EURASIP Journal on Applied Signal Processing 2004:1, 146–153


# Genomic Signal Processing: The Salient Issues

**Edward R. Dougherty**
*Department of Electrical Engineering, Texas A&M University, 3128 TAMU College Station, TX 77843-3128, USA*
*Email: e-dougherty@tamu.edu*

**Ilya Shmulevich**
*Department of Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA*
*Email: is@ieee.org*

**Michael L. Bittner**
*Molecular Diagnostics and Target Validation Division, Translational Genomics Research Institute, Tempe, AZ 85281, USA*
*Email: mbittner@tgen.org*

*Received 10 October 2003*

This paper considers key issues in the emerging field of genomic signal processing and its relationship to functional genomics. It focuses on some of the biological mechanisms driving the development of genomic signal processing, in addition to their manifestation in gene-expression-based classification and genetic network modeling. Certain problems are inherent. For instance, small-sample error estimation, variable selection, and model complexity are important issues for both phenotype classification and expression prediction used in network inference. A long-term goal is to develop intervention strategies to drive network behavior, which is briefly discussed. It is hoped that this nontechnical paper demonstrates that the field of signal processing has the potential to impact and help drive genomics research.

**Keywords and phrases:** functional genomics, gene network, genomics, genomic signal processing, microarray.

## 1. INTRODUCTION

Sequences and clones for over a million expressed sequence tagged sites (ESTs) are currently publicly available. Only a minority of these identified clusters contains genes associated with a known functionality. One way of gaining insight into a gene's role in cellular activity is to study its expression pattern in a variety of circumstances and contexts, as it responds to its environment and to the action of other genes. Recent methods facilitate large-scale surveys of gene expression in which transcript levels can be determined for thousands of genes simultaneously. In particular, expression microarrays result from a complex biochemical-optical system incorporating robotic spotting and computer image formation and analysis. Since transcription control is accomplished by a method that interprets a variety of inputs, we require analytical tools for expression profile data that can detect the types of multivariate influences on decision making produced by complex genetic networks. Put more generally, signals generated by the genome must be processed to characterize their regulatory effects and their relationship to changes at both the genotypic and phenotypic levels. Two salient goals of functional genomics are to screen for key genes and gene combinations that explain specific cellular phenotypes (e.g., disease) on a mechanistic level, and to use genomic signals to classify disease on a molecular level.

Genomic signal processing (GSP) is the engineering discipline that studies the processing of genomic signals. Owing to the major role played in genomics by transcriptional signaling and the related pathway modeling, it is only natural that the theory of signal processing should be utilized in both structural and functional understanding. The aim of GSP is to integrate the theory and methods of signal processing with the global understanding of functional genomics, with special emphasis on genomic regulation. Hence, GSP encompasses various methodologies concerning expression profiles: detection, prediction, classification, control, and statistical and dynamical modeling of gene networks. GSP is a fundamental discipline that brings to genomics the structural model-based analysis and synthesis that form the basis of mathematically rigorous engineering.

Application is generally directed towards tissue classification and the discovery of signaling pathways, both based on the expressed macromolecule phenotype of the cell. Accomplishment of these aims requires a host of signal processing approaches. These include signal representation relevant to transcription, such as wavelet decomposition and more general decompositions of stochastic time series, and system

modeling using nonlinear dynamical systems. The kind of correlation-based analysis commonly used for understanding pairwise relations between genes or cellular effects cannot capture the complex network of nonlinear information processing based upon multivariate inputs from inside and outside the genome. Regulatory models require the kind of nonlinear dynamics studied in signal processing and control, and in particular the use of stochastic dataflow networks common to distributed computer systems with stochastic inputs. This is not to say that existing model systems suffice. Genomics requires its own model systems, not simply straightforward adaptations of currently formulated models. New systems must capture the specific biological mechanisms of operation and distributed regulation at work within the genome. It is necessary to develop appropriate mathematical theory, including optimization, for the kinds of external controls required for therapeutic intervention as well as approximation theory to arrive at nonlinear dynamical models that are sufficiently complex to adequately represent genomic regulation for diagnosis and therapy while not being overly complex for the amounts of data experimentally feasible or for the computational limits of existing computer hardware.

## 2. BACKGROUND

A central focus of genomic research concerns understanding the manner in which cells execute and control the enormous number of operations required for normal function and the ways in which cellular systems fail in disease. In biological systems, decisions are reached by methods that are exceedingly parallel and extraordinarily integrated, as even a cursory examination of the wealth of controls associated with the intermediary metabolism network demonstrates. Feedback and damping are routine even for the most common activities, such as cell cycling, where it seems that most proliferative signals are also apoptosis priming signals, with the final response to these signals resulting from successful negotiation of a large number of checkpoints, which themselves involve further extensive cross checks of cellular conditions.

Traditional biochemical and genetic characterizations of genes do not facilitate rapid sifting of these possibilities to identify the genes involved in different processes or the control mechanisms employed. Of course, when methods do exist to focus genetic and biochemical characterization procedures on a smaller number of genes likely to be involved in a process, progress in finding the relevant interactions and controls can be substantial. The earliest understandings of the mechanics of cellular gene control were derived in large measure from studies of just such a case, metabolism in simple cells. In metabolism, it is possible to use biochemistry to identify stepwise modifications of the metabolic intermediates and genetic complementation tests to identify the genes responsible for catalysis of these steps, and those genes and *cis*-regulator elements involved in the control of their expression. Standard methods of characterization guided by some knowledge of the connections could thus be used to identify process components and controls. Starting from the basic outline of the process, molecular biologists and biochemists have been able to build up a very detailed view of the processes and regulatory interactions operating within the metabolic domain.

In contrast, for most cellular processes, general methods to implicate likely participants and to suggest control relationships have not emerged. The resulting inability to produce overall schemata for most cellular processes has meant that gene function is, for the largest part, determined in a piecemeal fashion. Once a gene is suspected of involvement in a particular process, research focuses on the role of that gene in a very narrow context. This typically results in the full breadth of important roles for well-known, highly characterized genes being slowly discovered. A particularly good example of this is the relatively recent appreciation that oncogenes such as Myc can stimulate apoptosis in addition to proliferation [1].

Recognition of this bottleneck has stimulated the field's appetite for methods that can provide a wider experimental perspective on how genes interact. High-throughput microarray technology, which facilitates large-scale surveys of gene expression, can now provide enormous data sets concerning transcriptional levels [2, 3, 4, 5]. As these measurements are snapshots of the types of levels of transcripts required to achieve or maintain the cell state being observed, they constitute a de facto source of information about transcript interactions involved in gene regulation.

Analysis of this data can take two routes: gene-by-gene analysis or multivariate analysis of interactions among many genes simultaneously. Correlation and other similarity measures can identify common elements of a cell's response to a particular stimulus and thus discern some groups of genes; however, correlation does not address the fundamental problem of determining the sets of genes whose actions and interactions drive the cell's decision to set the transcriptional level of a particular gene. Because transcriptional control is accomplished by a complex method that interprets a variety of inputs [1, 6, 7], the development of analytical tools that detect multivariate influences on decision-making present in complex genetic networks is essential. To carry out such an analysis, one needs appropriate analytical methodologies.

As a discipline, signal processing involves the construction of model systems. These can be composed of various mathematical structures, such as systems of differential equations, graphical networks, stochastic functional relations, and simulation models. By its nature, signal processing draws upon many related disciplines, including estimation, classification, pattern recognition, control, information, networks, computation, statistics, imaging, coding, and artificial intelligence. These in turn draw upon signal processing to the extent that their application involves processing signals.

Numerous mathematical and computational methods have been proposed for construction of formal models of genetic interactions. Many of these models have the following general characteristics:

(1) the models essentially represent *systems* in that they

(a) characterize an interacting group of components forming a whole,
(b) can be viewed as a process that results in a transformation of signals,
(c) generate outputs in response to input stimuli;

(2) the models are *dynamical* in that they
(a) capture the time-varying quality of the physical process under study,
(b) can change their own behavior over time;

(3) the models can be considered generally *nonlinear* in that the interactions within the system yield behavior more complicated than the sum of the behaviors of the agents.

The preceding characteristics are representatives of nonlinear dynamical systems. These are composed of states, input and output signals, transition operators between states, and output operators. In their most abstract form, they are very general. More mathematical structure is provided for particular application settings. For instance, in computer science they can be structured into the form of dataflow graphical networks that model asynchronous distributed computation, a model that is very close to genomic regulatory models. There have been many attempts to model gene regulatory networks including probabilistic graphical models, such as Bayesian networks [8, 9, 10, 11], neural networks [12, 13], differential equations [14], Boolean [15] and probabilistic Boolean networks [16, 17], and models including stochastic components on the molecular level [18].

As we look towards medical applications based on functional genomics, dynamical modeling is at the center. Somogyi and Greller [19] give the following areas in which dynamical modeling will play a "pivotal role":

(i) stimulus-response interactions,
(ii) prediction of new targets based on pathway context,
(iii) potential use of combinatorial therapies,
(iv) pathway responses including the understanding of reactive or compensatory behavior,
(v) stress and toxic response mechanisms,
(vi) off-target effects of therapeutic compounds,
(vii) pharmacodynamics,
(viii) characterization of disease states by dynamical behavior,
(ix) gene expression and protein expression signatures for diagnostics,
(x) design of optimized time-dependent dosing regimens.

As we consider the salient issues of GSP, it should become evident that the preceding list offers a call for a major effort on the part of the signal processing community to apply its store of knowledge to genetic science and medicine.

## 3. TECHNOLOGY

A cell relies on its protein components for a wide variety of its functions, including energy production, biosynthesis of component macromolecules, maintenance of cellular architecture, and the ability to act upon intra- and extra-cellular stimuli. Each cell in an organism contains the information necessary to produce the entire repertoire of proteins the organism can specify. Since a cell's specific functionality is largely determined by the genes it is expressing, it is logical that transcription, the first step in the process of converting the genetic information stored in an organism's genome into protein, would be highly regulated by the control network that coordinates and directs cellular activity. A primary means for regulating cellular activity is the control of protein production via the amounts of mRNA expressed by individual genes. The tools to build an understanding of genomic regulation of expression will involve the characterization of these expression levels. Microarray technology, both cDNA and oligonucleotide, provides a powerful analytic tool for genetic research. Since our concern in this paper is to articulate the salient issues for GSP, and not to delve deeply into microarray technology, we confine our brief discussion to cDNA microarrays.

Complementary DNA microarray technology combines robotic spotting of small amounts of individual, pure nucleic acid species on a glass surface, hybridization to this array with multiple fluorescently labeled nucleic acids, and detection and quantitation of the resulting fluor-tagged hybrids by a scanning confocal microscope. A basic application is quantitative analysis of fluorescence signals representing the relative abundance of mRNA from distinct tissue samples. Complementary DNA microarrays are prepared by printing thousands of cDNAs in an array format on glass microscope slides, which provide gene-specific hybridization targets. Distinct mRNA samples can be labeled with different fluors and then co-hybridized onto each arrayed gene. Ratios (or sometimes the direct intensity measurements) of gene expression levels between the samples can be used to detect meaningfully different expression levels between the samples for a given gene. Given an experimental design with multiple tissue samples, microarray data can be used to cluster genes based on expression profiles, to characterize and classify disease based on the expression levels of gene sets, and for other signal processing tasks.

A typical glass-substrate and fluorescent-based cDNA microarray detection system is based on a scanning confocal microscope, where two monochrome images are obtained from laser excitations at two different wavelengths. Monochrome images of the fluorescent intensity for each fluor are combined by placing each image in the appropriate color channel of an RGB image. In this composite image, one can visualize the differential expression of genes in the two cell types: test sample typically placed in red channel, and the reference sample in the green channel. Intense red fluorescence at a spot indicates a high level of expression of that gene in the test sample with little expression in the reference sample. Conversely, intense green fluorescence at a spot indicates relatively low expression of that gene in the test sample compared to the reference. When both test and reference samples express a gene at similar levels, the observed array spot is yellow. Assuming that specific DNA products from two samples have an equal probability of hybridizing to the specific target, the fluorescent intensity measurement

is a function of the amount of specific RNA available within each sample, provided that samples are well mixed and there is sufficiently abundant cDNA deposited at each target location.

When using cDNA microarrays, the signal must be extracted from the background. This requires image processing to extract signals arising from tagged reverse-transcribed cDNA hybridized to arrayed cDNA locations [20], and variability analysis and measurement quality assessment. The objective of the microarray image analysis is to extract probe intensities or ratios at each cDNA target location and then cross-link printed clone information so that biologists can easily interpret the outcomes and high-level analysis can be performed. A microarray image is first segmented into individual cDNA targets, either by manual interaction or by an automated algorithm. For each target, the surrounding background fluorescent intensity is estimated, along with the exact target location, fluorescent intensity, and expression ratio.

In a microarray experiment, there are many sources of variation. Some types of variation, such as differences of gene expressions, may be highly informative as they may be of biological origin. Other types of variation, however, may be undesirable and can confound subsequent analysis, leading to wrong conclusions. In particular, there are certain systematic sources of variation, usually due to specific features of the particular microarray technology, that should be corrected prior to further analysis. The process of removing such systematic variability is called normalization. There may be a number of reasons for normalizing microarray data. For example, there may be a systematic difference in quantities of starting RNA, resulting in one sample being consistently over-represented. There may also be differences in labeling or detection efficiencies between the fluorescent dyes (e.g., Cy3 or Cy5), again leading to systematic overexpression of one of the samples. Thus, in order to make meaningful biological comparisons, the measured intensities must be properly adjusted to counteract such systematic differences.

## 4. SALIENT ISSUES FOR GSP

In this section we address what we consider to be the salient issues for GSP: phenotype classification and genetic regulatory networks, which include expression prediction and network intervention and control. Other topics, including image processing, signal extraction, data normalization, quantization, compression, expression-based clustering, and signal processing methods for sequence analysis play necessary and supportive roles.

### *4.1. Classification*

An expression-based classifier provides a list of genes whose product abundance is indicative of important differences in cell state, such as healthy or diseased, or one particular type of cancer or another. Among such informative genes are those whose products play a role in the initiation, progression, or maintenance of the disease. Two central goals of molecular analysis of disease are to use such information to directly diagnose the presence or type of disease and to produce therapies based on the disruption or correction of the aberrant function of gene products whose activities are central to the pathology of a disease. Correction would be accomplished either by the use of drugs already known to act on these gene products or by developing new drugs targeting these gene products.

Achieving these goals requires designing a classifier that takes a vector of gene expression levels as input and outputs a class label that predicts the class containing the input vector. Classification can be between different kinds of cancer, different stages of tumor development, or many other such differences. Classifiers are designed from a sample of expression vectors. This requires assessing expression levels from RNA obtained from the different tissues with microarrays, determining genes whose expression levels can be used as classifier variables, and then applying some rule to design the classifier from the sample microarray data. Design, performance evaluation, and application of classifiers must take into account randomness arising from both biological and experimental variability. To rapidly move from expression data to diagnostics that can be integrated into current pathology practice or to useful therapeutics, expression patterns must carry sufficient information to separate sample types.

Classification using a variety of methods has been used to exploit the class-separating power of expression data in cancer: leukemias [21], various cancers [22], small, round, blue-cell cancers [23], hereditary breast cancer [24], colon cancer [25], breast cancer [4], melanoma [26], and glioma [27].

Three critical statistical issues arise for expression-based classification [28, 29]. First, given a set of variables, how does one design a classifier from the sample data that provides good classification over the general population? Second, how does one estimate the error of a designed classifier when data is limited? Third, given a large set of potential variables, such as the large number of expression level determinations provided by microarrays, how does one select a set of variables as the input vector to the classifier? The problem of small-sample error estimation impacts variable selection in a devilish way. An error estimator may be unbiased but have a large variance, and therefore often be low. This can produce a large number of gene (variable) sets and classifiers with low error estimates. For a small sample, one can end up with thousands of gene sets for which the error estimate from the data at hand is zero. In the other direction, a small sample size enhances the possibility that a designed classifier will perform worse than the optimal classifier. Combined with a high error estimate, the result will be that many potentially good diagnostic gene sets will be pessimistically evaluated.

Not only is it important to base classifiers on small numbers of genes from a statistical perspective, but there are also compelling biological reasons for small classifier sets. As previously noted, correction of an aberrant function would be accomplished by the use of drugs. 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. Small gene sets are necessary to allow construction of a practical immunohistochemical diagnostic panel. In sum, it is important to develop classification algorithms specifically tailored for small samples [27].

While clustering algorithms do not produce the specificity and quantitative predictability of classification procedures, they can provide the means to group expression patterns that are coexpressed over a range of experiments in order to detect common regulatory motifs in an unsupervised manner. Moreover, by considering expression profiles over various tissue samples, clustering these samples based on the expression levels for each sample helps to develop techniques that offer the potential to discriminate pathologies and to recognize various forms of cancers or cell types. Clustering constitutes a supporting methodology for classification and prediction.

Many clustering approaches, such as $K$-means [30], self-organizing maps [31], hierarchical clustering [32], and others, have been applied to gene expression data analysis. One difficulty is that the selection of various algorithm parameters and other choices (e.g., type of linkage), initial conditions, and distance measures can all critically impact the results of clustering. Moreover, the number of clusters must often be chosen in advance. Therefore, comparison of results and analysis of the inference capability of clustering algorithms is important [33]. A good overview of clustering algorithms, as applied to gene expression data, including cluster validation, is available in [34].

### 4.2. Networks

A model of a genetic regulatory network is intended to capture the simultaneous dynamical behavior of all elements, such as transcript or protein levels, for which measurements exist. Needless to say, it is possible to devise theoretical models, for instance based on systems of differential equations, that are intended to represent as faithfully as possible the joint behavior of all of these constituent elements. The construction of the models, in this case, can be based on existing knowledge of protein-DNA and protein-protein interactions, degradation rates, and other kinetic parameters. Additionally, some measurements focusing on small-scale molecular interactions can be made, with the goal of refining the model. However, global inference of network structure and fine-scale relationships between all the players in a genetic regulatory network is still an unrealistic undertaking with existing genome-wide measurements produced by microarrays and other high-throughput technologies.

Thus, if we take the pragmatic viewpoint that models are intended to predict certain behavior, be it steady-state expression levels of certain groups of genes or simply the functional relationships between a group of genes, we must then develop them with the awareness of the types of data that are available. For example, it may not be prudent to attempt inferring dozens of continuous-valued rates of change and other parameters in differential equations from only a few discrete-time measurements taken from a population of cells that may not be synchronized with respect to their gene activities (e.g., cell cycle) and with a limited knowledge and understanding of the sources of variation due to the measurement technology and the underlying biology. What we should rather strive for is obtaining the simplest model that is capable of "explaining" the data at some chosen level of "coarseness" (Ockham's Razor). That is, we must strike the right balance between goodness-of-fit and model complexity.

Recently, a new class of models, called probabilistic Boolean networks (PBNs), has been proposed for modeling gene regulatory networks [16]. PBNs inherently capture the dynamics of gene regulation and activity, are probabilistic in nature, thus being able to absorb some of the uncertainty intrinsic to the data, are rule-based, and can be inferred from gene expression data sets in a straightforward manner. This class of models constitutes a probabilistic generalization of the well-known Boolean network model [35]. The PBN can be constructed so as to involve many simple but good predictors of gene activity. Just as importantly, it can include the situation where the structure of the model network changes in accord with the activity of latent variables outside the model, in effect, thereby resulting in a model composed of a family of constituent classical Boolean networks [17].

#### 4.2.1. Prediction

The study of gene interaction and the concomitant behavioral changes due to signals external to the genome itself fits into the classical theories of nonlinear filtering, stochastic control, and nonlinear dynamical systems. Central to both analysis and design is prediction. With microarray technology, the gene expression measurements compose a random vector over time. They have a stochastic nature on account of both inherent biological variability and experimental noise. Genetic changes over time concern this random vector as a temporal process. Questions regarding the interrelation between genes at a given moment of time concern this vector at that moment. Comparison of two cell lines, say tumorigenic and nontumorigenic, involves two random processes and their cross probabilistic characteristics.

The genome is not a closed system. It is affected by intracellular activity, which in turn is affected by external factors. At a very general level, we might represent the situation by a pair of vectors, $X$ denoting the gene expression time process and $Z$ being a vector of variables external to the genome, either cellular or otherwise. In any practical situation, these will only include variables that are observable, measurable, and of interest. In a laboratory setting, $Z$ might be composed of several components decided upon by the experimenter. Ultimately, our concern is with temporal transitions of $X$, affected by both the current states of $X$ and $Z$. The most critical problem is the prediction of $X$ at a future time from a current observation of $X$ and knowledge of $Z$.

A predictor must be designed from data, which ipso facto means that it is an approximation of the predictor whose action one would actually like to model. The precision of the approximation depends on the design procedure and the sample size. Even for a relatively small number of predictor genes, good design can require a very large sample; however,

one typically has a small number of microarrays. There is also the computational problem inherent in the vast number of possible combinations of genes that can be involved in prediction. The problems of classifier design apply essentially unchanged when inferring predictors from sample data. To be effectively addressed, they need to be approached within the context of constraining biological knowledge, since prior knowledge significantly reduces the data requirement.

Even in the context of limited data, there are modest approaches that can be taken. One general statistical approach is to discover associations between the expression patterns of genes via the coefficient of determination [36, 37, 38]. This coefficient measures the degree to which the transcriptional levels of an observed gene set can be used to improve the prediction of the transcriptional state of a target gene relative to the best possible prediction in the absence of observations. The method allows incorporation of knowledge of other conditions relevant to the prediction, such as the application of particular stimuli or the presence of inactivating gene mutations, as predictive elements affecting the expression level of a given gene. Using the coefficient of determination, one can find sets of genes related multivariately to a given target gene. No causality is inferred. It may be that the target is controlled by a function of the predictive genes, or they predict well the behavior of the target because it is a switch for them. The relationship may involve intermediate genes in a complex pathway.

Another approach for finding groups of genes or factors that are likely to determine the activity of some target gene is the minimal description length (MDL) principle, which has been applied in the context of gene expression prediction [39]. This approach essentially seeks flexible classes of models with good predictive properties and considers the complexity of the models as a penalizing factor. With the fundamental goal being to improve the predictive accuracy or generalizability of the model [40], the MDL principle attempts to select the model that achieves the shortest code length describing both the data and the model. A related approach, called normalized maximum likelihood (NLM), has also been recently used for gene-expression-based prediction and classification [41].

#### *4.2.2. Intervention*

One reason for studying regulatory models is to develop intervention strategies to help guide the time evolution of the network towards more desirable states. Three distinct approaches to the intervention problem have been considered in the context of probabilistic Boolean networks by exploiting their Markovian nature. First, one can toggle the expression status of a particular gene from ON to OFF or vice versa to facilitate transition to some other desirable state or set of states. Specifically, by using the concept of the mean first passage time, it has been demonstrated how the particular gene, whose transcription status is to be momentarily altered to initiate the state transition, can be chosen to "minimize" in a probabilistic sense the time required to achieve the desired state transitions [42]. A second approach has aimed at changing the steady-state (long-run) behavior of the network by minimally altering its rule-based structure [43]. A third approach has focused on applying ideas from control theory to develop an intervention strategy, using dynamic programming, in the general context of Markovian genetic regulatory networks whose state transition probabilities depend on an external (control) variable [44].

## 5. CONCLUDING REMARKS

Computational genomics has been greatly influenced by data mining, partly due to the availability of large data sets and databases. Although data mining, as a discipline, is quite broad and lies at the intersection of statistics, machine learning, pattern recognition, and artificial intelligence, there are a number of challenging and important problems in computational genomics that can benefit from the application of engineering principles and methodologies, the latter being characterized by systems-level modeling and simulation.

Modern signal processing, though encompassing many of the same subject areas, has had a different history and background. As such, the applications around which the field has developed have been of a substantially different nature than those in data mining. While data mining problems are often centered around visualization and exploratory analysis of large high-dimensional data sets, finding patterns in data, and discovering good feature sets for classification, some common tasks in signal processing include removal of interference from signals, transforming signals into more suitable representations for various purposes, and analyzing and extracting some characteristics from signals.

Of importance in signal processing is the optimal design of operators under various criteria and constraints. That is, given a "true" signal and its noise-corrupted version, the goal is to find an optimal estimator, from some class of estimators (constraint), such that when it is applied to the noisy signal, some error (criterion) between its output and the true signal is minimized. Alternatively, if a representative signal is not available for training, armed with only the knowledge of the noise characteristics and a class of operators, the goal is to select an optimal estimator under a different criterion, such as minimizing the variance of the noise at its output.

Though these approaches have much in common with machine learning and statistical estimation theory, the nature of the constraints and criteria, and consequently the ensuing theory and algorithms, are guided by application-specific needs, such as detail and edge preservation, robustness to outliers, and other statistical and structural constraints. At the same time, much of the theory behind signal processing, in particular nonlinear digital filters, is tightly intertwined with dynamical systems theory, involving constructs such as finite and cellular automata.

It is clear that signal processing theory, tools, and methods can make a fundamental contribution to gene-expression-based classification and network modeling. Needless to say, traditional signal processing approaches, such as transform theory, can play an important role in other genomic applications, such as DNA or protein sequence analysis [45, 46, 47]. It is our belief that researchers with a background in

signal processing have the potential to make significant contributions and bring their unique perspectives to this exciting and important field.

## REFERENCES


[1] G. Evan and T. Littlewood, "A matter of life and cell death," *Science*, vol. 281, no. 5381, pp. 1317–1322, 1998.

[2] J. L. DeRisi, L. Penland, P. O. Brown, et al., "Use of a cDNA microarray to analyse gene expression patterns in human cancer," *Nature Genetics*, vol. 14, no. 4, pp. 457–460, 1996.

[3] J. L. DeRisi, V. R. Iyer, and P. O. Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science*, vol. 278, no. 5338, pp. 680–686, 1997.

[4] C. M. Perou, T. Sorlie, M. B. Eisen, et al., "Molecular portraits of human breast tumours," *Nature*, vol. 406, no. 6797, pp. 747–752, 2000.

[5] L. Wodicka, H. Dong, M. Mittmann, M. H. Ho, and D. J. Lockhart, "Genome-wide expression monitoring in *Saccharomyces cerevisiae*," *Nature Biotechnology*, vol. 15, no. 12, pp. 1359–1367, 1997.

[6] H. H. McAdams and L. Shapiro, "Circuit simulation of genetic networks," *Science*, vol. 269, no. 5224, pp. 650–656, 1995.

[7] C.-H. Yuh, H. Bolouri, and E. H. Davidson, "Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene," *Science*, vol. 279, no. 5358, pp. 1896–1902, 1998.

[8] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian networks to analyze expression data," *Journal of Computational Biology*, vol. 7, no. 3-4, pp. 601–620, 2000.

[9] A. J. Hartemink, D. K. Gifford, T. S. Jaakkola, and R. A. Young, "Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks," in *Proc. 6th Pacific Symposium on Biocomputing*, pp. 422–433, Mauna Lani, Hawaii, USA, January 2001.

[10] E. J. Moler, D. C. Radisky, and I. S. Mian, "Integrating naive Bayes models and external knowledge to examine copper and iron homeostasis in *S. cerevisiae*," *Physiological Genomics*, vol. 4, no. 2, pp. 127–135, 2000.

[11] K. Murphy and S. Mian, "Modelling gene expression data using dynamic Bayesian networks," Tech. Rep., Computer Science Division, University of California, Berkeley, Calif, USA, 1999.

[12] M. Wahde and J. A. Hertz, "Coarse-grained reverse engineering of genetic regulatory networks," *Biosystems*, vol. 55, pp. 129–136, 2000.

[13] D. C. Weaver, C. T. Workman, and G. D. Stormo, "Modeling regulatory networks with weight matrices," in *Proc. Pacific Symposium on Biocomputing*, vol. 4, pp. 112–123, Mauna Lani, Hawaii, USA, January 1999.

[14] T. Mestl, E. Plahte, and S. W. Omholt, "A mathematical framework for describing and analysing gene regulatory networks," *Journal of Theoretical Biology*, vol. 176, no. 2, pp. 291–300, 1995.

[15] S. A. Kauffman, "Metabolic stability and epigenesis in randomly constructed genetic nets," *Journal of Theoretical Biology*, vol. 22, no. 3, pp. 437–467, 1969.

[16] I. Shmulevich, E. R. Dougherty, S. Kim, and W. Zhang, "Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks," *Bioinformatics*, vol. 18, no. 2, pp. 261–274, 2002.

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

[18] A. Arkin, J. Ross, and H. H. McAdams, "Stochastic kinetic analysis of developmental pathway bifurcation in phage $\lambda$-infected *Escherichia coli* cells," *Genetics*, vol. 149, no. 4, pp. 1633–1648, 1998.

[19] R. Somogyi and L. D. Greller, "The dynamics of molecular networks: applications to therapeutic discovery," *Drug Discovery Today*, vol. 6, no. 24, pp. 1267–1277, 2001.

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

[21] T. R. Golub, D. K. Slonim, P. Tamayo, et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[22] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini, "Tissue classification with gene expression profiles," *Journal of Computational Biology*, vol. 7, no. 3-4, pp. 559–583, 2000.

[23] J. Khan, J. S. Wei, M. Ringner, et al., "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature Medicine*, vol. 7, no. 6, pp. 673–679, 2001.

[24] I. Hedenfalk, D. Duggan, Y. Chen, et al., "Gene-expression profiles in hereditary breast cancer," *New England Journal of Medicine*, vol. 344, no. 8, pp. 539–548, 2001.

[25] U. Alon, N. Barkai, D. A. Notterman, et al., "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 12, pp. 6745–6750, 1999.

[26] M. Bittner, P. Meltzer, J. Khan, et al., "Molecular classification of cutaneous malignant melanoma by gene expression profiling," *Nature*, vol. 406, no. 6795, pp. 536–540, 2000.

[27] S. Kim, E. R. Dougherty, I. Shmulevich, et al., "Identification of combination gene sets for glioma classification," *Molecular Cancer Therapeutics*, vol. 1, no. 13, pp. 1229–1236, 2002.

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

[29] E. R. Dougherty, "Small sample issues for microarray-based classification," *Comparative and Functional Genomics*, vol. 2, no. 1, pp. 28–34, 2001.

[30] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church, "Systematic determination of genetic network architecture," *Nature Genetics*, vol. 22, no. 3, pp. 281–285, 1999.

[31] P. Tamayo, D. Slonim, J. Mesirov, et al., "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 6, pp. 2907–2912, 1999.

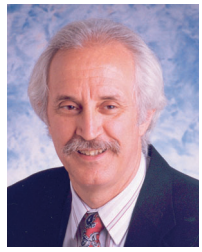
[32] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 25, pp. 14863–14868, 1998.

[33] E. R. Dougherty, J. Barrera, M. Brun, et al., "Inference from clustering: application to gene-expression time series," *J. Comput. Biol.*, vol. 9, no. 1, pp. 105–126, 2002.

[34] Y. Moreau, F. de Smet, G. Thijs, K. Marchal, and B. de Moor, "Functional bioinformatics of microarray data: from expression to regulation," *Proceedings of the IEEE*, vol. 90, no. 11, pp. 1722–1743, 2002.

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

[36] E. R. Dougherty, S. Kim, and Y. Chen, "Coefficient of determination in nonlinear signal processing," *Signal Processing*, vol. 80, no. 10, pp. 2219–2235, 2000.
[37] S. Kim, E. R. Dougherty, M. L. Bittner, et al., "General nonlinear framework for the analysis of gene interaction via multivariate expression arrays," *Biomedical Optics*, vol. 5, no. 4, pp. 411–424, 2000.
[38] S. Kim, E. R. Dougherty, Y. Chen, et al., "Multivariate measurement of gene expression relationships," *Genomics*, vol. 67, no. 2, pp. 201–209, 2000.
[39] I. Tabus and J. Astola, "On the use of MDL principle in gene expression prediction," *EURASIP Journal on Applied Signal Processing*, vol. 2001, no. 4, pp. 297–303, 2001.
[40] I. Shmulevich, "Model selection in genomics," *EHP Toxicogenomics*, vol. 111, no. 6, pp. A328–A329, 2003.
[41] I. Tabus, J. Rissanen, and J. Astola, "Normalized maximum likelihood models for Boolean regression with application to prediction and classification in genomics," in *Computational and Statistical Approaches to Genomics*, W. Zhang and I. Shmulevich, Eds., Kluwer Academic Publishers, Boston, Mass, USA, 2002.
[42] I. Shmulevich, E. R. Dougherty, and W. Zhang, "Gene Perturbation and intervention in probabilistic Boolean networks," *Bioinformatics*, vol. 18, no. 10, pp. 1319–1331, 2002.
[43] I. Shmulevich, E. R. Dougherty, and W. Zhang, "Control of stationary behavior in probabilistic Boolean networks by means of structural intervention," *Journal of Biological Systems*, vol. 10, no. 4, pp. 431–445, 2002.
[44] A. Datta, A. Choudhary, M. L. Bittner, and E. R. Dougherty, "External control in Markovian genetic regulatory networks," *Machine Learning Journal*, vol. 52, no. 1-2, pp. 169–191, 2003.
[45] D. Anastassiou, "Frequency-domain analysis of biomolecular sequences," *Bioinformatics*, vol. 16, no. 12, pp. 1073–1081, 2000.
[46] P. D. Cristea, "Large scale features in DNA genomic signals," *Signal Processing*, vol. 83, no. 4, pp. 871–888, 2003.
[47] K. M. Bloch and G. R. Arce, "Analyzing protein sequences using signal analysis techniques," in *Computational and Statistical Approaches to Genomics*, W. Zhang and I. Shmulevich, Eds., pp. 113–124, Kluwer Academic Publishers, Boston, Mass, USA, 2002.



**Edward R. Dougherty** is a Professor in the Department of Electrical Engineering at Texas A&M University in College Station. He holds an M.S. degree in computer science from Stevens Institute of Technology in 1986 and a Ph.D. degree in mathematics from Rutgers University in 1974. He is the author of eleven books and the editor of other four books. He has published more than one hundred journal papers, is an SPIE Fellow, and has served as an Editor of the Journal of Electronic Imaging for six years. He is currently Chair of the SIAM Activity Group on Imaging Science. Prof. Dougherty has contributed extensively to the statistical design of nonlinear operators for image processing and the consequent application of pattern recognition theory to nonlinear image processing. His current research focuses on genomic signal processing, with the central goal being to model genomic regulatory mechanisms. He is Head of the Genomic Signal Processing Laboratory at Texas A&M University.

**Ilya Shmulevich** received his Ph.D. degree in electrical and computer engineering from Purdue University, West Lafayette, Ind, USA, in 1997. From 1997 to 1998, he was a Postdoctoral Researcher at the Nijmegen Institute for Cognition and Information at the University of Nijmegen and National Research Institute for Mathematics and Computer Science at the University of Amsterdam in the Netherlands, where he studied computational models of music perception and recognition. From 1998 to 2000, he worked as a Senior Researcher at Tampere International Center for Signal Processing in the Signal Processing Laboratory at Tampere University of Technology, Tampere, Finland. Presently, he is an Assistant Professor at Cancer Genomics Laboratory at The University of Texas MD Anderson Cancer Center in Houston, Tex. He is an Associate Editor of Environmental Health Perspectives: Toxicogenomics. His research interests include computational genomics, nonlinear signal and image processing, computational learning theory, and music recognition and perception.

**Michael L. Bittner** was initially trained as a biochemical geneticist, studying phage replication and bacterial transposition with a variety of biochemical and bacterial genetic methods at Princeton University, where he received his Ph.D. degree from Washington University School of Medicine, and the Population and Molecular Genetics Department of the University of Georgia, where he carried out his postdoctoral researches. Since that time, his efforts was concentrated on the practical application of knowledge about the control systems operating in prokaryotes and eukaryotes. At Monsanto Corporation in St. Louis, Dr. Bittner was involved in developing technology for the biologic production of peptides and proteins useful in human medicine and agriculture. At Amoco Corporation in Downers Grove, Illinois, he played a central role in developing methods for producing, in yeast, small molecule precursors of vitamins of human and veterinary pharmacologic interest. He collaborated in the development of cytogenetic molecular diagnostics based on in-situ hybridization that produced a series of technologies leading to the founding of Vysis Corporation, also in Downers Grove. His recent efforts in the National Institutes of Health and the Translational Genomics Research Institute focus on developing ways of making accurate measures of the transcriptional status of cells and analytic tools that allow inferences to be drawn from these measures that provide insight into the cellular processes operating in healthy and diseased cells.