Edward R. Dougherty

Genomic Signal Processing

Signal processing has played a major auxiliary role in medicine via the array of technologies available to physicians. Only a rapidly diminishing proportion of the population can recall medicine without computer tomography, magnetic resonance imaging, and ultrasound. In this capacity, signal processing serves only a supporting function. The future will be different. Like a factory, regulatory logic defines the cell as an operational system [1]: "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." *Ipso facto*, the mathematical foundations of biology, and therefore its translational partner, medicine, reside in the mathematics of systems theory. Hence, the roles of signal processing and the closely related theories of communication, control, and information will play constitutive functions as medicine evolves into a translational science resting on a theoretical framework.

This article illustrates these basic-science roles with diagnostic and therapeutic models involving logical circuits for combinatorial drug analysis, Karnaugh maps in the construction of gene regulatory networks, Markov chain perturbation theory for determining therapeutic action, queuing theory in analyzing the effects of gene copy number alterations (CNAs) on gene expression, and the use of minimum-mean-square-error estimation in the design of biomarkers for disease. My aim is simple: attract engineers into theoretical medicine, where their expertise can improve the human condition.

Recognition of the fundamental role of systems theory for the life sciences is not a recent phenomenon. In the original 1948 edition of *Cybernetics: or Control and Communication in the Animal and Machine*, Norbert Wiener wrote, "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" [2].

Wiener's insights were in accord with the contemporaneous thinking of one of the greatest biologists of the 20th century, Conrad Waddington, who in his 1935 book, *How Animals Develop*, stated, "The processes which keep an animal alive have to be quite as highly organized as the operations in the most complicated mass-production factory. . . . 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." [3] These comments, and those of Wiener, are basic epistemological statements about the nature of biological science and medicine [1].

PROBABILISTIC BOOLEAN NETWORKS

Gene regulatory networks describe the manner in which cells execute and control functioning. They are central to systems medicine, for which a basic aim is to develop therapies based on the disruption or mitigation of aberrant gene function contributing to the pathology of a disease, mitigation being accomplished by the use of drugs to act on gene products. Owing to the role played by Boolean networks and their generalization to probabilistic Boolean networks (PBNs), we begin with a brief section on these, deferring to the literature for a general discussion [4].

A PBN consists of a set of genes with discrete expression levels, often, although not necessarily, one (expressed) and zero (not expressed). Each gene is regulated by some subset of "predictor" genes and the prediction rules depend on the "context" of the cell, which is determined by latent variables external to the network. Latent variables are inevitable because the model network contains only a limited number of genes and the cell is open to extracellular signals. One might think of a multiplexor determining the operative logic at any time point.

Formally, a PBN consists of a set $V = \{x_1, x_2, \ldots, x_n\}$ of n nodes, $x_i \in \{0, 1, \ldots, d-1\}$, and a set $\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_k\}$ of vector-valued functions. For gene regulation, x_i represents the expression value of the ith gene, but it is common to refer to x_i as the ith gene. The vector $\mathbf{x}(t) = (x_1(t), \ldots, x_n(t))$ is called a "gene activity profile" (GAP) and represents the network state at time t. The vector function $\mathrm{f}_l = (f_{l1}, \ldots, f_{ln})$ defines a constituent network, or "context," of the PBN. The function $f_{li}: \{0, \ldots, d-1\}^n \to \{0, \ldots, d-1\}$ is the predictor of gene i in context l. At each time point, there is a probability q of selecting a new context. If the context remains unchanged, then the values of all genes are updated synchronously according to the current context. If a switch is called for, then a context is randomly selected according to a selection-probability distribution and gene values are updated using the new context rules. A single-context PBN without perturbation is a classical Boolean network.

Digital Object Identifier 10.1109/MSP.2012.2185868
Date of publication: 9 April 2012

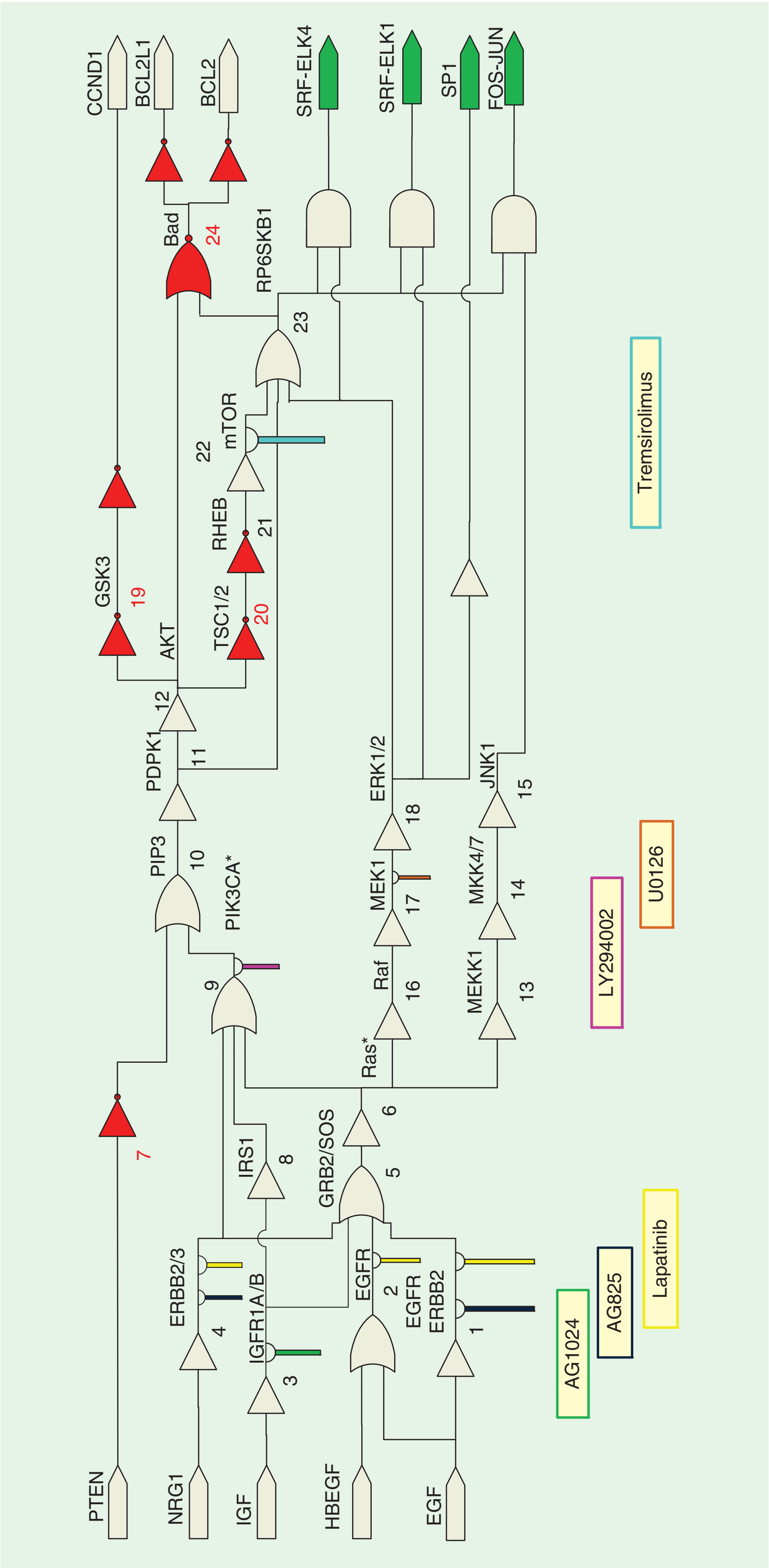
1053-5888/12/$31.00©2012IEEE

The dynamic behavior of a PBN is modeled by a Markov chain whose states are ordered pairs consisting of a context and a GAP For a PBN with perturbation, each gene may randomly change its value with small probability p at each epoch. Assuming $p > 0$, the Markov chain (network) possesses a steady-state distribution.

COMBINATORIAL DRUG ANALYSIS

Cell regulation involves nonlinear multivariate relations among many genes, involving extensive parallelism, redundancy, feedback, and distributed control. Cancer is typically a disease of several network faults, a fault being a structural error in the system. The accumulation of DNA mutations may cause the signaling pathways to behave erratically, leading to proliferation (unregulated cell growth) or the survival of cells that should undergo apoptosis (programmed cell death) on account of DNA damage. In the context of a Boolean network, "stuck-at" faults correspond to frozen cell logic, where a component is either stuck at zero (OFF) or stuck at one (ON). A key goal is to find drug combinations that intervene so as to alter aberrant behavior leading to cancerous phenotypes. We consider signaling pathways associated with growth factors, these being external signals directing a cell to divide. The goal is to model these signaling pathways as an input-output Boolean circuit and to use the model for 1) enumerating the different fault (or malfunction) possibilities, 2) carrying out fault classification, and 3) designing the appropriate corrective action (therapy) [5].

In the Boolean circuit shown in Figure 1, there are five inputs to the network forming the binary vector [EGF, HBEGF, IGF, NRG1, PTEN]. There are seven outputs, transcription factors (TFs) (signaling proteins), marked in green, and the activation status of some key proteins, not colored, forming the binary vector [FOS-JUN, SP1, SRF-ELK1, SRF-ELK4, BCL2, BCL2L1, CCND1]. The figure shows intervention points for six drugs and possible locations of stuck-at-1 (black) and stuck-at-0 (red) errors that can induce proliferation and stop apoptosis—for instance, Points 6 and 7 refer to Ras and PTEN being stuck ON and OFF, respectively.



[FIG1] Growth factor logic circuit with drug interventions.

Consider the input 00001, in which PTEN is activated. Absent faults, the output is 0000000, which is nonproliferative; however, with faults, the outputs will be different. The goal of drug intervention is to produce an output signal close to the nonproliferative output 0000000 and away from the most proliferative output 1111111.

Given the logic diagram of Figure 1, one can develop software that, given an input vector, computes output vectors under different fault scenarios and different drug combinations [5]. From these, therapy would be chosen based on trying to achieve maximum reduction of proliferation with a minimum of drug intervention.

INFERRING REGULATORY NETWORKS FROM GENETIC PATHWAYS

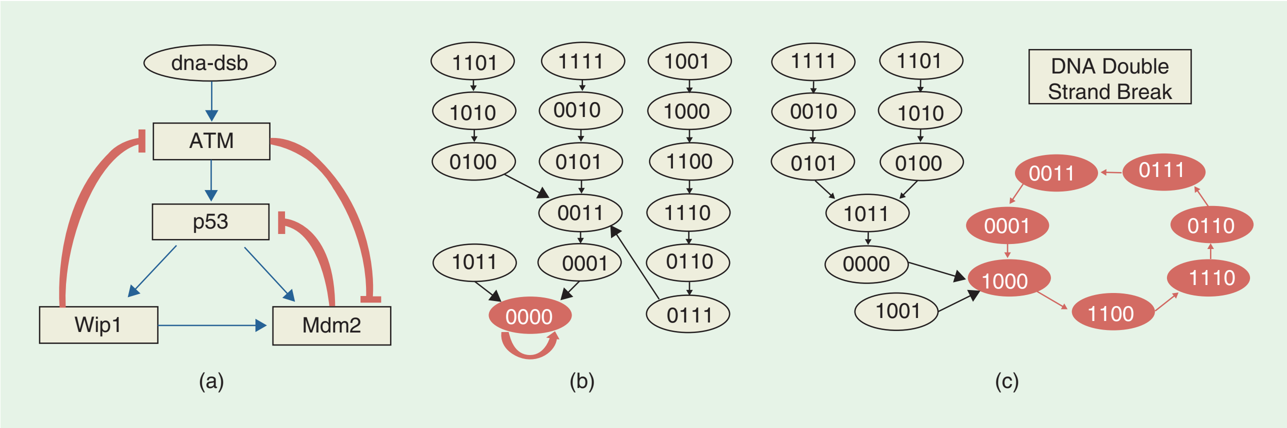
A pathway diagram (for instance, Ras → Raf → MEK1 in Figure 1) represents a portion of a gene regulatory network. It is missing the full topology (interaction) between all genes in the diagram and any multivariate regulation. Since many networks involving the same genes might manifest the same pathways, given a pathway diagram the inverse problem is to construct an uncertainty class of networks, each manifesting the given pathways. The approach taken in [6] is to list the simple pathways $A \rightarrow B$, the arrow indicating activation or deactivation, in a pathway diagram and build a set of Karnaugh maps where each map represents the next state of a gene and the constituents of a map are the predictors for the next gene value. The pathway knowledge partially fills in the Karnaugh maps. Since it is common for a pathway diagram to be built of pathways corresponding to different cell lines or different cellular contexts, there can be contradictions and these must be resolved via some protocol. At the end of the procedure, the collection of Karnaugh maps will not represent a single network, but rather an uncertainty class of networks.

Figure 2(a) shows a pathway diagram involving the gene p53, whose primary role in mammalian genomes is to function as a transcription factor for downstream genes whose expression can modulate cell cycle progression, repair damaged DNA, and induce senescence and apoptosis. The state space is composed of binary vectors [ATM, p53, Wip1, Mdm2]. At the top of the diagram is the external signal dna-dsb, the DNA damage input. An uncertainty class of Boolean networks has been constructed [6]. Figure 2(b) and (c) shows two potential state spaces, with the allowed state transitions. Parts (b) and (c) show the state-space transition diagram under normal conditions ($\text{dna}-\text{dsb} = 0$) and in the presence of DNA damage ($\text{dna}-\text{dsb} = 1$), respectively. Under normal conditions, there is a single steady state in which p53 is inactive; with DNA damage, there is a cycle in the steady state in which p53 switches between activation and deactivation. These possible steady-state scenarios are implicit in the diagram of Figure 2(a), but only become explicit in the inferred networks.

STRUCTURAL INTERVENTION IN GENE REGULATORY NETWORKS

In the context of gene regulatory networks, mainly PBNs, two basic intervention approaches have been considered: stationary control and structural intervention. Stationary control employs Markov decision processes on the network Markov chain and is generally based on flipping (or not flipping) the value of a control gene over time with the goal of beneficially altering network dynamics away from undesirable states [7], [8]. Structural intervention involves a one-time change of the network structure (wiring) to move the steady-state mass away from states considered to be undesirable [9].

In a normal mammalian cell cycle, 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). If gene p27 is mutated and permanently unexpressed, it is possible for both CycD and Rb to be simultaneously inactive and, consequently, for the cell to cycle in the absence of any growth factor, a cancerous scenario. The mutated Boolean functions for eight genes, Rb, E2F, CycE, CycA, Cdc20, Cdh1, UbcH10, and CycB, are given in Table 1 [9]. 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, reflects the state of the growth



[FIG2] p53 networks: (a) pathway diagram involving p53, (b) state transition diagram under normal conditions, and (c) state transition diagram in the presence of DNA damage.

[TABLE 1] LOGICAL REGULATORY FUNCTIONS FOR MUTATED BOOLEAN CELL CYCLE NETWORK.

ORDER	GENE	REGULATING FUNCTION
X1	$CycD$	EXTRA-CELLULAR SIGNALS
X2	Rb	$\overline{CycD} \wedge \overline{CycE} \wedge \overline{CycA} \wedge \overline{CycB}$
X3	$E2F$	$\overline{Rb} \wedge \overline{CycA} \wedge \overline{CycB}$
X4	$CycE$	$E2F \wedge \overline{Rb}$
X5	$CycA$	$(E2F \vee CycA) \wedge (\overline{Rb} \wedge \overline{Cdc20} \wedge \overline{(Cdh1 \wedge UbcH10)})$
X6	$Cdc20$	$CycB$
X7	$Cdh1$	$(\overline{CycA} \wedge \overline{CycB}) \vee Cdc20$
X8	$UbcH10$	$\overline{Cdh1} \vee (Cdh1 \wedge UbcH10 \wedge (Cdc20 \vee CycA \vee CycB))$
X9	$CycB$	$\overline{Cdc20} \wedge \overline{Cdh1}$

factor, and is not part of the network. Depending on the expression status of CycD (CycD = 0 or CycD = 1), one of two context Boolean networks is obtained. The PBN model is completed by defining a probability for switching contexts and a small probability that a gene may randomly flip its value.

In the cancerous scenario, the goal is to avoid states with CycD = Rb = 0. We consider structural intervention involving a 1-b perturbation of the gene logic to reduce the long-run probability of being in a state with CycD = Rb = 0; indeed, our aim is to find the optimal perturbation, the one minimizing the total steady-state probability mass in such states. The method employs Markov chain perturbation theory to express the steady-state distribution of the perturbed network in terms of the steady-state distribution and fundamental matrix of the original network.

Table 2 shows the undesirable steady-state mass for the most beneficial logic perturbation for each gene. Flipping the output expression in the truth table for the second Boolean network leads to the minimum undesirable steady-state mass. Hence, in practice, we choose to intervene, if possible, in the function regulating Rb to shift the steady-state mass away from undesirable states.

EFFECT OF GENE COPY NUMBER ON GENE EXPRESSION

CNAs, abnormal numbers of gene copies, are major causes of genetic diseases. Studying CNAs and their effects on gene expression is important for understanding the pathogenesis of cancer. Gene regulation involves TFs being assembled into multiprotein complexes for specific regulatory functions. To characterize the effect of CNAs on transcription, one should consider gene expression as not only a function of DNA copy number but also of TF quantities. Since transcription can be viewed as a mechanism constituted by a series of stochastic processes, queuing theory can be applied to model temporal TF binding activities and thereby describe the effect of copy number changes on gene expression.

The transcriptional model for Endo16, a gene encoding a secreted protein of a sea urchin's embryonic and larval midgut, provides a good system for studying the relationship between copy number and expression because a computational model for the logical operations at TF binding sites is well established. Figure 3 shows three binding sites, CG2, CG3, and CG4, for a TF called CG, in module A, which functions like an AND gate connecting to the switch of an amplifier. When all binding sites are bound by CG proteins, the switch turns on; otherwise, it is off and the amplifier stops working. A similar computational function is found in module B: binding sites for CY and CB1, two different TFs that operate like an AND gate controlling the output of another regulatory signal.

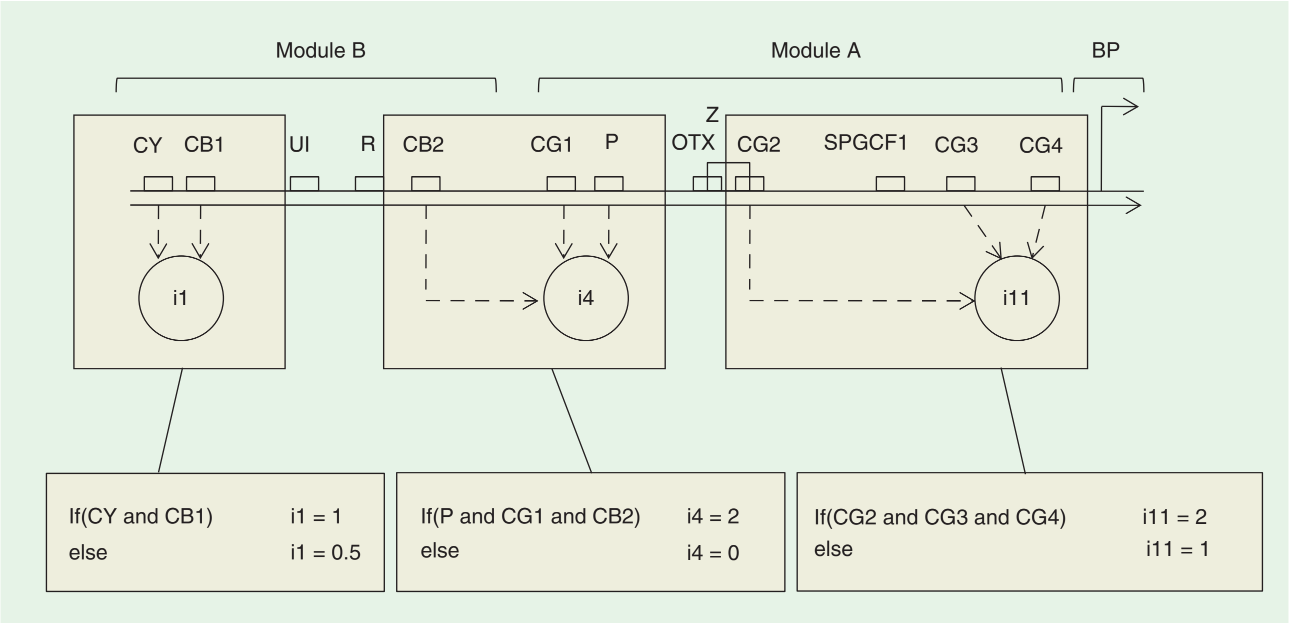
Based on the computational model, two simplified regulatory apparatuses are applied to evaluate the effect of CNAs: 1) single TF, in which all binding sites connect to a single AND logic gate; 2) general cases with multiple TFs, in which multiple AND gates modeled in 1) connect to another AND gate controlling the transcription switch [10]. Using queuing theory, gene expression values under different DNA copy numbers are derived based on various TF arrival/departure assumptions. A key conclusion is that the relationship between copy number and gene expression is generally nonlinear, but approaching linearity when the ratio of TF arrival rate to TF departure rate is large. This would explain low correlations between copy number and expression found in the literature.

MINIMUM-MEAN-SQUARE-ERROR BIOMARKER ERROR ESTIMATION

A salient hope for the new high-throughput genomic and proteomic technologies is the construction of gene/protein biomarkers for diagnosis and prognosis, in particular, for early detection and providing molecular-based decisions for therapeutic alternatives. Although a vast amount of data is gathered, we are actually faced with a paucity of measurements. It is commonplace in phenotypic classification studies for there to be tens of thousands of features, say, gene expressions, and to have sample sizes less than 100, and often less than 50. This is a striking reversal of the situation that makes for good classifier design and error estimation: a small number of features and a large sample.

[TABLE 2] UNDESIRABLE STEADY-STATE MASS FOR THE MOST BENEFICIAL LOGIC PERTURBATION FOR EACH GENE.

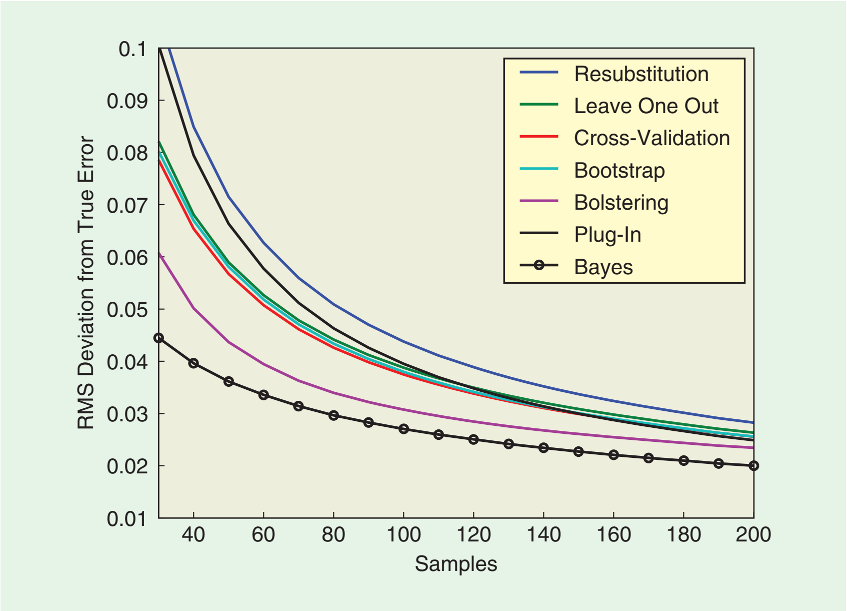
GENE	Rb	E2F	CycE	CycA	Cdc20	Cdh1	UbcH10	CycB
BN1	0.1901	0.2903	0.2534	0.2484	0.2071	0.2576	0.2587	0.2532
BN2	0.0413	0.2198	0.2529	0.2543	0.2568	0.2576	0.2587	0.2550



[FIG3] **Logic of the Endo 16 cis-regulatory system.**

With regard to error estimation, the small sample sizes mean that the data cannot be split and error estimation must take place on the same data as used for training, with cross-validation error estimation being ubiquitous. While it is true that cross-validation permits error estimation using the training data, cross-validation is usually unacceptably inaccurate on small samples owing to large variance, meaning that the estimate is scientifically vacuous [1].

A surprising historical fact concerning cross-validation is that, although it has been around for more than 40 years, originally in the form of leave-one-out, and has been used extensively, until several years ago there were no analytic studies on its relationship with the true error. That relationship is completely characterized by the joint distribution of the true and estimated errors and characterized to a lesser extent by their mixed second-order moment, which is required for the mean-square error (MSE) between them. Only in 2006 was the joint distribution found, via complete enumeration, for the true and leave-one-out estimated errors for the multinomial distribution. Exact representation for the mixed second-order moment was discovered in 2010. For leave-one-out in the Gaussian model, the exact joint distribution was found in 2010 for the univariate model [11]. A double-asymptotic (dimension and sample size increase at a fixed proportional rate) representation of the second-order moments for the multivariate model was discovered in 2011 [12]. Taking together the simulation and theoretical studies, it is seen that cross-validation can be reasonably accurate for small samples if the optimal (Bayes) error for the classification is very small; that is, if one makes appropriate modeling assumptions.



[FIG4] **RMS deviation for linear classification in a Gaussian model.**

If one is going to make modeling assumptions on the feature-label distribution, then why not take a classical engineering approach and find the minimum-mean-square-error (MMSE) error estimate? To wit, assume that the true feature-label distribution belongs to an uncertainty class, $\mathcal{V}$, of distributions and there is a prior distribution governing the parameters of the distributions in $\mathcal{V}$ [13]. For instance, in the Gaussian model, there are class-conditional distributions for the two classes to be discriminated, each parameterized by its mean vector and covariance matrix, and these parameters are assumed to satisfy some prior distribution. According to Bayesian theory, the optimal MMSE error estimate is given by the expected value of the true error relative to the posterior distribution arising from the sample data. Analytic representation of this estimate has been found for multinomial discrimination [13] and for the Gaussian model with linear classification [14]. This estimate possesses best average performance, and is unbiased, over all distributions in $\mathcal{V}$ and all samples (of a given size). Figure 4 compares the average root mean square (RMS) (square root of MSE) over all samples and distributions as a function of sample size for the MMSE estimate with some distribution-free estimates (resubstitution, leave-one-out, five-fold cross-validation, .632 bootstrap, bolstering, and plug-in) in a certain Gaussian model [14]. Once again, a signal processing methodology provides an optimal solution to a salient medical problem.

CONCLUSION

To the extent that medicine concerns interventions in biological systems and, concomitantly, decisions regarding optimal interventions, its theoretical knowledge must be constituted in terms of mathematical models formalizing human knowledge regarding the systems, operations on those models characterizing physical interventions, and objective criteria corresponding to the benefits of intervention. Given the primary roles of communication and control as distinguishing features of a biological system, as opposed to merely a collection of complex molecules, the mathematical foundations of medicine naturally fall within signal processing and systems theory. This article has presented examples of how various biomedical problems fit into classical engineering paradigms so that their solutions are obtained within standard mathematical frameworks.

AUTHOR

Edward R. Dougherty (edward@ece.tamu.edu) is a professor of electrical and computer engineering at Texas A&M University in College Station. He is the director of the Computational Biology Division of the Translational Genomics Research Institute in Phoenix, Arizona, and is an adjunct professor in the Department of Bioinformatics and Computational Biology at the University of Texas M.D. Anderson Cancer Center in Houston.

REFERENCES

[1] E. R. Dougherty and M. L. Bittner, *Epistemology of the Cell: A Systems Perspective on Biological Knowledge* (IEEE Press Series in Biomedical Engineering). Hoboken, NJ: Wiley, 2011.

[2] N. Wiener, *Cybernetics: Or Control and Communication in the Animal and Machine*. Cambridge, MA: MIT Press, 1948.

[3] C. H. Waddington, *How Animals Develop*. London: Allen & Unwin, 1935.

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

[5] R. Layek, A. Datta, M. L. Bittner, and E. R. Dougherty, "Cancer therapy design based on pathway logic," *Bioinformatics*, vol. 27, no. 4, pp. 548–555, 2011.

[6] R. Layek, A. Datta, and E. R. Dougherty, "From biological pathways to networks," *Mol. BioSyst.*, vol. 7, no. 3, pp. 843–851, 2011.

[7] R. Pal, A. Datta, and E. R. Dougherty, "Optimal infinite horizon control for probabilistic Boolean networks—Part 2," *IEEE Trans. Signal Processing*, vol. 54, no. 6, pp. 2375–2387, 2006.

[8] B. Faryabi, J.-F. Chamberland, G. Vahedi, A. Datta, and E. R. Dougherty, "Optimal intervention in asynchronous genetic regulatory networks," *IEEE J. Select. Topics Signal Processing*, vol. 2, no. 3, pp. 412–423, 2008.

[9] X. Qian and E. R. Dougherty, "Effect of function perturbation on the steady-state distribution of genetic regulatory networks: Optimal structural intervention—Part 1," *IEEE Trans. Signal Processing*, vol. 56, no. 10, pp. 4966–4975, 2008.

[10] F.-H. Hsu, E. Serpedin, Y. Chen, and E. R. Dougherty, "Stochastic modeling of the relationship between copy number and gene expression based on transcriptional logic," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 1, pp. 272–280, 2012.

[11] A. Zollanvari, U. M. Braga-Neto, 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 Trans. Inform. Theory*, vol. 56, no. 2, pp. 784–804, 2010.

[12] A. Zollanvari, U. M. Braga-Neto, and E. R. Dougherty, "Analytic study of performance of error estimators for linear discriminant analysis," *IEEE Trans. Signal Processing*, vol. 59, no. 9, pp. 4238–4255, 2011.

[13] L. Dalton and E. R. Dougherty, "Bayesian minimum mean-square error estimation for classification error—Part I: Definition and the Bayesian MMSE error estimator for discrete classification," *IEEE Trans. Signal Processing*, vol. 59, no. 1, pp. 115–129, 2011.

[14] L. Dalton and E. R. Dougherty, "Bayesian minimum mean-square error estimation for classification error—Part II: Linear classification of Gaussian models," *IEEE Trans. Signal Processing*, vol. 59, no. 1, pp. 130–144, 2011.

[SP]