

Chapter 1

Introduction

Computational systems bioinformatics is a new and rapidly developing field of research that focuses on understanding the structures and processes of biological systems at molecular, cellular, tissue, and organ levels by computational modeling and novel information theoretic data- and image-analysis methods. With the breakthrough in deciphering the human genome and by the use of modern experimental biology and the most up-to-date computational approaches, it has become possible to understand the structure and function of biomolecules, DNA, RNA, nucleic acids' expressions into proteins, protein structure proteomics, metabolic pathways and networks, intracellular and intercellular signaling, and the physical and chemical mechanisms involved in all of these. By applying information theoretic and computational modeling methodologies to experimental genotype and phenotype data obtained through the application of microarrays, gels, and mass spectroscopy to proteins as well as through molecular and cellular high-content imaging and microscopy, it is possible to understand the structure and function of biosystems.

Generally speaking, computational systems bioinformatics focuses on either information processing of biological data or modeling of physical and chemical processes. Through this type of quantitative systems approach, computational systems bioinformatics can play a central role in disease prediction and preventive medicine, in gene technology and pharmaceuticals, and in other biotechnology

fields. In this book, we introduce the necessary mathematical, statistical, and data mining principles, and then present step-by-step bioinformatics technology for microarray and sequence analysis. We also present protein structure informatics as well as molecular and bioimaging informatics for subcellular and cellular environments.

What is systems bioinformatics? The emerging area of systems bioinformatics is a whole-silico approach to understanding biology.¹ It aims at a systems-level understanding of biology. It examines the structure and dynamics of cellular and organism functions instead of the characteristics of isolated parts of a cell or organism.^{3,4} Many properties of life arise at the systems level only, as the behavior of the system as a whole cannot be explained by its constituents alone.

Biological knowledge is growing very rapidly and data analysis can hardly keep pace. In the bioinformatics area, tools have been developed — and will become more advanced — that can handle the huge and rapidly growing amount of data stored in biological databases. One main effort is aimed at grouping and comparing data in order to gain information about individual molecules in comparison with molecules that are similar. In structural biology, there have been efforts to predict the three-dimensional structures of proteins and DNA by homology and *de novo* methods. In the latter case, the interactions between atoms within a molecule are modeled and advances in computer hardware are making computationally intensive simulations possible. Simulation for *de novo* structure prediction has had some limited successes so far; for example, it has been successfully applied in the refinement of experimentally defined structures.

Systems bioinformatics takes a different approach: it tries to integrate biological knowledge and understand how molecules act together within the network of interactions which comprise life. Again, model building promises to be the key in advancing understanding. The exploding amount of biological data and the predicament that it cannot be understood by simply drawing lines between interacting molecules demonstrate the demand for a systematic approach. The expectations around this are highlighted by numerous articles in leading scientific journals⁵; by the opening of numerous centers worldwide devoted to systems bioinformatics; and by the funding

of many research collaborations which bring together expertise in mathematics, information science, and biology.

However, most biological data so far are qualitative rather than quantitative; and many additional breakthroughs in experimental devices, advanced software, and analytical methods will be required before the achievements of systems bioinformatics can live up to its much-touted potential.³ Whereas many of the systems bioinformatics approaches still focus on data mining, the integration, processing, and representation of large and heterogeneous metadata have to be considered, as a great deal of further development is needed at each level.

Nevertheless, systems bioinformatics is much needed and its modeling approaches are, in principle, quite powerful. The quantity of data encountered in systems bioinformatics is familiar to engineers, such as those designing the control systems for modern passenger jets.² Model building to aid the understanding of complex systems is familiar and is the method of choice in areas like ecology and economics. Therefore, the lessons from systems analyses of other advanced technologies and from engineering theory suggest that systems can be divided into subsystems so that one does not have to tackle and solve the whole system all at once, and that work on systems bioinformatics promises to be worth the effort.

Cells, tissues, organs, and organisms can be considered as complex, interacting systems of molecules. At all levels, these systems have constantly been defined, refined, and optimized through billions of years of evolution. Modern biological research seeks to simplify, categorize, and study the components and interactions that make up what we define as a “biological system.” However, advances in biological knowledge will require integration of the biological data on many information levels. Computational systems bioinformatics aims at an integrative, systems-level understanding of biological systems by analyzing quantities of experimental biological data through the use of computational techniques such as model building. The field studies interacting systems by defining the basic structures of the biological network in a living cell, by examining how biological systems respond to changing conditions so as to maintain robustness

and stability, and by making predictions based on our modeling results.

Molecular biology has until now focused mainly on the investigation of individual molecules, i.e. on their properties as isolated entities or as complexes in many simple systems. However, biological molecules in living systems participate in complex networks, including regulatory networks for gene expression, intracellular metabolic networks, and intracellular and intercellular communication networks. Such networks are involved in the maintenance (homeostasis) and differentiation of the cellular systems of which we have at present an incomplete understanding.

Nevertheless, the progress of molecular biology has already made possible detailed descriptions of the components that constitute living systems, notably in the areas of genes and proteins. Large-scale genome sequencing means that we can delineate all genomic components of a given cellular system; and microarray experiments as well as large-scale proteomics will soon give us large amounts of experimental data on gene regulation, molecular interactions, and cellular networks. The challenge of the 21st century will be to understand how these individual components integrate into complex systems and to understand the function and evolution of these systems, thus allowing the understanding to scale up from molecular biology to systems biology.

In Chapter 2, we will review some mathematical, statistical, signal processing, pattern recognition, and optimal control principles that are often used in data mining for bioinformatics.

In Chapter 3, we will introduce missing value estimation using the K -nearest neighbor (KNN) and Bayesian methods. Data from microarray experiments are usually in the form of large matrices of expression levels of genes under different experimental conditions. Owing to various reasons, certain experimental values are frequently missing. Estimating these missing values is important because they affect downstream analysis, including clustering, classification, and network design. Several methods of missing value estimation are in use. The problem has two parts: (1) the selection of genes for estimation, and (2) the design of an estimation rule. We will first

review some existing methods, including a row average method, a singular value decomposition method (SVDimpute), and a weighted K -nearest neighbor method (KNNimpute). We will then present our Bayesian variable selection method for choosing the genes to be used for the estimation, and we will present both linear and nonlinear regression methods for the estimation rule itself. Fast implementation issues for these methods will be discussed, including the use of QR decomposition for parameter estimation. Our methods will be tested on data sets arising from the study of hereditary breast cancers and small round blue cell tumors.

In Chapter 4, we will discuss normalization, scaling, and discretization. We will first review the well-known normalization algorithms and scaling algorithms such as log-, mean-, and median-based normalization and scaling methods as well as Z -score-based normalization and optimal-method-based scaling. Then, we will present our discretization algorithm. Although gathered as continuous data, expression measurements from gene microarrays may be quantized prior to downstream analysis and modeling. This is especially useful for modeling gene prediction and genetic regulatory networks. Coarse quantization results in lower computational requirements, lower data requirements for model inference, and easier conceptualization. We will discuss our mixture model for binarization of microarray: for each gene, the model composed of a Gaussian mixture is fit to the expression data for that gene, and data points are binarized according to the model. This mixture model is based on the assumption of multiplicative upregulation. This method will be compared to mean and median binarizations by comparing these methods' classification performances on binary data from the different experiments. Classification examples will be given for simulated data generated from a previously studied microarray model as well as for cancer data arising from two studies involving hereditary breast cancer and small round blue cell tumors of childhood.

In Chapter 5, we will present a cross-platform comparison. The amount of microarray data is increasing at a tremendous rate, and many different platforms exist for the generation of expression data.

A problem arises because the expression data taken by different platforms is quite different, even for the same genes and the same tissues. The questions are thus: Are there significant differences between the same platform's plates? Are there significant differences between different platforms? How can one perform meta-analysis on cross-platform microarray data? We will address the problems of cross-platform normalization, comparison, and meta-analysis not just separately, but also all three in combination. We will examine repeatability within one platform by use of quartile normalization. We will first discuss normalization for cross-platform comparison. In the cross-platform comparison, we will compare the traditional Pearson correlation and rank correlation, as well as coinertia analysis, and we will discuss our principal component analysis (PCA)- and independent component analysis (ICA)-based comparison methods. These methods allow us to compare different aspects of different platforms. In meta-analysis, we will demonstrate how to employ a two-component model to detect differential genes from cross-platform microarray data. As an example, we will apply these methods to brain research data from six different microarray platforms.

In Chapter 6, we will present an overview of the discovery of differential biomarkers. DNA microarrays are a new and promising biotechnology that allow the simultaneous monitoring of the expression levels in cells of thousands of genes. An important and common goal of microarray experiments is the identification of differentially expressed genes, i.e. genes whose expression levels are associated with a response or a covariate of interest. The biological question of differential expression can be restated as a problem in multiple hypothesis testing — the simultaneous test for each gene of the null hypothesis (of no association between the expression level and the response or covariate). As a typical microarray experiment measures expression levels for thousands of genes simultaneously, large-multiplicity statistical problems are routinely generated. We will review different approaches to multiple hypothesis testing in the context of microarray experiments. The methods to be discussed will be based on p -values, false discovery rates,

t-tests, analyses of variance (ANOVAs), and significance analyses. Examples of these technologies as applied to cancer research will be given.

In Chapter 7, we will discuss gene selection and classification. Given the thousands of genes and the small number of data samples involved in microarray-based classification, gene selection is a critical issue. We will review several existing, well-known methods based on familiar algorithms, including support vector machine (SVM), genetic algorithm, perceptron, Bayesian variable selection, and minimum description length for model selection. Then, we will present several methods that we have developed: gene-selection-based and classification-based linear probit regression, nonlinear probit regression, and logistic regression; mutual-information-based selection; and SVM and fuzzy SVM selection and classification. These algorithms will then be applied to cancer prediction and diagnoses.

In Chapter 8, gene clustering and function analysis will be discussed. Cluster analysis of genewide expression data from DNA microarray hybridization studies has proven to be a useful tool for the identification of biologically relevant groupings of genes and the construction of gene regulatory networks. We will review several existing methods, including linkage, *K*-means, fuzzy *c*-means, self-organization map, and graphic theory-based clustering. Then, we will present a clustering strategy of ours that is based on minimizing the mutual information between gene clusters. Simulated annealing will be employed to solve the optimization problem. Bootstrap techniques will be employed to get accurate estimates of the mutual information when the data sample size is small. Then, we will combine mutual information criteria and traditional distance criteria, such as the Euclidean distance and the fuzzy membership metric, in this clustering algorithm. Its performance will be compared with that of some existing methods, using both synthesized data and experimental data, and we will see that clustering based on a combined metric of mutual information and fuzzy membership achieves the best performance. This algorithm will then be applied to some cancer microarray data sets.

In Chapter 9, gene prediction and function prediction will be discussed. A critical issue in the construction of genetic regulatory networks is the identification of network topologies from data. In the context of deterministic and probabilistic Boolean networks as well as in their extension to multilevel quantization, this issue is related to the more general problem of expression prediction, in which we intend to find small subsets of genes to be used as predictors of target genes. Even given some maximum number of predictors to be used, a full search of all possible predictor sets is combinatorially prohibitive except for small predictor sets, and even then may require supercomputing-class processing power. Hence, suboptimal approaches to finding predictor sets and network topologies are desirable. We will first review some existing methods, such as the perceptron model using coefficients of determination, linear regression, and neural network. Then, we will present a Bayesian variable selection method for prediction using a multinomial probit regression model with data augmentation to turn the multinomial problem into a sequence of smoothing problems. The probit regressor discussed will be approximated as a linear combination of the genes, and a Gibbs sampler will be employed to find the strongest genes. Numerical techniques to speed up the computation will be discussed. Two predictor models — the estimated probit regressors and the optimal full-logic predictor based on the selected strongest genes — will be compared to optimal prediction without feature selection based on their performance on malignant melanoma microarray data.

Chapter 10 talks about reverse engineering: the construction of gene regulatory networks and their behavior. This is an extremely challenging topic on which many scientists are working. We will review common gene regulatory network models, including those that employ linear predictor, neural network, differential equation, Bayesian network, Boolean network, and probabilistic Boolean network (PBN) methods. Then, we will introduce two of our methods.

The first method addresses the construction of the PBNs. First, we will discuss how the number of possible parent gene sets and input sets of gene variables corresponding to each gene can be

determined by a clustering technique based on mutual information minimization; simulated annealing will be discussed as a solution to the optimization problem. After obtaining initial knowledge about each gene, we discuss the restriction of different functions from the possible parent gene sets to each target gene. Second, we will discuss how to model each function by a two-layer perceptron with a linear term plus a nonlinear term. A reversible-jump Markov chain Monte Carlo (MCMC) annealing method will then be used to calculate the model order and the parameters. Finally, coefficients of determination (CoDs) will be employed to compute the probability of selecting different predictors for each gene. A “leave-one-out cross-validation” algorithm will be discussed for the adjustment of the partition found by mutual information. Examples of this approach to the construction of PBNs will be given using data from known gene response pathways, including ionizing radiation and downstream targets of inactivating gene mutations.

The second method hypothesizes that the construction of transcriptional regulatory networks by the optimization of connectivity will lead to regulations that are consistent with biological expectations. A key expectation will be that the hypothetical networks should produce a few, very strong attractors which are highly similar to the original observations and which mimic the true biological-state stability and determinism. Another expectation will be that, since it is expected that biological controls are distributed and mutually reinforcing, interpretation of the observations should lead to a very small number of connection schemes. We will present a fully Bayesian method of constructing probabilistic gene regulatory networks (PGRNs) that emphasizes network topology. This method computes the possible parent sets of each gene, the corresponding predictors, and the associated probabilities by a nonlinear perceptron model using a reversible-jump MCMC technique; an MCMC method will be employed to search the network configurations to find those with the highest Bayesian scores in the construction of PGRNs. This Bayesian method will then be used to construct a PGRN based on the observed behavior of a set of genes whose expression patterns vary across a set of melanoma samples

exhibiting two very different phenotypes with respect to cell motility and invasiveness.

In Chapter 11, we will discuss motif detection and transcription factor binding site identification. Since the basic local alignment search tool (BLAST) was invented for sequence alignment, there have been numerous local sequence alignment algorithms developed for motif identification, promoter prediction, and transcription factor binding site identification. For motif detection, we will review two very famous algorithms: MEME and Gibbs sampler. For promoter prediction, we will review some of the existing software (these software can be found by simply typing their name in Google): the position–weight matrix (PWM)-based programs PromoterScan (EPD), TATA, and GeneID; the artificial neural network (ANN)-based programs NN Promoter Prediction (NNPP), Dragon Promoter Finder (and its CG+ version), and Promoter 2.0; the discriminant analysis (hyperplane)-based programs FirstEF (which is nonlinear and was first developed by M. Zhang), CorePromoter (which is also nonlinear and was first developed by M. Zhang), CpG Promoter (which is, again, nonlinear and was first developed by M. Zhang), and TSSG and TSSW (which are linear and LDA-based); the Markov chain–based and HMM-based program McPromoter (which is CpG-based); the threshold-based program CpGProD; and two programs whose features are defined by equivalence classes which allow a fuzzy description of the promoter context, PromFind and PromoterInspector. Although there are many existing methods for promoter prediction, few people have really studied how to integrate the three features: CpG Islands, PWM, and interpolated Markov chain. An example of this will be presented, showing how to apply this technology to the study of microRNA functions in mammalian neuronal development.

In Chapter 12, we will examine gene interaction through the combination of microarray and sequence technologies. A central goal of molecular biology is to discover the regulatory mechanisms governing the expression of genes in cells. The expression of a gene is controlled by many mechanisms. A key factor in these mechanisms is mRNA transcription regulation by various proteins known

as transcription factors (TFs), which bind to specific sites in the promoter region of a gene that activate or inhibit transcription. The target genes of TFs have been determined in several different ways. We will discuss the identification of TF targets and motifs by the well-known free software MEME and by matching to TF databases. We will focus on mining gene expression data, since these data provide a direct measurement of the transcriptional procedure in cells. Standard methods have predicted gene and protein function and interaction by clustering genes with similar expression profiles. However, the gene expression relationship between a TF and its target is complex; in most cases, their expression profiles do not correlate over time. We will start with genes that have already been selected (by the methods discussed in other chapters), and we will refine the clustering results of those genes by integrating their transcriptional factor binding site and gene expression data using Bayesian networks.

In Chapter 13, we will discuss protein structure informatics. We will first review some traditional protein structure comparison methods such as root mean square deviation (RMSD); class, architecture, topology, and homologous superfamily (CATH); and structure classification of proteins (SCOP). CATH and SCOP use Euclidean distances to compare protein structures. Since the traditional methods only consider the absolute distance between the C-alpha pairs of two proteins and do not consider proteins' spatial geometries and topological structures, their performance is very limited. Recently, researchers have employed the knot theory to measure proteins' geometrical invariants, but this mainly considers the orientation. Motivated by the knot theory, we will discuss two methods that we have developed to extract proteins' geometrical invariants from their secondary structure by considering the protein's spatial and geometrical properties. We can measure the structural similarity between proteins by correlating the principal components of their secondary structure interaction matrix. In this approach, referred to as principal component correlation (PCC), the (symmetric) matrix for an individual protein is constructed with the relationship parameters between secondary elements that can take the form of distance,

orientation, or other meaningful structural invariants. When using a distance-based construction encoded with N- to C-terminal sense, there are strong correlations in the principal components of the interaction matrix among topologically similar proteins. This finding has been extensively tested for protein structures and for domains that belong to the same topological class, but are identified by RMSD as being significantly different. Moreover, the correlation by RMSD is poor for proteins of similar shapes but different topological trajectories. In an example, we will compare a set of results from PCC analyses with those from CATH and from a knot-theory feature extraction method. The numerical results of this example will show that the PCC method is highly flexible in adopting various structural parameters as a general means of pairwise structure comparison.

Mass spectrometry is being increasingly used to detect disease-related biomarkers from some tissue samples for early diagnosis, prognosis, and monitoring of disease progression or response to treatment. It can be applied to differentiate patient samples from one another, such as diseased from normal, or to identify which patients are most likely to benefit from particular treatments. Surface-enhanced laser desorption/ionization–time of flight (SELDI-TOF) is one of such mass spectrometry technologies. Patterns of masses rather than actual protein identifications are produced by SELDI analysis. Chapter 14 provides an application example of the development of proteomics biomarkers via mass spectrometry. Due to the high dynamic range of the protein concentration in human blood, the application of proteomics technology for protein profiling can generate large arrays of data for the development of optimized clinical biomarker panels. The objective of this chapter is to review the existing methods and then give our approach to discover an optimized panel of biomarkers for predicting the risk of major adverse cardiac events (MACEs) in subjects.

In Chapter 15, we will discuss bioimaging informatics. During the past decade, cell biologists have acquired large numbers of light microscopy images from cells and tissues as a way of studying

cellular dynamics at different biological levels of complexity and resolution. Images are of such activities as cell movement; changes in cell shape in response to environmental changes; intracellular traffic of vesicles; responses to pathogens; cell shape changes in nucleic acids, proteins, and lipids; biogenesis of organelles; and, recently, movement and behavior of single molecules in a cell. There are, however, significant challenges in such high-content bioimaging, such as the accurate segmentation and tracking of the dynamic behavior of cells in a large population as well as the segmentation and tracking of thousands of moving particles/molecules within a cell. Traditional methods and the bioimage analysis tools derived from these methods are extremely limited in scope and capacity when tasked with the analysis of high-content live cell imaging. Currently, scientists have to resort to slow, manual analysis to extract information from such image series. Image processing and object modeling have become the rate-limiting factor in realizing the potential of dynamic cellular and molecular imaging studies.

Finally, we will conclude with a discussion on the progression, perspectives, and future of computational systems bioinformatics, including such topics as modeling issues for systems bioinformatics and quantitative analyses of biomolecular systems. The emerging field of systems bioinformatics involves the application of experimental, theoretical, and modeling techniques to the study of biological organisms at all levels — from the molecular, through the cellular, to the behavioral. Its aim is to understand biological processes as whole systems instead of as isolated parts. Developments in this field have been made possible by advances in molecular biology, in particular new technologies for determining DNA sequences, gene expression profiles, protein–protein interactions, and so on.

This book covers the central topics of computational systems bioinformatics: comprehensive and automated measurements, reverse engineering of genes and metabolic networks from experimental data, software issues, modeling and simulation, and systems-level analysis.

References

1. Chong L, Ray LB. Introduction to special issue Whole-istic Biology. *Science* **295**:1661, 2002.
2. Csete ME, Doyle JC. Reverse engineering of biological complexity. *Science* **295**:1664–1669, 2002.
3. Kitano H. Systems biology: a brief overview. *Science* **295**(5560): 1662–1664, 2002.
4. Kitano H. Computational systems biology. *Nature* **420**(6912):206–210, 2002.
5. *Science Special*, March 1, 2002.