation on data availability, we would like to have large samples for both classifier design and error estimation. This is because classifier design usually improves with larger training samples and error estimation improves with larger test samples. In practice, holding out data from training is often infeasible owing to insufficient data. For instance, in expression-based classification, data are often severely limited, so that holding out test data results in unacceptably poor classifier design. Thus, one has to use a different approach, that being to use all the available data for design and then to apply an *error estimation rule* using the same data. A simple error estimation rule is to count the errors made on the sample data by the designed classifier and estimate its error by the proportion of errors on the sample. This method is called *resubstitution* and, although simple, tends to be biased low, often very low, for small samples, precisely when one wants to use it. Various other error estimation rules for when the data are not split are given in the literature, all with their own benefits and drawbacks for small samples.

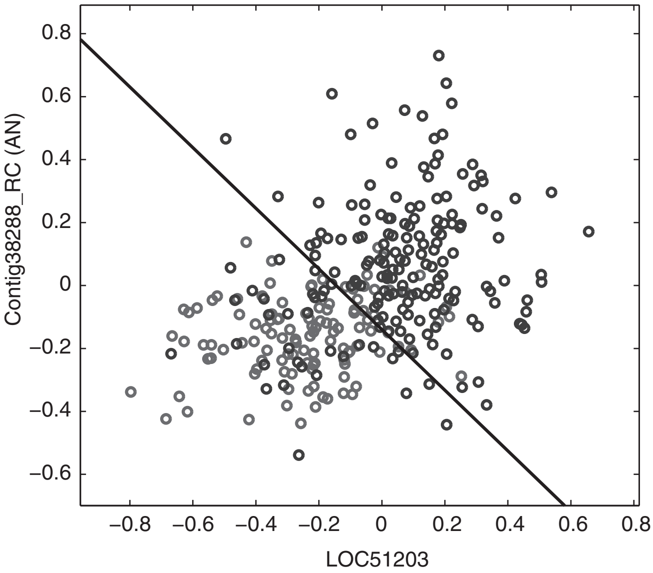

Figure 6.5 Linear classifier separating breast cancer patients with good (gray) and bad (black) prognosis using two genes, LOC51203 and Contig38288_RC (AN). [Braga-Neto, U. M., "Fads and Fallacies in the Name of Small-Sample Microarray Classification," *IEEE Signal Processing Magazine*, 24(1), 91–99, January, 2007. © 2009 IEEE].

We illustrate classification with two real-world examples based on gene expression. The first uses patient data from a study involving microarrays prepared with RNA from breast tumor samples from 295 patients (van de Vijver et al., 2002). Of the 295 microarrays, 115 belong to the good-prognosis class and 180 belong to the poor-prognosis class. From the original published data set, the expression profiles of 70 genes were found to be the most correlated with disease outcome (van't Veer et al., 2002). From among these 70 genes, two genes, LOC51203 and Contig38288_RC (AN), have been found to be the most discriminating for linear classification. Figure 6.5 shows the data for the two genes (gray and black dots representing good and bad prognosis, respectively), along with the LDA boundary for the designed classifier, for which the reported estimated misclassification error is 0.0582 (Braga-Neto, 2007). The classifier and estimated error have been found by classification and error estimation rules, respectively, without splitting the data into training and test data (see Braga-Neto, 2007, for details).

In this case, given such a simple classifier and a fairly large sample being used to design the classifier and estimate the error, there is warranted hope that the designed classifier will have approximately the same error on the ideal feature-label distribution as it does on the sample, where one must keep in mind the primary assumption that the real data, already observed and yet to be observed, follow some ideal feature-label distribution. A key aspect of classification epistemology is to formalize this intuitive notion of approximation.

In practice, much smaller samples are commonplace. We consider a study using expression data from microarrays for 597 genes to identify gene combinations for use as glioma classifiers (Kim et al., 2002). Gliomas are the most common malignant primary brain tumors. These tumors are derived from neuroepithelial cells and can be divided into two principal lineages: astrocytomas and oligodendrogliomas. Using a test set of 25 patients, single genes and two- to three-gene combinations have been identified for distinguishing four types of glioma: oligodendroglioma (OL), anaplastic oligodendroglioma (AO), anaplastic astrocytoma (AA), and glioblastoma multiforme (GM). Figure 6.6 shows a

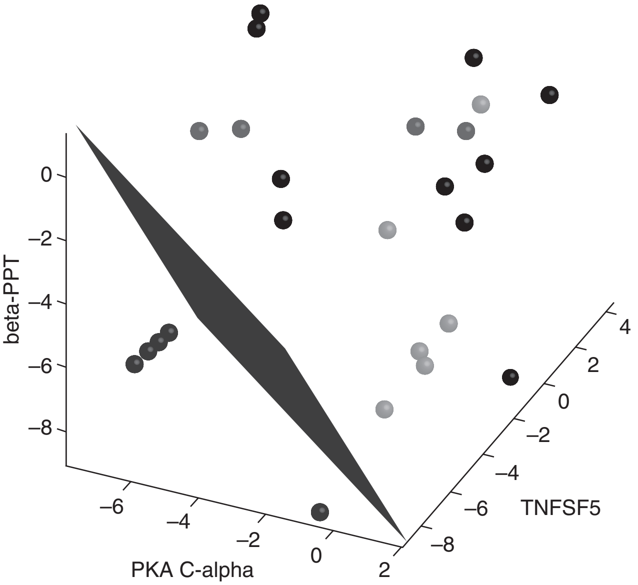

Figure 6.6 Three-gene linear classifier separating AO (dark gray) from three other types of glioma, AA (middle gray), GM (black), and OL (light gray). (Kim et al., 2002, by permission of *Molecular Cancer Therapeutics*).

linear classifier for discriminating AO from the others using the genes PKA C-alpha, TNFSF5, and beta-PPT. The error for the classifier is estimated to be 0.04 using the method of leave-one-out error estimation, which will be discussed shortly. The figure looks promising and the error estimate is small; however, as we will see, leave-one-out error estimation is not reliable for small samples, and 25 patients constitute a very small sample.

Previously, we remarked that it may be better to use a simple model with fewer variables than a more complex model. This situation occurs in classification relative to the number of features. If one has a large number, m, of features available and selects any subset of these containing $k < m$ features, then the Bayes error for the full set of m features cannot exceed the Bayes error for the subset of k features; however, when designing a classifier from sample data, the situation can be quite different. For instance, it is not uncommon for the expected classifier error to decrease as the number of features increases up to a point and then increase as the number of features grows beyond that point. This is known as the *peaking phenomenon* (Hughes, 1968).

We consider data generated from Gaussian class-conditional distributions from which to design the classifiers (see Hua et al., 2005, for details). The maximum number of features is 30, so that the peaking phenomenon can only appear in the graphs when peaking occurs with less than 30 features. The surface in Fig. 6.7(a) shows expected errors for LDA using a model in which the class-conditional distributions share a common covariance matrix and features are slightly correlated. The black line shows the optimal number of features for each sample size. Peaking occurs with very few features for sample sizes below 30, but exceeds 30 features for sample sizes above 90. In Fig. 6.7(b), assuming common covariance matrix and highly correlated the features, even with a sample size of 200, the optimal number of features is only 8. The concave behavior and increasing number of optimal features in parts (a) and (b) of the figure correspond to the usual understanding of peaking. But one must beware of easy generalizations. Figure 6.7(c) shows results using a different classification rule (a linear support vector machine), which also constructs a linear boundary, based on data generated when the covariance matrices are not the same. Not only does the optimal number of features not increase as a function of sample size, for fixed sample size the error curve is not concave. For some sizes the error decreases, increases, and then decreases again as

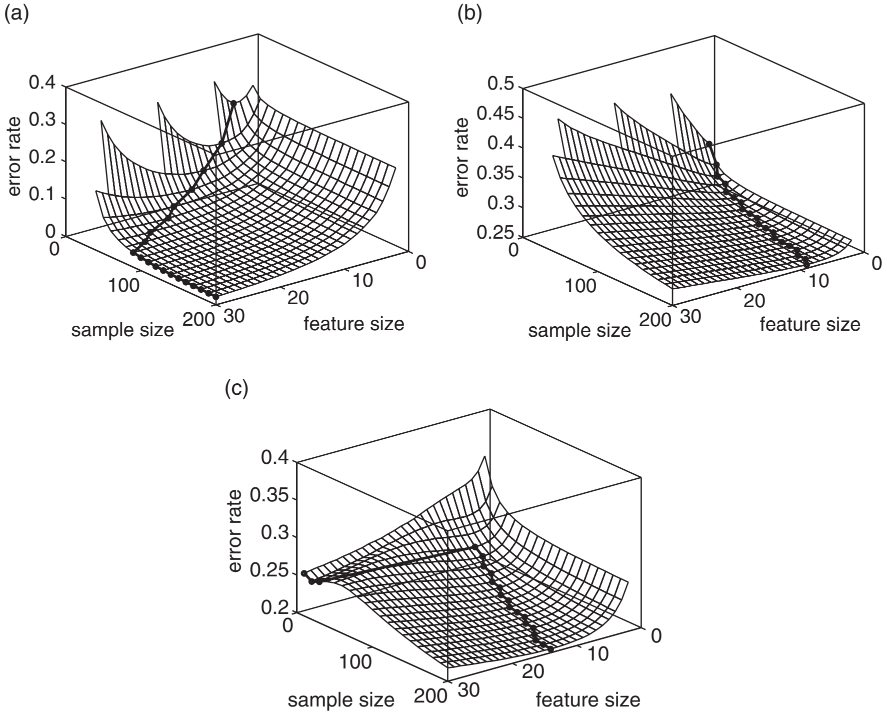

Figure 6.7 Optimal number of features: (a) LDA in model with common covariance matrix and slightly correlated features; (b) LDA in model with common covariance matrix and highly correlated features; and (c) linear support vector machine in model with unequal covariance matrices. [Hua, J., Xiong, Z., Lowey, J., Suh, E., and E. R. Dougherty, "Optimal Number of Features as a Function of Sample Size for Various Classification Rules," *Bioinformatics*, 21(8), 1509–1515, 2005, by permission of The International Society for Computational Biology].

the feature set size grows, thereby forming a ridge across the expected error surface. This behavior is not rare.

Having discussed some general classification issues, we now consider model validity in detail. Formally, we define a *classifier model* $\mathcal{M} = (\psi, \xi)$ to be a pair composed of a $\{0, 1\}$-valued function $\psi(\mathbf{x})$ and a number ξ between 0 and 1. ψ and ξ are called the *classifier* and *error*, respectively, of the model $\mathcal{M}$. The mathematical form of the model is abstract, with ξ not specifying an actual error probability corresponding to ψ. $\mathcal{M}$ becomes a scientific model when it is applied to a feature-label distribution, at which point the question of model validity arises.

There are two issues regarding model validity. The first, which we will refer to as *error validity*, concerns the accuracy of ξ as an estimate of the error of ψ relative to the feature-label distribution. This problem arises because ψ is a decision function that will make the right decision with probability $1 - \varepsilon[\psi]$ and the wrong decision with probability $\varepsilon[\psi]$, where $\varepsilon[\psi]$ is the actual error for ψ on the feature-label distribution. ξ has been obtained in some manner to estimate the actual error rate. The difficulty is that we do not know the feature-label distribution and therefore we cannot find the actual error; indeed, if we knew the feature-label distribution we would simply take ψ to be a Bayes classifier and ξ to be the Bayes error. Thus, we cannot define error validity in terms of the difference between $\varepsilon[\psi]$ and ξ.

In practice, ψ and ξ are estimated from sample data via a classification rule Ψ and an estimation rule Ξ, which together constitute a *rule model* $\mathcal{L} = (\Psi, \Xi)$. The rule model is applied to sample data to obtain a scientific model $\hat{\mathcal{M}} = (\psi, \hat{\varepsilon}[\psi])$, where ψ and $\hat{\varepsilon}[\psi]$ have been determined by the rules Ψ and Ξ, respectively. Since we do not know the actual error, $\varepsilon[\psi]$, of the designed classifier on the feature-label distribution, we cannot measure error validity by the absolute difference $|\hat{\varepsilon}[\psi] - \varepsilon[\psi]|$, between the estimated and actual errors of the designed classifier. Instead, we consider the performance of the error estimation rule.

Based on the primary epistemological assumption, the particular data set, say of size n, from which the classifier and error estimate have been derived via the rule model, is a sample drawn from the ideal feature-label distribution and we will assume that it has been drawn randomly (leaving the precise definition of this to the statistical literature). Consider all possible samples of size n drawn from the feature-label distribution and the corresponding classifiers (constructed via Ψ), actual classifier errors (found from the feature-label distribution), and estimated classifier errors (found via Ξ). Given the feature-label distribution, we can compute for each sample the difference $\hat{\varepsilon}[\psi] - \varepsilon[\psi]$. The population of these differences comprises the *deviation distribution*, which depends on the feature-label distribution, the classification rule, the error estimation rule, and the sample size. A good error estimation rule, within the framework of a given feature-label distribution and classification rule, is one for which the deviations are tightly distributed about 0. Figure 6.8 shows three deviation distributions. The one with the dotted line is tightly concentrated, but not around 0. It suffers from

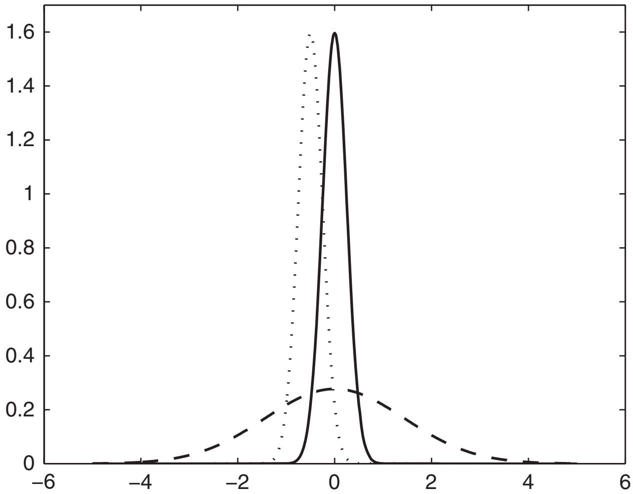

Figure 6.8 Three deviation distributions.

low bias of the error estimation rule, meaning that $\hat{\varepsilon}[\psi]$tends to be less than $\varepsilon[\psi]$. The solid line represents a good performing error estimation rule: It is tightly concentrated about 0. The dashed line depicts an error estimation rule which is unbiased, on average it gives the correct error, but which is widely spread so that individual error estimates tend to be much too large or much too small.

A measure is needed to quantify the tightness of the deviation distribution around 0. One way is to take the average value of the absolute deviations across all possible samples. This gives the expected value of the absolute deviation, $E[|\hat{\varepsilon}[\psi]-\varepsilon[\psi]|]$. A related measure, known as the *root mean square* (*RMS*) difference, is the square root of the expected squared deviation:

$$\mathrm{RMS}=\sqrt{E\left[(\hat{\varepsilon}[\psi]-\varepsilon[\psi])^2\right]} \tag{6.3}$$

The RMS has the useful mathematical property that it can be decomposed into the bias,

$$\mathrm{Bias}[\hat{\varepsilon}]=E[\hat{\varepsilon}[\psi]-\varepsilon[\psi]], \tag{6.4}$$

of the error estimator relative to the true error, and the deviation variance,

$$\mathrm{Var}_{\mathrm{dev}}[\hat{\varepsilon}] = \mathrm{Var}[\hat{\varepsilon}[\psi] - \varepsilon[\psi]], \tag{6.5}$$

according to

$$\mathrm{RMS} = \sqrt{\mathrm{Var}_{\mathrm{dev}}[\hat{\varepsilon}] + \mathrm{Bias}[\hat{\varepsilon}]^2} \tag{6.6}$$

Other measures are possible. Since our interest is epistemological, not the specifics of this or that measure, for the sake of the discussion we will restrict our attention to the RMS. The RMS is a characteristic of the rule model, but it is also dependent on the feature-label distribution, the classification rule, and the sample size. The validity of any classifier model $\hat{\mathcal{M}} = (\psi, \hat{\varepsilon}[\psi])$ relative to the accuracy of the error estimate is judged by the RMS. The basic idea is that, since we cannot know the accuracy of any specific error estimate, we will judge validity based on a relevant characteristic of the error estimation rule. In this way, the RMS provides a measure of validity. If it is small, say 0.05, then this indicates that the error estimation rule is accurate, meaning credence can be given to errors resulting from it; if it is large, say 0.4, then this indicates that the error estimation rule is inaccurate and little credence should be given to any classifier model for which it has been employed in the rule model.

Since we do not know the feature-label distribution, the RMS cannot be computed exactly. All that can be hoped for is that we can find an upper bound on the RMS, this taking the form RMS $\le B$, keeping in mind that the bound also depends on the classification rule. For most error estimators (when such bounds are known), the bound goes to 0 as the sample sizes increase, but it may require a very large sample size to make the bound small.

If hold-out is used, that is, if the data are split into training and independent test data, then there is a very simple bound on the RMS, namely,

$$\mathrm{RMS} \le \frac{1}{2\sqrt{m}}, \tag{6.7}$$

where m is the number of points in the test set (Devroye et al., 1996). This bound gives good results. If the test set has 100 sample points, then RMS ≤ 0.05. For 400 points, the bound gives RMS ≤ 0.025. Moreover, the bound depends on neither the feature-label distribution nor the classification rule. The problem is that for small samples one cannot hold out 100 sample points. If only 25 sample points are held out, then RMS ≤ 0.1, which may be acceptable in some cases. If, however, one has only 75 data points, then holding out 25 points would leave only 50 points for classifier design, thereby significantly increasing the risk of poor classifier design. The trade-off with hold-out is clear: Hold out more points to get better error estimation and get a poorer classifier; hold out fewer points to get a better classifier and get poor error estimation.

The hold-out RMS bound of Eq. 6.7 is very general, in that it makes no assumptions on the feature-label distribution or the classification rule. This is possible because the classifier is designed on the training data and the error is estimated on independent test data. The situation is not so simple when the same data are used for design and error estimation. In addition, with small samples, RMS bounds are often much larger when the data are not split.

Consider the problem of trying to predict whether gene g is activated based on the activations of a set of genes, $g_1, g_2, \ldots, g_b$. By activation and deactivation we mean that there is a threshold value above which a gene is considered to be activated (1) and below which a gene is considered deactivated (0). The general classification problem in this regard is known as *multinomial discrimination*. We use the following classification rule: for each activation-deactivation pattern $(x_1, x_2, \ldots, x_b)$ of 0s and 1s observed in the data, define $\psi(x_1, x_2, \ldots, x_b) = 1$ if $x = 1$ more than $x = 0$, where x is the $\{0, 1\}$-value of g, and $\psi(x_1, x_2, \ldots, x_b) = 0$ otherwise. If a pattern is never observed in the data, the value of $\psi(x_1, x_2, \ldots, x_b)$ can be assigned in any manner whatsoever. For this classification rule we have the following RMS bound, where n is the sample size and resubstitution is the error estimation rule:

$$\mathrm{RMS} \le \sqrt{\frac{6b}{n}} \tag{6.8}$$

(Devroye et al., 1996). If there are six predictor genes and the sample size is $n = 100$, then this bound becomes RMS ≤ 0.6, which is useless. If we increase the sample size to $n = 400$, then the bound becomes RMS ≤ 0.3, which is still not good. For $n = 10{,}000$, the bound becomes RMS ≤ 0.06. To use this bound, a very large sample is needed to insure good error estimation. But this begs the question as to why to use resubstitution at all in this instance. With a sample size of 10,000 the obvious choice would be to split the data into training and test data and not use resubstitution. Indeed, holding out 400 points for testing still leaves 9600 points for training and via Eq. 6.7 results in an RMS bound of 0.025.

The resubstitution error estimation rule, although converging to the true error as the sample size increases, is usually optimistically biased, and severely so when samples are small. An error estimation rule that possesses very little bias is *leave-one-out cross-validation*. This estimate is computed by removing one point from the sample, designing the classifier on the remaining sample points, and applying it to the left-out point. This is done repeatedly for each point in the sample. The error estimate is the fraction of errors made on the left-out points. Like resubstitution, the leave-one-out estimate converges to the true error as the sample size increases. While the leave-one-out error estimation rule is known to have small bias, it is also known to have a large deviation variance for small samples, which means large RMS. Again considering multinomial discrimination, we have the following RMS bound for leave-one-out error estimation:

$$\mathrm{RMS} \le \sqrt{\frac{1+6e^{-1}}{n} + \frac{6}{\sqrt{\pi(n-1)}}} \tag{6.9}$$

(Devroye et al., 1996). If the sample size is $n = 100$, then the bound is approximately 0.601, which is approximately the same as resubstitution with six predictors. If $n = 10{,}000$, then the RMS is approximately 0.184, which is greater than the resubstitution bound of Eq. 6.8.

Although we do not know the feature-label distribution, as discussed in the transcription factor illustration above, we may make assumptions regarding it. Generally, the more we assume about it, the tighter the bound will be on the RMS. In the best-case scenario, the feature-label distribution is modeled in terms of a small number of parameters and the RMS is exactly expressed as a function of these parameters. For instance, if there is a single parameter ω, then the RMS takes the form RMS(ω). Although we do not know the feature-label distribution, we would then know that the RMS is bounded by the maximum value of RMS(ω) over all possible values of ω. The RMS for both resubstitution and leave-one-out with multinomial

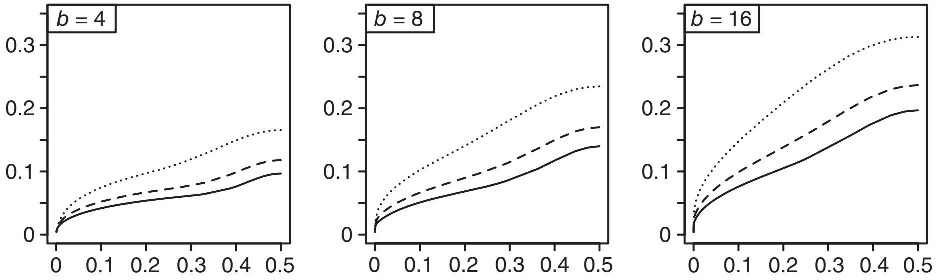

Figure 6.9 RMS versus Bayes error for resubstitution in the Zipf model: $n = 20$ (dotted line); $n = 40$ (dashed line); $n = 60$ (solid line).

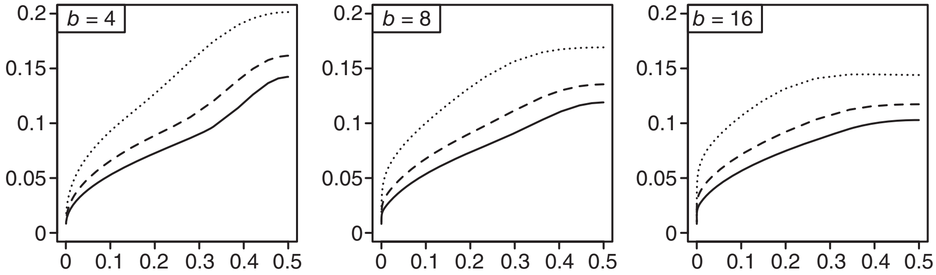

Figure 6.10 RMS versus Bayes error for leave-one-out in the Zipf model: $n = 20$ (dotted line); $n = 40$ (dashed line); $n = 60$ (solid line).

discrimination has been expressed exactly for a Zipf model (Braga-Neto and Dougherty, 2005). Figures 6.9 and 6.10 show the RMS for resubstitution and leave-one-out, respectively, as a function of the Bayes error for $b = 4$, 8, 16, and sample sizes $n = 20$, 40, 60. Note that the RMS is much greater for leave-one-out when $b = 4$, the RMS is much greater for resubstitution for $b = 16$, and there is little difference in the RMS when $b = 8$. Under the Zipf modeling assumption with sample size $n = 40$, for $b = 4$, RMS ≤ 0.12 for resubstitution and RMS ≤ 0.17 for leave-one-out; for $b = 16$, RMS ≤ 0.23 for resubstitution and RMS ≤ 0.12 for leave-one-out. Notice how much better the bounds are under the Zipf modeling assumption as opposed to the model-free bounds of Eqs. 6.8 and 6.9. In practice, the model-free bounds are essentially useless.

When one derives a classifier model from a rule model without a quantitative mathematical statement regarding the validity of the error estimation, for instance, the RMS, then the model is scientifically vacuous. Absent relevant properties of the error estimation rule, the rule is simply a meaningless computation and the scientific requirement of validation is missing. Admittedly, one might have to qualify the classifier model with some assumptions regarding the ideal feature-label distribution, but such qualification is not uncommon in science. After all, Newton assumed constant mass in his second law of motion. The problem lies not in the epistemology but in our knowledge regarding the RMS. The RMS depends upon the joint distribution of the error and the error estimate, both of these being random variables relative to the random sampling process. While one need not know the full joint distribution, one needs to know at least the second moments of the joint distribution. Until very recently, there were no results on the joint distribution in the literature. To date, the joint distribution of the error and error estimate has been discovered for the multinomial model with resubstitution or leave-one out (Braga-Neto and Dougherty, 2005; Xu et al., 2006) and for the model in which both class-conditional densities are Gaussian under the further assumption that they share a common known covariance matrix and the classification rule is LDA (Zollanvari et al., 2010).

There is a price to be paid for making distributional assumptions to arrive at a useful RMS bound—that is, to be able to claim model validity. The price is that model validity is conditional with respect to the modeling assumptions. While omitting distributional assumptions might seem desirable so as not to limit the scope of the conclusions, this is generally a vain hope because the absence of distributional assumptions usually renders the entire study vacuous. Scientifically sound model-free classification is virtually impossible with small data sets. One might have reassuring theorems concerning the behavior of classification and error estimation rules for very large samples, but these say nothing when samples are small. Consider the comment by Ronald A. Fisher in 1925 on the limitations of large-sample methods:

> Little experience is sufficient to show that the traditional machinery of statistical processes is wholly unsuited to the needs of practical research. Not only does it take a cannon to shoot a sparrow, but it misses the sparrow! The elaborate mechanism built on the theory of infinitely large samples is not accurate enough for simple laboratory data. Only by systematically tackling small sample problems on their merits does it seem possible to apply accurate tests to practical data. (Fisher, 1925)

In reading Fisher's statement, one should recognize that laws of large numbers, that is, theorems concerning the convergence to zero of the difference between an estimate of a distributional parameter and the parameter as the sample size "increases to infinity," go back to Jacob Bernoulli, and that central limit theorems, that is, theorems that the Gaussian distribution is the limit of a sequence of other distributions, go back to De Moivre and Laplace.

What if the modeling assumptions do not hold? Indeed, they will almost surely not hold for the feature-label distribution. In any event, we will never know. This is science, not metaphysics. We sincerely doubt whether Galileo believed that he would ever come across a "frictionless" plane. There are no propositions absent hypotheses. Even if one has a very large sample and applies some limit-at-infinity theorem, theorems such as laws of large numbers and central limit theorems do not generally tell us how large the number has to be and, to obtain how-large statements, conditions have to be imposed. Prior to that, limit theorems generally require randomization and other assumptions, such as independence, none of these being open to empirical verification. The scientist lives in a doubly uncertain world. The rigor of science depends on specifying assumptions together with quantifying the uncertainty of propositions given the assumptions, which are uncertain. The uncertainty of the assumptions was known to Galileo and Newton; the quantification of propositional uncertainty is a product of modern probability theory and statistics.

As previously stated, there are two validation issues regarding classifier models; to this point we have only considered one, that being the validity of the error estimate, given a designed classifier. The other issue is the validity of the classifier as an optimal discriminator between the two classes. Just as the conditional expectation provides the optimal prediction of the value of one random variable based on the observation of another and thereby serves as the best scientific model for this prediction, a Bayes classifier provides the optimal discriminator between two classes because it possesses the minimum classification error. Thus, the validity of a designed classifier as a scientific model for the natural class discrimination inherent in the feature-label distribution depends on the closeness of its error to the Bayes error, which we do not know.

Analogous to the situation with error estimation, we fall back upon the expected performance of the classification rule to characterize *classification validity*. Together, a data sample and a classification rule Ψ

yield a classifier ψ. Letting ε_{bay} denote the Bayes error, we define the *design cost* $\Delta = \varepsilon[\psi] - \varepsilon_{\text{bay}}$, which gives the increase in error above the Bayes error resulting from using ψ instead of deriving the Bayes error from the feature-label distribution. If we average the design cost over all possible samples, we obtain the expected design cost, $E[\Delta]$, which can be used to characterize the validity of the classification rule. Qualitatively, a classification rule is valid if $E[\Delta]$ is small. A small design error does not mean that the classifier has small error; only that its error is close to the Bayes error, which is the intrinsic classification error for the feature-label distribution. Although we will not go into detail, a well-studied problem in classification theory is to obtain bounds on $E[\Delta]$ based upon the classification rule and feature-label distribution (Vapnik and Chervonenkis, 1974). As with error estimation, one can obtain smaller bounds by making assumptions on the feature-label distribution and, without such assumptions, the bounds are generally useless unless the sample is very large.

Taking a different perspective on classifier performance, one might be concerned with deciding the better classifier between two classifiers, ϕ and ψ, designed from different classification rules. The better classifier is the one with smaller error, thereby being closer to the Bayes error. Ostensibly, the issue resolves to a classical hypothesis test; decide between $\varepsilon[\phi] \leq \varepsilon[\psi]$ (ϕ is better) and $\varepsilon[\phi] > \varepsilon[\psi]$ (ψ is better); however, once again we are faced with a dichotomy between hold-out error estimation and training data-based error estimation. Before considering this problem, we say a few words about hypothesis testing.

The basic form of a one-sided hypothesis test is H_0: $\theta_1 \geq \theta_2$ (the null hypothesis) and H_1: $\theta_1 < \theta_2$ (the alternative hypothesis), which can be rewritten as

$$\begin{aligned} &H_0\colon \theta_2 - \theta_1 \leq 0 \\ &H_1\colon \theta_2 - \theta_1 > 0. \end{aligned} \tag{6.10}$$

A test statistic ρ is chosen, a critical value c is chosen, the null hypothesis is accepted if $\rho \leq c$, and the alternative hypothesis is chosen if $\rho > c$. The test statistic should reflect the difference $\theta_2 - \theta_1$. There are two errors associated with the hypothesis test. Type I error occurs if the null hypothesis reflects the true state of affairs but is rejected. Type II error occurs if the null hypothesis does not reflect the true state of affairs but is accepted. Type I and type II errors are typically denoted by α and β, respectively. A balanced approach to the two hypotheses would be to let $c = 0$, so that $\rho > 0$ would mean a rejection of the null hypothesis. This would lead to α being the probability that $\rho > 0$ and $\theta_2 - \theta_1 \leq 0$, and β being the probability that $\rho \leq 0$ and $\theta_2 - \theta_1 > 0$. With this approach, α and β tend not to be small, so that the probability of error is not small. The standard approach is to formulate the hypotheses so that the objective of the scientist is to provide empirical support for the alternative hypothesis. Hence, c is chosen sufficiently large that type I error is small, meaning that, if the null hypothesis is rejected, then there is small probability that the alternative hypothesis has been erroneously accepted. Of course, if α is chosen small, then this pushes up the value of β, meaning that, if the null hypothesis is accepted, then the probability of error need not be small.

A difficulty is that α is a function of the actual value of $\theta_2 - \theta_1$, so that $\alpha = \alpha(\theta_2 - \theta_1)$, and β is a function of the actual value of $\theta_2 - \theta_1$, so that $\beta = \beta(\theta_2 - \theta_1)$. Given that the scientific focus is on rejecting the null hypothesis, the usual way of proceeding is to be conservative and choose the value of $\theta_2 - \theta_1$ satisfying the null hypothesis that maximizes $\alpha(\theta_2 - \theta_1)$ and this value is given by $\theta_2 - \theta_1 = 0$, so that the hypothesis test of Eq. 6.10 becomes

$$\begin{aligned} H_0 &: \theta_2 - \theta_1 = 0 \\ H_1 &: \theta_2 - \theta_1 > 0. \end{aligned} \tag{6.11}$$

This still leaves type II error as a function of $\theta_2 - \theta_1$, but, as stated, the goal is providing support for acceptance of the alternative hypothesis and type I error measures the probability of incorrectly accepting the alternative hypothesis. In this scenario, acceptance of the null hypothesis is not taken as providing scientific validity to the null hypothesis; however, if the null hypothesis is rejected, validity is ascribed to the accepted alternative hypothesis. This validity is measured by the size of α, which in this setting is called the *level of significance*. The choice of $\alpha = 0.05$ is commonplace, but lower values, say $\alpha = 0.01$, are used to engender more confidence in the decision.

The test statistic is chosen by the scientist because the scientist believes that it reflects the difference $\theta_2 - \theta_1$. It provides a measure of validity: The greater the test statistic, the greater the validity of the scientific theory, in this case, the theory saying that $\theta_2 > \theta_1$ for large values of the test statistic. However, here one must be careful. Owing to the randomness of the test statistic, a large value of the test statistic with a small sample does not say very much. How large is large enough? Both the distribution of the test statistic and the sample size must be taken into account. A hypothesis test incorporates these issues.

Consider the null hypothesis H_0: $\mu_2 - \mu_1 \leq 0$, where μ_1 and μ_2 are the means of two different distributions, F_1 and F_2, respectively. The alternative hypothesis asserts that the mean of F_2 lies to the right of the mean of F_1. An obvious test statistics is $\rho = \overline{X}_2 - \overline{X}_1$, where $\overline{X}_1$ and $\overline{X}_2$ are the sample means for F_1 and F_2, respectively. A large value of ρ supports the claim of the alternative hypothesis and in this sense quantifies the validity of the claim. Putting matters in the form of a hypothesis test and stating a level of significance puts the test of validity into a quantitative form, with the critical value depending on the distribution of ρ and the sample size.

Rather than setting a value of α, finding the corresponding critical value c, obtaining the test statistic ρ from the data, and seeing if $\rho > c$, another approach is to obtain the value of ρ and find the lowest level of significance for which the value of ρ results in rejection of the null hypothesis. This value is called the p value. Since α is the probability that the null hypothesis is true and $\rho > c$, if ρ is very large, then the maximum value of c for which $\rho > c$ is very large and the corresponding α is very small. Basically, the p value is a transformation of the test statistic and provides an alternative measure of validity to the test statistic. A p value transforms the test statistic in such a way as to provide a probabilistic measure: The smaller the p value, the greater the validity of the theory. Rather than giving an "accept" or "reject" decision, the p value provides a measure of strength. In addition, the p value does not involve the units of the test statistic. It also takes into account the distribution of the test statistic and the sample size. There are downsides to the p value when taken alone: (1) it hides the relationship of the test statistic to the scientific theory; (2) it hides the value of the test statistic; (3) it hides the sample size; and (4) it relates to type I error but not to type II error.

A major pitfall with using hypothesis tests, p values in particular, concerns the problem of multiple hypothesis tests. No matter the level of significance and the truth or falsity of the hypotheses involved, as the number of hypothesis tests performed tends to infinity, so does the number of accepted alternative hypotheses. Suppose type I error is 0.05 and one performs k independent hypothesis tests, then the expected number of erroneously rejected null hypotheses is $(0.05)k$ and the probability of at least one test statistic leading to a rejection of the null hypothesis when the null hypothesis is true is $1 - (0.95)^k$. For instance, if $k = 100$, the probability of erroneously accepting at least one alternative hypothesis is approximately 0.994. A classic problem in statistics is to adjust the level of significance to take into account this "multiple comparisons" problem. But this requires careful experimental reporting since it is not uncommon for a researcher to consider many hypotheses before settling on one. In this situation, validation is compromised unless all considered hypotheses are reported so that interaction between them can be assessed. Given the tens of thousands of features being considered in high-throughput biology and the ability to run thousands of hypothesis tests with current computational capability, assessing the scientific import of reported results can be highly problematic. On a closely related issue, it is common for one to try a large number of classifiers, find one with a low error estimate, and positively report on that classifier. Unfortunately, variation in the error estimator means that may of the estimates will be optimistic, so that the reported low error estimate may simply be a consequence of estimator variance. In Chapter 8, we will consider a variant of this multiple comparisons problem that arises from testing a classifier on many data sets.

A second issue concerning hypothesis tests is that the alternative hypothesis might be of little value in regard to what one is attempting to demonstrate. For instance, suppose one wishes to demonstrate that classifier ψ is an effective classifier on a certain feature-label distribution. Consider the hypothesis test consisting of H_0: $\varepsilon[\psi] \geq 0.5$ and H_1: $\varepsilon[\psi] < 0.5$. No matter the value of the test statistic, concluding the alternative provides no useful information because, as stated, conclusion of the alternative hypothesis allows the classifier error to be arbitrarily close to 0.5.

We now return to the problem of comparing two classifiers, first under the assumption that the classifiers are given and data are to be collected to perform the test. This problem is logically equivalent to designing two classifiers from training data and then using independent test data to perform the hypothesis test. We consider the hypothesis test

$$\begin{aligned} H_0&: \varepsilon[\phi] - \varepsilon[\psi] \leq 0 \\ H_1&: \varepsilon[\phi] - \varepsilon[\psi] > 0, \end{aligned} \tag{6.12}$$

where rejection of the null hypothesis leads to the conclusion that ψ is better than ϕ. If we let X denote the random variable that is 1 when ϕ

makes a correct decision and 0 otherwise, and we let Y denote the random variable that is 1 when ψ makes a correct decision and 0 otherwise, then the hypothesis test concerns the means of X and Y. If we split the test data into two independent samples, S_ϕ and S_ψ, and let $\hat{p}_\phi$ and $\hat{p}_\psi$ be the proportions of correct classifications for ϕ on S_ϕ and ψ on S_ψ, respectively, then $\hat{p}_\phi$ and $\hat{p}_\psi$ are the sample means for X and Y, respectively. The test statistic $\hat{p}_\phi - \hat{p}_\psi$ is an unbiased estimator of $\varepsilon[\phi] - \varepsilon[\psi]$ and this is the classical hypothesis test concerning two proportions. The test can be performed by taking the standardized version of $\hat{p}_\phi - \hat{p}_\psi$, which is approximately Gaussian assuming sufficiently large test sets, and performing an ordinary Z-test.

The preceding reasoning breaks down when the classifiers and error estimates are obtained from the same data; indeed, there is no hypothesis test in the ordinary sense because there are no classifiers prior to using the data and therefore no errors to form the hypotheses. However, suppose we change our focus to the classification rules, not the particular classifiers produced by the rules. In this case, we have two classification rules, Φ and Ψ, and to these there corresponds two random variables, $\varepsilon[\phi]$ and $\varepsilon[\psi]$, giving the errors of the designed classifiers across all possible samples of the given size. In this case, the hypothesis test involves the expected values (means) of $\varepsilon[\phi]$ and $\varepsilon[\psi]$ and takes the form

$$\begin{aligned} &H_0\colon E[\varepsilon[\phi]] - E[\varepsilon[\psi]] > 0 \\ &H_1\colon E[\varepsilon[\phi]] - E[\varepsilon[\psi]] > 0. \end{aligned} \tag{6.13}$$

An obvious test statistic is $\rho = \hat{\varepsilon}[\phi] - \hat{\varepsilon}[\psi]$, where $\hat{\varepsilon}$ is some error estimator on the training data. But what do we know about the distribution of $\hat{\varepsilon}[\phi] - \hat{\varepsilon}[\psi]$? In general, we do not know much. As discussed in reference to the RMS, there are only two models for which estimated error distributions are known, and only for resubstitution and leave-one-out error estimation. Even if one were to make the unwarranted assumption that the distribution of the test statistic is Gaussian and assume that the mean is zero under the null hypothesis, which would be a good approximation for leave-one-out error estimation, the variances of the error estimators are not known and cannot be estimated from a single sample. Based on our current state of knowledge, not only is it impossible to compare classifiers without independent test data, but it is also not possible to compare classification rules.

CHAPTER 7

Stochastic Validation: Networks

> The conventions we adopt must somehow work: they must serve us in our coping with Nature.
>
> —*William Barrett*

Classifiers are very simple mathematical structures and their biological content is minimal; on the other hand, networks are very complex mathematical structures and they constitute the basic modeling mechanism for biological state spaces and regulation. As might be expected, network validation is much more complicated than classifier validation and the theory is much less developed. While it is not our intention to delve deeply into the validation of network models, we believe it is important to at least point out some of the fundamental issues. The manner of classification validation is straightforward because a classifier is simply a decision rule and therefore validation concerns the error rate for the decision rule. Networks are an entirely different manner. They possess much more structure and therefore operational definitions must be related to this structure, thereby compelling validation procedures that take into account corresponding structure in the data.

If we make the primary epistemological assumption that there is a hypothesized ideal network, then validation requires a method to compare two networks, in this case, the model and the ideal. Given

Epistemology of the Cell: A Perspective on Biological Knowledge, First Edition.
Edward R. Dougherty, Michael L. Bittner.

networks $\mathcal{H}$ and $\mathcal{M}$, we need a *distance function*, $\mu(\mathcal{M}, \mathcal{H})$, quantifying the difference between them. Following Dougherty (2007), we require that μ be a *semimetric*, meaning that it satisfies the following four properties:

1. $\mu(\mathcal{M}, \mathcal{H}) \geq 0$,
2. $\mu(\mathcal{M}, \mathcal{M}) = 0$,
3. $\mu(\mathcal{M}, \mathcal{H}) = \mu(\mathcal{H}, \mathcal{M})$ [*symmetry*],
4. $\mu(\mathcal{M}, \mathcal{H}) \leq \mu(\mathcal{M}, \mathcal{N}) + \mu(\mathcal{N}, \mathcal{H})$ [*triangle inequality*].

If μ should satisfy a fifth condition,

5. $\mu(\mathcal{M}, \mathcal{H}) = 0 \Rightarrow \mathcal{M} = \mathcal{H}$,

then it is a *metric*. A distance function is often defined in terms of some characteristic, by which we mean some structure associated with a network, such as its regulatory graph or steady-state distribution. We do not require a distance function to be a metric because distinct networks may share common characteristics.

By the term "network" we mean a mathematical model that involves a multivariate state space evolving over time. Unfortunately, it has become common parlance in the biological literature to refer to any graphical model as a network. For the most part, these so-called "networks" are simply graphical visualizations of lists of relational pairs between genes or proteins, or both, and have nothing to do with dynamical behavior. They do not describe biological state trajectories through time. We are not saying that the discovery of correlation or codetermination between genes is unimportant; indeed, these kinds of relationships may represent important characteristics of a biological network; however, in and of themselves they do not constitute a network model in the dynamical sense.

Since our purpose is to highlight basic issues concerning network model validity and not to provide a general theoretical discussion of networks, we will consider a restricted class of networks that contains most of the networks thus far used in genomics and proteomics. We assume the underlying network structure is composed of a finite set, $V = \{X_1, X_2, \ldots, X_n\}$, of nodes (for instance, gene or protein expressions), with each node taking discrete values between 0 and $d - 1$. The corresponding state space possesses $N = d^n$ states, which we denote by $\mathbf{x}_1$, $\mathbf{x}_2, \ldots, \mathbf{x}_N$. We express the state $\mathbf{x}_j$ in vector form by $\mathbf{x}_j = (x_{j1}, x_{j2}, \ldots, x_{jn})$.

The corresponding dynamical system is based on discrete time, $t = 0, 1, 2, \ldots$, with the state-vector transition $\mathbf{X}(t) \to \mathbf{X}(t + 1)$ at each time instant.

We assume that the process $\mathbf{X}(t)$ is a *Markov chain*, meaning that the probability of $\mathbf{X}(t)$ conditioned on $\mathbf{X}$ at $t_1 < t_2 < \ldots < t_s < t$ is equal to the probability of $\mathbf{X}(t)$ conditioned on $\mathbf{X}(t_s)$, the most recent observation. We assume the chain is *homogeneous*, meaning that the transition probabilities depend only on the time difference, that is, for any t and u, the u-step transition probability,

$$p_{jk}(u) = P(\mathbf{X}(t+u) = \mathbf{x}_k \mid \mathbf{X}(t) = \mathbf{x}_j), \tag{7.1}$$

depends only on u. We are not asserting that the Markov property and homogeneity are necessary assumptions for biological regulatory networks; rather we make these assumptions to facilitate mathematically tractable modeling. Under these assumptions, we need only consider the transition probability matrix defined in Eq. 4.8, where the one-step transition probability, p_{jk}, is given by $p_{jk} = p_{jk}(1)$. Besides the state one-step probabilities, we can consider the node one-step probabilities,

$$p_i(j, r) = P(X_i(t+1) = r \mid \mathbf{X}(t) = \mathbf{x}_j). \tag{7.2}$$

These give the probabilities of the nodes at time $t + 1$ given the state at time t. Suppose that node X_i at time $t + 1$ depends only on values of nodes in a *regulatory set*, $R_i \subset V$, at time t, the dependency being independent of t. Then the node one-step probabilities are given by

$$p_i(j, r) = P(X_i(t+1) = r \mid X_l(t) = x_{jl} \text{ for } X_l \in R_i). \tag{7.3}$$

In this form, we see that the dependencies are restricted to regulatory nodes, in the case of gene networks these being the regulatory genes for the gene corresponding to X_i. The network has a *regulatory graph* (*connectivity graph*) consisting of the n nodes and a directed edge from x_i to x_j if $x_i \in R_j$. There is also a *state-transition graph* whose nodes are the N state vectors. There is a directed edge from state $\mathbf{x}_j$ to state $\mathbf{x}_k$ if and only if $\mathbf{x}_j = \mathbf{X}(t)$ implies $\mathbf{x}_k = \mathbf{X}(t + 1)$.

A homogeneous, discrete-time Markov chain with state space $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ possesses a steady-state distribution $(\pi_1, \pi_2, \ldots, \pi_N)$ if, for all pairs of states $\mathbf{x}_k$ and $\mathbf{x}_j$, $p_{jk}(u) \to \pi_k$ as $u \to \infty$. If there exists a steady-state distribution, then, regardless of the state $\mathbf{x}_k$, the probability of the Markov chain being in state $\mathbf{x}_k$ in the long run is π_k. Not all

Markov chains possess steady-state distributions. As mentioned in Chapter 4, a probabilistic Boolean network has an associated Markov chain that possesses a steady-state distribution under the assumption of a positive perturbation probability.

Various distance functions can be defined, depending upon which network characteristic is of interest. Here we describe three distance functions in the language of gene regulatory networks (although their use is certainly not restricted to gene regulation). For two Boolean networks with perturbation possessing the same gene set, with gene expressions $X_1, X_2, \ldots, X_n$, a rule-based distance is given by the proportion of incorrect rows in the function-defining truth tables. Denoting the state functions for networks $\mathcal{H}$ and $\mathcal{M}$ by $\mathbf{f} = (f_1, f_2, \ldots, f_n)$ and $\mathbf{g} = (g_1, g_2, \ldots, g_n)$, respectively, since there are n truth tables consisting of 2^n rows each, this distance is given by

$$\mu_{\text{fun}}(\mathcal{M}, \mathcal{H}) = \frac{1}{n2^n} \sum_{i=1}^{n} \sum_{k=1}^{N} I[g_i(\mathbf{x}_k) \neq f_i(\mathbf{x}_k)], \tag{7.4}$$

where I denotes the indicator function, $I[A] = 1$ if A is a true statement and $I[A] = 0$ otherwise. If we wish to give more weight to those states more likely to be observed in the steady state, then we can weight the inner sums in Eq. 7.4 by the corresponding terms in the steady-state distribution, $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$.

If our main interest is in the regulatory graph of a network, then we can apply adjacency matrices. Given an n-gene network, for $i, j = 1, 2, \ldots, n$, the (i, j) entry in the matrix is 1 if there is a directed edge from the ith to the jth gene; otherwise, the (i, j) entry is 0. If $\mathbf{A} = (a_{ij})$ and $\mathbf{B} = (b_{ij})$ are the adjacency matrices for networks $\mathcal{H}$ and $\mathcal{M}$, respectively, where $\mathcal{H}$ and $\mathcal{M}$ possess the same gene set, then the *hamming* distance between the networks is defined by

$$\mu_{\text{ham}}(\mathcal{M}, \mathcal{H}) = \sum_{i,j=1}^{n} |a_{ij} - b_{ij}|. \tag{7.5}$$

Alternatively, the hamming distance may be computed by normalizing the sum, such as by the number of genes or the number of edges in one of the networks, for instance, when one of the networks is considered as representing ground truth. The hamming distance is a coarse measure since it contains no steady-state or dynamical information. Two networks can be very different and yet have $\mu_{\text{ham}}(\mathcal{M}, \mathcal{H}) = 0$.

Since steady-state behavior is of particular interest, for instance, being associated with phenotypes in gene regulatory networks, a natural choice for a network distance is to measure the difference between steady-state distributions. If $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_m)$ and $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_m)$ are the steady-state distributions for networks $\mathcal{H}$ and $\mathcal{M}$, respectively, then a network distance is defined by

$$\mu_{\text{ss}}(\mathcal{M}, \mathcal{H}) = \sum_{i=1}^{n} |\pi_i - \omega_i|. \tag{7.6}$$

Other norms can be used to define distance functions based on the steady-state distribution.

The previous examples of network distance functions demonstrate a common scenario: A network semimetric is defined by a metric on some network characteristic, for instance, its regulatory graph and steady-state distribution. The metric requirement fails because distinct networks possess the same characteristic. To formalize the situation, let λ_M and λ_H denote the characteristic λ corresponding to networks $\mathcal{M}$ and $\mathcal{H}$, respectively. If ν is a metric on a space of characteristics (directed graphs, matrices, probability densities, etc.), then a semimetric μ_ν is induced on the network space according to

$$\mu_\nu(\mathcal{M}, \mathcal{H}) = \nu(\lambda_{\mathcal{M}}, \lambda_{\mathcal{H}}). \tag{7.7}$$

This is quite natural if our main interest is with the characteristic, not the full network itself.

Focus on network characteristics leads to the identification of networks possessing the same characteristic. Given any set, U, a relation ~ between elements of U is called an *equivalence relation* if it satisfies the following three properties for a, b, $c \in U$:

1. $a \sim a$ [reflexivity],
2. $a \sim b \Rightarrow b \sim a$ [symmetry],
3. $a \sim b$ and $b \sim c \Rightarrow a \sim c$ [transitivity].

If $a \sim b$, then a and b are said to be *equivalent*. An equivalence relation on U induces a partition of U. The subsets forming the partition are defined according to a and b lie in the same subset if and only if $a \sim b$. The subsets are called *equivalence classes*. The equivalence class of elements equivalent to a is denoted by $[a]^\sim$. According to the definitions, $[a]^\sim = [b]^\sim$ if and only if $a \sim b$. If ν is a semimetric on a set U and we define $a \sim b$ if and only if $\nu(a, b) = 0$, then

$$\mu([a]^\sim, [b]^\sim) = \nu(a, b) \tag{7.8}$$

defines a metric on the space of equivalence classes, which means that $\mu([a]^\sim, [b]^\sim) = 0$ implies $[a]^\sim = [b]^\sim$.

If we define $\mathcal{M} \sim \mathcal{H}$ if $\lambda_{\mathcal{M}} = \lambda_{\mathcal{H}}$, then this is a network equivalence relation. If we focus on equivalence classes of networks rather than the networks themselves, we are in effect identifying equivalent networks. For instance, if we are only interested in steady-state distributions, then it may be advantageous to identify networks possessing the same steady-state distribution. In this case, if the steady-state distributions of $\mathcal{M}$ and $\mathcal{H}$ are $F_{\mathcal{M}}$ and $F_{\mathcal{H}}$, respectively, then we can replace $\mu_{\text{ss}}(\mathcal{M}, \mathcal{H})$ by $\nu_{\text{ss}}(F_{\mathcal{M}}, F_{\mathcal{H}})$.

While the idea of dealing with equivalence classes of networks might at first appear abstract, it is epistemologically beneficial and forms the basis for model validation. Suppose we wish to validate a proposed network, say, a probabilistic Boolean network, $\mathcal{M}$ via its steady-state distribution. In this case, the problem is to mathematically derive the steady-state distribution, $F_{\mathcal{M}}$, from the network model and then compare this distribution to an empirical steady-state distribution F_{emp}, which is a histogram formed from state frequencies of empirical data that have resulted from an experiment designed to produce appropriate steady-state data. The primary epistemological assumption in this case is that there is a hypothesized ideal network from which the steady-state data have been drawn. Based on the fact that the experimental validation only reflects the steady state, model validation corresponds to the equivalence class of networks possessing the steady-state distribution $F_{\mathcal{M}}$, not to $\mathcal{M}$ itself. We might possess strong theoretical evidence that the behavior of the genes composing $\mathcal{M}$ can be modeled by the proposed probabilistic Boolean network, but the validation has only been with respect to the steady-state distribution (and relative to whatever test of validation for the steady-state distribution has been employed). The proposed network model $\mathcal{M}$ remains contingent, as are all scientific models, but observation has supported it, at least in regard to the equivalence class of networks possessing the steady-state distribution $F_{\mathcal{M}}$. Here we should keep in mind Einstein's words that "it is only necessary that enough propositions of the conceptual system be firmly enough connected with sensory experiences." Is the steady-state validation enough? Perhaps yes if the steady state is our main interest; perhaps no if our interests run deeper.

This procedure leads to an obvious question: Why would one only partially validate the proposed network in this manner? The answer is straightforward. It is much easier to design an experiment to produce steady-state data in order to evaluate $\nu_{ss}(F_{\mathcal{M}}, F_{emp})$ than to design an experiment to evaluate $\mu_{fun}(\mathcal{M}, \mathcal{H}_{emp})$, where $\mathcal{H}_{emp}$ is an empirically inferred network. Indeed, if we could reliably infer $\mathcal{H}_{emp}$ from data, why bother with $\mathcal{M}$ at all. Just replace it by $\mathcal{H}_{emp}$. This is analogous to the problem faced in classifier design: If we can reliably infer a feature-label distribution from data, why bother with the proposed classifier when we could simply take a Bayes classifier for the inferred feature-label distribution? The whole point of validating a model via characteristics is that the corresponding empirical characteristics can be more reliably inferred from data than the full model itself, which presumably is a creation of the intellect in conjunction with the inference of some small number of parameters. Moreover, if our interest is with the characteristic, for instance, phenotype determination via the steady-state distribution, then validation of the characteristic is salient.

Once again, the issue is hypothesis testing. We can formulate a hypothesis test

$$\begin{aligned} &H_0\colon \lambda_{\mathcal{M}} = \lambda_{\mathcal{H}} \\ &H_1\colon \lambda_{\mathcal{M}} \neq \lambda_{\mathcal{H}}, \end{aligned} \tag{7.9}$$

where $\lambda_{\mathcal{H}}$ is the characteristic for the ideal network, $\nu(\lambda_{\mathcal{M}}, \lambda_{emp})$ plays the role of a test statistic, and λ_{emp} is computed from independent test data. Given the hypothesis test, a critical value ν_0 is chosen such that the null hypothesis is accepted if $\nu(\lambda_{\mathcal{M}}, \lambda_{emp}) \leq \nu_0$ and the alternative hypothesis accepted if $\nu(\lambda_{\mathcal{M}}, \lambda_{emp}) > \nu_0$. Two problems immediately arise. First, evaluation of type I error requires knowledge of the distribution of the test statistic under the null hypothesis, knowledge we almost certainly lack, and evaluation of type II error would require some assumption as to a specific competing hypothesis and knowledge of the distribution of the test statistic under that competing hypothesis. Hence, formulating validation in a rigorous statistical fashion is not likely. Second, the critical value determines the orientation, whether we make it difficult to reject the null hypothesis or difficult to accept it. Suffice it to say that the choice of the critical value is an epistemological choice.

Model construction typically involves estimating some parameters from data. In this case, a model can be expressed in the form $\mathcal{M}(\mathbf{a})$, where **a** is a parameter vector. For instance, for a Boolean network, the model structure is created but the truth tables defining the rule structure are often inferred from data. By assuming that the model is known prior to testing (validation) and that validation is independent of the manner in which the model is conceived, we are in fact assuming that there are both training and testing data. If data are limited, then training and testing may be done on the same data, which affects the distribution of the test statistic. Letting λ denote the characteristic for the hypothetical model to be tested and z denote the corresponding empirical characteristic, since the characteristic depends on the model parameters and some of these have been estimated from the sample S, the characteristic takes the form $\lambda(S)$ and the corresponding observation is of the form $z(S)$. In the case of independent training and testing data, the test distance takes the form $\nu(\lambda, z(S))$; when training and testing are done on the same data, it becomes $\nu(\lambda(S), z(S))$. As in the case of classification, this kind of resubstitution estimate can be expected to suffer from optimistic bias: if Σ is the point process generating the sample points, then it is very likely that

$$E_{\Sigma}[\nu(\lambda(\Sigma), z(\Sigma))] < E_{\Sigma}[\nu(\lambda, z(\Sigma))], \tag{7.10}$$

where the expectation is taken with respect to Σ. The extent of such bias depends on the model, test characteristic, and sample size. Whereas substantial effort has gone into studying these kinds of problems in classification, there appears to have been little effort in the case of networks.

EXAMPLE: A DIFFERENTIAL EQUATION MODEL FOR PROTEIN CONCENTRATIONS IN *ESCHERICHIA COLI*

Validation of dynamical networks has not been commonplace in genomics; however, it has not been totally absent and here we describe a validation effort in the case of a model based on piecewise-linear differential equations (Batt et al., 2005). The underlying theory rests on a class of piecewise-linear differential equations long used to model gene regulation (Glass and Kauffman, 1973) and validation is with reference to a quantization of the continuous piecewise-linear differential equation model (de Jong et al., 2003). Given the differential equation model for protein concentrations based upon the derivatives and various parameters, including threshold concentrations, synthesis parameters, and degradation parameters, knowing the relative order of certain parameters and quotients of parameters results in a partitioning of the phase space into a system of domains. The domain system includes transitions between domains and a labeling of each domain in terms of the signs of the derivatives of the concentration variables and a marker as to whether the domain is persistent or instantaneous. Prediction of the signs of concentration derivatives can be obtained from the model, thereby facilitating model validation.

To illustrate the modeling scheme, we consider the following two-equation network modeling the concentrations x and y of the proteins A and B (Batt et al., 2005):

$$\frac{dx}{dt} = \kappa_x s^-(x, \theta_x^2) s^-(y, \theta_y) - \gamma_x x \tag{7.11}$$

$$\frac{dy}{dt} = \kappa_y s^-(x, \theta_x^1) - \gamma_y y, \tag{7.12}$$

where θ_x^1, θ_x^2, and θ_y are threshold concentrations, κ_x and κ_y are synthesis parameters, γ_x and γ_y are degradation parameters, and $s^-(x, \theta)$ is a step function with $s^-(x, \theta) = 1$ if $x < \theta$ and $s^-(x, \theta) = 0$ if $x > \theta$. According to Eq. 7.11, protein A is produced if and only if neither $x > \theta_x^2$ nor $y > \theta_y$, and when produced it is produced at rate κ_x. According to Eq. 7.12, protein B is produced at rate κ_y if and only if $x < \theta_x^1$. It can be shown that flow in the phase space can be determined from knowledge of the relative order of the threshold parameters and the quotients of the synthesis and degradation parameters. Figure 7.1(a) illustrates the flow in the phase space (dots marking equilibrium points) and Fig. 7.1(b) shows the domain partition of the phase space. The domains interior to the rectangular partitions are "persistent" and those forming the boundaries of the partition are "instantaneous." Figure 7.1(c) shows the state transition graph corresponding to the domain space (dots representing self-transitions).

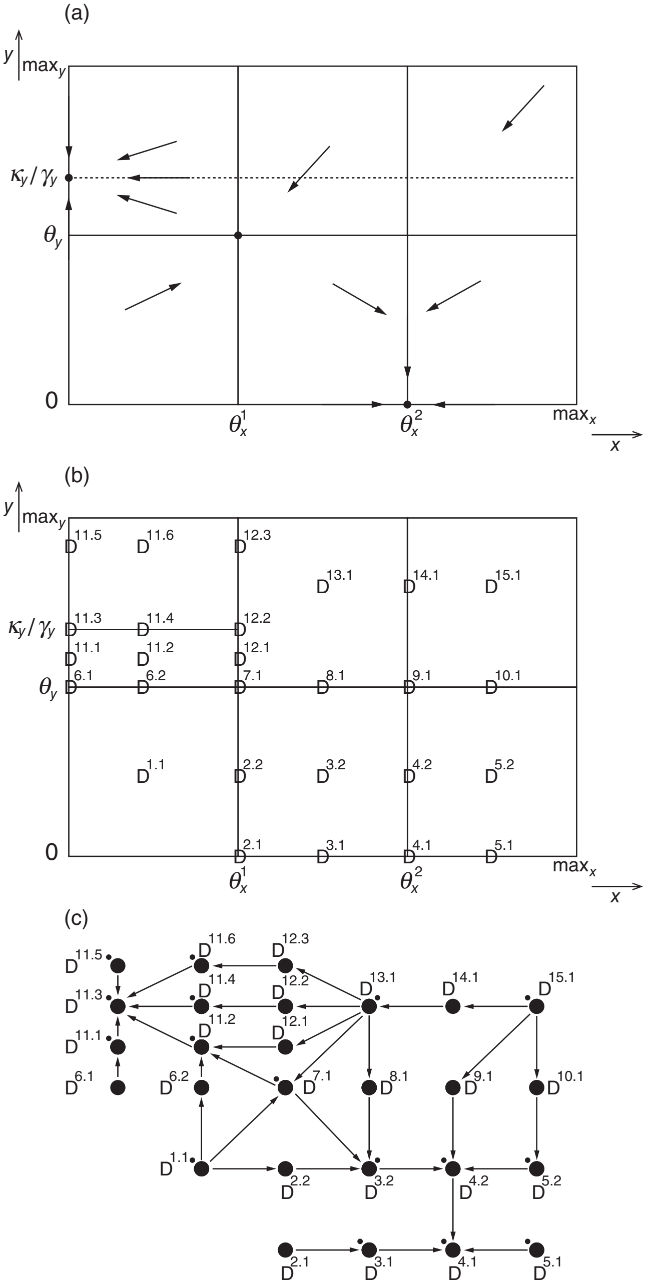
(a)
y
max$_y$
κ_y/γ_y
θ_y
0
θ_x^1
θ_x^2
max$_x$
x
(b)
y
max$_y$
$D^{11.5}$ $D^{11.6}$ $D^{12.3}$
$D^{13.1}$ $D^{14.1}$ $D^{15.1}$
κ_y/γ_y
$D^{11.3}$ $D^{11.4}$ $D^{12.2}$
$D^{11.1}$ $D^{11.2}$ $D^{12.1}$
θ_y
$D^{6.1}$ $D^{6.2}$ $D^{7.1}$ $D^{8.1}$ $D^{9.1}$ $D^{10.1}$
$D^{1.1}$ $D^{2.2}$ $D^{3.2}$ $D^{4.2}$ $D^{5.2}$
0
$D^{2.1}$ $D^{3.1}$ $D^{4.1}$ $D^{5.1}$
θ_x^1
θ_x^2
max$_x$
x
(c)
$D^{11.5}$ $D^{11.6}$ $D^{12.3}$
$D^{11.3}$ $D^{11.4}$ $D^{12.2}$ $D^{13.1}$ $D^{14.1}$ $D^{15.1}$
$D^{11.1}$ $D^{11.2}$ $D^{12.1}$
$D^{6.1}$ $D^{6.2}$ $D^{7.1}$ $D^{8.1}$ $D^{9.1}$ $D^{10.1}$
$D^{1.1}$ $D^{2.2}$ $D^{3.2}$ $D^{4.2}$ $D^{5.2}$
$D^{2.1}$ $D^{3.1}$ $D^{4.1}$ $D^{5.1}$

Figure 7.1 Dynamics of the two-gene network: (a) dynamics in the phase space, dots marking equilibrium points; (b) domain partition of the phase space, each label signifying a domain (interior, boundary, or boundary intersection: a label in the interior denotes the interior; a label on a boundary denotes the boundary, and a label on a boundary intersection denotes the point at the intersection; (c) state transition diagram, dots indicating self-transitions. [Batt, G., Ropers, D., de Jong, H., Geiselmann, J., Mateescu, R., Page, M., and D. Schneider, "Validation of Qualitative Models of Genetic Regulatory Networks by Model Checking: Analysis of the Nutritional Stress Response in *Escherichia coli*," *Bioinformatics*, 21(supp 1), 19–28, 2005, by permission of The International Society for Computational Biology].

A sequence of domains constitutes a path. A key property is that every solution of the original differential equation model corresponds to a path in the state transition diagram (however, the converse is not true). Consequently, paths corresponding to predicted regulatory behavior can be compared with experimental data. From the perspective of the original differential equation model, the paths correspond to characteristics of the model and the original model can be validated via these characteristics.

Turning to the biological application of immediate interest, an *E. coli* population transitions from exponential growth to a nongrowth "stationary phase" when nutritionally stressed. In Batt et al. (2005), based on experimental literature, a seven-variable piecewise-linear differential equation model is constructed to characterize nutritional stress response in *E. coli*. The model consists of six protein concentrations corresponding to six genes and one input variable denoting the presence or absence of a carbon starvation signal. The genes are involved in the cell's response to carbon starvation: *crp* and *cya* (transduction of the carbon starvation signal); *fis* (metabolism); *rrn* (cell growth); *topA* and *gyrAB* (modulation of gene expression). The model has seven differential equations and 40 inequality constraints.

Absent the carbon starvation signal, the domain system reaches a single equilibrium state (domain) corresponding to exponential growth. Starting from this equilibrium state, flipping the starvation signal yields a 66-state transition graph possessing a single equilibrium state corresponding to stationary-phase conditions. Validation in Batt et al. (2005) surrounds the question as to whether predictions obtained from this model are concordant with experimental data. For instance, the concentration of Fis has been experimentally shown to decrease at the end of the exponential phase and become steady in the stationary phase, which agrees with model predictions. However, preliminary data indicate that the level of DNA supercoiling decreases during and after the transition to the stationary phase, which implies that the concentration of GyrAB must decrease or the concentration of TopA must increase, neither of which is predicted by the model since in all paths the TopA concentration remains constant and the GyrAB concentration increases. Thus, there are model predictions not in agreement with experimental observations and the model.

Here we are once again reminded that a scientific model is a contingent hypothesis perpetually open to rejection should its predictions fail. A model may be validated by existing data as they relate to certain aspects of the model, but refuted by new data. In the words of Popper, "The acceptance by science of a law or a theory is tentative only; which is to say that all laws and theories are conjectures, or tentative hypotheses . . . We may reject a law or theory on the basis of new evidence, without necessarily discarding the old evidence which originally led us to accept it" (Popper, 1963).

Before leaving this example, we wish to emphasize the salient role played by reduction of the full model to characteristics for the purpose of validation. In particular, the state transition diagram in the domain space is a consequence of the full piecewise-linear differential equation model. There are two technical issues with the validation that should be pointed out. First, there is no quantitative specification of the relations between model variables and the experimental measurements; second, the relation between model design and experimental data—for instance, their independence—is not clarified. However, these would be easy to address and should not cause one to lose sight of the fundamental methodology employed.

Just as one can evaluate the performance of classification rules, one can evaluate network inference procedures. An inference procedure operates on data generated by a network $\mathcal{H}$ and constructs an inferred network $\mathcal{M}$ to serve as an estimate of $\mathcal{H}$, or it constructs a characteristic to serve as an estimate of the corresponding characteristic of $\mathcal{H}$. For full network inference, the inference procedure is a mathematical mapping from a space of samples to a space of networks, and it must be evaluated as such. There is a generated data set S and the inference procedure is of the form $\psi(S) = \mathcal{M}$. If a characteristic is being estimated, then $\psi(S)$ is a characteristic.

Focusing on full network inference, the goodness of an inference procedure ψ is measured relative to some distance, μ, specifically, $\mu(\mathcal{M}, \mathcal{H}) = \mu(\psi(S), \mathcal{H})$, which is a function of the sample S. In fact, S is a realization of a random process, Σ, governing data generation from $\mathcal{H}$. In general, there is no assumption on the nature of Σ. It might be directly generated by $\mathcal{H}$ or it might result from directly generated data corrupted by noise of some sort. Moreover, ψ might include filtering to reduce noise, missing value estimation, quantization, or normalization in time, space, or quantification. If a particular application is in mind, one should generate the synthetic data in such a way as to reflect real-world data and apply the appropriate filtering scheme when validating inference (e.g., see Husmeier, 2003). $\mu(\psi(\Sigma), \mathcal{H})$ is a random variable and the performance of ψ is characterized by the distribution of $\mu(\psi(\Sigma), \mathcal{H})$, which depends on the distribution of Σ. The salient statistic regarding the distribution of $\mu(\psi(\Sigma), \mathcal{H})$ is its expectation, $E_\Sigma[\mu(\psi(\Sigma), \mathcal{H})]$, with respect to Σ.

Rather than considering a single network, we can consider a distribution, H, of random networks, where, by definition, the occurrences of realizations $\mathcal{H}$ of H are governed by a probability distribution. This is precisely the situation with regard to the classical study of random Boolean networks. Averaging over the class of random networks, our interest focuses on the expectation, $E_\mathrm{H}[E_\Sigma[\mu(\psi(\Sigma), \mathrm{H})]]$, with respect to H. It is natural to define the inference procedure ψ_1 better than the inference procedure ψ_2 relative to the distance μ, the random network H, and the sampling procedure Σ if

$$E_\mathrm{H}[E_\Sigma[\mu(\psi_1(\Sigma), \mathrm{H})]] < E_\mathrm{H}[E_\Sigma[\mu(\psi_2(\Sigma), \mathrm{H})]]. \tag{7.13}$$

Whether an inference procedure is "good" is not only relative to the distance function, but it is also relative to how one views the value of the expected distance. Indeed, it is not really possible to determine an absolute notion of goodness.

In practice, the expectation is estimated by an average,

$$\overline{E_\mathrm{H}[E_\Sigma[\mu(\psi(\Sigma), \mathrm{H})]]} = \sum_{j=1}^{m} \mu\left(\psi\left(S_j\right), \mathcal{H}_j\right), \tag{7.14}$$

where $S_1, S_2, \ldots, S_m$ are sample point sets generated according to Σ from networks $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_m$ randomly chosen from H.

The preceding analysis applies virtually unchanged when a characteristic is being estimated. One need only replace $\mathcal{H}$ and H by λ and

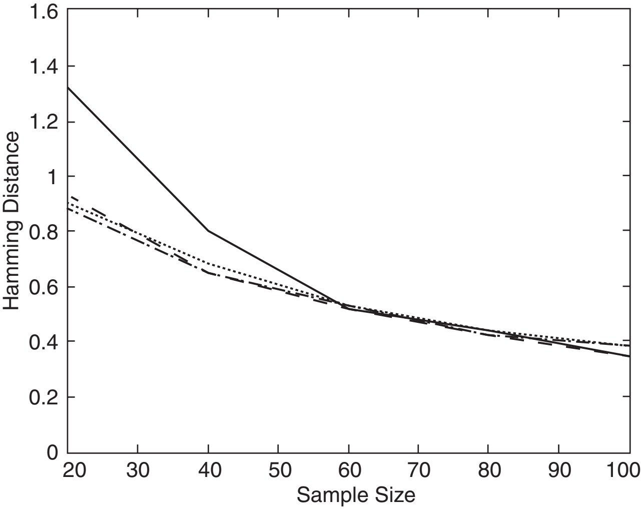

Figure 7.2 Hamming distance performance for inferring regulatory graphs using information theory: REVEAL—solid line, $\Gamma = 0.2$—dotted line, $\Gamma = 0.3$—dashed line, $\Gamma = 0.4$—dash-dot line. [Zhao, W., Serpedin, E., and E. R. Dougherty, "Inferring Gene Regulatory Networks from Time Series Data Using the Minimum Description Length Principle," *Bioinformatics*, 22(17), 2129–2135, 2006, by permission of The International Society for Computational Biology].

Λ, where λ and Λ are a characteristic and a random characteristic, respectively, and replace the network distance μ by the characteristic distance. A good deal of bioinformatics effort has gone into inferring regulatory graphs (see Werhli et al., 2006, and Marbach et al., 2010, for comparative reviews).

As an illustration, we consider inference of a genetic regulatory graph. There have been a number of papers addressing the inference of regulatory graphs using information-theoretic approaches. In a study proposing using the minimum description length (MDL) principle to infer regulatory graphs (Zhao et al., 2006), the hamming distance was used to compare the performance of the newly proposed algorithm with an earlier information-theoretic algorithm, called REVEAL (Liang et al., 1998). Figure 7.2 compares the hamming distances between the inferred networks and the corresponding synthetic networks that generated the data relative to increasing sample size. It does so for the REVEAL algorithm and the MDL algorithm using three different settings for a user-defined parameter (Γ). The performance measures are obtained by averaging over 30 randomly generated networks, each containing 20 genes and 30 edges, with the distance function being normalized over 30, the number of edges in the synthetic networks.

It is important to recognize that inference performance is a mathematical issue concerning operators on random samples. Performance of a particular inference procedure depends upon the class of networks being considered. Accurate performance analysis requires that the sample data be generated from the network class under consideration. For this reason, performance analysis using real data is problematic. If real data are employed, then the inferred network is compared, not with the unknown random network generating the data, but with a model network that has been human-constructed from the literature (and implicitly assumed to approximate the data-generating network). A network $\mathcal{H}$ (or characteristic) is constructed from relations found in the literature and $\mu(\psi(\Sigma), \mathcal{H})$ is computed. The aim is to compare the result of the inference procedure with some network related to existing biological knowledge. The problem is that the constructed network may not be a good approximation to the regulatory graph for the system generating the data. This can happen because the literature is incomplete, there are insufficiently validated connections reported in the literature, or the conditions under which connections have been discovered, or not discovered, in certain papers are not compatible with the conditions under which the current data have been derived. As a result of any of these situations, the overall validation procedure is confounded by the precision (or lack thereof) of the approximation (see Dougherty, 2007, for a mathematical characterization of the problem).

The validation of stochastic models is a subtle business and requires the solutions to difficult statistical problems. Assumptions are inevitable, not the least of which is the commonplace assumption that sample points are independent, something that can never be absolutely empirically verified. Naïve intuition has no role in scientific validation. The validation criteria must be carefully articulated, both the assumptions on which they depend and the mathematical characterization of their satisfaction. In all cases, the nature of our knowledge rests with the mathematical theory we have concerning the measurements. That cannot be simplified. If either the available theory or one's familiarity with the theory is limited, then one's appreciation of the scientific content of a model is limited.

The epistemology of science is inextricably bound up with the stochastic nature of scientific models. It is not simply a matter of relating a model operationally to empirical phenomena, running an experiment, and observing whether the predicted phenomena occur, in particular, whether the measurements are sufficiently close to the predicted measurements so that any difference can be attributed to experimental error. The stochastic nature of the model insures that the variables within the model are random variables possessing some probability distribution and the measurements must be interpreted as sample points of a probability distribution. One may observe an entire trajectory in time so that validation involves the manner in which one or more time processes occur with respect to their predicted behavior based on the model. In this book, we have avoided continuous-time processes and therefore have avoided the measure-theoretic issues surrounding them; nonetheless, the epistemological issues are analogous. The key point is that it is insufficient to propose a scientific model and give supporting empirical measurements without explicitly stating the validation criteria and characterization of the extent to which those criteria are satisfied. It is this explicitness that makes scientific enquiry intersubjective. Two investigators possessing sufficient training to understand the mathematical model and the experimental protocol, sufficient sensory capability to observe the experimental measurements, and sufficient mathematical training to understand the validation criteria and their quantification will of necessity agree with the mathematical and operational meanings of the model and the degree of its validation. However, they may not agree on whether the model should be accepted. That will depend on the degree of validation demanded by each. Before that, they might not agree on the validation criteria themselves. Intersubjectivity does not apply to agreement on scientific propositions except to the extent that there is agreement on the overall methodology.

As Reichenbach has noted, scientific truth is based on a “functional conception of knowledge,” and this means that agreement on a scientific theory depends first of all on agreement on the criteria of functionality. Agreement on the validity of a regulatory model indicates agreement on the functionality of that model relative to certain agreed-upon validation criteria; it does not indicate agreement on the totality of actual interactions in Nature. What is actually “out there in the world,” none of us knows. As Kant so well understood, the nature and forms of the things-in-themselves are not for us to know. Einstein states the matter in terms of a metaphor of a man trying to form a mental picture of the mechanism of a clock with closed case:

> Physical concepts are free creations of the human mind, and are not, however it may seem, uniquely determined by the external world. In our endeavor to understand reality we are somewhat like a man trying to understand the mechanism of a closed watch. He sees the face and the moving hands, even hears its ticking, but he has no way of opening the case. If he is ingenious he may form some picture of a mechanism which could be responsible for all the things he observes, but he may never be quite sure his picture is the only one which could explain his observations. He will never be able to compare his picture with the real mechanism and he cannot even imagine the possibility or the meaning of such a comparison. (Einstein and Infeld, 1967)

Not only must the scientist's model "explain" (fit) current observations, it must be capable of predicting new observations, with its scientific truthfulness depending on the accuracy of these predictions—all to be determined without opening the watch. Should it be surprising, then, that validation is such a subtle issue, to the point that the criteria of validation are themselves "free creations of the human mind"? True, this leaves us with a radical uncertainty as regards Nature, but it does not leave us disconnected from Nature. Elsewhere, Einstein writes, "For even if it should appear that the universe of ideas cannot be deduced from experience by logical means, but is, in a sense, a creation of the human mind, without which no science is possible, nevertheless this universe of ideas is just as little independent of the nature of our experiences as clothes are of the form of the human body" (Einstein, 1922). The creations are free but they are not independent of experience.

Centuries of scientific struggle have led to an epistemology grounded on prediction formalized within the theory of statistics. In the end, because they are grounded on agreements as to how they are manifested in experiments, our theories embody the kinds of models we can make based on the mathematical descriptions of relationships available to us. Even though the mathematical descriptions may not have any similarity in form to the actual relationships, these models are still rigorously tied to Nature, since when the models are tested by experimentation using the agreed operational definitions they reliably produce accurate predictions. We can hardly expect our science to be free of uncertainty when the foundations of our mathematics themselves are not free of paradox. When discussing difficulties in mathematics, Barrett writes, “The conventions we adopt must somehow work: they must serve us in our coping with Nature” (Barrett, 1979). What better words could one apply to the scientific enterprise? The theories must “work,” and that means they must satisfy human requirements regarding their predictive capabilities in regard to Nature.

CHAPTER 8

Sola Fides

> The objectivity of scientific statements lies in the fact that they can be intersubjectively tested.
>
> —*Karl Popper*

There are strong constraints on scientific knowledge and these impose a great burden on scientific research. It is not easy to predict the future and prediction is the basis for scientific validity. Moreover, prediction is uncertain and must be characterized within a rigorous theory of probability. Not only do the phenomena not conform to our commonsense intuitions; neither does the probability theory that must characterize uncertainty in our prediction of phenomena. Whether it is because we are raised from childhood with a naïve causal-deterministic mindset or that our mental faculties are insufficient, our intuition regarding probability is often mistaken. When confronted with the subtlety and complexity of multivariate nondeterministic systems, one may turn to univariate deterministic "insight," thereby leading to ridiculous conclusions.

The transformation from causality to expectation as the basis of science is paralleled by a student's maturation. A wet-behind-the-ears undergraduate reads Hume's *An Enquiry Concerning Human Understanding* and a metamorphosis begins: The silliness of childhood becomes starkly apparent and all the pylons supporting everyday youthful beliefs are permanently shattered. Shortly thereafter the student

Epistemology of the Cell: A Perspective on Biological Knowledge, First Edition.
Edward R. Dougherty, Michael L. Bittner.

reads Kant's *Prolegomena to Any Future Metaphysics* and feels a kinship with his famous confession: "I readily confess, the reminder of David Hume was what many years ago first broke my dogmatic slumber, and gave my researches in the field of speculative philosophy quite a different direction" (Kant, 1977). Kant is awakened from his comfortable rest in the uncritical pre-Humean world and there is no way back, either for him or the shaken undergraduate. But whatever solace the student takes in Hume's jolt to Kant quickly dissipates as Kant redefines the meaning of phenomena through the lens of his categories of understanding. The mind reels as Kant turns reason upon itself. "Do I really need to suffer through this?" the student might ask. Barrett provides an affirmative response:

> Kant . . . has more than a century of the new science to reflect upon, and he is the first philosopher to understand what has happened. The whole of his *Critique of Pure Reason* is not primarily an attempt to set up a system of idealistic philosophy; it is the effort, stubborn and profound, to grasp the meaning of the new science and its consequences for human understanding generally. . . . What has happened is nothing less than the transformation of human reason itself. (Barrett, 1979)

If the undergraduate wants to take science seriously, then he or she must experience this transformation. But this is only the first step in the cleavage from one's childhood moorings. Close to the same time, the student is introduced to the conundrums of quantum mechanics and the brain-teasing problems of a first course on probability theory. From that point on, it is a lifetime's work to gain probabilistic intuition.

If one goes on to study stochastic dynamical systems, the very lifeblood of biological science, one is forced to cling ever tighter to the mathematics while recognizing, as Kolmogorov did, the epistemological barriers to application. Surely one must pay strict adherence to mathematics when playing on a field where stochasticity and massive complexity leave intuition notoriously wanting, but there is a danger here. Cloaked in the security of mathematics, one might lose sight of the science. Perhaps this is what Fisher was thinking when he warned against relying on "the elaborate mechanism built on the theory of infinitely large samples" and not coming to grips with the statistical issues that need be addressed "to apply accurate tests to practical data." It is not simply mathematics that is required, but the appropriate mathematics for the scientific issues at hand, and this mathematics might not be sitting on the shelf.

It is not surprising that science is often unpopular, even among those who would call themselves scientists. For some, the epistemology is too demanding; for others, it is to limiting. In either case, the urge to circumvent the constraints leads to a rejection of careful predictive experiments and to the substitution of verbal explanations in place of precise mathematical language, the latter degradation resulting in "empty talk," to use Einstein's expression. The rejection of precise mathematical models and predictive experiments leads to a disconnection between the conceptual system and the phenomena, as well as a loss of intersubjectivity, the ultimate consequence being a subjectivism that is antithetical to legitimate scientific enquiry. We believe that subjectivism has grown stronger throughout the last half-century, to the point where today it is ubiquitous in certain areas of science, in particular, the one that concerns us in the present book. We are not waxing philosophical; indeed, we will use classification in conjunction with what is occurring in the genomic literature to support our contention and we will give concrete examples from among the studies we have conducted.

As discussed in Chapter 6, classification validity focuses on the quality of the error estimate associated with a classifier. Since the purpose of an error estimate is to approximate the true error on the feature-label distribution, we would like the estimated and true errors to be strongly correlated. From the perspective of RMS, the desire for strong correlation can be seen in the following equation, which shows that high correlation mitigates the deviation variance caused by the individual variances:

$$\mathrm{Var}_{\mathrm{dev}}[\hat{\varepsilon}] = \mathrm{Var}[\hat{\varepsilon} - \varepsilon] = \mathrm{Var}[\hat{\varepsilon}] + \mathrm{Var}[\varepsilon] - 2\mathrm{Corr}[\varepsilon, \hat{\varepsilon}]\sqrt{\mathrm{Var}[\hat{\varepsilon}]\mathrm{Var}[\varepsilon]}, \tag{8.1}$$

where Corr denotes the correlation coefficient. If the sample is large, then the individual variances tend to be small, so that the deviation variance is small; however, when the sample is small, the individual variances tend to be large, so that a large correlation coefficient is needed to offset these variances. Thus, the correlation between the true and estimated errors plays a vital role in assessing the goodness of the error estimator.

The correlation is related to the linear regression between the true and estimated errors. Whereas nonlinear regression of the true error on the estimated error corresponds directly to the conditional expectation $E[\varepsilon \mid \hat{\varepsilon}]$, which gives the best prediction of the true error given the estimated error, the linear regression of the true on the estimated error arises from approximating the conditional expectation by a linear model,

$$E[\varepsilon \mid \hat{\varepsilon}] \approx a\hat{\varepsilon} + b. \tag{8.2}$$

Since we would like to have the conditional expectation of the true error given the estimated error to be close to the true error, meaning $E[\varepsilon \mid \hat{\varepsilon}] \approx \hat{\varepsilon}$, we would like a to be close to 1. Geometrically, we would like the linear regression to be close to a 45° line through the origin. A perfectly flat linear regression would mean that we obtain the same prediction of the error no matter the estimated error.

Let us first consider a model-based situation in which the class-conditional distributions have been modeled by Gaussian distributions and linear discriminant analysis has been used as the classification rule. In general, a model-based approach to the study of error estimation allows one to study the effects of changing different structural parameters within the model and exact computation of the true error. A large number of samples of size 60 have been randomly drawn from the feature-label distribution and, in each case, five features have been selected from 200 features by the *t*-test, a classifier has been designed, and its error estimated by leave-one-out. Figure 8.1 shows the scatter plot for the true-estimated error pairs and the linear regression of the true error on the leave-one-out estimated error. Not only is the scatter plot widely dispersed, indicating poor correlation and high variance, but the regression line is virtually horizontal.

The downside of using a model-based approach to study error estimation is that real biological data can usually be expected to be far less structured than that resulting from standard models, thereby prompting the use of real data in error analysis. Using real data to evaluate error estimator performance is problematic from two perspectives, one methodological and the other epistemological. Methodologically, the data set has to be sufficiently large so that it can be randomly split into two subsets: (a) a training sample on which to train a classifier and estimate its error using a training sample-based

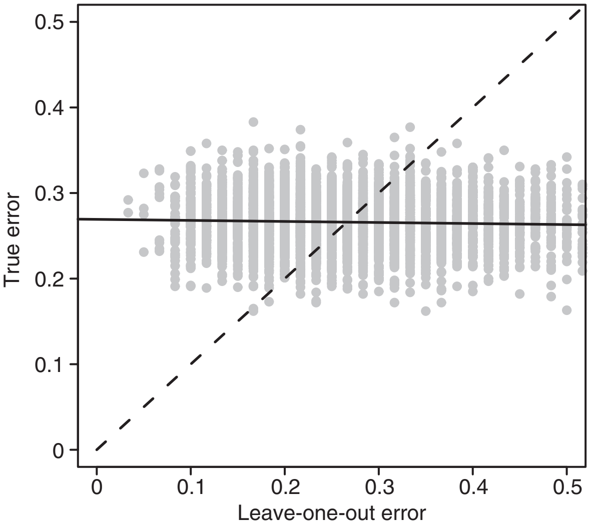

Figure 8.1 Linear regression for the true error on the leave-one-out error for sample size 60 in a Gaussian model using LDA.

error estimator; and (b) a disjoint test set on which to obtain a precise estimate of the true error. The data set has to be large enough that the disjoint test set is large enough for precise estimation and that the training and test sets obtained in successive random splits are close to being independent, which can only be done for large data sets because for them the dependence is small.

Epistemologically, one supposes that the data set has been generated from some unknown feature-label distribution and that the estimated error rate corresponds to the true error of the classifier on this unknown distribution; indeed, the whole idea of error estimation presupposes an error to estimate. Practically, the data have been obtained from some measurement process, say microarray readings on a sample of cancer patients in which the patients are divided into two classes. If the data set is large enough to satisfy the methodological demands, then the epistemological ground of the overall procedure is that, when applied to future cancer patients in the two defined classes, the classifier will have a fixed error rate and the error rate computed on the test sample is a very accurate estimate of this error rate. There is a serious epistemological problem here: What is meant by the "defined classes"? If these classes are generated by a tightly controlled experimental protocol, then one has some confidence that the classes are meaningful; however, if they are merely defined by some set of measurements across a widely diverse population, they may, in fact, not be well defined because, given a patient, it may not be decidable with a strong degree of certainty as to which class the patient belongs. To the degree that there is a tight experimental protocol resulting in two labeled classes, hypothesizing an ideal theoretical distribution from which the data have been drawn has justification and error estimation is meaningful; to the extent that data are grouped in some ad hoc manner, labeling loses its justification and, not only is the notion of an ideal feature-label distribution problematic, but the entire notion of prediction, and therefore prediction accuracy, also loses its meaning.

Supposing that the data set we have justifies performance evaluation, we will illustrate error regression and correlation using microarray data. We use the same 295-patient breast cancer data set as used in Fig. 6.5. The data set is reduced to a selection of the 2000 genes with highest variance, these are reduced to 50 by using *t*-test feature selection, and a classifier is designed using the LDA classification rule. In the simulations, we divide the data into two sets. The first set consists of 50 examples drawn without replacement from the full data set. It is used for both training and training sample-based error estimation. The remaining examples are used as a hold-out test set to get an accurate estimate of the true error, which is taken as the true error. This procedure is repeated 10,000 times. Figure 8.2 shows the scatter plot for the true-estimated error pairs and the linear regression of the true error on the leave-one-out estimated error. As in Fig. 8.1, there is wide dispersion and a virtually horizontal regression line. Figures 8.1 and 8.2 should provide a stark warning about cross-validation.

The flat regression lines in Figs. 8.1 and 8.2, which have been arrived at by Monte Carlo simulation, raise an obvious question: Can there actually be negative correlation and a negatively sloped regression line between the true error and the leave-one-out cross-validation estimate? In the case of multinomial discrimination, it has been demonstrated analytically that negative correlation arises between the true error and leave-one-out error estimate for small samples. Figure 8.3 is taken from Braga-Neto and Dougherty (2010), which gives analytic formulas for the correlation between the true and estimated errors for the discrete histogram rule. The *x*- and *y*-axes correspond to the sample size and correlation coefficient, respectively. A Zipf distribution is employed and the figure corresponds to the model with Bayes error

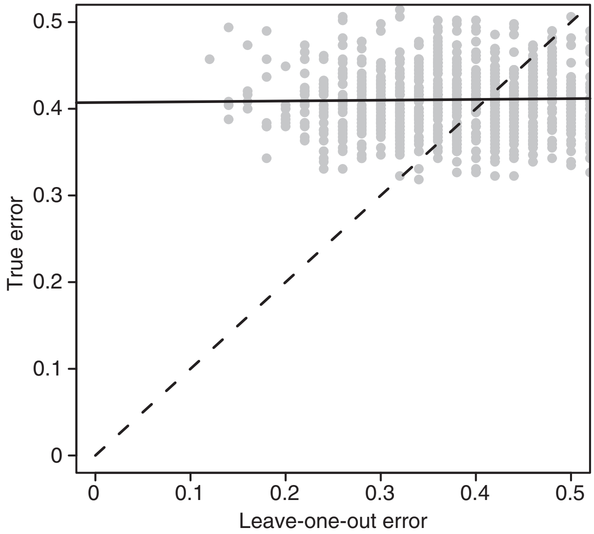

Figure 8.2 Linear regression for the true error on the leave-one-out error for sample size 50 using breast-cancer data and LDA.

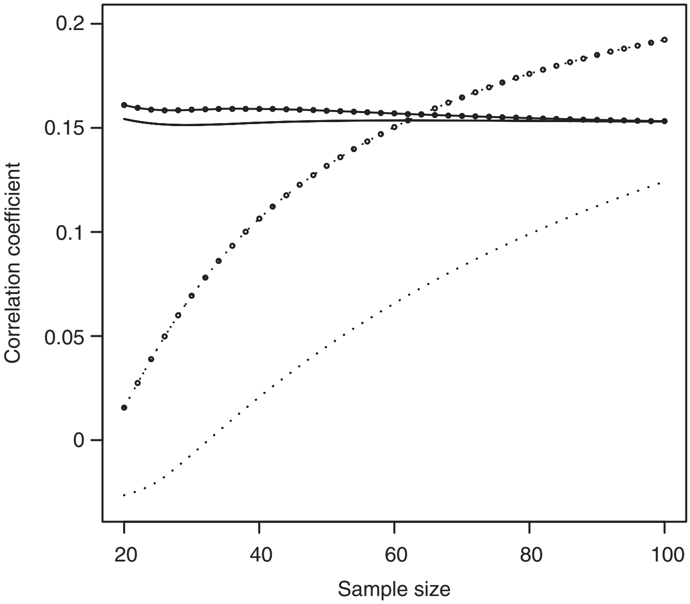

Figure 8.3 Correlation coefficient for true error with both resubstitution and leave-one errors as a function of sample size in multinomial discrimination: solid lines are for resubstitution and dotted lines are for leave-one-out. The lines with and without the circles correspond to bin sizes $b = 16$ and $b = 32$, respectively. (Braga-Neto and Dougherty, 2010, by permission of *Pattern Recognition Letters*).

0.10. The solid and dotted lines are for resubstitution and leave-one-out, respectively. The lines with and without the circles correspond to bin sizes $b = 16$ and $b = 32$, respectively. For $b = 32$, the leave-one-out estimate is negatively correlated with the true error for $n < 30$.

A popular way to demonstrate the efficacy of a proposed classification rule in biomedical applications is to use ROC curves (receiver operating characteristic curves). Leaving the mathematical details to the literature, we simply note that the more the curve extends above the 45° line, the better the classifier performance, and this is quantified by the area under the ROC curve (AUC). When using leave-one-out error estimation, variance problems can even be worse than for error estimation. The risk of using ROC curves with small samples is exemplified in Fig. 8.4. It comes from a model-based study (Hanczar et al., 2010) and shows sample sizes 50 and 100. The gray and black curves correspond to leave-one-out error estimation and the true error, respectively. The solid lines indicate the mean ROC curves. The dashed lines provide the corresponding 95% confidence bounds. The black dashed lines show the effect of random sampling. The gray dashed lines show the combined effect of error estimation and random sampling. Leave-one-out has significantly widened the confidence bands, to the point

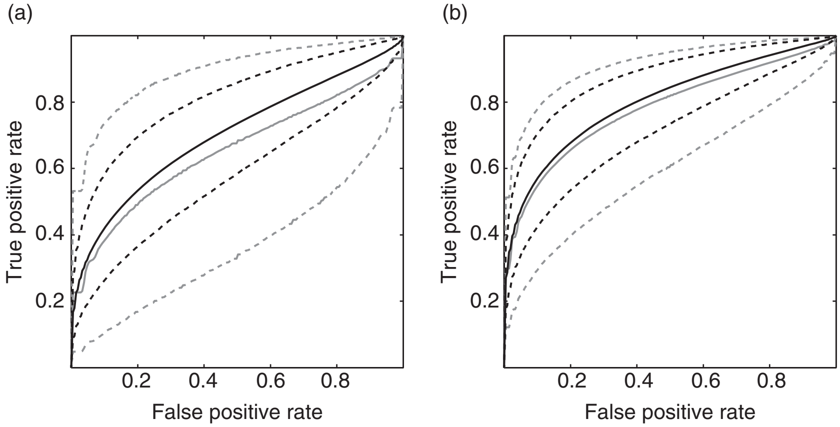

Figure 8.4 ROC confidence intervals: (a) $n = 50$; (b) $n = 100$. [Hanczar, B., Hua, J., Sima, C., Weinstein, J., Bittner. M. L., and E. R. Dougherty, “Small-Sample Precision of ROC-related Estimates,” *Bioinformatics*, 26(6), 822–830, 2010, by permission of The International Society for Computational Biology].

where, for sample size 50, they contain the line $y = x$. Even with sample size 100, the lower 95% confidence bound is barely above the line $y = x$. One can only imagine the medical “benefits” accruing from evaluating classifier performance using small-sample ROC curves.

Given the kind of results observed in Figs. 8.1 through 8.3 and demonstration of the excessive variance of cross-validation resampling methods across a wide array of models (Braga-Neto and Dougherty, 2004), the large number of papers in the scientific literature using cross-validation with such small samples is remarkable. It is certainly not the case that the variance problems with cross-validation have only recently been discovered. In a classic 1978 paper, Ned Glick considered LDA classification for one-dimensional Gaussian class-conditional distributions possessing unit variance, with means μ_0 and μ_1, and a sample size of $n = 20$ with an equal number of sample points from each distribution (Glick, 1978). Figure 8.5 is based on Glick’s paper; however, we have increased the Monte Carlo repetitions from 400 to 20,000 for increased accuracy. The x-axis is labeled with $m = |\mu_0 - \mu_1|$, with the parentheses

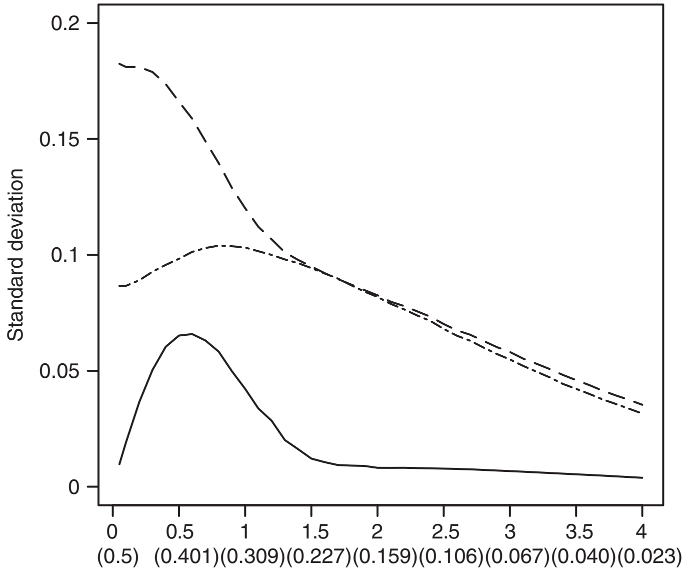

Figure 8.5 Standard deviation plots as a function of the distance between the means (and Bayes error) for LDA discrimination in a one-dimensional Gaussian model: true error (solid); resubstitution error (dots); and leave-one-out error (dashes).

containing the corresponding Bayes error. The figure shows standard deviation plots of the true (solid), resubstitution (dots), and leave-one-out (dashes) errors as functions of *m*. When the Bayes error is small (large *m*), the standard deviations of the leave-one-out and the resubstitution errors are close, but when the Bayes error is large, the leave-one-out error has a much greater standard deviation. Glick was sufficiently concerned that, with regard to the leave-one-out estimator, he wrote, "I shall try to convince you that one should not use this modification of the counting estimator"—not even for LDA in the Gaussian model.

Recognizing the risks of small-sample classifier design, authors have sometimes proposed using additional computational analyses to support the validity of a classifier. Unfortunately, the supporting methods themselves may not have been demonstrated to be informative. For instance, some papers suggest the use of permutation-based *p* values for obtaining information regarding the selection of relevant genes or for assessing the quality of classification. Essentially, a statistic relating to class discrimination is computed from the data, the class labels are randomized some large number of times, the statistic is computed for each relabeling, a histogram is formed from these relabeled statistics, and the *p* value of the statistic corresponding to the actual labeling is computed. The issue is whether this *p* value is informative. If the *p* value gives insight into the distribution of the error or the reliability of the estimated error, then an argument can be made for using the *p* value to assess classifiers. Since the randomly relabeled data contain little or no information on the true joint distribution of the features and the labels, any insight based on the *p* value must come solely from the estimated error.

Because the formulation of the relevant hypothesis test is somewhat technical, we leave a precise mathematical description to the literature (Hsing et al., 2003). Suffice it to say that, intuitively, the null hypothesis H_0 is taken to mean that the classifier does not discriminate and the alternative hypothesis H_1 is taken to mean that it does discriminate. If ε_0 and ε_1 are the error estimates for the randomized and actual data, respectively, then *p* is the probability that $\varepsilon_0 \leq \varepsilon_1$. The top part of Fig. 8.6 gives the *p* value as a function of the estimated error for the actual data and the bottom part gives the distribution of the error estimates, these being for three-nearest-neighbor classification, sample size 40, leave-one-out error estimation, and a Gaussian model for which the

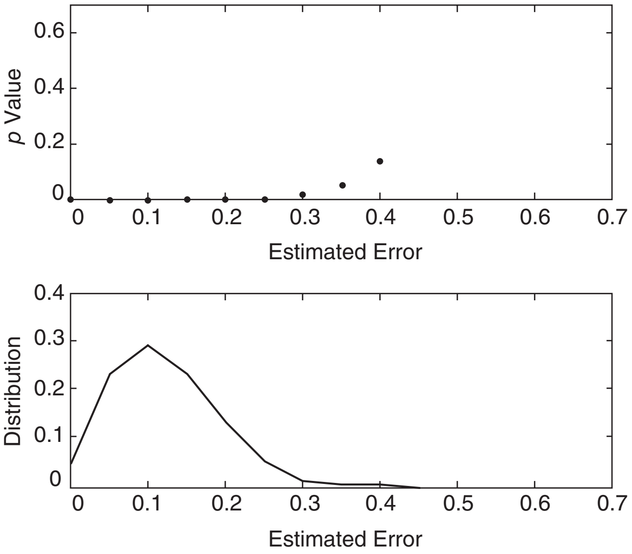

Figure 8.6 Regression of the permutation *p* value on the estimated error: Top part: *p* value as a function of the estimated error for actual data. Bottom part: error distribution.

optimal classifier has error 0.10 (Hsing et al., 2003). Comparing the two parts of the figure, we see that, for the region where the mass of the error estimates lie, there is virtually no regression of the *p* value on the error estimate. Thus, the *p* value says essentially nothing about the error and is therefore useless as a classifier performance measure. The problem is that the hypothesis test is irrelevant to the problem at hand, which is the validity of the classifier model, the key issue being the accuracy of the error estimate.

Given the massive amount of energy, time, and money spent on algorithmically designing classifiers and obtaining error estimates using estimation procedures lacking any semblance of mathematical justification—indeed, which have been demonstrated to perform very poorly in precisely the kind of environments in which they are being used—one is compelled to question the basis on which the results are being justified by those producing them. Given the absence of any scientific ground or even a half-hearted attempt to supply appropriate epistemological grounds, we come to the conclusion that it is justification by faith alone. Scientific epistemology is being abandoned in favor of the great cry of the Protestant Reformation: *Sola fides*! Of course, Martin Luther was speaking of justification and faith in the context of religion, but if we check *Webster's* dictionary, we see that the first definition of faith is "unquestioned belief"; it is only in the second and third definitions does that unquestioned belief refer to God or religion. One who presents a classifier and error estimate computed from some data set absent any validating criteria to support the conclusions is certainly proceeding with unquestioned belief. Had he or she stopped to question belief in the results, then the entire matter would have been put on hold until there was solid validation. In fact, what we behold is exactly the opposite: absolutely no effort at validation. What else can one conclude except that the entire study and its conclusion have been justified by faith alone?

The desire for *sola fides* is not simply tacit; rather, it has open adherents. We quote Julian Simon in the preface to his book, *Resampling: The New Statistics*:

> Monte Carlo resampling simulation takes the mumbo-jumbo out of statistics and enables even beginning students to understand completely everything that is done. . . . Resampling refers to the use of the observed data or of a data generating mechanism (such as a die) to produce new hypothetical samples, the results of which can then be analyzed.… Even many experts are unable to understand intuitively the formal mathematical approach to the subject. Clearly, we need a method free of the formulas that bewilder almost everyone. (Simon, 1997)

Taken as a whole, this set of statements affirms a "scientific" method without a mathematically characterized relation to empirical observation. *Sola fides*! Simon denigrates a rigorous scientific epistemology based on mathematical statistics as "mumbo-jumbo" that cannot be intuitively understood my "many experts" and should therefore be abandoned in favor of something that even beginning students can understand. The notion that a beginning student should understand statistics is preposterous. It is a difficult subject requiring a strong background in mathematical analysis and probability theory. Statistical theory is often counterintuitive to everyday thinking, especially when small samples are involved or the underlying distributions are highly complex. Only through rigorous mathematical training can one hope to achieve a proper understanding when it comes to difficult issues like representation of the joint distribution of the true error and estimated error of a classifier. No doubt, there are situations where resampling has benefits, but these can only be determined by rigorous probabilistic characterization of re-sampling, for instance, rates of convergence or bounds on approximations. We have already illustrated in this chapter that resampling is quite useless and, in fact, leads to misleading results when used in the wrong setting. In general, the conditions under which producing "hypothetical samples" is beneficial must be understood and this can only be done using fundamental theorems—for instance, by bounding the RMS in the case of cross-validation error estimation.

Simon undercuts his own position in this regard when he writes,

> This leads to valuable student discussion about whether the probability of a girl is exactly half (there are about 105 males born for each 100 females), whether .5 is a satisfactory approximation, whether four coins flipped once give the same answer as one coin flipped four times, and so on. Soon the class decides to take actual samples of coin flips. And students see that this method quickly arrives at estimates that are accurate enough for most purposes. Discussion of what is "accurate enough" also comes up, and that discussion is valuable, too. (Simon, 1997)

The epistemological issue concerns the meaning of "accurate enough" and this can only be addressed via characterization of the binomial distribution, which models the repeated coin flips. Moreover, Simon is depending on Bernoulli's law of large numbers, the oldest such law in probability theory, to assure convergence of the probability estimates to the probability. Thus, while eschewing theory, he is depending on theory to have confidence in convergence, in addition to the fact that he needs to define a suitable notion of convergence to even discuss convergence.

The kind of loose talk surrounding probabilistic convergence exemplified by Simon's "accurate enough" statement has proven, and continues to prove, detrimental to biology. Another example of this kind is the loose assumption of normality based on central limit theorems. Early central limit theorems were proven by De Moivre and Laplace, but it was Liapunov who first provided very general conditions under which a sum of independent random variables will be guaranteed to converge to a Gaussian distribution, with more general conditions subsequently provided by Lindbergh. Fine, but these are limit theorems. Only by imposing distributional assumptions can one get at the rate of convergence and, therefore, applicability to real-world problems that do not suppose “infinitely” large samples. Again we return to Fisher’s concern with relying on “the elaborate mechanism built on the theory of infinitely large samples.” One might posit a categorical assumption that a limiting condition is presupposed for application of the conclusions and this is the only assumption required for the resulting analysis; however, this kind of assumption is epistemologically unacceptable because it is not an assumption on the biological variables. An assumption on the expression-phenotype (feature-label) distribution is a biological assumption that can be phenomenally tested and therefore leads to biologically interpretable limiting conditions, whereas just to make a blanket statement regarding the appropriateness of limiting condition is unrelated to the phenomena and rests on faith alone.

The desire for theory-free science is exemplified by the use of data sets, rather than probability distributions, to evaluate the performance of proposed classification rules. It is common practice to apply a proposed classification rule on a number of data sets and compare its performance with one or more other classification rules. If the proposed classification rule is applied to a number of random samples of a certain size from a particular feature-label distribution, then the error of each can be accurately assessed and the expected error of the classification rule accurately determined. One could go further and randomize the sample sizes and obtain an accurate average error. The key point is that knowledge of the feature-label distribution permits precise error computation. On the other hand, if real data sets are used and these are not sufficiently large to provide a very small RMS for whatever error estimation procedure is being employed, then the error estimates across the data sets vary widely from their true values. The result is that, when an ordered list of estimated errors is created, the better estimates in the list tend to be biased optimistically and the worst tend to be biased pessimistically. Thus, reporting the better end of the list creates reporting bias (Yousefi et al., 2010). For instance, if one is testing a classification rule to be used on various types of cancer data, one might mistakenly assume that the rule works better for some kinds of cancer types than others, but this might simply be due to randomness in the error estimate.

The effect of reporting bias is shown in Fig. 8.7, which comes from a simulation study in which *m* samples are drawn from a feature-label distribution, an LDA classifier is designed on each sample, the estimated is computed by cross-validation on each sample, and the error

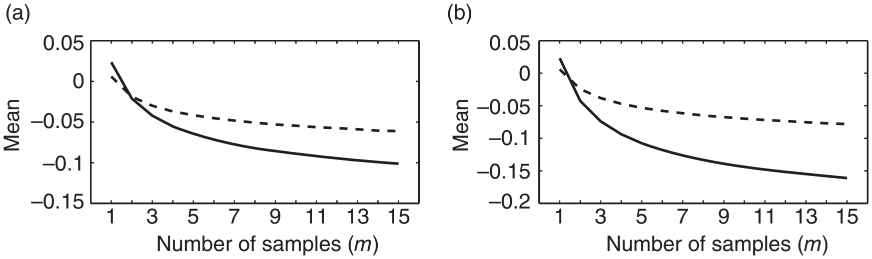

Figure 8.7 Reporting bias: (a) average deviation of the estimated and true errors for the sample showing the least estimated error; (b) average deviation of the estimated error for the sample showing the least estimated error and the expected true error across all the samples.

estimates are ordered from the lowest to the highest. Since the true errors can be found from the feature-label distribution, the deviation, $\delta(m)$, between the lowest error estimate and true error for that sample can be computed as a function of the number of samples. This process is repeated a large number of times and the average deviation, μ_{dev}, computed. These average deviations are plotted in Fig. 8.7(a), the dashed and solid lines for samples of sizes 60 and 120, respectively. The negative bias is seen in the increasingly negative values for μ_{dev} as the number of samples grows. Figure 8.7(a) does not reveal the full extent of the problem because it only measures the difference in estimated and true errors for the sample on which the classier error estimate is the lowest. Since the samples are all randomly drawn from the full feature-label distribution, the expected performance of the classifier should be the same for all samples. Figure 8.7(b) shows the average deviation between the smallest estimated error and the expected error across all the samples. This gives the true extent of the bias and we see that it is quite large even when only five samples are drawn. The problem with testing on real data is clear: The samples need to be sufficiently large so that variation in error estimation does not give the illusion of good performance, and only a model-based theoretical analysis can determine sufficiency.

Corresponding to the bias resulting from applying a single classification rule on many data sets is the bias arising from applying many classification rules (Boulesteix and Strobl, 2009); in particular, trying many feature sets (Zhao et al., 2010). We consider LDA in a Gaussian model and select two features out of 50. In Fig. 8.8, the dashed, solid,

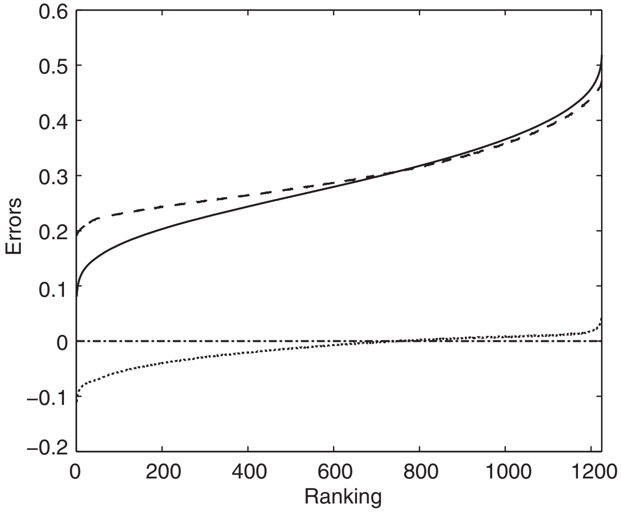

Figure 8.8 Error versus ranking curves for a Gaussian model: estimated error—solid line; true error—dashed line; error difference—dotted line. (Zhao et al., 2010, by permission of *Cancer Informatics*).

and dotted curves show the true error, estimated error, and difference between the true and estimated errors, respectively, for all possible feature sets, ranked according to their cross-validation estimated errors. At the top end of the ranking the estimated errors are optimistic; at the low end, they are pessimistic. We only care about the top end. If we do an exhaustive search and select the feature set with the lowest estimated error, then we can expect that the estimated error is severely low biased. Even if one takes a list of feature sets with the lowest estimated errors, if the list is too short, then, there may not be any good features sets in the list; if the list is too long, then there will likely be good feature sets but the list will be of little practical value.

A half-century ago, the cavalier attitude regarding scientific epistemology expounded by Simon (and others) would have been remarkable, but today it appears to be commonplace, as noted by various of papers in regard to biology (Dougherty and Brun, 2004; Mehta et al., 2004; Keller, 2005; Braga-Neto, 2007; Dupuy and Simon, 2007; Dougherty, 2008; Boulesteix, 2010; Jelizarow et al., 2010). Much contemporary literature seems to support the abandonment of quantifiable relations between theory and observation in favor of faith alone. How else can one explain the proliferation of the unjustified used of cross-validation, permutation tests, and bootstrap methods? The scientific literature is now home to a cadre of "experts" who "completely understand everything that is done" and whose work is "free of the formulas that bewilder almost everyone." These new "experts" are unencumbered by deep knowledge and have been spared the humbling rigors of the "mumbo jumbo" in books like Loeve's *Probability Theory* or Cramer's *Mathematical Methods of Statistics*. Biological science is particularly vulnerable to these new "experts" because biological knowledge inheres within the theory of stochastic dynamical systems, so that its mathematical foundations and validation are fully dependent on random processes and the statistical issues surrounding application of those processes.

Once experiments have been put aside in favor of data mining (fast groping in the dark) and predictions are not characterized in a precise mathematical statistical framework, there is not much alternative to *sola fides* if one is going to believe that the resulting models are not simply pure mathematical systems. To believe there is meaning to a scientific theory entails that the theory is related to sensory perceptions. In eschewing mathematically rigorous experimental predictions, while simultaneously believing a theory contains knowledge, one is expressing an unquestioned belief because, ipso facto, any questioning of the belief must involve questioning the relation of the model to phenomenal observations. Surely, there is no scientific questioning of a mathematical system absent questioning the empirical viability of the system.

Sola fides in science is inseparable from a radical subjectivism with regard to Nature. Popper writes, "The objectivity of scientific statements lies in the fact that they can be intersubjectively tested" (Popper, 1959). The evisceration of the experimental methodology manifested in a lack of concern for rigorous statistical estimation represents a rejection of intersubjectivity because it leaves the interpretation of the experimental results completely subjective.

Perhaps during ancient or medieval times one could conceive of science based on the physical truth of human conceptions such as space, time, and causality, but that time is past and we cannot go back. Causality in science cannot be resurrected from Hume's death-dealing critique, nor can the "truth" of Euclidean geometry be resurrected following Einstein's conception of space-time. Even before Hume's devastating attack on causality and the existence of a physically intuitive understanding of natural phenomena, Galileo and Newton had set forth on an inexorable course to modernity by bracketing causality, a bracketing from which it has never emerged. Pure reason has been put in her place in regard to the knowledge of Nature. If prediction is rejected as the fundamental epistemological requirement for relating conceptualizations and phenomena, then faith alone is all that remains.

In part, the breakdown of scientific thinking embodied in theoretically ungrounded algorithmic approaches to model construction is a consequence of an overzealous, even slavish, infatuation with computation. Ever newer technologies produce data in ever growing orders of magnitude and ever faster computers process the data through ever more complicated algorithms to produce ever more complex models. Each step in the algorithm seems to make sense. Perhaps each step in isolation has some mathematical foundation. Maybe so, but the overall algorithm is never shown to converge to the solution that it is purported to either reach or approximate and the complex model it produces is never validated.

The lure of contemporary high-throughput technologies is that they can measure tens, or even hundreds, of thousands of variables simultaneously, thereby spurring the hope that complex patterns of interaction can be sifted from the data; however, two limiting problems immediately arise. First, the vast number of variables implies the existence of an exponentially greater number of possible patterns in the data, the majority of which likely have nothing to do with the problem at hand and a host of which arise spuriously on account of variation in the measurements, where even slight variation can be disastrous owing to the number of variables being considered. A second problem is that the mind cannot conceptualize the vast number of variables. Sound experimental design constrains the number of variables to facilitate finding meaningful relations among them. Recall Einstein's comment that, for science, "the truly creative principle resides in mathematics." The creativity of which Einstein speaks resides in the human mind. There appears to be an underlying assumption to data mining that the mind is inadequate when it comes to perceiving salient relations among phenomena and that machine-based pattern searching will do a better job. This is not a debate between which can grope faster, the mind or the machine, for surely the latter can grope much faster. The debate is between the efficacy of mind in its creative synthesizing capacity and pattern searching, whether by the mind or the machine. Barrett notes, "The absence of an intelligent idea in the grasp of a problem cannot be redeemed by the elaborateness of the machinery one subsequently employs" (Barrett, 1986).

Reflecting on Kant's previously quoted passage, "It is only the principles of reason which can give to concordant phenomena the validity of laws, and it is only when experiment is directed by these rational principles that it can have any real utility," Barrett writes,

> The scientist's mind is not a passive mirror that reflects the facts as they are in themselves (whatever that might mean); the scientist constructs models, which are not found among the things given him in his experience, and proceeds to impose those models upon Nature. And he must often construct those models conceptually before they are translated at any point into the material constructions of his apparatus in the laboratory. . . . The imprint of mind is everywhere on the body of this science, and without the founding power of mind it would not exist. (Barrett, 1986)

Not only does Barrett's comment reflect Kant's thinking at the end of the Enlightenment, it is consistent with Einstein's notion of creativity in the twentieth century. Does anyone really believe that data mining could produce the general theory of relativity?

No doubt, for some the resurgence of groping in the dark is simply a consequence of the mystification of technique, a blindness caused by the marvels of computation, not an intentional desire to return to pre-Galilean science. In the case of data mining, perhaps it is the collective amnesia of a horde of technicians drunk on technique. But for others, the call to return to pre-Galilean times has been made explicit, and not simply as a response to growing data sets and more powerful computers. Strangely enough, this call to medieval (or earlier) science has been made in reference to the complexity of biological systems, this in the face of nearly a century of progress demonstrating that the representation and analysis of complex systems is best, if not only, handled in the framework of modern stochastic processes. Following in the footsteps of Einstein, Levy, Kolmogorov, Wiener, and other luminaries leading the development of stochastic processes in physics and systems theory have been the successes of signal processing, communications theory, and control theory, all dependent on rigorous mathematical stochastic analysis. Calls on the part of biologists to revert to medieval science in the face of the successes of systems theory, a theory so suitable to biology, are at least as astonishing as the call of computational technicians to abandon the mathematical and experimental bases of scientific validation.

An explicit call for a return to the medieval past has been made by Werner Callebaut and Manfred Laubichler in a editorial entitled, "Biocomplexity as a Challenge for Biological Theory," in which they comment on a statement of Schrödinger that we have quoted previously and which reads, "The relation of cause and effect, as Hume pointed out long ago, is not something that we find in Nature but is rather a characteristic of the way in which we regard Nature" (Schrödinger, 1957). Referring to this quote appearing in Dougherty and Braga-Neto (2006), they write, "Causation is still regarded here, with Hume and Kant, as 'a characteristic of the way in which we regard Nature' rather than intrinsic to Nature, which amounts to nothing less than renouncing knowledge of Nature itself" (Callebaut and Laubichler, 2007). So we are to go back to Aristotle! It is astonishing that anyone could approach biocomplexity and stochastic dynamical systems with a pre-Galilean mindset.

Yet biologist Stuart Kaufman explicitly advocates just such a position in an essay entitled, "Breaking the Galilean Spell." Kaufman writes,

> Galileo rolled balls down incline planes and showed that the distance traveled varied as the square of the time elapsed. From this he obtained a universal law of motion. Newton followed with his *Principia*, setting the stage for all of modern science. With these triumphs, the Western world came to the view that all that happens in the universe is governed by natural law. . . . The Galilean spell that has driven so much science is the faith that all aspects of the natural world can be described by such laws. Perhaps my most radical scientific claim is that we can and must break the Galilean spell. Evolution of the biosphere, human economic life, and human history are partially indescribable by natural law. This claim flies in the face of our settled convictions since Galileo, Newton, and the Enlightenment. (Kauffman, 2008)

Certainly one can criticize Kauffman for a confusion of metaphysics and science, the latter making no claims concerning "all that happens in the universe." Moreover, it is certainly not a "settled conviction" that science can fully describe the biosphere—quite the opposite, since the strong limitations of science constitute one of its key epistemological traits. Nonetheless, while Kaufman may not demonstrate a sound understanding of scientific epistemology, his main point is clear: He is dissatisfied with the limitations of science. Compelling evidence for this is contained in a statement he made in a different article:

> What we think of as natural law may not suffice to explain nature. We now know for example, that evolution includes Darwinian pre-adaptations—unused features of organisms that may become useful in a different environment and thus emerge as novel functionalities, such as our middle ear bones, which arose from the jaw bones of an early fish. Could we pre-state all the possible Darwinian pre-adaptations even for humans, let alone predict them? It would seem unlikely. And if not, the evolution of the biosphere, the economy and civilization are beyond natural law. If this view holds, then we will undergo a major transformation of science. Partially beyond law, we are in a co-constructing, ceaselessly creative universe whose detailed unfolding cannot be predicted. Therefore, we truly cannot know all that will happen. (Kauffman, 2007)

Kauffman is dissatisfied with science on several accounts, Science cannot explain Nature. Science cannot predict outcomes in a set that is not well defined. Science cannot provide human beings with the knowledge of all that will happen! To alleviate these impediments, he claims that "we will undergo a major transformation of science." This is a remarkable call to return to a time when scientists thought that they could explain Nature, mathematicians did not understand set theory, and scientists believed that they could predict the entire future of the universe. It is difficult to say what this means; however, the transformation he desires would bring us at least to a time when the sun was believed to revolve around the Earth.

The confusion in the epistemological thicket that Kauffman would have us enter is illustrated by his own words. He says that he "knows" that "evolution includes Darwinian pre-adaptations," but he gives no idea of the epistemological ground of that knowledge. He specifically states that the knowledge is not scientific because he says that pre-adaptations are not predictive. So what is the ground of his knowing? The answer seems clear: *Sola fides*.

CHAPTER 9

Model-based Experimentation in Biology

> An experiment is a question. A precise answer is seldom obtained if the question is not precise; indeed, foolish answers—i.e., inconsistent, discrepant or irrelevant experimental results—are usually indicative of a foolish question.
>
> —*Arturo Rosenblueth and Norbert Wiener*

From the discussion so far, it might seem as if the building of mathematical and logical models to serve as a basis for experimental design to test and validate predictions concerning biological behavior is completely alien to biology. This is certainly not true. Biology has advanced in different ways depending on what aspect of biology is being considered. We will discuss two forms that represent approaches from different levels, macromolecular components, a bottom-up approach, and cells, a top-down approach. In this discussion, our intent is to focus on both the types of experimentation that result when starting from either perspective and on the availability of constraints that can focus the experimentation so that it becomes more practical to carry out.

Within biology, the understanding of which design approach has been taken in previous work is itself controversial. Much of molecular biology, which focuses on how particular macromolecules carry and process information, is often thought to be the result of an entirely

Epistemology of the Cell: A Perspective on Biological Knowledge, First Edition.
Edward R. Dougherty, Michael L. Bittner.

bottom up approach. It is certainly true that discoveries in the 1950s and 1960s concerning the ways in which nucleic acids code information focused a great deal of attention on the various mechanisms that allowed this information to be used to control and direct the synthesis of other macromolecules. Nonetheless, the foundational work on which the coding work was based had its origins in a much higher-level formulation of basic biological problems. The theory of evolution implied the existence of numerous capabilities that cells would need in order to implement heredity, the passing on of the information specifying an organism's type over generations. Work over the 90 years following publication of the theory provided both validation of a physical means of cell-to-cell passage of information and, ultimately, the implication that DNA was the heritable agent. All of this work was directly driven by a model that supposed that a descendent of an individual organism or of a mating pair of organisms received components carrying information from the parent(s) that would have a guiding effect on the development of the new individual. The model further specified that the information transfer was not perfect, so that there were ways in which some fraction of the information could be different from that of the source individual(s). A well-known example of mathematical modeling and experimental testing of the model is Mendel's studies of heredity. These explicitly tested some of the core concepts of transmission, modeling the transmission of each parent's contribution in a simple mathematical way and introducing the possibility of the contributions being of unequal potency in driving the production of characteristic phenotypes, trait dominance, and recessivity.

Thus, the higher level, systems view of the implied capabilities required for heredity led directly to the generation of the types of questions and experiments that came to be recognized as the province of the fields of genetics and molecular biology. The rapid acceleration of progress in these fields following the first understandings of the mechanics of coding had the effect of making the field seem relatively unlinked from the past and has had the effect of blinding many investigators to the ways in which the original questions played a necessary role in guiding the field inexorably toward the mechanisms by which molecules carry information and by which the information can be used to allow cells to carry out the varied functions required for the survival and specialization of cells and of the organism of which they are parts.

During this same time period another biological field, development, was also experiencing considerable growth and success. Development focused more directly on the top-down aspects of cell operations at the systems level. Studies in this field were driven to this higher-level view because the processes guiding the many intricate changes in cell activities supporting the development from egg to embryo were initially unapproachable from the molecular level and experiments could only be carried out at the cell and organism levels. As this complication is still well remembered by many developmental researchers, molecular biology is seen as a useful tool for trying to understand the components of the developmental processes rather than a way of discovering developmental processes.

A number of organisms, such as frogs, newts, and fruit flies, produce eggs and developing organisms of suitable size and ease of manipulation for direct visual studies of early through late developmental events. Frog eggs provided one of the most widely exploited experimental platforms for studying early embryonic development. Amphibian eggs were a particularly well-suited starting point for following the dynamic series of cell multiplications and movements that occur very early in embryogenesis and many early thoughts concerning the manner of regulating factors and events that might underlie the development of the body plan from a single cell were developed during this time. These included ideas and clever experiments, where eggs were mechanically divided along different axes, demonstrating that the visible differences in the architecture of the egg cell were the result of asymmetric deposition of components important in guiding successful development. Simple cell marking and tracking experiments demonstrated that differentiation of cells into three particular types forming the ectodermal, mesodermal, and endodermal cell layers was key to the development of the tubular body plan. Close observations of finer patterning in insects also stimulated early thoughts on what could possibly produce these highly regular partitions of embryonic space, leading theorists such as Alan Turing to suggest the presence of "morphogen" molecules whose diffusion through a region of the body could provide a gradient that could be sensed and acted upon to produce a wide variety of patterns (Turing, 1952). Based on a large number of experimental observations following the dynamics of change of cell positioning in the embryo and the associated types of cell differentiation happening as the cells experienced different interactions, developmental biologists accurately predicted cell compartments that would follow a specific developmental trajectory involving both ultimate spatial placement and state of differentiation in the adult organism. They could also accurately predict what local interactions at what point in development were required to achieve this trajectory, together with the results of artificially forced interactions of a type not normally experienced. The campaign of experimentation carried out in developmental studies is the best example of the benefits of considering single cells as systems and organisms as systems of cells as the primary level of experimental focus that has yet been achieved. Excellent reviews of the subject, such as that of J. M. W. Slack (Slack, 1983), have much to offer anyone seriously interested in the design of biological experimentation aimed at systems level understanding.

These studies of the dynamics of development at the whole-organism level have had much the same stimulatory effect on biology that the theory of evolution had, and for many of the same reasons. Researchers looking to find the ways by which the clearly defined steps in development could be implemented by cells were searching for molecular machinery capable of guiding the cells in very specific ways. The constraints of having to find system components involved in driving very particular types of system processes provided the basis for constructing experiments that could identify macromolecules that were key to the processes for which a physical description of the starting and ending states of the system were available. Molecular biology and genetics were used to carry out mutational screens that identified gene mutations leading to the disruption of development at a very precise step, and molecular biology and biochemistry were used to gain insight into the kinds of activities these molecules exhibited.

The extraordinarily salutary effects of combining system knowledge that considerably constrains the types of components that might serve to guide and carry out embryonic partitioning with knowledge derived from genetics, molecular biology, and biochemistry concerning genes whose function appears to fit the requirements produced many experiments that efficiently revealed the key players and many of the key relationships that produce the drosophila body plan. As one expects of biology, the methods used ultimately rely on the actions and interactions of many macromolecules, making the processes reliant on simple components to achieve robustness, while still achieving complex results through the multiplicity of factors providing inputs into the final result.

The current synthesis of understanding on this subject stands soundly on the foundations built from the initial cell centric characterizations of development. Much of what was predicted from consideration of the early observations, the maternal contributions to the architecture of the egg, gradients formed by diffusion of proteins responsible for marking out particular tissue territories, and the participation of transcription cascades driven by localized transcription factors, was ultimately shown to be correct. A clear picture of the key starting points and the evolution of the responses that result in a properly developed drosophila embryo is presented in an overview by D. St Johnston and C. Nusslein-Volhard (St Johnston and Nusslein-Volhard, 1992).

The principal difference between an experiment designed with a focus on the system components rather than a focus on the system itself is that, with a component focus, one is usually trying to make two inferences at once, the first being that there are components identifiably connected with some cell state and the second being that some or all of those components are part of the process that dictates that cell state. This differs significantly from starting with the knowledge that there is an observationally supported process that drives the action leading to a particular cell state and then developing an experiment to identify components of that process. As previously described, prior characterization of the dynamics of a process at the cell or organism level provides natural constraints on the design of the experiment to identify the process components.

In the bottom-up case, where the process of interest is undefined, investigators typically use what are called “unbiased” methods, where an extensive number of measurements are made on some quantifiable gene characteristic, presence or absence, mutational status, allelic state, expression level, protein level, and so on, and then the measured behavior of the analyte is analyzed with regard to some characteristic of the cell or organism. The questions asked in a genetics study might revolve around whether or not the distribution of a given genetic marker is enriched in a set of samples taken from patients who have been found to be susceptible to a particular disease relative to a matched set of patients who are not susceptible. A molecular biological question might revolve around whether a set of colon cancer patients who are responsive to a drug have higher or lower expression levels of a particular protein or set of proteins than is seen in a matched set of patients who do not respond to the drug. In these cases, there is no attempt to explicitly follow the dynamics of the cells responsible for the phenotype in question. In the genetics case, one either has the disease or not. The history of the evolution of the diseased state in the cells affected is not a consideration. In the molecular biology case, there is likewise no attempt to get a meaningful picture of the dynamics of the sensitive and resistant cells as they respond to the drug, only to identify a gene that can serve as a predictor of response. In developmental studies, to study a drug that distorts embryogenesis, one would apply the drug to a model system at various times during embryogenesis and then follow the consequences to see when and where the drug causes exceptions from the normal, along with how these initial abnormalities perturb the subsequent developmental sequences. A focus on the consequences to the processes throughout the evolution of the response to the perturbation is a key difference in studies aimed at the system level rather than at the component level.

The question of the relative merits of the different approaches has recently been receiving increasing scrutiny as a result of the 10th anniversary of the sequencing of the human genome. Some of those involved in the project made very strong claims at the completion of the project in terms of the medical benefits we could expect to see within the next decade. Many reviews of the progress over the last 10 years mention the stimulatory effect that the project has had on other types of molecular measurements now being made on proteins, RNA, metabolites and other analytes; however, most also have multiple disclaimers about the overall success of the efforts. Harold Varmus writes,

> As several articles in this series will describe, detailed maps of genetic markers of human variation, mostly single-nucleotide polymorphisms (SNPs), have facilitated many remarkable genomewide efforts to associate known SNPs with disease predisposition. But this approach has usually failed to reveal strongly influential haplotypes, and in general, other implicated susceptibility haplotypes collectively account for only a small fraction of the apparent heritable risk.… Still, genomics and related disciplines are more closely aligned with modern science than with modern medicine. They produce knowledge that is broad in its scope, but only a few selected items of that new information are now widely used as guides to risk, diagnosis, or therapy. Physicians are still a long way from submitting their patients' full genomes for sequencing, not because the price is high, but because the data are difficult to interpret. (Varmus, 2010)

Given these considerations, what differences in understanding of functional connections in a system are actually observed in cases where genes are studied within the component context and within the system context? One example where both types of studies have been employed at different times concerns the gene Myc. The Myc gene was originally identified in 1982 as a homolog of a viral oncogene (Vennstrom et al., 1982). The action of the Myc gene was linked by this and other experiments to stimulation of proliferation and this continued to be the area in which functional research on the gene was centered. Screening assays can frequently be tightly enough designed that the majority of identified genes do indeed exhibit a function of the type the screen was designed to find. Thus, it is typical that a screen-identified gene is heavily characterized in terms of the function facilitating its detection and not with regard to alternative functions. A further description of Myc's function was published a decade later (Evan et al., 1992). This work was driven by the cellular system level observation that, although Myc-transformed cells were indeed proliferating at an extremely high rate, they were also dying at a rate only somewhat slower. This key observation was made by simple microscopic examination of the cell population over time, meeting the fundamental requirement for studying a system, which is to observe the system and the cells, and to follow the dynamics of the system, taking observations over time. Had such observations been made as an integrated part of the screening process, it would have been possible to immediately identify both the proliferative and the death-inducing functions of Myc. Such information would have been valuable, as it indicates that Myc by itself is unlikely to be a high-potency driver of cancer and that attempting to control cancer by altering Myc function would carry high risk. To date, none of the efforts to control cancer based on controlling Myc has proven successful.

Early gene association study results illustrate a variety of problems that can ensue from taking this approach as a means of arriving at methods for controlling a disease. Two discoveries that were heralded as indicators that effective therapy was immanent were Huntington's chorea and cystic fibrosis (CF). The gene identified for Huntington's disease, huntingtin (Htt), was found to be a gene in which a region of the sequence coding for the gene product consisted of repeats of the codon for glutamine (The Huntington's Disease Collaborative Research Group, 1993). This region is susceptible to expansion owing to errors in cellular DNA processing. Consequently, a protein containing long tracts of highly charged glutamine could be produced and inherited. After a long study, it is still unclear how the presence of this protein in the brain mediates the degeneration that leads to chorea and dementia. It is certainly possible that the effect is not particularly specific and involves physical interactions with many genes that can interact with the long polyglutamine tract. A recent study of whether genes that physically associate with Htt can modify the severity of neural degeneration in a drosophila model suggests that quite a large number of genes may contribute to the disease (Kaltenbach et al., 2007). The assumption that a component will have limited functional capabilities, thereby making it easy to determine what capability should be targeted to treat the disease, is implicit in the bottom-up strategy.

CF is a disease caused by malfunction of the exocrine glands, causing the production of thick mucus. When this mucus is produced in the lungs, it can block the bronchi and lead to more frequent and severe respiratory infections. The gene identified for this disease is the CFTR gene (Rommens et al., 1989), which is a sodium pump for the cell. No gene-targeted therapy has been devised for CF. Direct gene function intervention would require that the incapacitation of the pump caused by mutation of the gene be repaired or that a functioning gene replacement be installed. To date, such manipulations cannot be reliably carried out in tissue within a patient. A form of therapy that has greatly lengthened the life span of many CF patients is based upon thinking at a higher level of system abstraction than the cell. The therapy targets action to the organ level. The treatment consists of strapping on a backpack that delivers percussive blows to the back that help dislodge the thick mucus, so that it can be coughed up and removed from the lung.

Given the limitations to producing effective disease interventions without systems-level knowledge, are there instances in which one can succeed, and how common are these likely to be? Two examples from cancer, acute promyelocytic leukemia (APML) and chronic lymphocytic leukemia (CML), are in line with the expectation that it is possible to act without systems-level knowledge. Both tumors arise from partially differentiated white blood cell precursors and both involve a translocation, the joining of two regions of the genome not normally adjacent to each other. Each translocation alters the control of the gene product production and the structure of the gene product. CML arises from a translocation involving chromosomes 9 and 22, fusing the BCR and ABL genes (Nowell and Hungerford, 1960). Cells bearing the fusion express the fused gene and the activity of the ABL kinase portion of the gene has been shown to be responsible for the cancer phenotype. CML can be treated with an inhibitor of the kinase activity. APML arises predominantly from a fusion between chromosomes 15 and 17, fusing the PML and RAR genes (Chen et al., 1991). Cells bearing the fusion genes express a fused gene, which contains large regions of both genes. Cancer induction is thought to rely on aberrant nuclear localization of the fusion gene that allows disruption of the formation of normal protein complexes containing the PML gene that regulate apoptosis (Koken et al., 1994). APML can be treated with all *trans*-retinoic acid, which simultaneously reduces nuclear localization of the fusion gene and induces terminal differentiation of the proliferating cells, thereby terminating the cancer.

Considerable work has been expended in the case of CML to understand the range of cell types that can effectively be transformed by the presence of the BCR:ABL translocation. Characterization of the induction of this tumor in mouse models by the introduction of a promoter driven BCR:ABL fusion gene into mouse blood cells at various degrees of differentiation has revealed that circulating blood cells are extremely hard to turn into cancers, whereas blood cells from bone marrow, which are rich in less-differentiated precursors, are rapidly and quite easily transformed into a cancer that is much like CML (Pear et al., 1988). This pattern of differential susceptibility to the transforming effect of the BCR:ABL fusion product suggests that there is a precursor in the blood cell lineage possessing an existing set of differentiation-defined abilities that enable the alterations in genomic regulation produced by BCR:ABL to drive these cells to the cancerous state of CML. The existence of a "ready-to-use" regulatory network and a single defining change in how the network is operated allows the production of a cancer that will have low levels of genomic alteration and high similarity of expression phenotype across instances of this tumor type. While the mechanism of action of the PML:ATRA fusion is less clear, what is known suggests that APML also requires a "ready-to-use" regulatory network to be effective as a single, system-transforming agent.

The value of finding these genes in terms of medical utility is enormous. Both treatment types produce complete remissions at rates above 80% and lengthy overall survival at rates above 70%. In the case of APML, therapy with ATRA was in use on a wide range of leukemias prior to finding the gene; however, identification facilitated the development of a test for the fusion that would direct the treatment to those patients who would benefit. In the case of CML, a therapy aimed specifically against the target kinase was developed and its use was targeted to patients with the fusion using existing tests. These therapies are the most effective known treatments of cancer, which makes a compelling argument for the possibility of success using a component identification approach.

Why then are we arguing for systems approaches if component approaches can work so well? There are two clear problems with component approaches. First, for them to work, the cancer must be of a kind that arises from a particular cell type that can be transformed to a cancer cell by the action of a single gene. Such tumors are rare, APML and CML being the only known examples of such cancers. The requirement for a particular genomic event to occur in a cell type that is likely to only transiently be in the required state likely means that all such tumors are of much lower incidence than the tumors causing the most mortality. The second reason to argue for systems approaches is that all attempts to produce cancer drugs have been based on component level thinking and nearly all have failed to produce therapies of the caliber of those for APML and CML. Many genes, such as EGFR, ERBB2, and RAS, are frequently altered in gene product abundance or structure (or both) in a wide variety of high incidence cancers. Inhibitors of these gene products' activities have been generated and tested in clinical trials. While these drugs can frequently provoke tumor responses, such as shrinkage or cessation of further growth, they do not provide the high rates of either full remission or lengthy overall survival seen in CML and APML.

A common observation in the majority of tumors that are not readily treatable is the very high extent of change present in these tumors' genomes. Measurements of the type and abundance of mRNA species present in tumors, as well as measurements of regions of altered copy number, typically show wide variance between even those tumors arising from a single tissue or origin (breast, lung, colon, etc.). Recently initiated studies in which the entire genomes of tumors are sequenced and compared with the sequence of normal tissue from the same individual show great variety in the particular genes mutated in any given tumor. It is likely that what we are seeing is that there is a very large number of ways for the control exerted in a normal cell to fail and that, after early failures somewhat impair cellular regulation, the pace of further failures accelerates, thereby providing each tumor with a large repertoire of ways to proliferate and survive. It may be necessary to think of these tumors as having gone from the normal state of having redundant safety controls to now operating in a state where there are redundant risks. Such a picture is concordant with what we now observe when targeted cancer therapies are applied to patients. There is an initial period where the tumor responds, showing some to considerable shrinkage and cessation of growth over some period. This interval is then followed by a period of considerable tumor growth in a variety of places. Following this reversal, treatments with another drug will show the same sequence of events, but in a shorter time period. There is a growing consensus that, if most tumors already have a multiplicity of ways to continue to drive proliferative and survival processes, then treatments with combinations of drugs will have to be devised to address these multiple tumor capabilities. It would seem that the only way forward in this situation is to develop more knowledge of tumors as systems. It is unlikely that the sets of capabilities will be totally random in composition; rather, there will likely be more common sets where the set of later capabilities that arise are influenced by the types of earlier alterations. Ferreting out these complex relationships will require types of testing that specifically follow the sequence of events that occur when a tumor is challenged with a drug. What processes are curtailed and what processes are induced? How does the order and timing of drug application affect the response? What combinations and delivery modes produce the highest synergy? All of these questions can only be asked and answered within the context of following cell operations as the processes of a system.

Given the necessity of systems level cellular analyses to obtain sufficient knowledge for designing more effective disease interventions, how are the relevant experiments to be performed? Several constraints drive the choice of platforms for such work. Perhaps the most difficult requirement is the need to make response measurements at sufficiently frequent intervals to obtain an accurate view of the progression of the various processes being studied. Ideally, the measurements would be taken in a cell-by-cell basis on a population of cells representing the disease state before and during the drug response, so that the set of data points at each sampling would be a distribution based on each cell's behavior at the time of sampling. Many typical forms of measuring gene behavior currently in use do not produce such data sets because the processing of the cells to make the measurements requires that the whole population to be sampled undergo processing in which the cells are lysed and the macromolecules forming the analyte for the assay (DNA, RNA, protein, metabolites, etc.) are mixed and then subjected to a measurement that outputs a number representing the average relative abundance of the various species of that analyte. This kind of analysis provides only an accurate representation of cellular behavior if all cells subjected to the treatment are altering their processes quite synchronously, which is often not the case. Obviously, if one must sacrifice each set of cells subjected to the assay, the experiment would require a set of cells in treatment for each time point being sampled. This adds levels of technical variation that could be avoided if nondestructive testing could be applied that sampled from the same cohort of cells repetitively over time.

These demands on experimental design are challenging; however, advances in a variety of biological reagents and automated imaging devices make it possible to produce at least some types of systems-level measurements. As an example, we will describe an approach that we have devised that allows one to follow changes in gene regulation for a variety of genes and gather data in a nondestructive, gene-by-gene analysis. The most suitable nondestructive assay for transcription activity currently available is the use of fluorescent reporter genes under the transcriptional control of the promoter whose activity is to be assayed (Chalfie et al., 1994). The use of these reporters to allow knowledge of when and where in the developing body a gene is active has been remarkably informative in studying genes' roles in regulating processes in embryonic development (Carroll, 2005). These reporters can be readily delivered to mammalian cells using third-generation lentiviral delivery systems (Wu et al., 2000). This kind of assay is also convenient in that each reporter assay can be carried out in a small culture chamber, such as a single well of a 384-well culture plate. This allows many different promoter reporter derivatives of a single cell line to be simultaneously exposed to a stimulus and tracked. Commercial imaging systems (MDS ImageXpress Micro) are available that can capture fluorescent images of hundreds of cells per well for an entire 384-well plate in less than 20 minutes while maintaining an appropriate tissue culture environment. As transcriptional responses in mammalian cells are typically of sufficient magnitude to be readily detectable at the mRNA level 4–6 hours post treatment, a sampling rate of one per hour is sufficient to give reasonably detailed descriptions of how mRNA levels are rising and falling. Using levels of green fluorescent protein (GFP) to detect such changes adds a uniform delay in the detection dictated by the lengths of time required for translation to produce the protein and proteolysis to turn over the existing GFP protein.

An example of the kind of output that can be obtained from such a system is shown in Fig. 9.1, which shows the response of a colon cancer cell line, HCT116, to the drug lapatinib. The color version of the figure

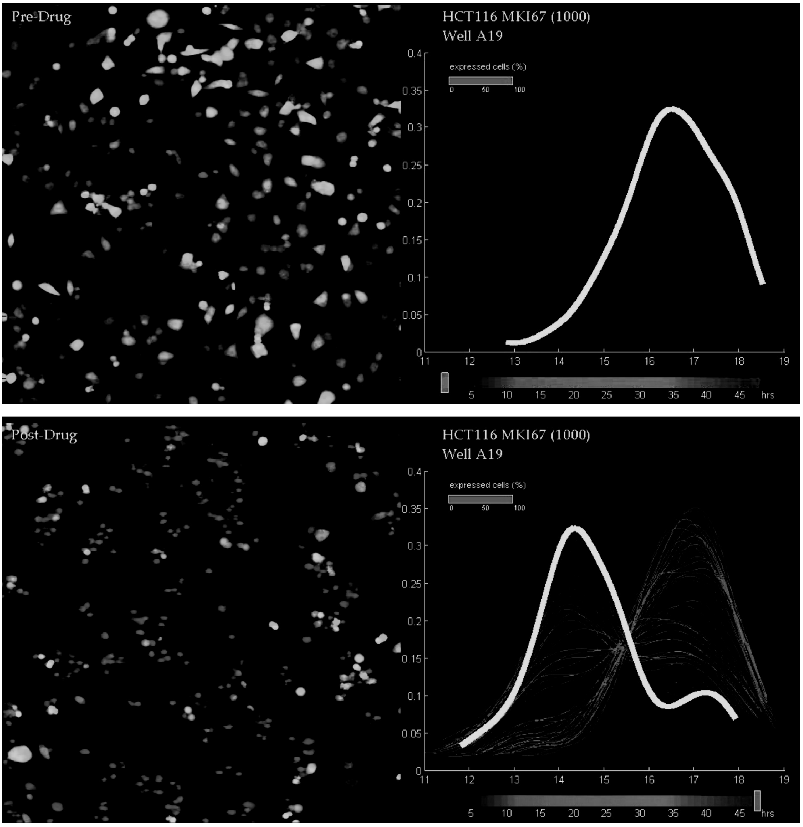

Figure 9.1 Time course of response to lapatinib by HCT116 with a promoter-reporter for gene MKI67. The left halves of the panels show the image at the current time point. The cell nuclei are blue (gray) and the green fluorescent protein is green (bright). The right halves show a histogram (wide yellow [bright] line) of the distribution of cell GFP intensities for the current image and all previous images (thin lines colored in a range from red to blue from the first to last point).

is on the front cover. A time course of response to lapatinib by HCT116 with a promoter-reporter for gene MKI67 is shown. The left halves of the panels show the image of the cells at the current time points, top initial and bottom final. The cell nuclei are blue (gray) and the green fluorescent protein in the cell cytoplasm is green (bright). The right halves of the panels show a histogram (wide yellow [bright] line) of the distribution of cell GFP intensities for the current image and all previous images (thin lines colored in a range from red to blue from the first to last point). The histogram presents the percent of cells at a given fluorescent intensity (*y*-axis) plotted against the observed GFP intensity (*x*-axis). The intensity units are arbitrary camera units of fluorescence per cell, indicated as the exponent of 2. The current time point is indicated by the slider on the "hours" strip. The transcriptional reporter being tracked produces GFP under the direction of the promoter for the human gene MKI67, a gene widely used as an index of the level of ongoing proliferation (Bryant et al., 2006). Data are extracted from these images by using the blue nuclear fluorescence to carry out segmentation to indicate the presence of a nucleus and then finding the adjacent location of the GFP in the cytoplasm of the cell through segmentation of those zones. The amount of GFP intensity per cell is determined and a histogram of the percent of cells showing a given $\log_2$ intensity is plotted. The set of histograms in panel (b) shows the history of the evolution of the transcriptional response of this gene. The cells do not show a continuous response to the treatment, but instead reflect switch-like discrete control. Over time the cell population shifts from a distribution centered at $2^{17.5}$ counts to a distribution centered at $2^{14.3}$ counts, indicating a large decrease in promoter activity. Cells without any GFP have a native fluorescence in the green channel of approximately 2^{14} counts.

While one could envision long-term studies that would allow collection of response data over many cell lines, reporters, and drugs, the cost and time to produce a specific study expands as the product of the numbers of cell lines, drugs, and reporters. Further technological improvements will no doubt increase experimental capacity, but currently the generation of this kind of data for a specific response question is limited to specific studies comparing the response of four to eight cell lines with 6–10 drugs using a set of 10–20 reporters. Operating at this scale, this type of experiment is quite well suited to asking questions that would address some key questions left frequently unanswered during the course of drug development: Does the drug produce the effects it was designed to bring about on targeted pathways and processes? Are there either compensatory alterations in the pathway downstream of the drug target or in independent competing pathways not affected by the drug that render the drug ineffective? Are there recognizable conditions in the disease cell that can indicate likely response to the drug? If the drug is ineffective even when its target is present and operating in the diseased cell, is it frequently due to another specific process so that a combination therapy could prove effective?

Most of these questions deal with the effects of multiple processes operating contemporaneously to produce the drug response. A fair number of specific ways in which processes that produce the general effects of supporting proliferation, survival, or both are known. By using as much knowledge as possible, one can make a hypothesis about what processes may be operating and develop a simple wiring diagram of those systems and their possible interactions. The library of available drugs that have a known effect on one or more parts of those processes is expanding quite rapidly, so one can use these drugs with or without a novel drug whose action is not well characterized to perturb points along the hypothesized wiring diagram and make predictions of the experimental outcome. The study could be carried out and compared with the predictions. Exceptions to the predictions could then be used to revise the hypothesis, make new predictions, and experimentally verify those.

The data generated in these systems studies readily allow one to make some simple interpretations. From the temporal sequence of changes, one gains knowledge of what changes in transcription precede others, making them possible initiators of sets of changes. Data on the magnitude of response can suggest which gene products may have changed their abundance positively or negatively enough to cross a threshold of effectiveness for driving some process. Data on the number of cells shifted from one abundance distribution to a discrete higher, or lower, level can serve as a measure of the likelihood of events happening at different times being related. If a large number of cells are involved in a shift at a late time point, that effect cannot have been mediated (by an effect that only acts within a cell) by an early effect that only shifted a small portion of the cells. While these simple ways of reasoning from this kind of data can be instructive, it is possible to use more powerful analytical methods to find the connections between elements of the processes.

The basic preceding experimental design and the kinds of questions suggested by it fall into the general category of stochastic processes long studied in systems theory and we will now provide a mathematical structure going back to the 1930s that captures many of the essentials of this design. It is not our intention here to go into any specific biological modeling or provide an extensive account of systems theory; however, it is important in a book on biological epistemology to at least outline basic underlying representation theory, especially in the context of discussing the relationship between experimental design and modeling cellular dynamical processes.

We consider n genes, $g_1, g_2, \ldots, g_n$, within a particular kind of cell evolving over time in a particular situation. The expressions form a vector of random variables,

$$\mathrm{X}(t) = (X_1(t), X_2(t), \ldots, X_n(t)), \tag{9.1}$$

where t is time and $X_k(t)$ is the expression level of gene g_k at time t. $X_k(t)$ describes the random fluctuations of gene k at time t and its distribution can be estimated via expression measurements taken at time t. Let us assume that the measurements are taken over discrete time with fixed sampling interval, which without loss of generality we take to be 1. Then the measurement times are given by 1, 2,…, m and a full probabilistic description of $\mathbf{X}(t)$ is given by the probability distribution function

$$P(X_1(1) \leq x_{11}, X_1(2) \leq x_{12}, \ldots, X_1(m) \leq x_{1m}, X_2(1) \leq x_{21}, \ldots, X_n(m) \leq x_{nm}) \tag{9.2}$$

for all values of $x_{11}, x_{12}, \ldots, x_{nm}$.

As should be clear from the preceding probability function, a random vector is a complicated mathematical entity and tractability is gained by bypassing a full probabilistic description and using only second-order information. The cost in doing so is that all higher-order information is omitted from consideration—for instance, there is no distinction between processes possessing identical second-order moments. However, if sufficiently good predictions can be had by using only second-order information, then the increase in mathematical tractability and experimental feasibility are worth the loss, especially when a more general approach is experimentally impractical. Among the mathematical advantages of using only second-order information is that when one wishes to filter a second-order process the optimal filter is linear, the formulation and derivation of this filter being at the center of the Wiener–Kolmogorov theory alluded to in Chapter 5. Should the higher-order moments be of sufficient importance to render worthless the predictions from a second-order model, then there is no alternative to taking on a more difficult modeling and experimental task.

The first moments constitute the mean vector,

$$\boldsymbol{\mu}_{\mathbf{X}}(t) = (\mu_1(t), \mu_2(t), \ldots, \mu_n(t)), \tag{9.3}$$

where $\mu_k(t)$ is the mean of the expression for gene g_k at time t.

The second moments are more interesting. In a static situation one could consider the covariances (correlations) between different genes without considering time; however, in a dynamical situation one must consider the covariance for a gene relative to its changes across time and the covariance for every pair of genes across time. If we let $K_{i,j}(r, s)$ denote the covariance between X_i at time r and X_j at time s, then the covariance matrix takes the form

$$\mathbf{K}_{\mathbf{X}}(r,s) = \begin{pmatrix} K_{1,1}(r,s) & K_{1,2}(r,s) & \cdots & K_{1,n}(r,s) \\ K_{2,1}(r,s) & K_{2,2}(r,s) & \cdots & K_{2,n}(r,s) \\ \vdots & \vdots & \ddots & \vdots \\ K_{n,1}(r,s) & K_{n,2}(r,s) & \cdots & K_{n.n}(r,s) \end{pmatrix} \tag{9.4}$$

for all time pairs (r, s), $r, s = 1, 2, \ldots, m$. The covariance matrix summarizes the second-order relations between genes across time.

In general, the covariances $K_{i,j}(r, s)$ and $K_{i,j}(u, v)$, relating genes g_i and g_j need not have any particular relation to each other, which means that the covariance between X_i at time r and X_j at time s, need not be the same as the covariance between X_i at time u and X_j at time v. From an experimental perspective, estimating the covariance between two genes between a given pair of time points, say, near the beginning of the process, in general says nothing about the covariance between the same two genes between a different pair of time points, say, near the end of the process. Thus, the experimentalist must take the necessary measurements across the entire time process and between all pairs of genes in the model. This instability among covariances across time has a second downside in that the mathematics involved with the model become more difficult, but not impossible.

In some engineering applications, the problem simplifies and a number of useful properties are satisfied. If the covariance $K_{i,j}(r, s)$ can be expressed as

$$K_{i,j}(r,s) = K_{i,j}(\tau), \tag{9.5}$$

where $\tau = s - r$, the time differential between s and r, and, in addition, if the mean vector is constant, then the random vector $\mathbf{X}(t)$ is said to be *wide-sense stationary*. Wide-sense stationary processes are easier to work with and possess advantageous properties. In particular, the covariance matrix simplifies to

$$\mathbf{k}(\tau) = \begin{pmatrix} k_{1,1}(\tau) & k_{1,2}(\tau) & \cdots & k_{1,n}(\tau) \\ k_{2,1}(\tau) & k_{2,2}(\tau) & \cdots & k_{2,n}(\tau) \\ \vdots & \vdots & \ddots & \vdots \\ k_{n,1}(\tau) & k_{n,2}(\tau) & \cdots & k_{n.n}(\tau) \end{pmatrix}, \tag{9.6}$$

where

$$k_{i,j}(\tau) = K_{i,j}(1, 1+\tau), \tag{9.7}$$

for $\tau = 0, 1, \ldots, m - 1$. The experimental advantage is evident: Rather than having to make sufficient measurements to estimate all covariances of the form $K_{i,j}(1, 1)$, $K_{i,j}(1, 2), \ldots, K_{i,j}(1, m)$, $K_{i,j}(2, 1), \ldots, K_{i,j}(m, m)$, one need only estimate $K_{i,j}(1, 1)$, $K_{i,j}(1, 2)$, $K_{i,j}(1, 3), \ldots, K_{i,j}(1, m)$. In a sense, time does not matter, only time differential. Thus, we need only estimate the covariance for two genes across one time interval, anywhere in the process. This is what is meant by stationarity: The behavior of the process does not depend on the actual time.

While wide-sense stationary processes may be useful for modeling cellular processes when cells are in a steady state, they are not appropriate for modeling cells under the influence of drugs; indeed, a cell's response to a drug over time varies markedly. Nonetheless, from the epistemological perspective, the key point is that the representation of Eq. 9.4 is suitable for making predictions, characterizing and answering critical questions concerning cell evolution over time, and designing therapeutic strategies for dealing with aberrant cell behavior.

The foregoing considerations regarding experimentation relating to cellular processes exhibit four general conditions for a sound experiment: (1) a question or set of questions to be addressed; (2) a concept of the physical process in whose framework the questions should be posed; (3) a mathematical structure to model the variables and the relationships that pertain to the questions; and (4) a physical experiment to elicit measurements relating to the questions. Without the starting point of a question, the remaining conditions are moot—and it had better be a carefully considered question. If not, then there is little likelihood of the experiment producing worthwhile results. On this point, Rosenblueth and Wiener are clear: "An experiment is a question. A precise answer is seldom obtained if the question is not precise; indeed, foolish answers—i.e., inconsistent, discrepant or irrelevant experimental results—are usually indicative of a foolish question" (Rosenblueth and Wiener, 1945). Biological knowledge, as it is manifested within the epistemology of the cell, requires precise questions.

References

Aristotle, Physics, *Great Books of the Western World*, Vol. 8, eds. R. M. Hutchins and M. J. Adler, Encyclopedia Britannica, Chicago, 1952.

Ayala, F. J., From Paley to Darwin: Design to Natural Selection, *Back to Darwin*, ed. J. B. Cobb, William B. Eerdmans Publishing Co., Grand Rapids, MI, 2008.

Bacon, F., Novum Organum, *Great Books of the Western World*, Vol. 35, eds. R. M. Hutchins and M. J. Adler, Encyclopedia Britannica, Chicago, 1952 (originally published 1620).

Barrett, W., *The Illusion of Technique*, Anchor Books, New York, 1979.

Barrett, W., *Death of the Soul*, Anchor Books, New York, 1986.

Batt, G., Ropers, D., de Jong, H., Geiselmann, J., Mateescu, R., Page, M., and D. Schneider, "Validation of Qualitative Models of Genetic Regulatory Networks by Model Checking: Analysis of the Nutritional Stress Response in *Escherichia coli*," *Bioinformatics*, 21(suppl. 1), 19–28, 2005.

Bertsekas, D. P., *Dynamic Programming and Optimal Control*, Vols. 1 and 2, Athena Scientific, Nashua, NH, 1995.

Bittner, M., Meltzer, P., Khan, J., Chen, Y., Jiang, Y., Seftor, E., et al., "Molecular Classification of Cutaneous Malignant Melanoma by Gene Expression Profiling," *Nature*, 406(3), 536–540, 2000.

Boulesteix, A. L., "Over-optimism in Bioinformatics Research," *Bioinformatics*, 26, 437–439, 2010.

Boulesteix, A.-L., and C. Strobl, "Optimal Classifier Selection and Negative Bias in Error Rate Estimation: An Empirical Study on High-dimensional Prediction," *BMC Medical Research Methodology*, 9, 85, 2009.

Braga-Neto, U. M., "Fads and Fallacies in the Name of Small-sample Microarray Classification," *IEEE Signal Processing Magazine*, 24(1), 91–99, 2007.

Braga-Neto, U. M., and E. R. Dougherty, "Is Cross-Validation Valid for Small-sample Microarray Classification," *Bioinformatics*, 20(3), 374–380, 2004.

Braga-Neto, U. M., and E. R. Dougherty, "Exact Performance of Error Estimators for Discrete Classifiers," *Pattern Recognition*, 38(11), 1799–1814, 2005.

Epistemology of the Cell: A Perspective on Biological Knowledge, First Edition.
Edward R. Dougherty, Michael L. Bittner.

Braga-Neto, U. M., and E. R. Dougherty, “Exact Correlation Between Actual and Estimated Errors in Discrete Classification,” *Pattern Recognition Letters*, 31, 407–413, 2010.

Bryant, R. J., Banks, P. M., and D. P. O’Malley, “Ki67 Staining Pattern as a Diagnostic Tool in the Evaluation of Lymphoproliferative Disorders,” *Histopathology*, 48(5), 505–515, 2006.

Callebaut, W., and M. D. Laubichler, “Biological Complexity as a Challenge for Biological Theory,” *Biological Theory*, 2, 1–2, 2007.

Carroll, S. B., *Endless Forms Most Beautiful: The New Science of Evo Devo and the Making of the Animal Kingdom*, 1st ed., Norton, New York, 2005.

Chalfie, M., Tu, Y., Euskirchen, G., Ward, W., and D. C. Prasher, “Green Fluorescent Protein as a Marker for Gene Expression,” *Science*, 263(5148), 802–805, 1994.

Chang, L., and M. Marin, “Mammalian MAP Kinase Signaling Cascades,” *Nature*, 410(6824), 37–40, 2001.

Chen, S. J., Zhu, Y. J., Tong, J. H., Dong, S., Huang, W., Chen, Y., et al., “Rearrangements in the Second Intron of the RARA Gene Are Present in a Large Majority of Patients with Acute Promyelocytic Leukemia and Are Used as Molecular Marker for Retinoic Acid-induced Leukemic Cell Differentiation,” *Blood*, 78(10), 2696–2701, 1991.

Datta, A., and E. R. Dougherty, *Introduction to Genomic Signal Processing with Control*, CRC Press, New York, 2007.

de Jong, H., Geiselmann, J., Hernandez, C., and M. Page, “Genetic Network Analyzer: Qualitative Simulation of Genetic Regulatory Networks,” *Bioinformatics*, 19(3), 336–344, 2003.

Devroye, L., Gyorfi, L., and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer-Verlag, New York, 1996.

Dougherty, E. R., “Validation of Inference Procedures for Gene Regulatory Networks,” *Current Genomics*, 8(1), 351–359, 2007.

Dougherty, E. R., “On the Epistemological Crisis in Genomics,” *Current Genomics*, 9(2), 69–79, 2008.

Dougherty, E. R., “Translational Science: Epistemology and the Investigative Process,” *Current Genomics*, 10(2), 102–109, 2009a.

Dougherty, E. R., Epistemology and the Role of Mathematics in Translational Science, *Festschrift in Honor of Jaakko Astola on the Occasion of His 60th Birthday*, eds. I. Tabus, K. Egiazarian, and M. Gabbouj, Tampere International Center for Signal Processing, Tampere, TICSP Series #47, 2009b.

Dougherty, E. R., and M. L. Bittner, “Causality, Randomness, Intelligibility, and the Epistemology of the Cell,” *Current Genomics*, 11(4), 221–237, 2010.

Dougherty, E. R., and U. M. Braga-Neto, “Epistemology of Computational Biology: Mathematical Models and Experimental Prediction as the Basis of Their Validity,” *Biological Systems*, 14, 65–90, 2006.

Dougherty, E. R., and M. Brun, “A Probabilistic Theory of Clustering,” *Pattern Recognition*, 37(5), 917–925, 2004.

Dougherty, E. R., Kim, S., and Y. Chen, “Coefficient of Determination in Nonlinear Signal Processing,” *EURASIP Journal on Applied Signal Processing*, 80(10), 2219–2235, 2000.

Dougherty, E. R., Brun, M., Trent, J. M., and M. L. Bittner, “A Conditioning-Based Model of Contextual Regulation,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6, 310–320, 2009.

Dupuy, A., and R. M. Simon, “Critical Review of Published Microarray Studies for Cancer Outcome and Guidelines on Statistical Analysis and Reporting,” *Journal of the National Cancer Institute*, 99, 147–157, 2007.

Einstein, A., *The Meaning of Relativity*, Princeton University Press, Princeton, NJ, 1922.

Einstein, A., *Herbert Spencer Lecture*, Oxford University Press, New York, 1933.

Einstein, A., In a letter to Robert A. Thornton, December, 1944a.

Einstein, A., Remarks on Bertrand Russell’s Theory of Knowledge, *The Philosophy of Bertrand Russell*, The Library of Living Philosophers, Vol. 5, ed. P. A. Schilpp, Tudor Publishers, Greensboro, NC, 1944b.

Einstein, A., Einstein’s Reply to Criticisms, *Albert Einstein: Philosopher-Scientist*, The Library of Living Philosophers Series, ed. P. A. Schilpp, Cambridge University Press, Cambridge, 1949.

Einstein, A., An Undated Letter to Maurice Solovine, *Letters to Solovine*, ed. M. Solovine, Carol Publishing Group, New York, 1993.

Einstein, A., and L. Infeld, *The Evolution of Physics*, Simon & Schuster, New York, 1967 (originally published 1938).

Evan, G. I., Wyllie, A. H., Gilbert, C. S., Littlewood, T. D., Land, H., Brooks, M., et al., “Induction of Apoptosis in Fibroblasts by c-myc Protein,” *Cell*, 69(1), 119–128, 1992.

Faryabi, B., Chamberland, J.-F., Vahedi, G., Datta, A., and E. R. Dougherty, “Optimal Intervention in Asynchronous Genetic Regulatory Networks,” *IEEE Journal of Selected Topics in Signal Processing*, 2, 412–423, 2008.

Faure, A., Naldi, A., Chaaouiya, C., and D. Thieffry, “Dynamical Analysis of a Generic Boolean Model for the Control of the Mammalian Cell Cycle,” *Bioinformatics*, 22(14), 124–131, 2006.

Feynman, R., *QED: The Strange Theory of Light and Matter*, Princeton University Press, Princeton, NJ, 1985.

Feynman, R., *The Meaning of It All: Thoughts of a Citizen Scientist*, Addison-Wesley, Reading, MA, 1998.

Fisher, R. A., *Statistical Methods for Research Workers*, Oliver and Boyd, Edinburgh, 1925.

Frank, P., *Modern Science and Its Philosophy*, Collier Books, New York, 1961.

Galen, On the Natural Faculties, *Great Books of the Western World*, Vol. 10, eds. R. M. Hutchins and M. J. Adler, Encyclopedia Britannica, Chicago, 1952.

Galileo, *Dialogues Concerning Two New Sciences*, Dover, New York, 1954 (originally published 1638).

Galileo, *Dialogue Concerning the Two Chief World Systems*, Modern Library, New York, 2001 (originally published 1632).

Glass, L., and S. A. Kauffman, "The Logical Analysis of Continuous Non-linear Biochemical Control Networks," *Journal of Theoretical Biology*, 39, 103–129, 1973.

Glick, N., "Additive Estimators for Probabilities of Correct Classification," *Pattern Recognition*, 10, 211–222, 1978.

Gomez-Lazaro, M., Fernandez-Gomez, F. J., and J. Jordan, "p53: Twenty-Five Years Understanding the Mechanism of Genome Protection," *Journal of Physiology and Biochemistry*, 60(4), 287–307, 2004.

Goutsias, J., and S. Kim, "A Nonlinear Discrete Dynamical Model for Transcriptional Regulation: Construction and Properties'," *Biophysical Journal*, 86, 1922–1945, 2004.

Hanczar, B., Hua, J., Sima, C., Weinstein, J., Bittner, M. L., and E. R. Dougherty, "Small-sample Precision of ROC-related Estimates," *Bioinformatics*, 26(6), 822–830, 2010.

Hsing, T., Attoor, S., and E. R. Dougherty, "Relation Between Permutation-test P Values and Classifier Error Estimates," *Machine Learning*, 52(1–2), 11–30, 2003.

Hua, J., Xiong, Z., Lowey, J., Suh, E., and E. R. Dougherty, "Optimal Number of Features as a Function of Sample Size for Various Classification Rules," *Bioinformatics*, 21(8), 1509–1515, 2005.

Hughes, G. F., "On the Mean Accuracy of Statistical Pattern Recognizers," *IEEE Transactions on Information Theory*, 14, 55–63, 1968.

Hume, D., *A Treatise of Human Nature*, Oxford University Press, Oxford, 1951 (originally published 1738).

Hume, D., An Enquiry Concerning Human Understanding, *Great Books of the Western World*, Vol. 35, eds. R. M. Hutchins and M. J. Adler, Encyclopedia Britannica, Chicago, 1952 (originally published 1751).

Husmeier, D., "Sensitivity and Specificity of Inferring Genetic Regulatory Interactions from Microarray Experiments with Dynamic Bayesian Networks," *Bioinformatics*, 19(17), 2271–2282, 2003.

Ivanov, I., and E. R. Dougherty, "Modeling Genetic Regulatory Networks: Continuous or Discrete?" *Biological Systems*, 14, 219–229, 2006.

James, W., *Pragmatism and Other Essays*, Washington Square Press, New York, 1963.

Jeans, J. H., *The Mysterious Universe*, Cambridge University Press, Cambridge, 1930.

Jelizarow, M., Guillemot, V., Tenenhaus, A., Strimmer, K., and A.-L. Boulesteix, "Over Optimism in Bioinformatics: An Illustration," *Bioinformatics*, 26(16), 1990–1998, 2010.

Kaltenbach, L. S., Romero, E., Becklin, R. R., Chettier, R., Bell, R., Phansaldar, A., et al., "Huntingtin Interacting Proteins Are Genetic Modifiers of Neurodegeneration," *PLoS Genetics*, 3(5), e82, 2007.

Kant, I., Critique of Pure Reason, *Great Books of the Western World*, 2nd ed., Vol. 42, eds. R. M. Hutchins and M. J. Adler, Encyclopedia Britannica, Chicago, 1952 (originally published 1787).

Kant, I., *Critique of Pure Reason*, Henry G. Bohn, London, 1855 (originally published 1787).

Kant, I., Prolegomena to Any Future Metaphysics, *Kant's Prolegomena and Metaphysical Foundations of Natural Science*, George Bell and Sons, London, 1891 (originally published 1783).

Kauffman, S. A., *The Origins of Order*, Oxford University Press, Oxford, 1993.

Kauffman, S. A., Does Science Make Belief in God Obsolete? John Templeton Foundation, http://www.templeton.org/belief/, 2007.

Kauffman, S. A., Breaking the Galilean Spell, *Edge*, http://www.edge.org/, 2008.

Keller, E. F., "Revisiting Scale-free Networks," *BioEssays*, 27, 1060–1068, 2005.

Kierkegaard, S., *Concluding Unscientific Postscript*, Princeton University Press, Princeton, NJ, 1941.

Kim, S., Dougherty, E. R., Shmulevich, I., Hess, K. R., Hamilton, S. R., Trent, J. M., et al., "Identification of Combination Gene Sets for Glioma Classification," *Molecular Cancer Therapeutics*, 1(13), 1229–1236, 2002.

Kline, M., *Mathematics and the Search for Knowledge*, Oxford University Press, Oxford, 1985.

Koken, M. H., Puvion-Dutilleul, F., Guillemin, M. C., Viron, A., Linares-Cruz, G., Stuurman, N., et al., "The t(15;17) Translocation Alters a Nuclear Body in a Retinoic Acid-reversible Fashion," *The EMBO Journal*, 13(5), 1073–1083, 1994.

Kolmogorov, A., "On the Analytical Methods of Probability Theory," *Mathematishe Anallen*, 104, 415–458, 1931.

Kolmogorov, A., Stationary Sequences in Hilbert Space. *Bulletin of the Mathematics University of Moscow*, 2, 1941.

Kumar, P. R., and W. Lin, "Optimal Adaptive Controller for Unknown Markov Chains," *IEEE Transactions on Automatic Control*, 27, 765–774, 1982.

Laplace, P.-S., *A Philosophical Essay on Probabilities*, Dover, New York, 1953 (originally published 1814).

Liang, S., Fuhrman, S., and R. Somogyi, "REVEAL A General Reverse Engineering Algorithm for Inference of Genetic Network Architectures," *Pacific Symposium on Biocomputing*, 3, 18–29, 1998.

Marbach, D., Prill, R. J., Schaffter, T., Mattiussi, C., Flreano, D., and G. Stolovitzky, "Revealing Strengths and Weaknesses of Methods for Gene Network Inference," *Proceedings of the National Academy of Sciences of the United States of America*, 107(14), 6286–6291, 2010.

Martins, D., Braga-Neto, U., Hashimoto, R., Bittner, M. L., and E. R. Dougherty, "Intrinsically Multivariate Predictive Genes," *IEEE Journal of Selected Topics in Signal Processing*, 2(3), 424–439, 2008.

Matheron, G., *Random Sets and Integral Geometry*, John Wiley, New York, 1975.

Maxwell, J. C., "On Faraday's Lines of Force," *Transactions of the Cambridge Philosophical Society*, 10, 155–229, 1855.

Mehta, T., Murat, T., and D. B. Allison, "Towards Sound Epistemological Foundations of Statistical Methods for High-dimensional Biology," *Nature Genetics*, 36, 943–947, 2004.

Mill, J. S., *A System of Logic, Ratiocinative and Inductive*, University Press of the Pacific, Honolulu, HI, 2002 (originally published 1843).

Moffat, J. W., *Reinventing Gravity: A Physicist Goes Beyond Einstein*, Smithsonian Books/Collins, New York, 2008.

Newton, I., Mathematical Principles of Natural Philosophy, *Great Books of the Western World*, Vol. 34, eds. R. M. Hutchins and M. J. Adler, Encyclopedia Britannica, Chicago, 1952 (originally published 1687).

Nilim, A., and L. El Ghaoui, "Robust Control of Markov Decision Processes with Uncertain Transition Matrices," *Operations Research*, 53, 780–798, 2005.

Noguchi, T., Metz, R., Chen, L., Mattéi, M. G., Carrasco, D., and R. Bravo, "Structure, Mapping, and Expression of ERP, a Growth Factor-inducible Gene Encoding a Nontransmembrane Protein Tyrosine Phosphatase, and Effect of ERP on Cell Growth," *Molecular and Cell Biology*, 13(9), 5195–5205, 1993.

Nowell, P. C., and D. A. Hungerford, "A Minute Chromosome in Human Chronic Granulocytic Leukemia," *Science*, 132(3438), 1497, 1960.

Pal, R., Datta, A., and E. R. Dougherty, "Optimal Infinite Horizon Control for Probabilistic Boolean Networks," *IEEE Transactions on Signal Processing*, 54, 2375–2387, 2006.

Pal, R., Datta, A., and E. R. Dougherty, "Robust Intervention in Probabilistic Boolean Networks," *IEEE Transactions on Signal Processing*, 56, 1280–1294, 2008.

Pear, W. S., Miller, J. P., Xu, L., Pui, J. C., Soffer, B., Quackenbush, R. C., et al., "Efficient and Rapid Induction of a Chronic Myelogenous Leukemia-like Myeloproliferative Disease in Mice Receiving P210 bcr/abl-transduced Bone Marrow," *Blood*, 92(10), 3780–3792, 1988.

Plato, Republic, *Great Books of the Western World*, Vol. 8, eds. R. M. Hutchins and M. J. Adler, Encyclopedia Britannica, Chicago, 1952.

Popper, K., *The Logic of Scientific Discovery*, Hutchinson, London, 1959.

Popper, K., *Conjectures and Refutations*, Routledge, London, 1963.

Pugachev, V. S., *Theory of Random Functions and Its Application to Control Problems*, Pergamon Press, Oxford, 1965.

Qian, X., and E. R. Dougherty, "Effect of Function Perturbation on the Steady-state Distribution of Genetic Regulatory Networks: Optimal Structural Intervention," *IEEE Transactions on Signal Processing*, 56, 4966–4975, 2008.

Qian, X., and E. R. Dougherty, "On the Long-run Sensitivity of Probabilistic Boolean Networks," *Journal of Theoretical Biology*, 257, 560–577, 2009.

Qian, X., Ivanov, I., Ghaffari, N., and E. R. Dougherty, "Intervention in Gene Regulatory Networks via Greedy Control Policies Based on Long-run Behavior," *BMC Systems Biology*, 3, 61, 2009.

Reichenbach, H., *The Rise of Scientific Philosophy*, University of California Press, Berkeley, 1971.

Rommens, J. M., Iannuzzi, M. C., Kerem, B., Drumm, M. L., Melmer, G., Dean, M., et al., “Identification of the Cystic Fibrosis Gene: Chromosome Walking and Jumping,” *Science*, 245(4922), 1059–1065, 1989.

Rosenblueth, A., and N. Wiener, “The Role of Models in Science,” *Philosophy of Science*, 12, 316–321, 1945.

Russell, B., “On the Notion of Cause,” *Proceedings of the Aristotelian Society*, 13, 1–26, 1913.

Schrödinger, E., *Science Theory and Man*, Dover, New York, 1957.

Shmulevich, I., and E. R. Dougherty, *Probabilistic Boolean Networks: The Modeling and Control of Gene Regulatory Networks*, SIAM Press, New York, 2010.

Simon, J. L., *Resampling: The New Statistics*, 2nd ed., Resampling Stats, 1997.

Slack, J. M. W., *From Egg to Embryo: Determinative Events in Early Development*, Cambridge University Press, Cambridge, 1983.

Smalley, K. S., “A Pivotal Role for ERK in the Oncogenic Behaviour of Malignant Melanoma?” *International Journal of Cancer*, 104(5), 527–532, 2003.

St Johnston, D., and C. Nusslein-Volhard, “The Origin of Pattern and Polarity in the Drosophila Embryo,” *Cell*, 68(2), 201–219, 1992.

Tabin, C. J., and R. A. Weinberg, “Analysis of Viral and Somatic Activations of the cHa-ras Gene,” *Virology*, 53(1), 260–265, 1985.

The Huntington’s Disease Collaborative Research Group, “A Novel Gene Containing a Trinucleotide Repeat That Is Expanded and Unstable on Huntington’s Disease Chromosomes,” *Cell*, 72(6), 971–983, 1993.

Turing, A. M., “The Chemical Basis of Morphogenesis,” *Philosophical Transactions of the Royal Society of London*, 237(641), 37–72, 1952.

Vahedi, G., Faryabi, B., Chamberland, J.-F., Datta, A., and E. R. Dougherty, “Intervention in Gene Regulatory Networks via a Stationary Mean-first-passage-time Control Policy,” *IEEE Transactions on Biomedical Engineering*, 15, 2319–2331, 2008.

van de Vijver, M. J., He, Y. D., van’t Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., et al., “Gene Expression Signature as a Predictor of Survival in Breast Cancer,” *The New England Journal of Medicine*, 347, 1999–2009, 2002.

van’t Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., et al., “Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer,” *Nature*, 415, 530–536, 2002.

Vapnik, V. N., and A. Chervonenkis, *Theory of Pattern Recognition*, Nauka, Moscow, 1974.

Varmus, H., “Ten Years On—The Human Genome and Medicine,” *The New England Journal of Medicine*, 362(21), 2028–2029, 2010.

Vennstrom, B., Sheiness, D., Zabielski, J., and J. M. Bishop, “Isolation and Characterization of c-myc, a Cellular Homolog of the Oncogene (v-myc) of Avian Myelocytomatosis Virus Strain 29,” *Virology*, 42(3), 773–779, 1982.

Waddington, C. H., *How Animals Develop*, Allen & Unwin, London, 1935.

Waddington, C. H., “Canalization of Development and the Inheritance of Acquired Characters,” *Nature*, 150, 563–565, 1942.

Waddington, C. H., *Principles of Development and Differentiation*, Macmillian, New York, 1966.

Webster’s New Twentieth Century Dictionary, Unabridged, 2nd ed., Collins, New York, 1978.

Weeraratna, A. T., Jiang, Y., Hostetter, G., Rosenblatt, K., Duray, P., Bittner, M. L., and J. M. Trent, “Wnt5a Signaling Directly Affects Cell Motility and Invasion of Metastatic Melanoma,” *Cancer Cell*, 1, 279–288, 2002.

Werhli, A. V., Grzegorczyk, M., and D. Husmeier, “Comparative Evaluation of Reverse Engineering Gene Regulatory Networks with Relevance Networks, Graphical Gaussian Models and Bayesian Networks,” *Bioinformatics*, 22(20), 2523–2531, 2006.

Wiener, N., *Cybernetics: or Control and Communication in the Animal and Machine*, MIT Press, Cambridge, MA, 1948.

Wiener, N., *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*, MIT Press, Cambridge, MA, 1949.

Wiener, N., and A. Rosenblueth, “The Mathematical Formation of the Problem of Conduction of Impulses in a Network of Connected Excitable Elements, Specifically in Cardiac Muscle,” *Archives of the Institute of Cardiology, Mexico*, 16, 204–265, 1946.

Windelband, W., *A History of Philosophy*, Harper and Brothers, New York, 1958.

Wu, X., Wakefield, J. K., Liu, H., Xiao, H., Kralovics, R., Prchal, J. T., and J. C. Kappes, “Development of a Novel Trans-lentiviral Vector That Affords Predictable Safety,” *Molecular Therapeutics*, 2(1), 47–55, 2000.

Xu, Q., Hua, J., Braga-Neto, U. M., Xiong, Z., Suh, E., and E. R. Dougherty, “Confidence Intervals for the True Classification Error Conditioned on the Estimated Error,” *Technology in Cancer Research and Treatment*, 5(6), 579–590, 2006.

Yousefi, M. R., Hua, J., Sima, C., and E. R. Dougherty, “Reporting Bias When Using Real Data Sets to Analyze Classification Performance,” *Bioinformatics*, 26(1), 68–76, 2010.

Zhao, W., Serpedin, E., and E. R. Dougherty, “Inferring Gene Regulatory Networks from Time Series Data Using the Minimum Description Length Principle,” *Bioinformatics*, 22(17), 2129–2135, 2006.

Zhao, C., Bittner, M. L., Chapkin, R., and E. R. Dougherty, “Characterization of the Effectiveness of Reporting Lists of Small Feature Sets Relative to the Accuracy of the Prior Biological Knowledge,” *Cancer Informatics*, 9, 49–60, 2010.

Zollanvari, A., Braga-Neto, U. M., and E. R. Dougherty, “On the Joint Sampling Distribution Between the Actual Classification Error and the Resubstitution and Leave-one-out Error Estimators for Linear Classifiers,” *IEEE Transactions on Information Theory*, 56(2), 784–804, 2010.

Index

Epistemology of the Cell: A Perspective on Biological Knowledge, First Edition.
Edward R. Dougherty, Michael L. Bittner.

BOOKS IN THE IEEE PRESS SERIES ON BIOMEDICAL ENGINEERING

The focus of our series is to introduce current and emerging technologies to biomedical and electrical engineering practitioners, researchers, and students. This series seeks to foster interdisciplinary biomedical engineering education to satisfy the needs of the industrial and academic areas. This requires an innovative approach that overcomes the difficulties associated with the traditional textbooks and edited collections.

Series Editor
Metin Akay, University of Houston, Texas

1. *Time Frequency and Wavelets in Biomedical Signal Processing*
   Metin Akay

2. *Neural Networks and Artificial Intelligence for Biomedical Engineering*
   Donna L. Hudson and Maurice E. Cohen

3. *Physiological Control Systems: Analysis, Simulation, and Estimation*
   Michael C. K. Khoo

4. *Principles of Magnetic Resonance Imaging: A Signal Processing Perspective*
   Zhi-Pei Liang and Paul C. Lauterbur

5. *Nonlinear Biomedical Signal Processing, Volume 1, Fuzzy Logic, Neural Networks, and New Algorithms*
   Metin Akay

6. *Fuzzy Control and Modeling: Analytical Foundations and Applications*
   Hao Ying

7. *Nonlinear Biomedical Signal Processing, Volume 2, Dynamic Analysis and Modeling*
   Metin Akay

8. *Biomedical Signal Analysis: A Case-Study Approach*
   Rangaraj M. Rangayyan

9. *System Theory and Practical Applications of Biomedical Signals*
Gail D. Baura

10. *Introduction to Biomedical Imaging*
Andrew G. Webb

11. *Medical Image Analysis*
Atam P. Dhawan

12. *Identification of Nonlinear Physiological Systems*
David T. Westwick and Robert E. Kearney

13. *Electromyography: Physiology, Engineering, and Non-Invasive Applications*
Roberto Merletti and Philip Parker

14. *Nonlinear Dynamic Modeling of Physiological Systems*
Vasilis Z. Marmarelis

15. *Genomics and Proteomics Engineering in Medicine and Biology*
Metin Akay

16. *Handbook of Neural Engineering*
Edited by Metin Akay

17. *Medical Image Analysis, Second Edition*
Atam P. Dhawan

18. *Advanced Methods of Biomedical Signal Processing*
Edited by Sergio Cerutti and Carlo Marchesi

19. *Models and Algorithms for Biomolecules and Molecular Networks*
Jie Liang and Bhaskar DasGupta

20. *Epistemology of the Cell: A Systems Perspective on Biological Knowledge*
Edward R. Dougherty and Michael L. Bittner

Forthcoming Titles
Biomedical Signal Analysis: A Case-Study Approach, Second Edition
Rangaraj M. Rangayyan