EURASIP Book Series on Signal Processing and Communications

# Genomic Signal Processing and Statistics

Edited by: Edward R. Dougherty, Ilya Shmulevich, Jie Chen, and Z. Jane Wang

# Genomic Signal Processing and Statistics

EURASIP Book Series on Signal Processing and Communications
Editor-in-Chief: K. J. Ray Liu
Editorial Board: Zhi Ding, Moncef Gabbouj, Peter Grant, Ferran Marqués, Marc Moonen, Hideaki Sakai, Giovanni Sicuranza, Bob Stewart, and Sergios Theodoridis

Hindawi Publishing Corporation
410 Park Avenue, 15th Floor, #287 pmb, New York, NY 10022, USA
Nasr City Free Zone, Cairo 11816, Egypt
Fax: +1-866-HINDAWI (USA toll-free)



ISBN 977-5945-07-0

EURASIP Book Series on Signal Processing and Communications, Volume 2

# Genomic Signal Processing and Statistics

*Edited by: Edward R. Dougherty, Ilya Shmulevich, Jie Chen, and Z. Jane Wang*

Hindawi Publishing Corporation
http://www.hindawi.com

# Contents


Genomic signal processing: perspectives, *Edward R. Dougherty, Ilya Shmulevich, Jie Chen, and Z. Jane Wang* ..... 1

**Part I. Sequence Analysis**

1. Representation and analysis of DNA sequences, *Paul Dan Cristea* ..... 15

**Part II. Signal Processing and Statistics Methodologies in Gene Selection**

2. Gene feature selection, *Ioan Tabus and Jaakko Astola* ..... 67
3. Classification, *Ulisses Braga-Neto and Edward R. Dougherty* ..... 93
4. Clustering: revealing intrinsic dependencies in microarray data, *Marcel Brun, Charles D. Johnson, and Kenneth S. Ramos* ..... 129
5. From biochips to laboratory-on-a-chip system, *Lei Wang, Hongying Yin, and Jing Cheng* ..... 163

**Part III. Modeling and Statistical Inference of Genetic Regulatory Networks**

6. Modeling and simulation of genetic regulatory networks by ordinary differential equations, *Hidde de Jong and Johannes Geiselmann* ..... 201
7. Modeling genetic regulatory networks with probabilistic Boolean networks, *Ilya Shmulevich and Edward R. Dougherty* ... 241
8. Bayesian networks for genomic analysis, *Paola Sebastiani, Maria M. Abad, and Marco F. Ramoni* ..... 281
9. Statistical inference of transcriptional regulatory networks, *Xiaodong Wang, Dimitris Anastassiou, and Dong Guo* ..... 321

**Part IV. Array Imaging, Signal Processing in Systems Biology, and Applications in Disease Diagnosis and Treatments**

10. Compressing genomic and proteomic array images for statistical analyses, *Rebecka Jörnsten and Bin Yu* ..... 341
11. Cancer genomics, proteomics, and clinic applications, *X. Steve Fu, Chien-an A. Hu, Jie Chen, Z. Jane Wang, and K. J. Ray Liu* ..... 367
12. Integrated approach for computational systems biology, *Seungchan Kim, Phillip Stafford, Michael L. Bittner, and Edward B. Suh* ..... 409

# Genomic signal processing: perspectives

**Edward R. Dougherty, Ilya Shmulevich, Jie Chen, and Z. Jane Wang**

No single agreed-upon definition seems to exist for the term *bioinformatics*, which has been used to mean a variety of things, ranging in scope and focus. To cite but a few examples from textbooks, Lodish et al. state that "bioinformatics is the rapidly developing area of computer science devoted to collecting, organizing, and analyzing DNA and protein sequences" [1]. A more general and encompassing definition, given by Brown, is that bioinformatics is "the use of computer methods in studies of genomes" [2]. More general still, "bioinformatics is the science of refining biological information into biological knowledge using computers" [3]. Kohane et al. observe that the "breadth of this commonly used definition of bioinformatics risks relegating it to the dustbin of labels too general to be useful" and advocate being more specific about the particular bioinformatics techniques employed [4].

Genomic signal processing (GSP) is the engineering discipline that studies the processing of genomic signals, by which we mean the measurable events, principally the production of mRNA and protein, that are carried out by the genome. Based upon current technology, GSP primarily deals with extracting information from gene expression measurements. The analysis, processing, and use of genomic signals for gaining biological knowledge constitute the domain of GSP. 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 [5]. 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.

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 [6, 7, 8, 9, 10]. Since transcription control is accomplished by a method that interprets a variety of inputs, we require analytical tools for the 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. Application is generally directed towards tissue classification and the discovery of signaling pathways.

Because transcriptional control is accomplished by a complex method that interprets a variety of inputs, 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. Perhaps the most salient aspect of GSP is that it is an engineering discipline, having strong roots in signals and systems theory. In GSP, the point of departure is that the living cell is a system in which many interacting components work together to give rise to execution of normal cellular functions, complex behavior, and interaction with the environment, including other cells. In such systems, the "whole" is often more than the "sum of its parts," frequently referred to as emergent or complex behavior. The collective behavior of all relevant components in a cell, such as genes and their products, follows a similar paradigm, but gives rise to much richer behavior, that is characteristic of living systems. To gain insight into the behavior of such systems, a systems-wide approach must be taken. This requires us to produce a model of the components and their interactions and apply mathematical, statistical, or simulation tools to understand its behavior, especially as it relates to experimental data.

In this introductory chapter, we comment on four major areas of GSP research: signal extraction, phenotype classification, clustering, and gene regulatory networks. We then provide brief descriptions of each of the contributed chapters.

## Signal extraction

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 complementary DNA (cDNA) and oligonucleotide, provides a powerful analytic tool for genetic research. Since our concern is GSP, not 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 with a scanning confocal microscope. cDNA 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

cohybridized onto each arrayed gene. Ratios or 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, the better choice depending on the sources of variation [11].

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: the test sample typically placed in the red channel, 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 samples are wellmixed 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, variability analysis, and measurement quality assessment [12]. 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 ratios.

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 owing to a 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 overrepresented. There may also be differences in labeling or detection efficiencies between the fluorescent dyes (e.g., Cy3, 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.

A major barrier to an effective understanding of variation is the large number of sources of variance inherent in microarray measurements. In many statistical analysis publications, the measured gene expression data are assumed to have multiple noise sources: noise due to sample preparation, labeling, hybridization, background fluorescence, different arrays, fluorescent dyes, and different printing locations. In attempting to quantify the noise level in a set of experiments, some studies employ ANOVA models in which the log-transformed gene expression signal is represented by true signal plus an additive noise [13, 14]. Other proposed models for expression signals include mixture models for gene effect [15], multiplicative model (not logarithm-transformed) [16, 17], ratio-distribution model [12, 18], binary model [19], rank-based models not sensitive to noise distributions [20], replicates using mixed models [21], quantitative noise analysis [22, 23], and design of reverse dye microarrays [24]. In addition to the many studies on noise estimation in microarrays, there is a large literature dealing with methods to isolate and eliminate the noise component from the measured signal. These studies suffer from the daunting complexity and inhomogeneity of the noise.

## Classification

Pattern classification plays an important role in genomic signal analysis. For instance, cDNA microarrays can provide expression measurements for thousands of genes at once, and a key goal is to perform classification via different expression patterns. This 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 a host of such differences. Early cancer studies include leukemias [25] and breast cancer [26, 27]. Classifiers are designed from a sample of expression vectors by 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.

An expression-based classifier provides a list of genes whose product abundance is indicative of important differences in a 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 mitigation of the aberrant function of gene products whose activities are central to the pathology of a disease. Mitigation 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.

Three critical statistical issues arise for expression-based classification [28]. 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 difficulty of successfully accomplishing these tasks is severely exacerbated by the fact that small samples are ubiquitous in studies employing expression microarrays, meaning that the potential number of variables (gene expressions) is huge in comparison to the sample size (number of microarrays) [29]. As with most studies, due to cost and patient availability, this investigation will be in the small-sample category. Three points must be taken into consideration: (1) to avoid overfitting, simple classifiers should be employed [28, 30, 31]; (2) again to avoid overfitting, small feature sets are required [32, 33, 34, 35]; and (3) because samples are small and error estimation must be performed using the training data, the choice of error estimation rule is critical [36, 37], with feature-set ranking being of particular importance in gene discovery [38].

The problem of small-sample error estimation is particularly troublesome. An error estimator may be unbiased but have a large variance, and therefore, often be low. This can produce a large number of feature sets and classifiers with low error estimates. 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, there are 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.

## Clustering

A classifier takes a single data point (expression vector) and outputs a class label (phenotype); a cluster operator takes a set of data points (expression vectors) and partitions the points into clusters (subsets). Clustering has become a popular data-analysis technique in genomic studies using gene-expression microarrays [39, 40]. Time-series clustering groups together genes whose expression levels exhibit similar behavior through time. Similarity indicates possible coregulation. Another way to use expression data is to take expression profiles over various tissue samples, and then cluster these samples based on the expression levels for each sample, the motivation being the potential to discriminate pathologies based on their differential patterns of gene expression. A host of clustering algorithms has been proposed in the literature and many of these have been applied to genomic data: $k$-means, fuzzy $c$-means, self-organizing maps [41, 42, 43], hierarchical clustering, and model-based clustering [44, 45].

Many validation techniques have been proposed for evaluating clustering results. These are generally based on the degree to which clusters derived from a set of sample data satisfy certain heuristic criteria. This is significantly different than classification, where the error of a classifier is given by the probability of an erroneous decision. Validation methods can be roughly divided into two categories (although this categorization can certainly be made finer)—*internal* and *external*.

Internal validation methods evaluate the clusters based solely on the data, without external information. Typically, a heuristic measure is defined to indicate the goodness of the clustering. It is important to keep in mind that the measure only applies to the data at hand, and therefore is not predictive of the worth of a clustering algorithm—even with respect to the measure itself. Since these kinds of measures do not possess predictive capability, it appears difficult to assess their worth—even what it means to be "worthy." But there have been simulation studies to observe how they behave [46].

External validation methods evaluate a clustering algorithm by comparing the resulting clusters with prespecified information [47]. Agreement between the heuristic and algorithm-based partitions indicates algorithm accuracy. It also indicates that the scientific understanding behind the heuristic partition is being reflected in the measurements, thereby providing supporting evidence for the measurement process.

With model-based clustering, a Bayesian approach can be taken to determine the best number of clusters. Two models can be compared relative to the sample data by a *Bayes factor* [48, 49].

To recognize the fundamental difference between clustering and classification, we note two key characteristics of classification: (1) classifier error can be estimated under the assumption that the sample data arise from an underlying feature-label distribution; and (2) given a family of classifiers, sample data can be used to learn the optimal classifier in the family. Once designed, the classifier represents a mathematical model that provides a decision mechanism relative to real-world measurements. The model represents scientific knowledge to the extent that it has predictive capability. The purpose of testing (error estimation) is quantifying the worth of the model. Clustering has generally lacked both fundamental characteristics of classification. In particular, lacking inference in the context of a probability model, it has remained essentially a subjective visualization tool. Jain et al. wrote, "Clustering is a subjective process; the same set of data items often needs to be partitioned differently for different applications. This subjectivity makes the process of clustering difficult" [50]. Duda et al. stated the matter radically, "The answer to whether or not it is possible in principle to learn anything from unlabeled data depends upon the assumptions one is willing to accept—theorems cannot be proved without premises" [51]. These criticisms raise the question as to whether clustering can be used for scientific knowledge. This issue has been raised specifically in the context of gene-expression microarrays by Kerr and Churchill when they wrote, "A great deal of effort has gone into identifying the best clustering techniques for microarray data. However, another question that is at least

as important has received less attention; how does one make statistical inferences based on the results of clustering?" [52]. Indeed, how is one going to judge the relative worth of clustering algorithms unless it is based on their inference capabilities?

For clustering to have a sound scientific basis, error estimation must be addressed in the context of an appropriate probabilistic model. Ipso facto, since a clustering algorithm partitions a set of data points, error estimation for clustering must assume that clusters resulting from a cluster algorithm can be compared to the correct clusters for the data set in the context of a probability distribution, thereby providing an error measure. The key to a general probabilistic theory of clustering, including both error estimation and learning, is to recognize that classification theory is based on operators on random variables, and that the theory of clustering needs to be based on operators on random points sets [53]. Once clustering has been placed into a probabilistic context, proposed clustering algorithms can be rigorously evaluated as estimators, rules can be developed from designing clustering algorithms from data (analogous to the design of classifiers via classification rules), and these rules can be evaluated based on the kinds of criteria used for classification rules, such as consistency, approximation, and sample size.

## Gene regulatory networks

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 [54]. 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 signaling pathways. A major problem for fine-scale modeling is its large data requirement. A second approach involves 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 probabilistic gene regulatory networks (PGRNs), which include the special case of probabilistic Boolean networks (PBNs).

Gene regulatory systems comprise an important example of a natural system composed of individual elements that interact with each other in a complex fashion, in this case, to regulate and control the production of proteins viable for cell function. Development of analytical and computational tools for the modeling and analysis of gene regulation can substantially help to unravel the mechanisms underlying gene regulation and to understand gene function [55, 56, 57, 58]. This, in turn, can have a profound effect on developing techniques for drug testing and therapeutic intervention for effective treatment of human diseases.

A model of a genetic regulatory network is intended to capture the simultaneous dynamical behavior of various elements, such as transcript or protein levels, for which measurements exist. There have been numerous approaches for modeling the dynamical behavior of genetic regulatory networks, ranging from deterministic to fully stochastic, using either a discrete-time or a continuous-time description of the gene interactions [54]. One way to proceed is to devise theoretical models, for instance, based on systems of differential equations intended to represent as faithfully as possible the joint behavior of all of these constituent elements [59]. 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 currently an unrealistic undertaking with existing genome-wide measurements produced by microarrays and other high-throughput technologies.

With the understanding that models are intended to predict certain behavior, be it steady-state expression levels of certain groups of genes or functional relationships among a group of genes, we must then develop them with an awareness of the types of available data. 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), with a limited knowledge and understanding of the sources of variation due to the measurement technology and the underlying biology. From an engineering perspective, a model should be sufficiently complex to capture the relations necessary for solving the problem at hand, and not so complex that it cannot be reliably estimated from the data. With the advent of microarray technology, a significant effort has been directed at 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 Bayesian networks [60], Boolean networks [61], and PBNs (and their extension to PGRNs) [62].

There are two important aspects of every genetic regulatory system that have to be modeled and analyzed. The first is the topology (connectivity structure), and the second is the set of interactions between the elements, the latter determining the dynamical behavior of the system [63, 64, 65]. Exploration of the relationship between topology and dynamics can lead to valuable conclusions about the structure, behavior, and properties of genetic regulatory systems [66, 67].

In a discrete-time functional network, the state of a gene at time $t + 1$ is considered to be a function of a set of genes in a *regulatory set* at time $t$. The connectivity of the network is defined by the collection of regulatory sets and the interactions are defined by the functions, which are often called *predictors*. 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. The problems of classifier design apply essentially unchanged when learning 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.

The oldest model for gene regulation is the Boolean network [61, 68, 69, 70, 71]. In a Boolean network, each gene is represented by a binary value, 0 or 1, indicating whether it is down- or up-regulated, and each gene value at the next time point is determined by a function of the gene values in its regulatory set. The action of the network is deterministic and after some finite time, it will settle into an attractor, which is a set of states though which it will endlessly cycle. The Boolean model has recently been extended so that instead of a single predictor function, each gene has a set of predictor functions, one of which is chosen at each time point. This extension results in the class of PBNs [62, 72]. In the early PBN papers, regulatory sets were chosen based on the coefficient of determination, which measures the degree to which the prediction of a target's random variable is improved by observation of the variables in the regulatory set relative to prediction of the target variable using only statistical information concerning the target variable itself [73, 74, 75]. If the predictor choice is random at each time point, then the network is said to be instantaneously random; the predictor is held fixed and only allowed to switch depending on some binary random variable, then the network is said to be context sensitive. The latter case results in a family of Boolean networks composing the PBN, with one of the constituent networks governing gene activity for some period of time. This reflects the effect of latent variables, not incorporated into the model. A PGRN has the same structure as a PBN except that each gene may take on a value within a discrete interval $[0, r]$, with $r$ not being constrained to 0 or 1.

A key objective of network modeling is to use the network to design different approaches for affecting the evolution of the gene state vector over time—for instance, in the case of cancer to drive the network away from states associated with cell proliferation. There have been a number of studies regarding intervention in the context of PBNs . These include resetting the state of the PBN, as necessary, to a more desirable initial state and letting the network evolve from there [76] and 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 [77, 78]. The latter approach is particularly promising because it involves the use of automatic control theory to derive optimal treatment strategies over time—for instance, using dynamic programming.

## Overview of the book

This edited book provides an up-to-date and tutorial-level overview of genomic signal processing (GSP) and statistics. Written by an interdisciplinary team of

authors, the book is accessible to researchers in academia and industry, who are interested in cross-disciplinary areas relating to molecular biology, engineering, statistics, and signal processing. Our goal is to provide audiences with a broad overview of recent advances in the important and rapidly developing GSP discipline.

In the following, we give a brief summary of the contents covered in this book. The book consists of twelve book chapters.

(i) In the first part, we focus on signal processing and statistics techniques in sequence analysis. In "Representation and analysis of DNA sequences," by Paul Dan Cristea, the author presents results in the analysis of genomic information at the scale of whole chromosomes or whole genomes based on the conversion of genomic sequences into genomic signals, concentrating on the phase analysis.

(ii) In the second part, we focus on signal processing and statistics methodologies in gene selection: classification, clustering, and data extraction. In "Gene feature selection," by Ioan Tabus and Jaakko Astola, the authors overview the classes of feature selection methods, and focus specially on microarray problems, where the number of measured genes (factors) is extremely large, in the order of thousands, and the number of relevant factors is much smaller. Classification plays an important role in genomic signal analysis. In "Classification," by Ulisses Braga-Neto and Edward Dougherty, the authors present various techniques in classification, including classifier design, regularization, and error estimation. In "Clustering: revealing intrinsic dependencies in microarray data," by Marcel Brun, Charles D. Johnson, and Kenneth S. Ramos, the authors address clustering algorithms, including interpretation, validation, and clustering microarray data. In "From biochips to laboratory-on-a-chip system," by Lei Wang, Hongying Yin, and Jing Cheng, the authors review various aspects related to biochips with different functionality and chip-based integrated systems.

(iii) In the third part, we focus on signal processing in genomic network modeling and analysis. In "Modeling and simulation of genetic regulatory networks by ordinary differential equations," by Hidde de Jong and Johannes Geiselmann, the authors review various methods for modeling and simulating genetic regulatory network and propose differential equations for regulatory network modeling. In "Modeling genetic regulatory networks with probabilistic Boolean networks," by Ilya Shmulevich and Edward R. Dougherty, the authors present a recently proposed mathematical rule-based model, the probabilistic Boolean networks (PBNs), to facilitate the construction of gene regulatory networks. In "Bayesian networks for genomic analysis," by Paola Sebastiani, Maria M. Abad, and Marco F. Ramoni, the authors show how to apply Bayesian networks in analyzing various types of genomic data, from genomic markers to gene expression data. In "Statistical inference of transcriptional regulatory networks," by Xiaodong Wang, Dimitris Anastassiou, and Dong Guo, the authors present parameter estimation methods for known network structures, including equation-based methods and Bayesian methods. They also discuss Bayesian techniques for inferring network structures.

(iv) In the last part of this book, we focus on microarray imaging, signal processing in systems biology, and applications in disease diagnosis and treatments. In "Compressing genomic and proteomic microarray images for statistical analyses," by Rebecka Jörnsten and Bin Yu, the authors propose a multilayer data structure as the principle for both lossless and lossy compression of microarray images. In "Cancer genomics, proteomics, and clinic applications," by X. Steve Fu, Chien-an A. Hu, Jie Chen, Jane Wang, and K. J. Ray Liu, the authors focus on genomics and proteomics of cancer, and discuss how cutting-edge technologies, like microarray technology and nanotechnology, can be applied in clinical oncology. In "Integrated approach for computational systems biology," by Seungchan Kim, Phillip Stafford, Michael L. Bittner, and Edward B. Suh, the authors address integrated approaches for computational systems biology including biological data and measurement technologies, systems for biological data integration, mathematical and computational tools for computational systems biology, and supercomputing and parallel applications.

Finally, the coeditors would like to thank the authors for their contributions. We hope that readers enjoy this book.

## Bibliography

[1] H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore, and J. E. Darnell, *Molecular Cell Biology*, W. H. Freeman, New York, NY, USA, 4th edition, 2000.

[2] T. A. Brown, *Genomes*, John Wiley & Sons, New York, NY, USA, 2nd edition, 2002.

[3] S. Drăghici, *Data Analysis Tools for DNA Microarrays*, Chapman & Hall/CRC, Boca Raton, Fla, USA, 2003.

[4] I. S. Kohane, A. Kho, and A. J. Butte, *Microarrays for an Integrative Genomics*, MIT Press, Cambridge, Mass, USA, 2003.

[5] E. R. Dougherty, I. Shmulevich, and M. L. Bittner, "Genomic signal processing: the salient issues," *EURASIP J. Appl. Signal Process.*, vol. 2004, no. 1, pp. 146–153, 2004.

[6] M. Schena, D. Shalon, R. W. 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.

[7] M. Schena, D. Shalon, R. Heller, A. Chai, P. O. Brown, and R. W. Davis, "Parallel human genome analysis: microarray-based expression monitoring of 1000 genes," *Proc. Natl. Acad. Sci. USA*, vol. 93, no. 20, pp. 10614–10619, 1996.

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

[9] 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.

[10] D. J. Duggan, M. L. Bittner, Y. Chen, P. S. Meltzer, and J. M. Trent, "Expression profiling using cDNA microarrays," *Nat. Genet.*, vol. 21, Suppl 1, pp. 10–14, 1999.

[11] S. Attoor, E. R. Dougherty, Y. Chen, M. L. Bittner, and J. M. Trent, "Which is better for cDNA-microarray-based classification: ratios or direct intensities," *Bioinformatics*, vol. 20, no. 16, pp. 2513–2520, 2004.

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

[13] M. K. Kerr, M. Martin, and G. A. Churchill, "Analysis of variance for gene expression microarray data," *J. Comput. Biol.*, vol. 7, no. 6, pp. 819–837, 2000.

[14] M. K. Kerr and G. A. Churchill, "Statistical design and the analysis of gene expression microarray data," *Genet. Res.*, vol. 77, no. 2, pp. 123–128, 2001.

[15] M. L. Lee, F. C. Kuo, G. A. Whitmore, and J. Sklar, "Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations," *Proc. Natl. Acad. Sci. USA*, vol. 97, no. 18, pp. 9834–9839, 2000.

[16] M. C. Yang, Q. G. Ruan, J. J. Yang, et al., "A statistical method for flagging weak spots improves normalization and ratio estimates in microarrays," *Physiol Genomics*, vol. 7, no. 1, pp. 45–53, 2001.

[17] R. Sasik, E. Calvo, and J. Corbeil, "Statistical analysis of high-density oligonucleotide arrays: a multiplicative noise model," *Bioinformatics*, vol. 18, no. 12, pp. 1633–1640, 2002.

[18] Y. Chen, V. Kamat, E. R. Dougherty, M. L. Bittner, P. S. Meltzer, and J. M. Trent, "Ratio statistics of gene expression levels and applications to microarray data analysis," *Bioinformatics*, vol. 18, no. 9, pp. 1207–1215, 2002.

[19] I. Shmulevich and W. Zhang, "Binary analysis and optimization-based normalization of gene expression data," *Bioinformatics*, vol. 18, no. 4, pp. 555–565, 2002.

[20] A. Ben-Dor, N. Friedman, and Z. Yakhini, "Scoring genes for relevance," Tech. Rep. AGL-2000-13, Agilent Laboratories, Palo Alto, Calif, USA, 2000.

[21] L. Wernisch, S. L. Kendall, S. Soneji, et al., "Analysis of whole-genome microarray replicates using mixed models," *Bioinformatics*, vol. 19, no. 1, pp. 53–61, 2003.

[22] Y. Tu, G. Stolovitzky, and U. Klein, "Quantitative noise analysis for gene expression microarray experiments," *Proc. Natl. Acad. Sci. USA.*, vol. 99, no. 22, pp. 14031–14036, 2002.

[23] H. M. Fathallah-Shaykh, M. Rigen, L. J. Zhao, et al., "Mathematical modeling of noise and discovery of genetic expression classes in gliomas," *Oncogene*, vol. 21, no. 47, pp. 7164–7174, 2002.

[24] K. Dobbin, J. H. Shih, and R. Simon, "Statistical design of reverse dye microarrays," *Bioinformatics*, vol. 19, no. 7, pp. 803–810, 2003.

[25] 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.

[26] 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.

[27] I. Hedenfalk, D. Duggan, Y. Chen, et al., "Gene-expression profiles in hereditary breast cancer," *N. Engl. J. Med.*, vol. 344, no. 8, pp. 539–548, 2001.

[28] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, vol. 31 of *Applications of Mathematics (New York)*, 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] V. N. Vapnik and A. Ya. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory Probab. Appl.*, vol. 16, no. 2, pp. 264–280, 1971.

[31] V. N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, New York, NY, USA, 1998.

[32] T. M. Cover and J. M. van Campenhout, "On the possible orderings in the measurement selection problem," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-7, no. 9, pp. 657–661, 1977.

[33] S. Raudys and A. K. Jain, "Small sample size effects in statistical pattern recognition: recommendations for practitioners," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 13, no. 3, pp. 252–264, 1991.

[34] A. K. Jain and D. Zongker, "Feature selection: evaluation, application, and small sample performance," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, no. 2, pp. 153–158, 1997.

[35] M. Kudo and J. Sklansky, "Comparison of algorithms that select features for pattern classifiers," *Pattern Recognition*, vol. 33, no. 1, pp. 25–41, 2000.

[36] U. M. Braga-Neto and E. R. Dougherty, "Is cross-validation valid for small-sample microarray classification?," *Bioinformatics*, vol. 20, no. 3, pp. 374–380, 2004.

[37] U. M. Braga-Neto and E. R. Dougherty, "Bolstered error estimation," *Pattern Recognition*, vol. 37, no. 6, pp. 1267–1281, 2004.

[38] C. Sima, U. Braga-Neto, and E. R. Dougherty, "Superior feature-set ranking for small samples using bolstered error estimation," to appear in Bioinformatics.

[39] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proc. Natl. Acad. Sci. USA*, vol. 95, no. 25, pp. 14863–14868, 1998.

[40] A. Ben-Dor, R. Shamir, and Z. Yakhini, "Clustering gene expression patterns," *J. Comput. Biol.*, vol. 6, no. 3-4, pp. 281–297, 1999.
[41] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biol. Cybern.*, vol. 43, no. 1, pp. 59–69, 1982.
[42] T. Kohonen, *Self-organizing Maps*, vol. 30 of *Springer Series in Information Sciences*, Springer-Verlag, Berlin, Germany, 1995.
[43] A. Flexer, "On the use of self-organizing maps for clustering and visualization," *Intelligent Data Analysis*, vol. 5, pp. 373–384, 2001.
[44] J. D. Banfield and A. E. Raftery, "Model-based Gaussian and non-Gaussian clustering," *Biometrics*, vol. 49, no. 3, pp. 803–821, 1993.
[45] C. Fraley and A. E. Raftery, "How many clusters? Which clustering method? Answers via model-based cluster analysis," *The Computer Journal*, vol. 41, no. 8, pp. 578–588, 1998.
[46] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 12, pp. 1650–1654, 2002.
[47] G. W. Milligan and M. C. Cooper, "A study of the comparability of external criteria for hierarchical cluster analysis," *Multivariate Behav. Res.*, vol. 21, pp. 441–458, 1986.
[48] R. E. Kass and A. E. Raftery, "Bayes factors," *J. Amer. Statist. Assoc.*, vol. 90, no. 430, pp. 773–795, 1995.
[49] K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo, "Model-based clustering and data transformations for gene expression data," *Bioinformatics*, vol. 17, no. 10, pp. 977–987, 2001.
[50] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Comput. Surveys*, vol. 31, no. 3, pp. 264–323, 1999.
[51] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley-Interscience, New York, NY, USA, 2001.
[52] M. K. Kerr and G. A. Churchill, "Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments," *Proc. Natl. Acad. Sci. USA*, vol. 98, no. 16, pp. 8961–8965, 2001.
[53] E. R. Dougherty and M. Brun, "A probabilistic theory of clustering," *Pattern Recognition*, vol. 37, no. 5, pp. 917–925, 2004.
[54] H. de Jong, "Modeling and simulation of genetic regulatory systems: a literature review," *J. Comput. Biol.*, vol. 9, no. 1, pp. 67–103, 2002.
[55] D. Endy and R. Brent, "Modelling cellular behaviour," *Nature*, vol. 409, no. 6818, pp. 391–395, 2001.
[56] J. Hasty, D. McMillen, F. Isaacs, and J. J. Collins, "Computational studies of gene regulatory networks: in numero molecular biology," *Nat. Rev. Genet.*, vol. 2, no. 4, pp. 268–279, 2001.
[57] T. Ideker, T. Galitski, and L. Hood, "A new approach to decoding life: systems biology," *Annu. Rev. Genomics Hum. Genet.*, vol. 2, pp. 343–372, 2001.
[58] H. Kitano, "Systems biology: a brief overview," *Science*, vol. 295, no. 5560, pp. 1662–1664, 2002.
[59] T. Mestl, E. Plahte, and S. W. Omholt, "A mathematical framework for describing and analyzing gene regulatory networks," *J. Theor. Biol.*, vol. 176, no. 2, pp. 291–300, 1995.
[60] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian networks to analyze expression data," *J. Comput. Biol.*, vol. 7, no. 3-4, pp. 601–620, 2000.
[61] S. A. Kauffman, "Homeostasis and differentiation in random genetic control networks," *Nature*, vol. 224, no. 215, pp. 177–178, 1969.
[62] 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.
[63] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Rev. Modern Phys.*, vol. 74, no. 1, pp. 47–97, 2002.
[64] M. E. J. Newman, "The structure and function of complex networks," *SIAM Rev.*, vol. 45, no. 2, pp. 167–256, 2003.
[65] S. Strogatz, "Exploring complex networks," *Nature*, vol. 410, no. 6825, pp. 268–276, 2001.
[66] T. Ideker, V. Thorsson, J. A. Ranish, et al., "Integrated genomic and proteomic analyses of a systematically perturbed metabolic network," *Science*, vol. 292, no. 5518, pp. 929–934, 2001.

[67] D. M. Wolf and F. H. Eeckman, "On the relationship between genomic regulatory element organization and gene regulatory dynamics," *J. Theor. Biol.*, vol. 195, no. 2, pp. 167–186, 1998.
[68] S. A. Kauffman, "The large scale structure and dynamics of gene control circuits: an ensemble approach," *J. Theor. Biol.*, vol. 44, no. 1, pp. 167–190, 1974.
[69] L. Glass and S. A. Kauffman, "The logical analysis of continuous, non-linear biochemical control networks," *J. Theor. Biol.*, vol. 39, no. 1, pp. 103–129, 1973.
[70] S. A. Kauffman, *The Origins of Order: Self-organization and Selection in Evolution*, Oxford University Press, New York, NY, USA, 1993.
[71] S. Huang, "Gene expression profiling, genetic networks, and cellular states: an integrating concept for tumorigenesis and drug discovery," *J. Mol. Med.*, vol. 77, no. 6, pp. 469–480, 1999.
[72] 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.
[73] E. R. Dougherty, M. L. Bittner, Y. Chen, et al., "Nonlinear filters in genomic control," in *Proc. IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing*, Antalya, Turkey, June 1999.
[74] E. R. Dougherty, S. Kim, and Y. Chen, "Coefficient of determination in nonlinear signal processing," *Signal Process.*, vol. 80, no. 10, pp. 2219–2235, 2000.
[75] S. Kim, E. R. Dougherty, M. L. Bittner, et al., "General nonlinear framework for the analysis of gene interaction via multivariate expression arrays," *J. Biomed. Opt.*, vol. 5, no. 4, pp. 411–424, 2000.
[76] 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.
[77] A. Datta, A. Choudhary, M. L. Bittner, and E. R. Dougherty, "External control in Markovian genetic regulatory networks," *Machine Learning*, vol. 52, no. 1-2, pp. 169–191, 2003.
[78] A. Datta, A. Choudhary, M. L. Bittner, and E. R. Dougherty, "External control in Markovian genetic regulatory networks: the imperfect information case," *Bioinformatics*, vol. 20, no. 6, pp. 924–930, 2004.

Edward R. Dougherty: Department of Electrical Engineering, Texas A&M University, 3128 TAMU, College Station, TX 77843-3128, USA

*Email*: edward@ee.tamu.edu

Ilya Shmulevich: The Institute for Systems Biology, 1441 North 34th Street, Seattle, WA 98103-8904, USA

*Email*: is@ieee.org

Jie Chen: Division of Engineering, Brown University, Providence, RI 02912, USA

*Email*: jie_chen@brown.edu

Z. Jane Wang: Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC V6T 1Z4, Canada

*Email*: zjanew@eee.ubc.ca

# 1 Representation and analysis of DNA sequences

Paul Dan Cristea

## 1.1. Introduction

Data on genome structural and functional features for various organisms is being accumulated and analyzed in laboratories all over the world, from the small university or clinical hospital laboratories to the large laboratories of pharmaceutical companies and specialized institutions, both state owned and private. This data is stored, managed, and analyzed on a large variety of computing systems, from small personal computers using several disk files to supercomputers operating on large commercial databases. The volume of genomic data is expanding at a huge and still growing rate, while its fundamental properties and relationships are not yet fully understood and are subject to continuous revision. A worldwide system to gather genomic information centered in the National Center for Biotechnology Information (NCBI) and in several other large integrative genomic databases has been put in place [1, 2]. The almost complete sequencing of the genomes of several eukaryotes, including man (*Homo sapiens* [2, 3, 4]) and "model organisms" such as mouse (*Mus musculus* [5, 6]), rat (*Rattus norvegicus* [7]), chicken (*Gallus-gallus* [8]), the nematode *Caenorhabditis elegans* [9], and the plant *Arabidopsis thaliana* [10], as well as of a large number of prokaryotes, comprising bacteria, viruses, archeia, and fungi [1, 2, 5, 11, 12, 13, 14, 15, 16, 17, 18, 19], has created the opportunity to make comparative genomic analyses at scales ranging from individual genes or control sequences to whole chromosomes. The public access to most of these data offers to scientists around the world an unprecedented chance to data mine and explore in depth this extraordinary information depository, trying to convert data into knowledge.

The standard symbolical representation of genomic information—by sequences of nucleotide symbols in DNA and RNA molecules or by symbolic sequences of amino acids in the corresponding polypeptide chains (for coding sections)—has definite advantages in what concerns storage, search, and retrieval of genomic information, but limits the methodology of handling and processing genomic information to pattern matching and statistical analysis. This methodological limitation

# 2 Gene feature selection

Ioan Tabus and Jaakko Astola

This chapter presents an overview on the classes of methods available for feature selection, paying special attention to the problems typical to microarray data processing, where the number of measured genes (factors) is extremely large, in the order of thousands, and the number of relevant factors is much smaller. The main ingredients needed in the selection of an optimal feature set consist in: the search procedures, the underlying optimality criteria, and the procedures for performance evaluation. We discuss here some of the major classes of procedures which are apparently very different in nature and goals: a typical Bayesian framework, several deterministic settings, and finally information-theoretic methods. Due to space constraints, only the major issues are followed, with the intent to clarify the basic principles and the main options when choosing one of the many existing feature selection methods.

## 2.1. Introduction

There are two major distinct goals when performing gene feature selection: the first is *discovering the structure* of the genetic network or of the genetic mechanisms responsible for the onset and progress of a disease; the second is eliminating the irrelevant genes from a classification (or prediction) model with the final end of *improving the accuracy* of classification or prediction. While there are many cases when both goals are equally relevant, there are others when only one of them is of primary focus.

This possible distinction of goals is certainly reflected at the methodological level, where the feature selection methods are usually split into two groups: filter methods and wrapper methods [1]. With *the filter methods* [2, 3], the genes are ranked according to some general properties (correlation, mutual information, discriminative power) that are relevant for the prediction or classification problem at hand (e.g., correlation with a disease type), but without making it explicit at this stage what is the particular prediction model that is going to be used subsequently. After ranking of the single genes or of the various groups of genes, a suitable set of genes is identified and proposed as the feature set to be used for all subsequent

# 3 Classification

Ulisses Braga-Neto and Edward R. Dougherty

## 3.1. Introduction

Classification plays an important role in genomic signal analysis. For instance, cDNA microarrays can provide expression measurements for thousands of genes at once, and a key goal is to perform classification via different expression patterns. This requires designing a classifier (decision function) 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 a host of such differences [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12] (see also the bibliography on microarray-based classification provided as part of the supplementary information to [13]). Classifiers are designed from a sample of expression vectors. This involves assessing expression levels from RNA obtained from the different tissues with microarrays, determining genes whose expression levels can be used as classifier features (variables), and then applying some rule to design the classifier from the sample microarray data. Expression values have randomness arising from both biological and experimental variability. Design, performance evaluation, and application of features must take this randomness into account. Three critical issues arise. 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 are limited? Third, given a large set of potential features, such as the large number of expression levels provided by each microarray, how does one select a set of features as the input to the classifier? Small samples (relative to the number of features) are ubiquitous in genomic signal processing and impact all three issues [14].

## 3.2. Classifier design

Classification involves a *feature vector* $\mathbf{X} = (X_1, X_2, \ldots, X_d)$ on $d$-dimensional Euclidean space $\mathbb{R}^d$, composed of random variables (*features*), a binary random

# 4 Clustering: revealing intrinsic dependencies in microarray data

Marcel Brun, Charles D. Johnson, and Kenneth S. Ramos

## 4.1. Introduction

Informal definitions for clustering can be found in the literature: the process of "unsupervised classification of patterns into groups" [1], the act of "partitioning of data into meaningful subgroups" [2], or the process of "organizing objects into groups whose members are similar in some way" [3]. In the context of pattern recognition theory, the objects are represented by vectors of *features* (the measurements that represent the data), called *patterns*. With these concepts in mind, clustering can be defined as the process of *partitioning the vectors into subgroups whose members are similar relative to some distance measure*. Therefore, two key questions that must be addressed prior to cluster implementation are about the distance to be used to measure the similarity of the objects and how to form the partitions that best group together these objects.

The answer to the first question depends on each particular problem where clustering is applied. The distance between patterns should reflect the relation that is considered significant for the analysis. The rationale for each distance measure will be addressed in this chapter. The second question relates to computational efficiency considerations and criteria to evaluate the quality of clustering. This too is dependent on the question being proposed.

The chapter is divided into four sections with several examples at the end. The section on Clustering Microarray Data introduces the application of clustering to microarray data, illustrating the practical aspects of these techniques. Measures of Similarity develops the topic of distance measures. The next section, Clustering Algorithms, presents the implementation of popular algorithms and their applicability to microarray data analysis. Lastly, the final section, Interpretation and Validation, discusses the available procedures to measure the validity of the resulting partitions, showing several examples of clustering applied to microarray data to solve specific biological questions.

## 4.2. Clustering microarray data

Data clustering has been used for decades in image processing and pattern recognition [4], and in the last several years it has become a popular data-analysis

# 5 From biochips to laboratory-on-a-chip system

Lei Wang, Hongying Yin, and Jing Cheng

Biochip-based systems have enjoyed impressive advancement in the past decade. A variety of fabrication processes have been developed to accommodate the complicated requirements and materials for making such a device. Traditional microfabrication processes and other newly developed techniques such as plastic molding and microarraying are being explored for fabricating silicon, glass, or plastic chips with diverse analytical functions for use in basic research and clinical diagnostics. These chips have been utilized to facilitate the total integration of three classic steps involved in all biological analyses, that is, sample preparation, biochemical reaction, and result detection and analysis, and finally construct fully integrated smaller, more efficient bench-top or even handheld analyzers—laboratory-on-a-chip system. Meanwhile, biochip-based analytical systems have demonstrated diversified use such as the analyses of small chemical compounds, nucleic acids, amino acids, proteins, cells, and tissues. In this chapter, aspects related to biochips with different functionality and chip-based integrated systems will be reviewed.

## 5.1. Technologies for fabricating biochips

Depending on the materials used, micromachining technologies employed for fabricating the biochips can be very different. Photolithographic processing techniques are by far the most commonly used methods for producing microchannels in the surface of a planar silicon or glass substrate. One advantage of using these materials is that their electrophoretic and chromatographic properties and surface derivatization chemistries are extensively studied in many cases. Another advantage is that many established microfabrication processes could be easily modified and applied. Injection-molding, casting, imprinting, laser ablation, and stamping processes represent another category of fabrication methods for machining plastic substrate. The advantage for using plastic as substrate is twofold. One is that plastic is less expensive and easier to manipulate than glass or silicon-based substrates. Another advantage is the easiness in disposing it after use. The third category of methods for fabricating one type of the most widely used biochips, that is, microarrays, is robotic station-based microdispensing methods.

# 6 Modeling and simulation of genetic regulatory networks by ordinary differential equations

Hidde de Jong and Johannes Geiselmann

A remarkable development in molecular biology has been the recent upscaling to the genomic level of its experimental methods. These methods produce, on a routine basis, enormous amounts of data on different aspects of the cell. A large part of the experimental data available today concern genetic regulatory networks underlying the functioning and differentiation of cells. In addition to high-throughput experimental methods, mathematical and computational approaches are indispensable for analyzing these networks of genes, proteins, small molecules, and their mutual interactions. In this chapter, we review methods for the modeling and simulation of genetic regulatory networks. A large number of approaches have been proposed in the literature, based on such formalisms as graphs, Boolean networks, differential equations, and stochastic master equations. We restrict the discussion here to ordinary differential equation models, which is probably the most widely used formalism. In particular, we compare nonlinear, linear, and piecewise linear differential equations, illustrating the application of these models by means of concrete examples taken from the literature.

## 6.1. Introduction

A remarkable development in molecular biology today is the upscaling to the genomic level of its experimental methods. Hardly imaginable only 20 years ago, the sequencing of complete genomes has become a routine job, highly automated and executed in a quasi-industrial environment. The miniaturization of techniques for the hybridization of labeled nucleic acids in solution to DNA molecules attached to a surface has given rise to DNA microarrays, tools for measuring the level of gene expression in a massively parallel way [1]. The development of proteomic methods based on two-dimensional gel electrophoresis, mass spectrometry, and the double-hybrid system allows the identification of proteins and their interactions on a genomic scale [2].

These novel methods in genomics produce enormous amounts of data about different aspects of the cell. On one hand, they allow the identification of interactions between the genes of an organism, its proteins, metabolites, and other small

# 7 Modeling genetic regulatory networks with probabilistic Boolean networks

Ilya Shmulevich and Edward R. Dougherty

## 7.1. Introduction

High-throughput genomic technologies such as microarrays are now allowing scientists to acquire extensive information on gene activities of thousands of genes in cells at any physiological state. It has long been known that genes and their products in cells are not independent in the sense that the activation of genes with subsequent production of proteins is typically jointly dependent on the products of other genes, which exist in a highly interactive and dynamic regulatory network composed of subnetworks and regulated by rules. However, discovering the network structure has thus far proved to be elusive either because we lack sufficient information on the components of the network or because we lack the necessary multidisciplinary approaches that integrate biology and engineering principles and computational sophistication in modeling. During the past several years a new mathematical rule-based model called probabilistic Boolean networks (PBN) has been developed to facilitate the construction of gene regulatory networks to assist scientists in revealing the intrinsic gene-gene relationships in cells and in exploring potential network-based strategies for therapeutic intervention (Shmulevich et al. [1, 2, 3, 4, 5, 6], Datta et al. [7, 8], Kim et al. [9], Zhou et al. [10], and Hashimoto et al. [11]). There is already evidence that PBN models can reveal biologically relevant gene regulatory networks and can be used to predict the effects of targeted gene intervention. A key goal of this chapter is to highlight some important research problems related to PBNs that remain to be solved, in hope that they will stimulate further research in the genomic signal processing and statistics community.

## 7.2. Background

Data comprised of gene expression (mRNA abundance) levels for multiple genes is typically generated by technologies such as the DNA microarray or chip. The role

# 8 Bayesian networks for genomic analysis

Paola Sebastiani, Maria M. Abad, and Marco F. Ramoni

Bayesian networks are emerging into the genomic arena as a general modeling tool able to unravel the cellular mechanism, to identify genotypes that confer susceptibility to disease, and to lead to diagnostic models. This chapter reviews the foundations of Bayesian networks and shows their application to the analysis of various types of genomic data, from genomic markers to gene expression data. The examples will highlight the potential of this methodology as well as the current limitations and we will describe new research directions that hold the promise to make Bayesian networks a fundamental tool for genome data analysis.

## 8.1. Introduction

One of the most striking characteristics of today's biomedical research practice is the availability of genomic-scale information. This situation has been created by the simultaneous but not unrelated development of "genome-wide" technologies, mostly rooted in the Human Genome Project: fast sequencing techniques, high-density genotype maps, DNA, and protein microarrays. Sequencing and genotyping techniques have evolved into powerful tools to identify genetic variations across individuals responsible for predispositions to some disease, response to therapies, and other observable characters known as phenotypes. Single-nucleotide polymorphisms (SNPs)—a single-base variation across the individuals of a population—are considered the most promising natural device to uncover the genetic basis of common diseases. By providing a high-resolution map of the genome, they allow researchers to associate variations in a particular genomic region to observable traits [1, 2]. Commercially available technology, such as the Affymetrix GeneChip Mapping 10 K Array and Assay Set (http://affymetrix.com), is able to simultaneously genotype 10 000 SNPs in an individual. Other technologies are able to interrogate the genomic structure of a cell on a genome-wide scale: CGH microarrays are able to provide genome-wide identification of chromosomal imbalances—such as deletions and amplifications—that are common rearrangements in most tumors [3]. These rearrangements identify different tumor types or stages and this technology allows us to dive into the mutagenic structure of tumor tissues.

# 9 Statistical inference of transcriptional regulatory networks

Xiaodong Wang, Dimitris Anastassiou, and Dong Guo

We give a general overview of modeling of gene regulatory networks and discuss various statistical inference problems related to these models. First various gene function modeling techniques are described, including qualitative models such as directed and undirected graphs, Boolean networks, and logic networks, and quantitative models, including differential equations, linear and nonlinear function models, and radial basis functions. Then parameter estimation methods are discussed for known network structures, including equation-based methods and Bayesian methods. Finally, Bayesian techniques for inferring network structures are discussed.

## 9.1. Introduction

A central theme of molecular biology is to understand the regulatory mechanism that governs gene expressions in cells. The gene expression is controlled at different levels by many mechanisms, among which a key mechanism is mRNA transcription regulated by various proteins, known as transcription factors, which are bound to specific sites in the promoter region of a gene that activate or inhibit transcription. Using advanced molecular biology techniques, it has become possible to measure the gene expression levels (mRNA levels) of most genes in an organism simultaneously, hence making it possible to understand gene regulation and interactions.

In general, inference of a gene regulatory network is composed of three principal components: function modeling of the effect of a group of genes on a specific target gene, parameter estimation for function modeling of a specific network, and topology inference of regulatory network. As most genetic regulatory systems of interest involve many genes connected through interlocking positive and negative feedback loops, function modelings of interactions are important to unambiguously describe the structure of regulatory systems while predictions of their behavior can be made in a systematic way. Formal methods for the function modeling can be roughly categorized into qualitative models (such as graph models [1], Boolean function models [2, 3], and extended logical function models [4, 5]),

# 10 Compressing genomic and proteomic array images for statistical analyses

Rebecka Jörnsten and Bin Yu

Information technology advancements are bringing about innovations for genomic and proteomic research. One such innovation is the array imaging technology based on which gene or protein expression levels are derived. These images have a fundamentally different purpose to serve than the traditional still images: they are for statistical information extraction, not for visual inspection or comparison. Due to the huge quantity of such images and the limited bandwidth for their sharing among different researchers, for both storage and transmission goals, these images need to be compressed. Dictated by the statistical analyses to follow, in this chapter we lay out a multilayer data structure as the principle for both lossless and lossy compression of array images. We illustrate this principle in the example of cDNA microarray image compression with results of an average of near 2 : 1 lossless compression ratio and an average of 8 : 1 lossy compression ratio. The lossless ratio is comparable with the off-the-shelf lossless compression scheme LOCO, but with the added benefit of a handy structure for statistical analysis; the lossy ratio is obtained with a quantization noise level comparable to that of the imaging technology or the variation between two replicate imaging experiments.

## 10.1. Introduction

We live in an exciting era of technology innovations with all their advantages (and disadvantages). These innovations are fueling, if not driving, the progresses in genomic research (the study of genetic material such as DNA and RNA), and the newer proteomics research (the study of proteins which are directly responsible for actions in cells).

A revolutionary innovation has been the DNA microarray imaging technology for genomic research and it takes different forms: cDNA (P. Brown, http://www-genome.stanford.edu/), Affymetrics gene chips (http://www.affymetrix.com/index.affx), and Inkjet (http://www.rii.com). It provides measurements of mRNA (messenger RNA) material existing in cells to develop an understanding of gene function, gene regulation, and gene interaction through a simultaneous study of expression levels of thousands of genes. Microarrays are also used extensively in

# 11 Cancer genomics, proteomics, and clinic applications

X. Steve Fu, Chien-an A. Hu, Jie Chen,
Z. Jane Wang, and K. J. Ray Liu

## Preface

Throughout the history of medicine, many advances are derived from important innovations in technology. For example, the invention of the X-Ray machine has revolutionized medicine and pioneered modern imaging. The invention of the microscope essentially redefined the field of pathology and microbiology. In the past few decades, "technology explosion" has created an immense impact on both biomedical research and clinical medicine. Tremendous strides were made with the aid of numerous new technologies such as recombinant DNA methods, DNA sequencing, magnetic resonance imaging (MRI), polymerase chain reaction (PCR), monoclonal antibodies, and so forth. Despite these, major hurdles remain. In the field of cancer medicine, limited successes are still overshadowed by the tremendous morbidity and mortality incurred by this devastating disease. It has become increasingly important to integrate new technologies into both cancer research and clinical practice if we hope to win the battle against cancer.

In this chapter, we will briefly review the molecular basis of cancer and our current understanding. We will focus our attention on genomics and proteomics of cancer. We believe that a thorough understanding of the DNA and protein compliments of cancers that dictate the subsequent disease phenotype would eventually lead to breakthroughs. The impact of modern technology on cancer diagnosis, prognosis, and treatment will also be discussed. We placed our emphasis on two of the cutting-edge technologies, microarray technology and nanotechnology, as they are clearly among the leading frontiers that will rapidly reshape biomedical sciences and clinical oncology. Finally, we will discuss our current active research to facilitate our understanding and management of cancer.

## 11.1. Understanding cancer

### 11.1.1. Overview

The financial and societal burden of common diseases such as cardiovascular, metabolic (e.g., diabetes), and neoplastic diseases (cancer) is very significant.

# 12 Integrated approach for computational systems biology

Seungchan Kim, Phillip Stafford,
Michael L. Bittner, and Edward B. Suh

## 12.1. Background

New technological advancements for the measurement of biological systems have given us much insight into genomic, transcriptomic, and proteomic views of a cell's behavior. Such recent advancements in the measurement technology include expression arrays [1], single nucleotide polymorphism (SNP) [2, 3], CpG island arrays [4], protein abundance and specialized glycoarrays [5, 6], and siRNA [7, 8, 9]. Different measurement techniques are meant to provide different kinds and resolutions of the information regarding target biological systems; therefore, choosing appropriate measurements for a given biological problem is considered fundamental in the solution of the problem. In addition to the technologies that provide a unique snapshot of different aspects of the cellular milieu, we now have the computational and data management challenge of storing, integrating, and analyzing data independently and when mixed. Data storage techniques become increasingly important when integration, and analysis are needed. Database design and planning are now as important as the analysis technologies that are being developed.

Biological problems of special importance now include the recognition of disease subtypes, identification of molecular markers for certain disease types, inference of regulatory mechanisms, discovery of new therapeutic targets for intervention and treatment of disease progression, and the development of novel single and additive drugs and therapeutics. Since the beginning of the modern biological era, the importance and applicability of mathematical, statistical and engineering tools has become quite clear. The Human Genome Project is a primary example. Numerous pattern recognition techniques have been applied to identify molecular markers for a specific disease as well as the identification of disease subtypes. Machine learning and Bayesian frameworks have proven to be effective in learning the mechanisms of genetic regulatory networks, and control theory is being applied to derive a better approach to therapeutic design. As the complexity of biological data increases, it is the combination, not a single specialized tool, which will be most efficacious to solving complex biological problems.