Journal of Biological Systems, Vol. 14, No. 1 (2006) 65–90

World Scientific
www.worldscientific.com

EPISTEMOLOGY OF COMPUTATIONAL BIOLOGY: MATHEMATICAL MODELS AND EXPERIMENTAL PREDICTION AS THE BASIS OF THEIR VALIDITY*

EDWARD R. DOUGHERTY

Department of Electrical Engineering, Texas A&M University
College Station, TX 77483, USA

Computational Biology Division, Translational Genomics Research Institute
Phoenix, AZ 85004, USA
edward@ee.tamu.edu

ULISSES BRAGA-NETO†

Virology and Experimental Therapy Laboratory
Aggeu Magalhães Research Center — CPqAM/FIOCRUZ
Recife, PE 50.670-420, Brazil
ulisses_braga@cpqam.fiocruz.br

Received 23 March 2005
Revised 4 July 2005

Knowing the roles of mathematics and computation in experimental science is important for computational biology because these roles determine to a great extent how research in this field should be pursued and how it should relate to biology in general. The present paper examines the epistemology of computational biology from the perspective of modern science, the underlying principle of which is that a scientific theory must have two parts: (1) a structural model, which is a mathematical construct that aims to represent a selected portion of physical reality and (2) a well-defined procedure for relating consequences of the model to quantifiable observations. We also explore the contingency and creative nature of a scientific theory. Among the questions considered are: Can computational biology form the theoretical core of biology? What is the basis, if any, for choosing one particular model over another? And what is the role of computation in science, and in biology in particular? We examine how this broad epistemological framework applies to important statistical methodologies pertaining to computational biology, such as expression-based phenotype classification, gene regulatory networks, and clustering. We consider classification in detail, as the epistemological issues raised by classification are related to all computational-biology topics in which statistical prediction plays a key role. We pay particular attention to classifier-model validity and its relation to estimation rules.

Keywords: Philosophy of Science; Epistemology; Computational Biology; Classification; Clustering; Regulatory Networks; Error Estimation.

*The opinions and views expressed in this paper are those of the authors and do not necessarily reflect the opinions of the Chief Editor or of other editors of the *Journal of Biological Systems*.
†Corresponding author.

1. Introduction

The advent of high-throughput technologies for genomics and proteomics has facilitated multivariate modeling in areas such as expression-based phenotype classification,[1–5] gene prediction,[6,7] gene regulatory networks,[8–12] and data clustering.[13–15] Given the large numbers of variables resulting from expression measurements, the relative paucity of data in comparison to the variable sets, and the many proposed inference algorithms for classifiers and networks, it is incumbent that the epistemological issue be raised in regard to computational biology: What is the meaning of computational biology in biological science? Computational biology utilizes statistics, computation, and mathematical structures such as directed graphs and differential equations, but these are more than an *ad hoc* collection of algorithms, methods, and theorems. In the context of computational biology, these purely mathematical entities play a scientific role and their meaning must be understood in scientific terms. Its epistemology determines how computational biology should be pursued and how it relates to biology.

We will examine the epistemology of computational biology on two levels: first, from the general perspective of modern science, the conception of which has evolved mainly in response to the monumental achievements in physics in the first half of the 20th century, and second, relative to the particular issue of classification, which holds great potential for the application of statistical decision making to molecular-based disease diagnosis. While a general scientific epistemology is important, it is also important to relate the general principles to specific situations. Not only is classification important in its own right; the epistemological issues raised by classification, such as the validity of a classifier model, are related to all topics in which statistical prediction plays a key role, including clustering and inference of gene regulatory networks, whose specific epistemological issues we also consider, though in less detail. We are particularly interested in providing a formal epistemological characterization of the validity of a classification model.

The need for careful epistemology in computational biology has been highlighted by the recent advances in high-throughput technologies and the consequent effort to deal with the extremely high-dimensional data sets produced by the technology, concurrent with very small samples. With the onset of microarray-based classification, one of us (E. Dougherty) considered the limiting impact of large numbers of genes and small numbers of microarrays on classifier design, feature selection, and error estimation, and wrote, "Owing to the limited number of microarrays typically used in these studies, serious issues arise with respect to the design, performance, and analysis of classifiers based on microarray data."[16] Mehta *et al.* have recently critiqued the application of statistical methods to high-dimensional biology, in particular, with respect to microarray-based analysis, and have cautioned, "Many papers aimed at the high-dimensional biology community describe the development or application of statistical techniques. The validity of many of these is questionable, and a shared understanding about the epistemological foundations of the statistical

methods themselves seems to be lacking."[17] They then go on to provide a set of basic recommendations regarding sound statistical methodology in the context of microarray data analysis.

The present article aims to go deeper into the epistemological foundation of computational biology and its role in constituting the theoretical core of biology, in analogy to the way that mathematical theories provided that core for physics in the 20th century. Except for supporting the dual roles of mathematical modeling and predictive experimentation as the epistemological foundation of computational biology, we do not advocate any specific philosophy of science. We provide the views of numerous scientists, among whom there are significant disagreements, but whose views converge with regard to the general necessity of the model-experiment duality. We expect that different readers will interpret what we say in different lights, depending on their own predilections, in particular with regard to the meaning of the model-experiment duality and how the duality is to be manifested in practice. These are deep, perplexing issues and differing views are inevitable. Like the introduction of any breakthrough observation technology, high-throughput, multimodal technologies for biology will change the conception of the subject. These changes will be accompanied by a debate as to their meaning and how they impact understanding. Our particular interest is translation of knowledge into medical application and in this regard we quote from a recent article resulting from the *Multimodal Bio-Medical Systems Workshop* held in 2004 at the National Library of Medicine:

> The life sciences rely on predictive models. Serious thought needs to be devoted to their epistemology. Models should include existing biological information, suitable mathematical formulas, and necessary data. Predictive mathematical models are necessary to move biology in the direction of a predictive science. They are also necessary to the application of engineering methods to translate biological knowledge into therapies with a mathematical and computational basis.[18]

The practical consequences of an inattention to epistemology are real and already showing themselves in questions regarding the efficacy of microarray-based classification.[19,20] Of great concern are conclusions based on error estimates arising from small samples, an issue fundamental to the epistemology of classification.[21,22] Our hope is that this paper will provoke practicing scientists to take the matter of epistemology seriously and to undertake earnest deliberations.

2. The Nature of a Scientific Theory

Galileo Galilei is usually considered the father of modern science. He proposed a conception of knowledge in which there is an economy of constructive terminology and a dependence of scientific propositions on judicious observation. To Galileo, science owes the concept of a mathematical model. Our understanding of models

and their relationship to nature is different today than in his time; nonetheless, the use of mathematical equations to describe abstract relationships between selected quantifiable variables is mainly due to Galileo.

A model is a logical, and therefore mental, construct in which the variables and relations between the variables represent a selected portion of physical reality. It is a skeleton that reflects the salient features of a physical situation of interest to the scientist. It is a conceptualization of a part of nature, a logical apparatus that bears a connection to nature through the scientist's ability to utilize it as a determining factor in the construction of experiments and the prediction of outcomes resulting from those experiments. The test of a model is its accuracy in the prediction of sensory events, such as the movement of a needle on a meter. At issue is the concurrence of observed and predicted measurements. A model gains its legitimacy from data. New data may reveal the inadequacy of a model. Thus, a model is a contingent hypothesis that perpetually stands open to rejection should its predictions fail. Writes Karl 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."[23] The epistemology and method of science are united in the model concept. Science ceases to be if the mathematical model is separated from the experimental method.

Among models there is a kind of survival of the fittest. Indeed, the terminology of struggle has been used in the philosophy of science. 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."[24] 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."[25]

Classically, the scientist worked with models whose fundamental terms referred to ideas whose origins lay in pre-scientific perceptual experience. Terms such as "particle," "wave," and "force" were of this genre. Moreover, the frames of experience, such as Euclidean three-dimensional space and linear time, and the underlying hypotheses concerning regularity, such as causality and continuity, had their origins in the commonplace perception of everyday phenomena. However, with the advent of quantum mechanics and general relativity, understanding of the mathematical apparatus changed. What became apparent was that the apparatus itself was of prime importance with regard to organization and prediction, and that any intuitive appreciation of this apparatus was secondary. In the words of James Jeans,

> 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.[26]

Non-mathematical reasoning is useful for the scientist in exploratory thinking, but it does not constitute the theoretical object of science, which is the mathematical model. One might use a metaphor of observers holding lights on approaching trains to make an intuitive point concerning relativity, but the theory lies properly within the equations. Any attempt to force a non-mathematical understanding creates the risk of having a diminished (or erroneous) scientific description. This results in the substitution of 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 commonplace understanding because commonplace understanding is inadequate for quantitative predictive models. Regarding the essential worth of a physical theory, Richard Feynman writes,

> 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.[27]

The absurdity of which Feynman speaks is not the absurdity of Nature in and of herself; rather, it is an absurdity relative to the relation between human rationality and Nature. Indeed, why should one expect natural phenomena to be describable in terms concordant with common-sense understanding? Human intuition and vocabulary have not developed with reference to any experience at the subatomic level or the speed of light, nor have they developed with reference to the kinds of massive nonlinear dynamical systems encountered in biology. The very recent ability to observe and measure complex, out of the ordinary phenomena necessitates scientific characterizations that go beyond what seems "reasonable" to ordinary understanding. As a product of human thinking, a mathematical model is not psychologically independent of human understanding; nevertheless, its validity rests solely with its ability to predict experimental outcomes, not its agreement with common sense.

One might object to the fundamental role of prediction by asking whether an investigation not be called "science" if one simply categorizes observations based on measurements. Certainly such categories represent a form of knowledge and their assembly, which can require great effort and ingenuity, is part of the scientific enterprise, but they do not constitute scientific knowledge unless they are utilized within some predictive framework. Scientific knowledge requires more. To use the language of pragmatism, it concerns knowledge with a *cash value*. Feynman writes, "Knowledge is of no real value if all you can tell me is what happened yesterday."[28] Scientific knowledge is worldly knowledge in that it points into the future by making

predictions about events that have yet to take place. It 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 states, "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."[29]

To place the entire matter into a practical clinical setting, consider a physician who applies the St. Gallen criteria to a patient with lymph-node-positive breast cancer to predict metastasis-free survival and based upon the prediction decides whether the patient would benefit from adjuvant systemic treatment. The truth of the physician's idea [the criteria constituting the mathematical model] depends upon events [survival or death] — or upon the statistics of the patient sample to which the idea is applied. Step ahead a few years. A gene-expression signature for the patient is put into a classifier model [idea] to predict metastasis-free survival and thereby predict the benefit of adjuvant systemic treatment.[30] The truth of the classifier model depends upon events. Now step into the not-too-distant future. Gene and protein expressions relating directly to the patient's tumor are measured, these measurements are input into the shell of a gene-protein network model to individualize the network for the particular patient, and a computer applies the theory of automatic control to the network to derive an optimal molecular-based treatment regimen.[31] This step ahead requires an explicit mathematical model possessing a level of complexity far beyond what a human can handle; nevertheless, like today's decision model, its truth depends upon events relating to patient outcomes. Moreover, with network models in hand that can predict disease dynamics, technology can make the application of scientific knowledge ever more productive: a device is embedded in the patient to monitor the relevant gene and protein expressions, this information is sent by wireless to a supercomputer that adjusts the network model to changing conditions in real time and applies control theory to obtain a therapeutic strategy, the details of the strategy are sent back to a nano-device embedded in the patient, and the device dispenses the required treatment with the precise composition and timing called for by the control algorithm. The greater scientific knowledge in the foreseeable future has a greater cash value than the knowledge of today, but in either case there is no cash value without prediction.

With all this emphasis on prediction, one might ask if there is more to prediction than functional relations between variables that agree with experiment. Is there causality? This is a subject with a long history, the philosophical debate over causality in the modern world beginning with David Hume and Immanuel Kant. Very briefly, Kant agrees with Hume that the law of causality is not a scientific law; however, whereas for Hume, habit underlies our belief in causality, for Kant causality is a form imposed on the data by the nature of the human mind. This is certainly not the place to delve into the issue. We limit ourselves to a statement by

Erwin Schroedinger, who 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." One is free to think, or not to think, causally. In the end, the verity of a scientific theory depends on whether it gives predictions that agree with experiment, not the way in which the scientist regards nature, either psychologically or metaphysically.

How is one to check if a scientific theory gives predictions that agree with experiment? Up until the 20th century the abstract symbols were assumed to be measurable in a straightforward manner. However, with the advent of Einstein's general theory of relativity, the terms of the purely mathematical structure no longer refer to the immediate phenomena of human perception as they earlier had. Verification of a system requires that the symbols be tied to observations by some semantic rules that relate not to the general principles of the mathematical model themselves but to conclusions drawn from the principles. In other words, the theory is checked by testing measurable consequences of the theory. These *operational definitions*, as they are called, are an intrinsic part of the theory, for without them there would be no connection between the principles and observation. The demand for operational definitions constitutes the *positivistic requirement* of science. The general principles 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 mean intensity of a spot on a cDNA microarray resulting from hybridized fluors. A scientific theory must have two parts: a structural model and a set of operational definitions for its symbols. It is not a straightforward matter to provide a suitable set of operational definitions, nor to even characterize what it means to be suitable; nevertheless, the two-part scheme provides a necessary general structure for a modern scientific theory.

Experimentation is no less important than the mathematical theory. 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 of a theory or a restriction on the symbols and operations constituting the theory. In a practical sense, the theorist and experimentalist must proceed in close connection. The theory, to be validated, must not exceed the experimentalist's ability to conceive and perform appropriate experiments, and the experimentalist cannot produce directly meaningful experiments unless they are designed with a symbolic structure in mind. In the context of the uncertainty principle, modern physics appears to have brought us beyond the situation of where the limitations on observation are owing to insufficient experimental apparatus to the point where

the limitations are unsurpassable in principle. In this vein, Erwin Schroedinger writes, "It really is the ultimate purpose of all schemes and models to serve as scaffolding for any observations that are at all conceivable."[32] He adds, "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." In other words, without observable effects due to an object, the object is not a subject of scientific inquiry. Charles Pierce goes so far as to say that an object of our thought is indistinguishable from its conceivable effects when he writes, "Consider what effects, that might conceivably have practical bearings, we conceive the object of our conception to have. Then our conception of these effects is the whole of our conception of the object."[33]

In his treatise on random geometrical measurements, Georges Matheron makes the following statement that binds the theorist and experimenter together:

> 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.[34]

The experimenter, in choosing the universe of relationships to be examined, frames at the outset the very kind of mathematical structure that can potentially result because the experimenter chooses the manner in which Nature is to be probed. The roots of this theorist-experimenter dialectic go back at least to Kant, who famously stated, "A concept without a percept is empty; a percept without a concept is blind."[35] An interpretation for the researcher might be, "A model without data is empty; data without a model is blind."

In the scientist's choice of how Nature is to be probed, subjectivity enters the scientific enterprise. A virtually unlimited number of experiments can be performed and those relatively few actually performed by the scientific community are somehow determined by psychological, cultural, and metaphysical considerations. Schroedinger 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."[32] The selection is influenced by the interests and goals of the investigator. These may be internal, such as ones desiring to alleviate the suffering of cancer or ones driving interest to unlock the secrets of Nature at the subatomic level, or they may be external, such as satisfying a granting agency or furthering the interest of a certain group. Schroedinger emphasizes the emotive

drive in scientific practice, as well as reinforcing 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."[32]

3. Computational Biology and "The Final Truth About Phenomena"

To the extent that computational biology provides mathematical models that can be validated by experimentation, it contains, in the words of James Jeans, "the final truth about [biological] phenomena." What does it mean for a model to be *biological*? Since the mathematical model consists of abstract symbols and relations between the symbols, the biological nature of the model does not inhere in it alone. Rather, it is phenomena that the model seeks to represent that determine the biological nature of a scientific theory. The biology inheres in the experiments, or perhaps more precisely, in the scientist's perception of the experiments.

For the moment, let us focus on gene regulatory networks. At once we are confronted by the need to define what we mean by a *gene regulatory network*. From the nature of the individual terms, such a thing would most likely concern a *network* of relations among strands of DNA (*genes*) and the *regulatory* activities related to this DNA. There are many mathematical systems that could be called "gene regulatory networks."

These are generally dynamical systems satisfying the aforementioned conditions and each is defined by a set of mathematical symbols and the relations between them. They are biologically important because describing the regulatory dynamics of a set of genes requires a mathematical model for the dynamical behavior of the expression vector for genes in the set. The critical epistemological role of gene regulatory dynamics is reflected in the outlook of Davidson *et al.*, who write, "The view taken here is that 'understanding' why a given development process occurs as it does requires learning the key inputs and outputs throughout the genomic regulatory system that controls the process as it unfolds."[36]

The goodness of a gene-regulatory model can be considered with respect to several criteria: the level of detailed description of the biochemical reactions involved in gene regulation, model complexity, model parameter estimation, and, most importantly, the predictive power of the model. The stochastic-differential-equation model is arguably the most detailed description of the dynamics of a gene-expression vector. It could imbed, at least in principle, all of the information about the known biochemical reactions involved in the gene interactions. At the same time, this kind of model has high complexity, and the estimation of its parameters cannot be done without reliable time series data, and a goodly amount of it.

Inevitably one looks for simpler, more pragmatic models. Perhaps the most extreme simplification is the Boolean model, originally proposed by Stuart Kauffman for gene regulatory modeling.[34] In the Boolean model, gene expression is quantized to two levels, ON and OFF, and the expression level of each gene

at time $t + 1$ is functionally related via logical rules to the levels of some other genes at time t. The basis of the Boolean model is that during regulation of functional states the cell exhibits switch-like behavior required for state transitions in normal growth or when a cell needs to respond to external signals. The model represents biological knowledge to the extent that it can predict observed binary vectors related to its own dynamical behavior as a nonlinear mathematical system. This knowledge is of a different sort than biochemical knowledge, but it is still knowledge.

Of the various forms of knowledge pertaining to genomics, one kind concerns 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 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.

As applied to gene expression, the Boolean model might best be described as a model of information dynamics. The functions that relate gene states at time $t + 1$ to those at time t only relate activity levels, and do not portray molecular transformations. This notion of activity relation is conveyed by the usual terminology for the state vector, which is called a *gene activity profile*. The functional relations between these profiles determine a state transition diagram that is used to study the long-run dynamics of the system. For a gene to be profitably included in a Boolean network, the distribution of its expression levels should be essentially bimodal, so that it can be reasonably modeled as ON-OFF and there is reason to suspect that it plays a switch-like role — for instance, as when two different transcription factors must bind to the cis-regulatory DNA to activate transcription, thereby exemplifying AND logic.

The greatest weakness in the Boolean model is not the binary nature of the state vectors, but its determinism. Whereas deterministic models can be used for phenomena not subject to consequential perturbations outside those internal to the system, they cannot model complex interactive physical systems subject to consequential external latent variables. Even the most ardent deterministic metaphysics does not dispute the necessity of stochastic scientific modeling. This is demonstrated by the words of the control theorist Vladimir Pugachev, who, after noting that the law of phenomenological inter-dependence is a fundamental law of dialectical materialism, states,

> 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.[38]

Deterministic phenomenology, be it Marxist or Laplacian, does not constrain science to deterministic models. Even if cell function were deterministic, it would be highly unlikely that this determinism would be reflected in a gene network since 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. The movement from Boolean networks to probabilistic Boolean networks[11] as models for genomic control is inevitable given the role of latent variables in the behavior of complex systems. In essence, a context-sensitive probabilistic Boolean network is a collection of Boolean (or finitely quantized) networks with the particular network governing the dynamical system at any point in time being determined by latent variables, and therefore being random relative to the model itself. Context sensitivity is dictated by the inevitability of variables outside the model system.

There is no doubt that for some purposes it would be better to employ more finely quantized models, or those incorporating protein interaction. And it is certainly more satisfying to possess a complex model from which a simpler model results under certain circumstances, as in the case of the excellent approximation given by Newton's laws to Einstein's laws at modest velocities. In the case of biological networks, a fuller description of the relations among the phenomena would result from networks involving DNA, RNA, and proteins; nevertheless, the extensive interaction between these constituents insures that a significant amount of the system information is available in each. In a recently proposed gene-regulatory-network model that includes both a genomic regulatory functional and a proteomic regulatory functional, it is shown that under certain conditions the model reduces to a purely genomic network; however, this can only occur in the steady state.[39] Similar reductions will no doubt occur in the future regarding the special conditions of steady-state behavior, as is typical for dynamical systems. To the degree that we are concerned with long-run behavior, these simpler systems can suffice.

Although we desire more complete descriptions of phenomena, or perhaps we should say mathematical models with richer sets of variables and relations between the variables, a key impediment to finer models is their need for more experimentation. As noted previously, limits on experimental capability place limits on model complexity. Expression microarrays make multivariate biological analysis feasible because they provide the ability to make the requisite experiments.[40,41] For instance, the historical approach of discovering relations between genes based on correlation is limited to finding linear univariate predictive relations. One of

the earliest uses of expression microarrays was to discover nonlinear multivariate predictive relationships.[6,42] Such predictive relationships can be employed as the functional relations governing state transitions in probabilistic Boolean networks. The new technology has provided the means to employ finer network models. Still, constraints on use of the new technology place limitations on the richness of the models. Owing to cost and the availability of RNA, sample sizes (the number of microarrays) tend to be small. This places restrictions on our ability to derive predictive relations from the data, in terms of both the complexity of predictor functions and the number of predictor variables. This restriction favors coarsely quantized networks with low connectivity and is reflected in the epistemological issues of predictive models. These issues are closely related to the corresponding epistemological issues for classification.

In general, one can understand a process in a rather deep way without exact knowledge of a particular mechanical implementation of the process. A process may be understood at different levels, with the appropriate level depending on ones intentions. For instance, if an algorithm is being implemented on a computer, the physical processing is taking place at the hardware level. If one is only concerned with the logical operations of the hardware, these are fully described by the machine code and the actual hardware can be ignored. Above the machine-code there are further levels of abstraction: assembly code, C code, Matlab code, and, finally, one can forego computer code altogether and describe the algorithm fully in mathematical terms. It is at this highest level that essential algorithmic properties such as convergence are best understood. In the area of information processing there is no reason to expect that biological decision-making represents a different process than any other kind of decision making, though it no doubt uses very different components to carry out the process. In analogy to computer science, one might say that there is a decision-making layer and a physical [chemical] layer. Boolean networks provide representation at the decision-making layer. Logical relations such as $MRC1 = VSNL1$ OR $HTR2C$ correspond to changes in the continuous data related to the up- and down-regulated character of the genes involved.[43] Such relations are predictive. It may not always be true that $MRC1 = 1$ when $\mathrm{V}SNL1 = 0$ and $HTR2C = 1$. Appreciating the validity of the logical relation depends on the probability of the relation holding under certain conditions, in much the same way as for classifiers, which themselves are decision functions.

Attaining a high complexity of form starting from a less complex form may itself be an example akin to computation, a type of process that has general rules and requirements and can be implemented on a wide variety of platforms. Knowledge of these rules and requirements may be necessary understandings that do not derive from ever-finer parameterization of mechanical events and may only be approached by studying a variety of the products of the process. In the case of computational processes, the existence of general requirements constrains the space of potential processes. Typical among such constraints are restrictions on

computational complexity. It is often the case that an optimal solution exists but one must look for a suboptimal solution owing to the complexity of a full search. Thus, the space of potential algorithms is reduced. If there are laws governing the evolution of complexity, then the space of potential developmental trajectories may be greatly diminished with respect to the unconstrained space of all possible random trajectories. In this vein, Kauffman writes,

> An effort to include the emergent self-organizing properties typical of large ensembles of systems in evolutionary theory must provoke a resonant set of questions and consequences. Not the least of these is an interesting epistemological implication. If we should find it possible to account for, explain, predict widespread features of organisms on the basis of the generic properties of underlying ensembles, then we would not need to carry out in detail the reductionistic analysis of organisms in order to explain some of their fundamental features. As the physicist explains ice formation as a typical phase transition in a general class of systems, so we might explain aspects of organisms as typical of their class.[44]

We close this section by asking what can be said about the reality of systems composed of symbols representing measurements and relations between those symbols, and is there something less real or less biological about a network representing information flow than one representing the chemical description of transcription factors binding to the DNA to seed the formation of a recognition site to which an RNA polymerase can bind and initiate copying from a DNA strand? Since a living system is of necessity an information processing system, it seems unreasonable to maintain that a genomic network corresponding specifically to control information is not biological. Indeed, the desire to understand information processing within the cell is a salient motivation for network construction. According to Davidson *et al.*,

> It seems no more possible to understand development from an informational point of view without unraveling the underlying regulatory networks than to understand where protein sequence comes from without knowing about the triplet code... The cis-regulatory systems at the nodes of the network in reality each process kinetic input information: the rise and fall of the activities of the transcription factors to which they respond.[36]

Regarding reality, the fact that a complete biochemical description of cellular activity would likely produce the corollary description of the information processing system does not denigrate the reality of the latter. Ultimately, scientific knowledge resides in the minds of scientists. Henri Poincaré states the matter well,

> Does the harmony which human intelligence thinks it discovers in Nature exist apart from such intelligence? Assuredly no. A reality completely independent of the spirit that conceives it, sees it or feels it, is an impossibility.

> A world so external as that, even if it existed, would be forever inaccessible to us. What we call "objective reality" is, strictly speaking, that which is common to several thinking beings and might be common to all; this common part, we shall see, can only be the harmony expressed by mathematical laws.[45]

4. Classifier Models

Classification and its related methodologies play a major role in the analysis of data from high-throughput technologies. A typical application is the use of microarray data to design an expression-based classifier to distinguish different types of glioma, for instance, to discriminate between anaplastic astrocytoma and anaplastic oligodendroglioma.[46] Suppose some classification rule, say a support vector machine, nearest-neighbor rule, or neural network, is used to obtain a classifier ψ from the data. The immediate question is this: What kind of knowledge is represented by ψ? It certainly provides a model relating gene expression to the categorization of glioma. If $g_1, g_2, \ldots, g_d$ are the genes whose expressions form the arguments for ψ, then upon providing values for these arguments, ψ produces a binary value, 0 or 1, representing either anaplastic astrocytoma or anaplastic oligodendroglioma. But to constitute scientific knowledge, ψ must be related to quantifiable observations, and not merely stand alone as a mathematical function.

One might naively approach the matter by saying that a classifier derived from data via a classification rule is *ipso facto* related to observations, those being the data from which it has been derived. But a scientific theory is not concerned merely with how a model is related to this or that particular observation, but how it is related to observations in general. In the extreme case, suppose there was a sound scientific principle that the expression levels of the genes $g_1, g_2, \ldots, g_d$ were completely determinative of the type of glioma. This would mean that there exists a classifier ϕ such that $\phi(\mathbf{x}) = 0$ whenever $\mathbf{x}$ represents the expression levels of the genes $g_1, g_2, \ldots, g_d$ coming from a patient suffering from anaplastic astrocytoma, and $\phi(\mathbf{x}) = 1$ whenever $\mathbf{x}$ represents the expression levels of the genes $g_1, g_2, \ldots, g_d$ coming from a patient suffering from anaplastic oligodendroglioma. Now, in the absence of a determinative scientific principle, suppose a designed classifier ψ perfectly classifies the sample data from which it has been derived. The fact that ψ is perfect on the sample data does not mean that it provides a perfect classifier in the sense that ψ provides perfect classification for future observations.

In fact, in the case of glioma, and in any other complex disease setting, there does not exist a classifier that provides perfect classification over all possible observation vectors. This means that every classifier will have an error rate (based on probabilistic considerations to be discussed subsequently). This error rate constitutes the goodness of the classifier and, absent its error rate, the classifier lacks a

quantifiable relationship with events. Notice that it is implicit that the error rate is relative to some population of events. Moreover, the error rate must be estimated from sample data. The quality of the estimate determines the validity of the (*classifier*, *error*) pair, and therefore the manner in which we measure this quality is a key epistemological factor in the scientific model. As with scientific epistemology in general, the procedure for estimating the error becomes paramount in this regard.

Having discussed classification issues in general, we will now consider its epistemological basis. This inevitably requires precise mathematical formulation.

Classification involves a *feature vector* $\mathbf{X} = (X_1, X_2, \ldots, X_d)$ on d-dimensional Euclidean space $\mathbb{R}^d$ composed of random variables (*features*), a binary random variable Y, and a function (*classifier*) $\psi : \mathbb{R}^d \to \{0, 1\}$ for which $\psi(\mathbf{X})$ is to predict Y. The values, 0 or 1, of Y are treated as class labels. Given a feature-label probability distribution $f_{\mathbf{X},Y}(\mathbf{x}, y)$, the error, $\epsilon_f[\psi]$, of ψ is the probability or erroneous classification, namely, $\epsilon_f[\psi] = P(\psi(\mathbf{X}) \neq Y)$. Classification accuracy, and thus the error, depends on how well the labels are separated by the variables used to discriminate them.

We consider a *classifier model* $\mathrm{M} = (\psi, \epsilon_\psi)$ as a pair composed of a function $\psi : \mathbb{R}^d \to \{0, 1\}$ and a real number $\epsilon_\psi \in [0, 1]$. ψ and ϵ_ψ are called the *classifier* and *error* of the model M. The mathematical form of the model is abstract, with ϵ_ψ not specifying an actual error probability corresponding to ψ. M becomes a scientific model when it is applied to a feature-label distribution. It is at this point that the validity of the model comes into question. The model is valid for the distribution $f_{\mathbf{X},Y}$ to the extent that ϵ_ψ approximates $\epsilon_f[\psi]$. Hence, quantification of model validity is relative to the absolute difference $|\epsilon_f[\psi] - \epsilon_\psi|$.

Since classification is inherently stochastic, the rate at which the classifier makes correct predictions is an inherent part of the model that measures our belief that the classifier will make the proper decision. In a deterministic model, the system is contingently validated each time an observation is in accord with its prediction. It stands perpetually open to invalidation if observations do not agree with predictions. In the case of a non-deterministic system, like a classifier, incorrect predictions are expected and therefore the system requires a probabilistic description of classifier accuracy. This description — the error rate in the case of decisions — must be part of the model, and the validity of the model corresponds to the accuracy of that probabilistic description.

What about the goodness of a model? It may be perfectly valid with $\epsilon_f[\psi] = \epsilon_\psi$, but with $\epsilon_f[\psi] = 0.5$, meaning it is no better than flipping a coin. In fact, the quality of goodness does not apply to the model M, but only to the classifier. Classifier ψ is *better* than classifier ϕ relative to the distribution f if $\epsilon_f[\psi] < \epsilon_f[\phi]$. If we have two models, $\mathrm{M}_\psi = (\psi, \epsilon_\psi)$ and $\mathrm{M}_\phi = (\phi, \epsilon_\phi)$, then it may well be that ψ is a better classifier than ϕ but that M_ϕ is more valid than M_ψ, in the sense that $|\epsilon_f[\psi] - \epsilon_\psi| > |\epsilon_f[\phi] - \epsilon_\phi|$. Regarding goodness, a classifier ψ is *optimal* (best) for a feature-label distribution $f_{\mathbf{X},Y}$ if $\epsilon_f[\psi] \leq \epsilon_f[\phi]$ for any classifier $\phi : \mathbb{R}^d \to \{0, 1\}$. An

optimal classifier, ψ_f, of which there may be more than one, and its error, $\epsilon_f[\psi_f]$, are deducible from the feature-label distribution. These are called a *Bayes classifier* and the *Bayes error*, respectively.

Putting together classifier goodness and model validity, and taking a standard approach to pairwise ordering, we say that model $\mathrm{M}_\psi = (\psi, \epsilon_\psi)$ is *better* than model $\mathrm{M}_\phi = (\phi, \epsilon_\phi)$ if $\epsilon_f[\psi] \leq \epsilon_f[\phi]$ and $|\epsilon_f[\psi] - \epsilon_\psi| \leq |\epsilon_f[\phi] - \epsilon_\phi|$, with strict inequality holding in at least one of the inequalities. Contingency of a (non-Bayes) model is manifest in the possibility that a better model may be discovered. Better models can be found by gathering more data or using a different procedure for model formation.

Model goodness and validity are problematic in practice. Given two models, $\mathrm{M}_\psi = (\psi, \epsilon_\psi)$ and $\mathrm{M}_\phi = (\phi, \epsilon_\phi)$, were we to know the feature-label distribution, we could then evaluate $\epsilon_f[\psi]$ and $\epsilon_f[\phi]$ directly to decide which classifier is better and which model is more valid. Of course, if we know the feature-label distribution, then we have the Bayes classifier and its error, and these would compose the model. In practice we do not know the feature-label distribution. If $\epsilon_\psi < \epsilon_\phi$, then ψ is apparently better than ϕ, but we cannot know if it is actually better because we cannot know $\epsilon_f[\psi]$ and $\epsilon_f[\phi]$. Comparing validity via $|\epsilon_f[\psi] - \epsilon_\psi|$ and $|\epsilon_f[\phi] - \epsilon_\phi|$ is impossible for the same reason. We will return to this issue shortly.

So far we have said nothing about how the model $\mathrm{M}_\psi = (\psi, \epsilon_\psi)$ has been discovered. In fact, discovery of the model is not part of the model. The scientist could have discovered it as "a bolt out of the blue," perhaps awakening with it out of a dream. One can imagine James Clerk Maxwell viewing the beautiful waters off the Isle of Skye, when suddenly Maxwell's equations pop into his head. So long as the equations are consistent with observation, whether he had been meditating on the problem for years, or had never thought of it before, is irrelevant. More to the immediate issue at hand, a researcher may have gathered some two-dimensional labeled data, plotted the data points on a graph, and based on the labels, by sight drawn a line to separate the points to some degree, thereby defining a classifier. The goodness of this effort depends on the error of the resulting classifier. In either event, whether it be the great physicist hypothetically vacationing on the scenic island or the researcher plotting points in the laboratory, the model is not arrived at by rational analysis. According to Karl Popper,

> The question of how it happens that a new idea occurs to a man — whether it is a musical theme, a dramatic conflict, or a scientific theory — may be of great interest to empirical psychology; but it is irrelevant to the logical analysis of scientific knowledge ... Every discovery contains "an irrational element," or "a creative intuition," in Bergson's sense. In a similar way, Einstein speaks of the "search for those highly universal laws ... from which a picture of the world can be obtained by pure deduction. There is no logical path," he says, "leading to these ... laws. They can only be

> reached by intuition, based on something like an intellectual love of the objects of experience."[25]

Albert Einstein accentuates the role of creativity when he states, "Experience, of course, remains the sole criterion for the serviceability of mathematical constructions for physics, but the truly creative principle resides in mathematics."[47] According to Feynman, "The laws are guessed; they are extrapolations into the unknown."[28] The veracity of a scientific model lies in experience, but its conception arises from the imagination. We must not interpret this to mean that there is creativity only in construction of the formal mathematical structure. The operational definitions that relate the model to the data of experience are an integral part of the scientific model and their formation is also a creative act, including the design of experiments that provide the data. This does not contradict the mathematical emphasis in Einstein's statement because the formal structure of the operational definitions is symbolic, including experimental design and the statistical procedures applied to the data, the latter necessarily being grounded within the mathematical theory of probability.

In practice, the scientist does not discover a classifier directly, but instead applies an algorithm that takes feature-label data as input and yields a classifier. Given a random sample $S_n = \{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \ldots, (\mathbf{X}_n, Y_n)\}$ of pairs drawn from a feature-label distribution $f_{\mathbf{X},Y}(\mathbf{x}, y)$, we desire a function on S_n that yields a good classifier. A *classification rule* is a mapping of the form $\Psi_n : [\mathbb{R}^d \times \{0,1\}]^n \to \mathrm{F}_d$, where F_d is the set of $\{0,1\}$-valued functions on $\mathbb{R}^d$. Given a specific sample s_n (a realization of S_n), we obtain a designed classifier $\psi_n = \Psi_n(s_n)$ according to the rule Ψ_n. Note that if feature selection is involved, then it is part of the classification rule. The rule still operates on d variables and the classifier is still a member of F_d, albeit with less than d essential variables.

A key epistemological point is that the classifier, which is part of the scientific model, is not obtained by a direct creative act (recalling Einstein's comment); rather, the creative act is in the choice of classification rule. It is in the choice of classification rule that the scientist brings to bear an accumulated understanding of the properties of classification rules and an appreciation of the phenomena under study. Once the scientist decides on a classification rule, classifier design is purely a deductive act via the machinery of mathematical operations. The data sample is entered into the operational machinery of the classification rule and a classifier results. In this deductive stage the computer plays a critical role: it facilitates the application of very complex and highly computational classification rules.

What about the discovery of the error term ϵ_ψ in the model $\mathrm{M} = (\psi, \epsilon_\psi)$? Suppose the scientist has observed some points in the plane and by sight drawn a straight line to separate them to some degree. Can ϵ_ψ be "picked out of the air?" Of course it can. But this kind of choice is not likely to produce a valid model, as judged by the comparison between ϵ_ψ and the true error of the classifier. A more likely scenario is that the scientist counts the number of points in the

data misclassified by the line. In this case, the scientist has actually proposed an estimation rule, Ξ_n. Although there is no logical necessity, we will assume that the classifier is part of the estimation rule (else one would be estimating the error independent of the classifier). Like the choice of classification rule, selection of an estimation rule is a creative act of the scientist. Also in analogy to the classification rule, once the scientist decides on an estimation rule, error estimation is a deductive act via the machinery of mathematical operations. The data sample is entered into the computational machinery of the estimation rule and an estimate results. This is a second deductive stage in which the computer plays a key role.

Altogether, we arrive at a scientific model $\mathrm{M} = (\psi, \epsilon_\psi)$ by a creative act that postulates a rule model $\mathrm{L} = (\Psi_n, \Xi_n)$ and then via deduction from a data sample. The rule model consists of two operators whose formal structure dictates the computations to be done on the data. This paradigm of rule-model formation, data acquisition, and deductive computation to arrive at a classifier model constitutes the scientific method in the context of classification. More generally, it constitutes the scientific method in "learning" models from data. It is crucial to computational biology, where model learning is critical in numerous contexts. Note that creativity is involved in three ways: forming the classification and estimation rules, experimental design, and the construction of computational methods from which to deduce the scientific models from the rule models.

One codicil must be amended to the possible nature of the rule model. The classification rule or estimation rule may contain a random component — for instance, random data perturbation or inclusion of random noise in classifier design, or random data selection in bootstrap and cross-validation error estimation. For randomized rules, Ψ_n and/or Ξ_n become random functions on $\mathbb{R}^d$ instead of deterministic functions on $\mathbb{R}^d$. The model $\mathrm{M} = (\psi, \epsilon_\psi)$ is still derived by a computation, but it is a randomized computation, with the computer performing both randomization (pseudo-randomization) and operator computation.

The goodness of a classifier relates to the precision of Ψ_n as an estimator of a Bayes classifier: if a classification rule is expected to yield a classifier whose error is close to that of a Bayes classifier, then we have confidence that a designed classifier will be close to being as good as possible for the feature-label distribution in question. In terms of a classifier model $\mathrm{M} = (\psi, \epsilon_\psi)$, we consider the goodness of ψ under the assumption that both ψ and ϵ_ψ have been arrived at via the rule model $\mathrm{L} = (\Psi_n, \Xi_n)$. Thus, we consider the model $\mathrm{M}_n = (\psi_n, \hat{\epsilon}[\psi_n])$, where $\psi_n = \Psi(S_n)$ and $\hat{\epsilon}[\psi_n] = \Xi_n(S_n)$ for sample data set S_n. As is typical in a probabilistic setting, we turn our attention to the performance of the rule model.

To compare two classifiers, ψ and ϕ, relative to a distribution $f_{\mathbf{X},Y}(\mathbf{x}, y)$, it is just as well to compare $\epsilon_f[\psi] - \epsilon_f[\psi_f]$ to $\epsilon_f[\phi] - \epsilon_f[\psi_f]$, as compare $\epsilon_f[\psi]$ to $\epsilon_f[\phi]$, both of which exceed the Bayes error $\epsilon_f[\psi_f]$. Hence, the relative goodness of a designed classifier ψ_n can be measured by its *design cost* $\Delta_f[\psi_n] = \epsilon_f[\psi_n] - \epsilon_f[\psi_f]$. From the perspective of the classification rule, $\Delta_f[\psi_n]$ and $\epsilon_f[\psi_n]$ are sample-dependent random variables. Thus the salient quantity for a classification rule is the expected

design cost, $E[\Delta_f[\psi_n]]$, the expectation being relative the random sample S_n. The expected error of the designed classifier is decomposed as

$$E[\epsilon_f[\psi_n]] = \epsilon_f[\psi_f] + E[\Delta_f[\psi_n]]. \tag{4.1}$$

A great deal of pattern recognition literature deals with finding classification rules for which $E[\Delta_f[\psi_n]]$ is small — qualitatively, a rule is good if $E[\Delta_f[\psi_n]]$ is small.

A classification rule can yield a classifier that makes very few errors on the sample data on which it is designed, but performs poorly on the distribution as a whole. This situation is exacerbated by complex classifiers and small samples. If the sample size is dictated by experimental conditions, such as cost or the availability of patient RNA for expression microarrays, then one only has control over classifier complexity. The problem is not necessarily mitigated by applying an error-estimation rule to the designed classifier to see if it "actually" performs well, since when there is only a small amount of data available, error-estimation rules are very imprecise, and the imprecision tends to be worse for complex classification rules. Thus, a low error estimate is not sufficient to mitigate the large expected design error owing to using a complex classifier with a small data set.

To alleviate the problem of overfitting, one may constrain classifier design by restricting the functions from which a classifier can be chosen to a class C. Constraint can reduce the expected design error, but at the cost of increasing the error of the best possible classifier. Since optimization in C is over a subclass of classifiers, the error of an optimal classifier, ψ_{C}, in C will typically exceed the Bayes error, unless $\psi_f \in \mathrm{C}$. This *cost of constraint* is $\Delta_f^{\mathrm{C}} = \epsilon_f[\psi_{\mathrm{C}}] - \epsilon_f[\psi_f]$. A classification rule yields a classifier $\psi_{n,\mathrm{C}} \in \mathrm{C}$ with error $\epsilon_f[\psi_{n,\mathrm{C}}]$, such that $\epsilon_f[\psi_{n,\mathrm{C}}] \geq \epsilon_f[\psi_{\mathrm{C}}] \geq \epsilon_f[\psi_f]$. Design error for constrained classification is $\Delta_{f,\mathrm{C}}[\psi_{n,\mathrm{C}}] = \epsilon_f[\psi_{n,\mathrm{C}}] - \epsilon_f[\psi_{\mathrm{C}}]$. For small samples, this can be much less than $\Delta_f[\psi_n]$, depending on C and the rule. The expected error of the designed classifier from C can be decomposed as

$$E[\epsilon_f[\psi_{n,\mathrm{C}}]] = \epsilon_f[\psi_f] + \Delta_f^{\mathrm{C}} + E[\Delta_{f,\mathrm{C}}[\psi_{n,\mathrm{C}}]]. \tag{4.2}$$

The constraint is beneficial if and only if $E[\epsilon_f[\psi_{n,\mathrm{C}}]] < E[\epsilon_f[\psi_n]]$, which is true if the cost of constraint is less than the decrease in expected design cost. The dilemma is that strong constraint reduces $E[\Delta_{f,\mathrm{C}}[\psi_{n,\mathrm{C}}]]$ at the cost of increasing Δ_f^{C}.

5. Classifier-Model Validity

We now consider the validity of a classifier model $\mathrm{M} = (\psi, \epsilon_\psi)$, which is the key epistemological issue. Again assuming that both ψ and ϵ_ψ have been arrived at via the rule model $\mathrm{L} = (\Psi_n, \Xi_n)$, we consider the model $\mathrm{M}_n = (\psi_n, \hat{\epsilon}[\psi_n])$, where $\psi_n = \Psi(S_n)$ and $\hat{\epsilon}[\psi_n] = \Xi_n(S_n)$ for sample data set S_n. Model validity relates to the precision of the error estimator $\hat{\epsilon}[\psi_n]$ in the model $\mathrm{M}_n = (\psi_n, \hat{\epsilon}[\psi_n])$, which can be considered random, depending on the sample. The precision of the estimator relates to the difference between $\hat{\epsilon}[\psi_n]$ and $\epsilon_f[\psi_n]$, and we require a probabilistic

measure of this difference. Here we use the root-mean-square error (square root of the expectation of the squared difference),

$$RMS(\Psi_n, \hat{\epsilon}, f, n) = \sqrt{E[|\hat{\epsilon}[\psi_n] - \epsilon_f[\psi_n]|^2]}. \tag{5.1}$$

Error-estimation precision depends on the classification rule Ψ_n, error estimator $\hat{\epsilon}$, feature-label distribution f, and sample size n. The RMS can be decomposed into the bias, $Bias[\hat{\epsilon}] = E[\hat{\epsilon}[\psi_n] - \epsilon_f[\psi_n]]$, of the error estimator relative to the true error, and the deviation variance, $Var_{dev}[\hat{\epsilon}] = Var[\hat{\epsilon}[\psi_n] - \epsilon_f[\psi_n]]$, namely,

$$RMS(\Psi_n, \hat{\epsilon}, f, n) = \sqrt{Var_{dev}[\hat{\epsilon}] + Bias[\hat{\epsilon}]^2}, \tag{5.2}$$

where we recognize that Ψ_n, f, and n are implicit on the right-hand side.

We consider two error estimators relative to model validity in the context of a classification rule for which quite a bit is known about the RMS. In *multinomial discrimination*, the feature components are random variables whose range is the discrete set $\{0, 1, \ldots, b-1\}$. This corresponds to choosing a fixed-partition in $\mathbb{R}^d$ with b cells. The *histogram rule* assigns to each cell the majority label in the cell. The *resubstitution estimator*, $\hat{\epsilon}^{res}$, is the fraction of errors made by the designed classifier on the sample. For histogram rules, it is biased low, meaning $E[\hat{\epsilon}^{res}[\psi_n]] \leq E[\epsilon_f[\psi_n]]$. For small samples, the bias can be severe, bias increasing for increasing complexity (increasing number of cells). Bias lessens for large samples. For the *leave-one-out estimator*, $\hat{\epsilon}^{loo}$, n classifiers are designed from sample subsets formed by leaving out one sample point, each is applied to the left-out point, and the estimator is $\frac{1}{n}$ times the number of errors made by the n classifiers. It is unbiased as an estimator of the expected error for samples of size $n-1$, meaning $E[\hat{\epsilon}^{loo}[\psi_n]] = E[\epsilon_f[\psi_{n-1}]]$. Thus, there is only a small bias component for the RMS; however, the leave-one-out estimator has a high variance component compared to resubstitution and the variance can be sufficiently severe to offset its bias advantage when it comes to model validity. There exist distribution-free bounds on the RMS for both resubstitution and leave-one-out in the context of the histogram rule for multinomial discrimination[48] and, given the feature-label distribution, there exist exact analytic formulations of the RMS for both resubstitution and leave-one-out.[49] The latter expressions show that the RMS decreases for decreasing b. They also show that for a wide range of distributions, resubstitution outperforms leave-one-out for fewer than eight cells (which corresponds to three binary predictor variables). Thus, we can expect greater model validity if we use resubstitution in these cases. For 16 cells (four binary predictor variables) and up, leave-one-out is superior.

In practice one does not know the feature-label distribution, so that distribution-free RMS bounds can be useful if they are suffciently tight; however, if exact validity measurements have been obtained for many representative models, then these provide an indication of what one might expect for similar distributions. As always, there is a creative step in choosing the rule model. There is no non-mathematical way to precisely describe knowledge regarding model validity. It depends on the

choice of validity measurement and the mathematical properties of that measurement as applied in different circumstances. Generally, model validity improves for large samples and less complex classifiers. 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 are limited, then one's appreciation of the scientific content of a model is limited.

Our considerations relating to model validity have practical consequences relative to experimental design. For instance, one can obtain bounds of the sort

$$E[\Delta_{f,\mathrm{C}}[\psi_{n,\mathrm{C}}]] \leq \lambda_{\mathrm{C}}\sqrt{\frac{\log n}{2n}}, \tag{5.3}$$

for the expected design cost of a constrained classifier, where λ_C is a constant depending on the complexity of the classification rule, and n is the sample size.[48,50] For small samples, the kind typically encountered with microarray experiments, the bound exceeds 1 even for relatively simple classification rules. One might argue that this is only a bound, and that in fact the design cost might be small for a complex classification rule even if the sample is not large. Certainly this is possible. One might further argue that the way to proceed is to go ahead and apply the complex classification rule and then estimate its error. If the estimated error is small, then conclude that classifier is good. The problem with this approach is that the model lacks validity. Given the small sample and high complexity, the error estimate will be poor. The situation is exacerbated if one applies several complex classification rules and then chooses the designed classifier with lowest estimated error. Here we confront a variant of the classical multiple-comparisons problem. Owing to the high variance of the error estimator, if one tries enough rules, one is bound to design a classifier possessing an optimistically small error estimate. To go further with these practical considerations, suppose it is actually true that a complex classifier has a low error for a small sample. Then, in fact, it is likely that the feature-label distribution is not complex and that a low-complexity classifier would have performed at least as well, or better. Although we have stated this principle absent mathematical formality, it can be given formality by defining a complexity measure for feature-label distributions.[51] The conclusion from these considerations is that one should apply a simple classification rule when the sample size is small.

From a general perspective of experimental design, one cannot decouple the mathematical model from the experiment or from the statistical methodology applied to the data. Douglas Montgomery puts the matter in a way that highlights the creative step in experimental design as well as its epistemological role:

> If an experiment is to be performed most efficiently, then a scientific approach to planning the experiment must be considered. By the statistical design of experiments we refer to the process of planning the experiment so

> that appropriate data will be collected, which may be analyzed by statistical methods resulting in valid and objective conclusions. The statistical approach to experimental design is necessary if we wish to draw meaningful conclusions from the data.[52]

It is implicit in the notion of appropriate data being analyzed in a way that results in valid conclusions that design must take into account the model to which the data are to be applied and the relation of the data to the variables in the model. In this way, Kant's concept-percept duality manifests itself in the design of experiments, in the process requiring a tight practical connection between theory and experiment.

6. Concluding Remarks

If we remain within the epistemology of modern science, then the mathematical role of computational biology is to form the theoretical content of biological knowledge, its computational role is to provide the algorithms to implement the complex functions necessary for modeling, and its statistical role is to provide the machinery to quantify the relation between the mathematical model and experimental data. It must always be remembered that science is not mathematics. There must be a procedure for relating consequences of the mathematical system to quantifiable observations. The theory must be predictive. It must have two parts: a mathematical system and a way of relating the system to observations. We understand the theory to the degree that we understand the mathematical system. We believe the theory to the degree to which observations confirm predictions of the mathematical system.

In discussing classification, which as we remarked is prototypical of statistical methods used in computational biology, we have shown how the theory conforms to the epistemological requirements of modern science. This is not an abstract discussion outside the purview of practicing scientists. The meaning of scientific knowledge lies within the scientific epistemology, and surely that meaning is important to any working scientist. Of particular importance is the predictive requirement and how it is satisfied by the theory of classification. It would be mistaken to think that epistemology is an afterthought, to be explained once a working theory — mathematical system and experimental method — are in place. Quite the opposite! Epistemology places demands on research if that research is to have scientific content.

To help make the point, we could first recall James Clerk Maxwell hypothetically sitting on the Isle of Skye dreaming of his equations, but instead we will imagine a young geneticist lying on shady grass outside the laboratory. The rest of the team is inside studying the recent microarray data and applying a neural-network classification rule to derive a classifier to discriminate between two types of glioma. The young geneticist is meditating on the molecular structure of certain genes and their roles in protein regulation. Eureka! Suddenly it crystallizes in the

young iconoclast's mind that the expression levels of CREB1 and RAB3A can be used as inputs to a linear classifier to discriminate between anaplastic astrocytoma and anaplastic oligodendroglioma. Not only that, but the meditations also reveal coefficients for the linear model. How does this model compare to one obtained by the team applying a neural-network classification rule to a typically small sample of microarray data? This will be determined by a fierce struggle for survival. For now, however, we can compare the error estimate of the designed neural network on the sample data with the error of the linear classifier on the sample data, keeping in mind the properties of error estimation. The truly better classifier is the one that will make fewer errors in the long run.

For another example of how epistemology places requirements on research, we will consider briefly data clustering. As historically applied, a set of data points is obtained and input to a clustering algorithm to partition the data into clusters. How the algorithm has been discovered is not a matter for clustering epistemology. Most likely it has been discovered by a researcher meditating on mathematical issues relating to grouping data of different types. For instance, the k-means algorithm is related to finding a collection of centers to minimize an empirical squared-Euclidean-distance error. The critical epistemological issue is prediction. Meditative model discovery is fine, but the scientific epistemology requires measurement of the model's predictive capability. This requires an error theory in which prediction can be evaluated, as with classification. Jain *et al.* get at the depth of the problem when they write, "Clustering is a subjective process; the same set of data items often needs to be partitioned differently for different applications."[53] The problem here is immediate; indeed, as stated by Karl Popper, "The objectivity of scientific statements lies in the fact that they can be inter-subjectively tested."[25] Subjective science is an oxymoron. But what else could one say when there is no theory of error? Going further, Duda *et al.* bring up the whole issue of *ad hoc* data manipulation in regard to clustering when they state, "The answer to whether or not it is possible in principle to learn anything from unlabeled data depends upon the assumptions one is willing to accept — theorems cannot be proved without premises."[54] These criticisms raise the question as to whether clustering can be used for scientific knowledge. What is to be done? Should we give up grouping data? Should we group data and give up on science? Neither option is acceptable. The answer is to place clustering into a proper mathematical framework so that the predictive capability of an algorithm can be measured.

Whereas a classifier operates on a point to produce a label, a clustering algorithm operates on a set of points to produce a partition of the point set. The probabilistic theory of classification, which provides its epistemological foundation, is based on a classifier being viewed as an operator on random points (vectors). A corresponding probabilistic theory of clustering would view a clustering algorithm as an operator on random point sets. Moreover, whereas the predictive capability of a classifier is measured by the decisions it yields regarding the labeling of random points, the

predictive capability of a clustering algorithm would be measured by the decisions it yields regarding the partitioning of random point sets. Once this is recognized, the path to the development of error estimators for clustering accuracy and rules to derive clustering operators from data is open and the entire issue can be placed on firm epistemological ground.[55] This does not close the matter; rather, new burdens are placed on creativity to develop the mathematical theory of clustering, invent clustering rule models, and devise experimental methods — all in the context of random sets, which is a much more challenging environment than that of random variables. Scientific epistemology is demanding, but it should never be looked upon as an impediment; rather, it should be seen as a guide to both theoretical and experimental research.

Acknowledgments

The first author acknowledges the support of the National Human Genome Research Institute (NHGRI), the National Cancer Institute (NCI), and the Translational Genomics Research Institute (TGen). The second author acknowledges the support of the National Institute of Allergy and Infectious Diseases (NIAID), under Grant U19 AI56541, and of the Brazilian National Research Council (CNPq), under Grant DCR 35.0382/2004.2.

References

1. Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M *et al.*, Tissue classification with gene expression profiles, *J Comput Biol* **7**:559–583, 2000.
2. Bittner M, Meltzer P, Khan J, Chen Y, Jiang Y *et al.*, Molecular classification of cutaneous malignant melanoma by gene expression profiling, *Nature* **406**:536–540, 2000.
3. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M *et al.*, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science* **286**:531–537, 1999.
4. Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M *et al.*, Gene expression profiles distinguish hereditary breast cancers, *New Engl J Med* **34**:539–548, 2001.
5. West M, Blanchette C, Dressman H, Huang E, Ishida S *et al.*, Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci USA* **98**:11462–11467, 2001.
6. Kim S, Dougherty ER, Bittner ML, Chen Y, Sivakumar K *et al.*, A general framework for the analysis of multivariate gene interaction via expression arrays, *Biomed Opt* **5**:411–424, 2000.
7. Zhou X, Wang X, Dougherty ER, Construction of genomic networks using mutual-information clustering and reversible-jump Markov-Chain-Monte-Carlo predictor design, *Signal Process* **83**:745–761, 2003.
8. de Jong H, Modelling and simulation of genetic regulatory systems: a literature review, *J Comput Biol* **9**:67–103, 2002.
9. Friedman N, Linial M, Nachman J, Pe'er D, Using Bayesian networks to analyze expression data, *J Comput Biol* **7**:601–620, 2000.

10. Huang S, Gene expression profiling, genetic networks, and cellular states: an integrating concept for tumorigenesis and drug discovery, *Mol Med* **77**:469–480, 1999.
11. Shmulevich I, Dougherty ER, Zhang W, From Boolean to probabilistic Boolean networks as models of genetic regulatory networks, *Proc IEEE* **90**:1778–1792, 2002.
12. Tegner J, Yeung MK, Hasty J, Collins JJ, Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling, *Proc Natl Acad Sci USA* **100**:5944–5945, 2003.
13. Ben-Dor A, Shamir R, Yakhini Z, Clustering gene expression patterns, *J Comput Biol* **6**:281–297, 1999.
14. Eisen MB, Spellman PT, Brown P, Botstein D, Cluster analysis and display of genome-wide expression patterns, *Proc Natl Acad Sci USA* **95**:14863–14868, 1998.
15. Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL, Model-based clustering and data transformation for gene expression data, *Bioinformatics* **17**:977–987, 2001.
16. Dougherty ER, Small sample issues for microarray-based classification, *Comp Funct Genomics* **2**:28–34, 2001.
17. Mehta T, Murat T, Allison DB, Towards sound epistemological foundations of statistical methods for high-dimensional biology, *Nat Genet* **36**:943–947, 2004.
18. Chen J, Dougherty ER, Demir S, Friedman C, Sheng Li C, Wong S, Grand challenges for multimodal bio-medical systems, *IEEE Circuits and Systems Magazine*, Spring Quarter, pp. 46–52, 2005.
19. Ioannidis JP, Microarrays and molecular research: noise discovery?, *Lancet* **365**:454–455, 2005.
20. Jenssen TK, Hovig E, Gene-expression profiling in breast cancer, *Lancet* **365**:634–635, 2005.
21. Michiels S, Koscielny S, Hill C, Prediction of cancer outcome with microarrays: a multiple random validation strategy, *Lancet* **365**:488–482, 2005.
22. Braga-Neto UM, Dougherty ER, Is cross-validation valid for small-sample microarray classification?, *Bioinformatics* **20**:374–380, 2004.
23. Popper K, *Conjectures and Refutations*, Routledge, London, 1963.
24. Frank P, *Modern Science and Its Philosophy*, Harvard University Press, Cambridge, 1949.
25. Popper K, *The Logic of Scientific Discovery*, Hutchinson, London, 1959.
26. Jeans JH, *The Mysterious Universe*, Cambridge University Press, Cambridge, 1930.
27. Feynman R, *QED The Strange Theory of Light and Matter*, Princeton University Press, Princeton, 1985.
28. Feynman R, *The Meaning of It All: Thoughts of a Citizen Scientist*, Addison-Wesley, Reading, 1998.
29. James W, *Pragmatism and Other Essays*, Washington Square Press, New York, 1963.
30. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AAM *et al.*, A gene expression signature as a predictor of survival in breast cancer, *New Engl J Med* **357**:1999–2009, 2002.
31. Datta A, Choudhary A, Bittner ML, Dougherty ER, External control in Markovian genetic regulatory networks, *Mach Learn* **52**:169–191, 2003.
32. Schroedinger E, *Science Theory and Man*, Dover, New York, 1957.
33. Pierce C, *The Collected Papers of Charles Sanders Pierce*, Harvard University Press, Cambridge, 1937.
34. Matheron G, *Random Sets and Integral Geometry*, John Wiley, New York, 1975.
35. Kant I, *Critique of Pure Reason*, Cambridge University Press, Cambridge, 1998 (original German publication, 1787).

36. Davidson E, Rast JP, Oliveri P, Ransick A, Calestani C *et al.*, A genomic regulatory network for development, *Science* **295**:1669–1678, 2002.
37. Kauffman SA, Metabolic stability and epigenesis in randomly constructed genetic nets, *Theor Biol* **22**:437–467, 1969.
38. Pugachev VS, *Theory of Random Functions and Its Application to Control Problems*, Pergamon Press, New York, 1962.
39. Goutsias J, Kim S, A nonlinear discrete dynamical model for transcriptional regulation: construction and properties, *Biophys J* **86**:1922–1945, 2004.
40. DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS *et al.*, Use of a cDNA microarray to analyse gene expression patterns in human cancer, *Nat Genet* **14**: 457–460, 1996.
41. Schena M, Shalon D, Davis RW, Brown PO, Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science* **270**:467–470, 1995.
42. Dougherty ER, Bittner M, Chen Y, Kim S, Sivakumar K *et al.*, *Proceedings of the IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing*, 1999.
43. Pal R, Datta A, Fornace AJ, Bittner ML, Dougherty ER, Boolean relationships among genes responsive to ionizing radiation in the NCI 60 ACDS. *Bioinformatics* **21**: 1542–1549, 2005.
44. Kauffman SA, *The Origins of Order*, Oxford University Press, New York, 1993.
45. Poincaré H, The Foundations of Science, The Science Press, Lancaster, 1946 (represents *Science and Hypothesis*, original French publication, 1901, and *The Value of Science*, original French publication, 1905).
46. Kim S, Dougherty ER, Shmulevich I, Hess KR, Hamilton SR *et al.*, Identification of combination gene sets for glioma classification, *Mol Cancer Ther* **1**:1229–1236, 2002.
47. Einstein A, *Herbert Spencer Lecture*, Oxford University Press, New York, 1933.
48. Devroye L, Gyorfi L, Lugosi G, *A Probabilistic Theory of Pattern Recognition*, Springer-Verlag, New York, 1996.
49. Braga-Neto U, Dougherty ER, Exact performance measures and distributions of error estimators for discrete classifiers, *Pattern Recognit* **38**:1799–1814, 2005.
50. Vapnik VN, Chervonenkis A, On the uniform convergence of relative frequencies of events to their probabilities, *Theory Probab Appl* **16**:264–280, 1971.
51. Attoor SN, Dougherty ER, Classifier performance as a function of distributional complexity, *Pattern Recognit* **37**:1629–1640, 2004.
52. Montgomery DC, *Design and Analysis of Experiments*, John Wiley, New York, 1976.
53. Jain AK, Murty MN, Flynn PJ, Data clustering: a review, *ACM Comput Surv* **31**: 264–323, 1999.
54. Duda R, Hart PE, Stork DG, *Pattern Classification*, John Wiley, New York, 2001.
55. Dougherty ER, Brun M, A probabilistic theory of clustering, *Pattern Recognit* **37**: 917–925, 2004.