Edward R. Dougherty
and Aniruddha Datta

Genomic Signal Processing: Diagnosis and Therapy

Genomics entails the study of large sets of genes with the goal of understanding collective gene function, rather than just that of individual genes. Genomic signal processing (GSP) is the engineering discipline that studies the processing of genomic signals. Since regulatory decisions within the cell utilize numerous inputs, analytical tools are necessary to model the multivariate influences on decision-making produced by complex genetic networks. Genomic signals must be processed to characterize their regulatory effects and their relationship to changes at both the genotypic and phenotypic levels. The aim of GSP is to integrate the theory and methods of signal processing with the global understanding of genomics, placing special emphasis on genomic regulation. GSP encompasses various methodologies related to signal profiles: detection, prediction, classification, control, and statistical and dynamical modeling of gene networks. In this article, we give an overview of GSP and describe how pattern recognition and network analysis are central to diagnosis and therapy for genetic diseases.

GENOMICS

Multicellular organisms, such as ourselves, are made up of approximately 100 trillion cells. A cell is the basic unit of life and, although each cell by itself is a completely functioning unit, in a multicellular organism the cells must coexist in harmony by obeying certain social controls. Cells replicate themselves by cell division, and irreparably damaged cells remove themselves by a process called *apoptosis*, which is essentially programmed suicide.

Each cell contains instructions necessary for its proper functioning. These instructions are written in the form of deoxyribonucleic acid (DNA) and must be replicated and handed down unchanged to the cell's progeny when it divides. Coded in the DNA are instructions that direct the cell to divide, undergo apoptosis, or perform a variety of other functions. Although all these instructions are present in almost every cell of an organism, they are not active at all times. The overall behavior of the cell arises from the manner in which the instructions are called. Control depends on complex interactions between the products of the cell and those of its environment. As might be expected from a highly complex system that must be both efficient and survivable, control is highly distributed and redundant.

When a particular instruction becomes active (in response to some internal or external stimulus), the corresponding gene (or stretch of DNA) is said to turn on or be expressed. What really happens is that the DNA strand (which serves as a master copy) is copied using a less stable molecule called ribonucleic acid or RNA, and one or more copies are made. The RNA strands are called messenger RNAs (mRNAs). The process by which they are produced is called *transcription*. The mRNAs are subsequently interpreted in accordance with a universal genetic code to produce the appropriate proteins, which are the molecules ultimately responsible for all cellular functions. The process of producing proteins from mRNAs is referred to as *translation*. This conduit of information flow from DNA to RNA to protein was one of the early central insights of molecular biology and is known as the *central dogma* of molecular biology.

GSP studies the many questions regarding cellular control mechanisms raised by the growing understanding of how information stored in DNA is converted into molecular machines with various capabilities. Such machines include those required to carry out the copying of DNA and the transformation of its code into RNA and protein. Regulation of transcription requires that transcription factors, which are proteins that recognize specific sequences on the DNA, bind to the DNA and seed the formation of protein complexes that constitute a recognition site. This is the site to which the complex of proteins that forms an RNA polymerase can bind and initiate copying from the DNA strand that serves as a template for the RNA. By means of such interactions among the proteins present in the cell and the interactions of these complexes with the DNA, intricate but reliable logical relations are produced. These then maintain highly varied patterns of gene expression among the differing cell types present in an organism.

The key point is that cellular control, and its failure in disease, results from multivariate activity among cohorts of genes and their products. Since all three levels in the central dogma—DNA, RNA, and protein—interact, it is not possible to fully separate them. Ultimately, information from all realms must be combined for full understanding; nevertheless, the high degree of interactivity between levels insures that a significant amount of the system information is available in each of the levels, so that focused studies provide useful insights. Efforts are currently focused at the RNA level owing to measurement considerations.

A central aspect of RNA-based genomic analysis is measurement of the

1053-5888/05/$20.00©2005IEEE

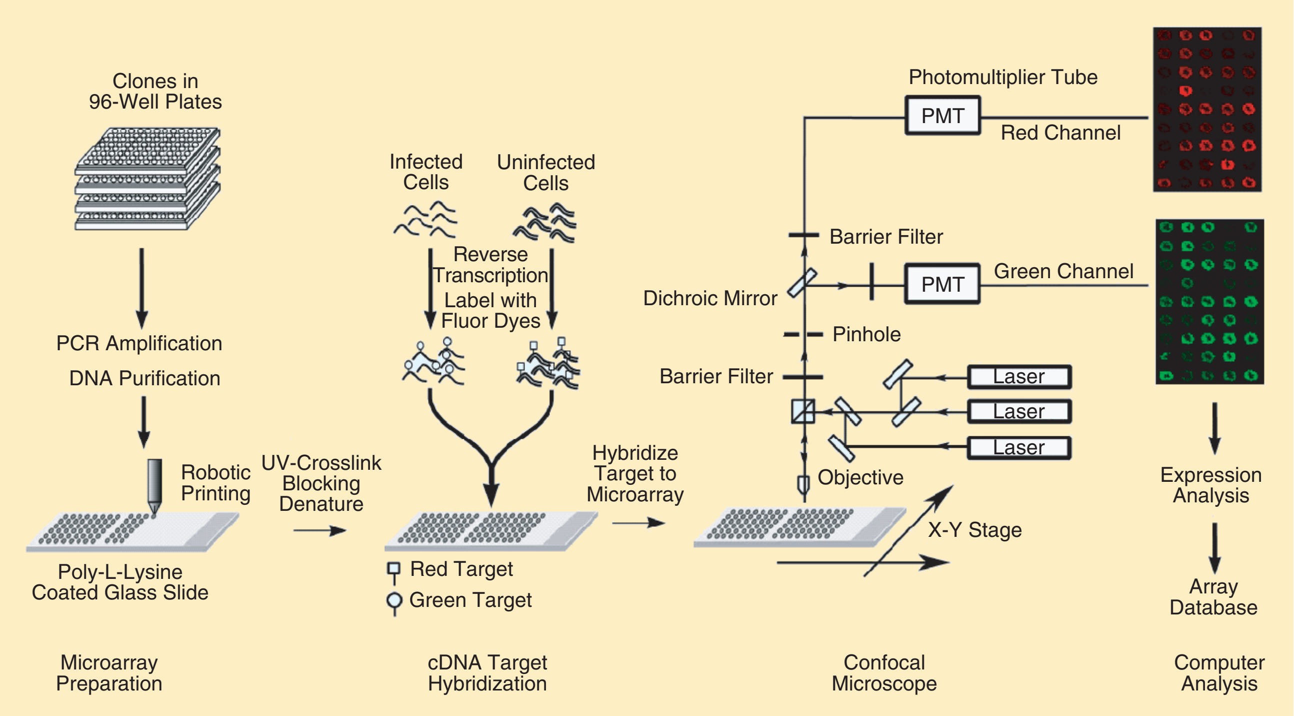
transcriptome, the collection of mRNAs in a cell at a given moment. Recently developed, high-throughput technologies make it possible to simultaneously measure the RNA abundances of thousands of mRNAs. In particular, expression microarrays result from a complex biochemical-optical system incorporating robotic spotting and computer image formation [1]. These arrays are grids of thousands of different single-stranded DNA molecules attached to a surface to serve as probes. Two major kinds include those using synthesized oligonucleotides and those using spotted cDNAs (complementary-DNA molecules). The basic procedure is to: 1) extract RNA from cells, 2) convert the RNA to single-stranded cDNA, 3) attach fluorescent labels to the different cDNAs, 4) allow the single-stranded cDNAs to hybridize to their complementary probes on the microarray, and then 5) detect the resulting fluor-tagged hybrids via excitation of the attached fluors and image formation using a scanning confocal microscope. Relative RNA abundance is measured via measurement of signal intensity from the attached fluors. This intensity is obtained by image processing and statistical analysis, with particular attention often paid to the detection of high- or low-expressing genes [2]. Figure 1 provides a schematic representation of the preparation, hybridization, image acquisition, and analysis for cDNA microarrays.

Two major goals of functional genomics are: 1) to use genomic signals to classify disease on a molecular level and 2) to screen for genes that determine specific cellular phenotypes (disease) and model their activity in such a way that normal and abnormal behavior can be differentiated. These goals correspond to diagnosing the presence or type of disease and to developing therapies based on the disruption or mitigation of the aberrant gene function contributing to the pathology of a disease. Mitigation would be accomplished by the use of drugs to act on the gene products. Creating diagnostic tools for use at the RNA level involves designing expression-based classifiers based on genes whose product abundances indicate key differences in cell state, such as one type of cancer or another. Creating therapeutic tools involves synthesizing nonlinear dynamical networks, analyzing these networks to characterize gene regulation, and developing intervention strategies to modify dynamical behavior. In this article, we briefly explain and give examples of the classification/diagnosis and network/therapy paradigms.

Considerable effort in GSP has been directed towards gaining an understanding of cancer and developing potential therapeutic approaches for treating it. Thus, it is appropriate at the outset to give a broad description of the disease with the intention of introducing some basic terminology. Essentially, cancer results when a cell divides uncontrollably and fails to undergo apoptosis. This can happen if damage to a cell's DNA perpetually turns on the instructions to divide or permanently switches off the instructions to undergo apoptosis. If that happens, then the cell and its progeny can experience uncontrolled growth, initially forming localized tumors. Further DNA mutation or rearrangement can provide the cell the ability to invade surrounding tissue, as well as the blood or lymph system. When malignant cells spread to distant organs, the cancer is said to have metastasized.

CLASSIFICATION FOR DIAGNOSIS

Classification for diagnosis involves designing a classifier that takes a vector of gene



[FIG1] The cDNA microarray technology. Schematic diagram showing slide preparation, hybridization, image acquisition, and analysis.

expression levels as input and outputs a class label, or decision. For cancer diagnosis, classification can be between different kinds of cancer, different stages of tumor development, or other such differences. Expression-based classification has been applied to many types of cancer, including leukemia, breast cancer, colon cancer, melanoma, and glioma [3]. Classifier design involves measuring expression levels from RNA obtained from the different tissues, determining genes whose expression levels can be used as features, applying a classification rule to construct the classifier, and applying an estimation rule to estimate the classifier error.

Critical issues arise due to the prevalence of small samples in microarray experiments: 1) limited classifier complexity, including a simple functional structure and a small number of features, to avoid overfitting the sample data; 2) error estimation using the training data from small samples; and 3) choosing a small set of genes as features from among thousands of genes on a microarray. These are difficult issues, and they are provoking a substantial amount of statistical and engineering research. In addition to the statistical reasons, small feature sets are advantageous from the diagnostic and therapeutic perspective, since sufficient information must be vested in gene sets small enough to serve as either convenient diagnostic panels or as candidates for the very expensive and time-consuming analysis required to determine if they could serve as useful targets for therapy.

We demonstrate classification for expression-based diagnosis by considering a glioma study performed at the University of Texas M.D. Anderson Cancer Center. This study uses expression data from microarrays for 597 genes to identify gene combinations for use as glioma classifiers [4]. Gliomas are the most common malignant primary brain tumors. These tumors are derived from neuroepithelial cells and can be divided into two principal lineages: astrocytomas and oligodendrogliomas. Using a group of 25 patients, gene combinations have been identified for distinguishing four types of glioma: oligodendroglioma (OL), anaplastic oligodendroglioma (AO), anaplastic astrocytoma (AA), and glioblastoma multiforme (GM). Linear classifiers, which have low complexity, have been derived using a form of analytic noise injection that serves as a regularization technique to improve classifier design for small samples. Error estimation has been done by bolstered resubstitution, a method that provides better results than cross validation on small samples and that is suitable for testing thousands of classifiers thanks to its speed of implementation. Thus, the first two small-sample issues mentioned previously have been addressed. For feature selection, the existence of only 597 genes on the microarray and the use of computationally efficient classifier design and error estimation have permitted the use of a supercomputer to test all possible feature sets. Figure 2 shows examples of hyperplanes for three-gene discriminators found by the method that yield low estimated errors for: (a) OL from others, (b) GM from others, (c) AO from others, and (d) AA from others. The axes give the names of the genes composing the classifiers.

Due to the difficulty of designing classifiers on small samples and the high variation of error estimators based on small training samples, it is imprudent to take a single classifier designed on a single set of microarray experiments as the final product of classification. We quote from a previous paper:

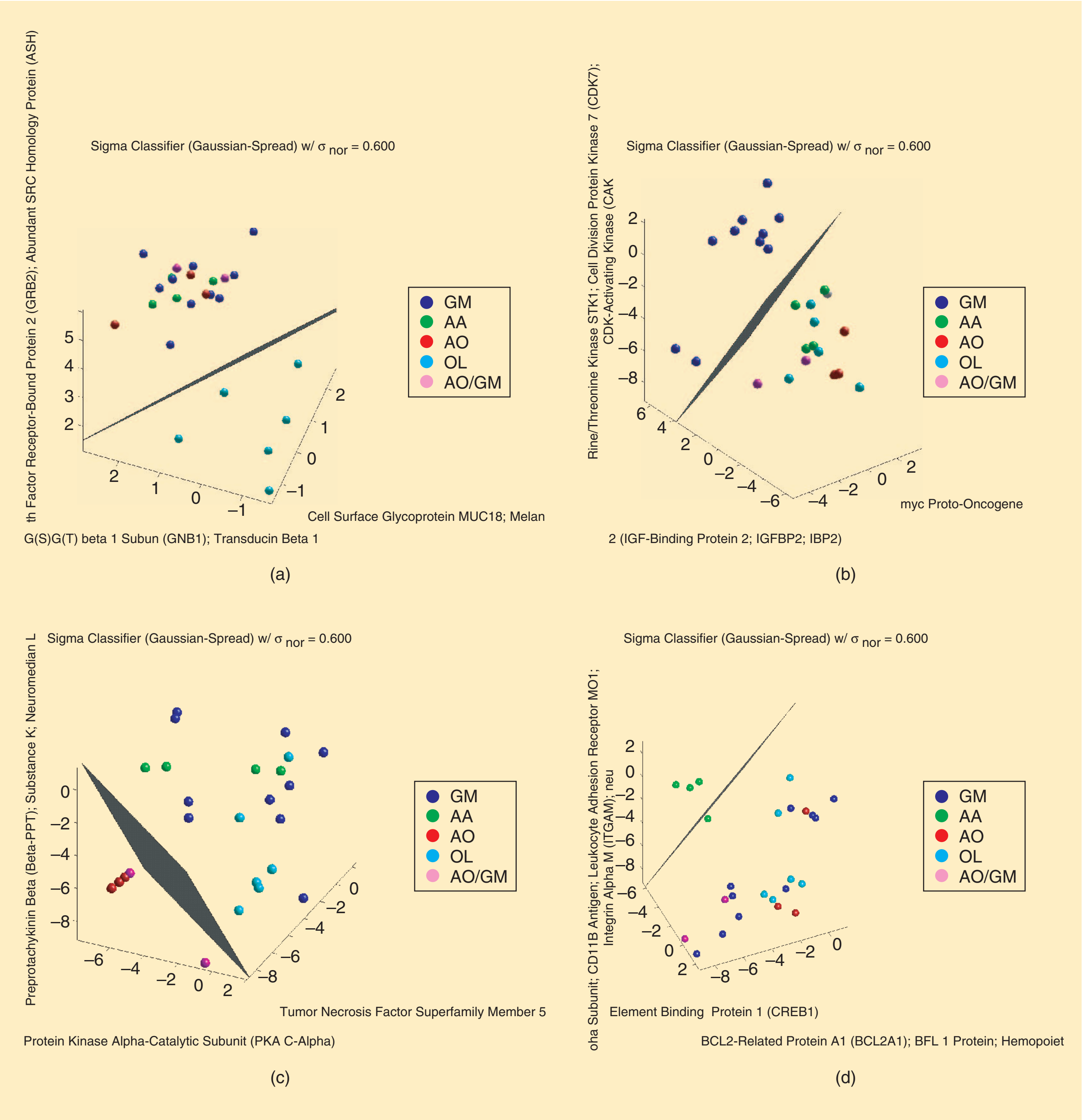
> Separation of the sample data by designed classifiers will likely have to be taken as evidence that the corresponding gene sets are potential variable sets for classification. Their effectiveness will have to be checked by large-replicate experiments designed to estimate their classification error, perhaps in conjunction with biological input or phenotype evidence. There may, in fact, be many gene sets that provide accurate classification of a given pathology. Of these, some sets may provide mechanistic insights into the molecular etiology of the disease, while other sets may be indecipherable. [5]

This is precisely the approach taken in the glioma study we have been discussing.

NETWORKS FOR THERAPY

Cellular control and its failure in disease result from multivariate activity among cohorts of genes. Thus, for therapeutic purposes, it is important to model this multivariate interaction. In the literature, two somewhat distinct approaches have been taken to carry out this modeling. The first approach is based on constructing detailed biochemical network models for particular cellular reactions of interest and makes use of ordinary differential equations, partial differential equations, and their variants [6]. While this method yields insights into the details of individual reaction pathways, it is not clear how the information obtained can be used to design a therapeutic regimen for a complex disease like cancer, which simultaneously involves many genes and many signalling pathways. A major problem for fine-scale modeling is its large data requirement. Consequently, here we focus on the second approach, which is geared towards building coarse models of genetic interaction using the limited amount of microarray gene expression data that is usually available. Paradigms that have been considered in this context include directed graphs, Bayesian networks, Boolean networks, generalized logical networks, and most recently, probabilistic Boolean networks. Here we will explicitly discuss Boolean and probabilistic Boolean networks since the therapy aspect is currently most developed within these two frameworks. The reader is referred to [6] and the references therein for in-depth discussions of other paradigms.

A Boolean network is defined by a set of nodes, $V = \{x_1, x_2, \ldots, x_n\}$, and a list of Boolean functions, $F = \{f_1, f_2, \ldots, f_n\}$. Each x_k represents the state (expression) of a gene, g_k, where $x_k = 1$ or $x_k = 0$, depending on whether the gene is expressed or not expressed. The Boolean functions represent the rules of regulatory interaction between genes. Network dynamics result from a synchronous clock with times $t = 0, 1, 2, \ldots$, and the value of gene g_k at time $t + 1$ is determined by $x_k(t+1) = f_k(x_{k1}, x_{k2}, \ldots, x_{k,m(k)})$, where the nodes in the argument of f_k form the regulatory set for x_k (gene g_k). The numbers of genes in the regulatory



[FIG2] Glioma classification: hyperplanes for three-gene discriminators: (a) OL from others; (b) GM from others; (c) AO from others; and (d) AA from others.

sets define the connectivity of the network, with maximum connectivity typically no more than three. At time point t, the state vector $\mathrm{x}(t) = (x_1(t), x_2(t), \ldots, x_n(t))$ is called the *gene activity profile* (GAP). The functions together with the regulatory sets determine the network wiring. A Boolean network is a very coarse model; nonetheless, it facilitates understanding of the generic properties of global network dynamics [7], [8], and its simplicity mitigates data requirements for inference.

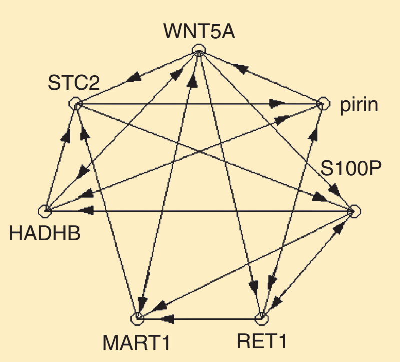
Microarray technology yields simultaneous measurements of expression status for thousands of genes and can be utilized for network inference. By viewing gene status across different conditions, it is possible to establish relationships between genes that show variable status across the conditions. Owing to limited replications, we assume that gene expression data is quantized using the methods in [2]. One way to establish multivariate relationships among genes is to quantify how the estimate for the expression status of a particular *target gene* can be improved by knowledge of the status of some other *predictor genes*. This is formalized via the *coefficient of determination* (CoD) [9], which is essentially a nonlinear, multivariate generalization of the familiar goodness of fit measure in linear regression. For our purposes, it is sufficient to note that the CoD measures the degree to which the best estimate for the transcriptional activity of a target gene can be improved using the knowledge of the transcriptional activity of some predictor genes, relative to the best estimate in the absence of any knowledge of the transcriptional activity of the predictors. The CoD is a number between zero and one, a higher value indicating a tighter relationship. Given a target gene, several predictor sets may provide equally good estimates of its transcriptional activity, as measured by the CoD. Moreover, one may rank several predictor sets via their CoDs. Such a ranking provides a quantitative measure to determine the relative ability of each predictor set to improve the estimate of the transcriptional activity of the particular target gene. While attempting to infer inter-gene relationships, it makes sense to not put all our faith in one predictor set; instead, for a particular target gene, a better approach is to consider a number of predictor sets with high CoDs. Considering each retained predictor set to be indicative of the transcriptional activity of the target gene with a probability proportional to its CoD represents feature selection for gene prediction.

Having inferred intergene relationships, this information can be used to model the evolution of the gene activity profile over time. It is unlikely that the determinism of the Boolean-network model will be concordant with the data. One could pick the predictor set with the highest CoD, but as noted previously, there are usually a number of almost equally performing predictor sets, and the CoDs we have for them are only estimates from the data. By associating several predictor sets with each target gene, it is not possible to obtain with certainty the transcriptional status of the target gene at the next time point; however, one can compute the probability that the target gene will be transcriptionally active at time $t+1$ based on the gene activity profile at time t. The time evolution of the gene activity profile then defines a stochastic dynamical system. Since the gene activity profile at a particular time point depends only on the profile at the immediately preceding time point, the dynamical system is Markovian. Such systems can be studied in the established framework of Markov chains and Markov decision processes. These ideas are mathematically formalized in probabilistic Boolean networks (PBNs) [10]. In a PBN, the transcriptional activity of each gene at a given time point is a Boolean function of the transcriptional activity of the elements of its predictor sets at the previous time point. The choice of a Boolean function and associated predictor set can vary randomly from one time point to another in accordance with the CoD-based selection probabilities associated with the different predictor sets. This defines an *instantaneously random* PBN.

An alternative approach is to take the view that the data on the microarrays come from distinct sources, each representing a *context* of the cell. That is, the data derive from a family of deterministic networks and were we able to separate the samples according to context, there would in fact be CoDs with value one, indicating deterministic biochemical activity for the wiring of a particular constituent network. Under this perspective, the only reason that it is not possible to find predictor sets with CoD equal (or very close) to one is that they represent averages across the various cellular contexts with their correspondingly various wirings. This perspective leads to the view that a PBN is a collection of Boolean networks in which one constituent network governs gene activity for a random period of time before another randomly chosen constituent network takes over, possibly in response to some random event, such as an external stimulus. Since the latter is not part of the model, network switching is random. This model defines a context-sensitive PBN. The probabilistic nature of the constituent choice reflects the fact that the system is open, not closed. The *context-sensitive* model reduces to the instantaneously random model by having network switching at every time point.

Given a Boolean network, one can partition the state space into a number of attractors along with their basins of attraction. The attractors characterize the long-run behavior of the Boolean network and have been conjectured by Kauffman to be indicative of the cell type and phenotypic behavior of the cell. For instance, it is thought that apoptosis and cell differentiation correspond to some singleton attractors and their basins, while cell proliferation corresponds to a cyclic attractor along with its associated basin [8]. Changes in the Boolean functions, via mutations or rearrangements, can lead to a rewiring in which attractors appear that are associated with tumorigenesis. This is likely to lead to a cancerous phenotype unless the corresponding basins are shrunk via new rewiring, so that the cellular state is not driven to a tumorigenic phenotype, or, if already in a tumorigenic attractor, the cell is forced to a different state by flipping one or more genes. The objective of cancer therapy would be to use drugs to do one or both of the above. These ideas for Boolean networks can be generalized to PBNs by noting that the dynamic behavior of PBNs can be described by Markov Chains, so that a PBN has equivalence classes of communicating states analogous to the basins of attraction for Boolean networks. Similarly, since all the states in an equivalence class communicate, there is a steady-state distribution local to each equivalence class so that the long-run behavior within that class can be studied. Furthermore, by assuming that each gene has a small probability of undergoing a random flip, we can make the overall Markov chain ergodic, which then guarantees the existence of a global steady-state distribution [10].

One objective of PBN modeling is to use the PBN to design different approaches for affecting the evolution of the gene activity profile of the network. To date, such intervention studies have used three different approaches: 1) resetting the state of the PBN, as necessary, to a more desirable initial state and letting the network evolve from there [10], 2) changing the steady-state (long run) probability distribution of the network by minimally altering its rule-based structure [10], and 3) manipulating external (control) variables that affect the transition probabilities of the network and can, therefore, be used to desirably affect its dynamic evolution over a finite time horizon [11].



[FIG3] The Seven Gene WNT5A Network.

We briefly describe the results in [11], where an intervention study was carried out using a PBN derived from gene expression data collected in a study of metastatic melanoma. In this expression profiling study, the abundance of mRNA for the gene WNT5A was found to be highly discriminating between cells with properties typically associated with high metastatic competence versus those with low metastatic competence. These findings were validated and expanded in a second study in which experimentally increasing the levels of the Wnt5a protein secreted by a melanoma cell line via genetic engineering methods directly altered the metastatic competence of that cell as measured by the standard in vitro assays for metastasis. Furthermore, it was found that an intervention that blocked the Wnt5a protein from activating its receptor (the use of an antibody that binds Wnt5a protein) could substantially reduce Wnt5a's ability to induce a metastatic phenotype. This suggests that a reasonable control strategy would be to use an intervention that reduces the WNT5A gene's action in affecting biological regulation, since the available data suggest that disruption of this influence could reduce the chance of a melanoma metastasizing, a desirable outcome.

To this end, a seven-gene network, including the activity of the WNT5A gene, was derived from the available gene expression data. This network, along with the multivariate relationships between the genes, is shown in Figure 3. For each gene in this network, the two best two-gene predictors were used and their associated CoDs computed. This information was used to obtain the transition probabilities for the Markov chain associated with the PBN. The intervention problem was then posed as a finite horizon optimal control problem. The performance index or cost function was chosen to reflect the tradeoffs between the intervention effort and the terminal penalty associated with ending up in an undesirable (bad) state at the end of the control horizon. Since the control objective here is to reduce the activity of the WNT5A gene, the entire state space was partitioned into good and bad regions, with bad regions being characterized by WNT5A overexpression. Bad states were assigned higher terminal penalties than the good ones, and the optimization problem was solved by Dynamic Programming. Two possible interventions were considered: intervening with Wnt5a directly (through its antibody) and intervening through another gene called pirin. In each case, it was found that the network with control performed better (in a probabilistic sense) than the network without control, so that the control objective was met. Furthermore, controlling WNT5A directly yielded better performance than trying to control it through pirin, which again is in agreement with intuitive expectations.

The intervention approaches 1) and 3) above do not attempt to alter the steady-state behavior of the network, while approach 2) attempts to increase the steady-state probability mass in the desirable states. However, all of these approaches are essentially first-cut solutions and will have to be improved upon. For instance, the approach in 2) uses a brute-force search algorithm, and a more systematic approach will have to be found through which one can increase the steady-state probability mass in the desirable set of states, while correspondingly decreasing the mass in the undesirable ones. Another aspect that merits further investigation is motivated by the fact that the currently available gene expression data comes from the *steady-state* phenotypic behavior and really does not capture any temporal history. Consequently, the process of inferring PBNs from the data will have to be modified, in the sense that it will have to be guided more by steady-state and limited connectivity considerations. Major research efforts in these directions are currently under way.

ACKNOWLEDGMENTS

This work was supported by the Translational Genomics Research Institute, the National Human Genome Research Institute, and the National Science Foundation under Grant ECS-0355227.

AUTHORS

Edward R. Dougherty and *Aniruddha Datta* are with the Department of Electrical Engineering and Genomic Signal Processing Laboratory, Texas A&M University, College Station.

REFERENCES

[1] M. Schena, D. Shalon, R. Davis, and P.O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science*, vol. 270, no. 5235, pp. 467–470, 1995.

[2] Y. Chen, E.R. Dougherty, and M. Bittner, "Ratio-based decisions and the quantitative analysis of cDNA microarray images," *Biomed. Optics*, vol. 2, no. 4, pp. 364–374, 1997.

[3] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[4] S. Kim, E.R. Dougherty, I. Shmulevich, K.R. Hess, S.R. Hamilton, J.M. Trent, G.N. Fuller, and W. Zhang, "Identification of combination gene sets for glioma classification," *Mol. Cancer Therapeutics*, vol. 1, no. 13, pp. 1229–1236, 2002.

[5] E.R. Dougherty, "Small sample issues for microarray-based classification," *Comp. Funct. Genomics*, vol. 2, no. 1, pp. 28–34, 2001.

[6] H. de Jong, "Modeling and simulation of genetic regulatory systems: A literature review," *Comput. Biol.*, vol. 9, no. 1, pp. 67–103, 2002.

[7] S.A. Kauffman, *The Origins of Order: Self-Organization and Selection in Evolution*. New York: Oxford Univ. Press, 1993.

[8] S. Huang, "Gene expression profiling, genetic networks, and cellular states: An integrating concept for tumorigenesis and drug discovery," *Mol. Med.*, vol. 77, no. 6, pp. 469–480, 1999.

[9] E.R. Dougherty, M. Bittner, Y. Chen, S. Kim, K. Sivakumar, J. Barrera, P. Meltzer, and J. Trent, "Nonlinear filters in genomic control," *Proc. IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing*, Antalya, Turkey, June 1999.

[10] I. Shmulevich, E.R. Dougherty, and W. Zhang, "From boolean to probabilistic boolean networks as models of genetic regulatory networks," *Proc. IEEE*, vol. 90, no. 11, pp. 1778–1792, 2002.

[11] A. Datta, A. Choudhary, M.L. Bittner, and E.R. Dougherty, "External control in markovian genetic regulatory networks," *Mach. Learn.*, vol. 52, no. 182, pp. 169–191, 2003.