PREFACE

Bioinformatics is the science of integrating, managing, mining, and interpreting information from biological data sets. Although tremendous progress has been made over the years, many of the fundamental problems in bioinformatics, such as protein structure prediction or gene finding, data retrieval, and integration, are still open. In recent years, high-throughput experimental methods in molecular biology have resulted in enormous amounts of data. Mining bioinformatics data is an emerging area of intersection between bioinformatics and data mining. The objective of this book is to facilitate collaboration between data mining researchers and bioinformaticians by presenting cutting-edge research topics and methodologies in the area of data mining for bioinformatics.

This book contains articles written by experts on a wide range of topics that are associated with novel methods, techniques, and applications of data mining in the analysis and management of bioinformatics data sets. It contains chapters on RNA and protein structure analysis, DNA computing, sequence mapping, genome comparison, gene expression data mining, metabolic network modeling, phyloinformatics, biomedical literature data mining, and biological data integration and searching. The important work of some representative researchers in bioinformatics is brought together for the first time in one volume. The topic is treated in depth and is related to, where applicable, other emerging technologies, such as data mining and visualization. The goal of the book is to introduce readers to the principal techniques of data mining in bioinformatics in the hope that they will build on them to make new discoveries of their own. The key elements of each chapter are summarized briefly below.

Progress in machine learning technology provided various advanced tools for prediction of protein secondary structure. Among the many machine learning

approaches, support vector machine (SVM) methods are the most recently applied in the structure prediction. It shows successful performance compared with other machine learning schemes. However, compared to other machine learning approaches, there is no systematic review about this SVM approach applied to secondary-structure prediction. In Chapter 1, H.-J. Hu, R. W. Harrison, P. C. Tai, and Y. Pan present methods for predicting secondary structure based on support vector machines. Evaluation of the performance of SVM methods, challenges of these SVM approaches, and efforts to overcome the problems are also discussed.

The problem of mining hidden links from complementary and noninteractive biomedical literature was exemplified by Swanson's pioneering work on the Raynaud's disease/fish oils connection. Two complementary and noninteractive sets of articles (independently created fragments of knowledge), when considered together, can reveal useful information of scientific interest not apparent in either of the two sets alone. In Chapter 2, X. Hu, X. Zhang, and X. Zhou discuss a comprehensive comparison of seven methods for mining hidden links among medical concepts. A series of experiments using these methods are performed and analyzed. Their research works present a comprehensive analysis for mining hidden links and how different weighting schemes plus semantic information affect the knowledge discovery procedure.

In Chapter 3, B. Jin and Y.-Q. Zhang discuss voting scheme–based evolutionary kernel machines used in drug activity comparisons. The performance of support vector machines is affected primarily by kernel functions. With the growing interest in biological data prediction and chemical data prediction, more complicated kernels are designed to measure data similarities: ANOVA kernels, convolution kernels, string kernels, tree kernels, and graph kernels. These kernels are implemented based on kernel decomposition properties. Experimental results show that the new kernel machines are more stable than the traditional kernels.

Microarrays are a well-established technology for measuring gene expression levels at a large scale. A tiling array is a set of oligonucleotide microarrays for the entire genome. Tiling array technology has several advantages over other microarray technologies. In Chapter 4, T. Joshi, J. Wan, C. J. Palm, K. Juneau, R. Davis, A. Southwick, K. M. Ramonell, G. Stacey, and D. Xu discuss the whole-genome tiling array design and techniques to analyzing these data to obtain a wide variety of genomic scale information using bioinformatics techniques. They also discuss ontological analyses and antisense identification techniques using tiling array data.

Identification of marker genes is of great importance to provide more accurate, cost-effective prediction, and a better understanding of genes' biological functions. In Chapter 5, J. Li, H. Su, and H. Chen discuss gene selection from high-dimensional microarray data. They present a framework for gene selection methods, focusing on optimal search–based gene subset selection. A comparative study of gene selection methods on three cancer data sets is presented with algorithmic details. Evaluation of both statistical and expert analysis is also presented.

In current practice, physicians assess the risk profile of a cancer patient based primarily on various clinical characteristics. However, these factors do not fully reflect the molecular heterogeneity of the disease, and treatment stratification is

difficult. In fact, in many cases, patients with a similar diagnosis respond very differently to the same treatment. In Chapter 6, H. Liu, L. Wong, and Y. Xu introduce a number of methods for predicting cancer survival based on gene expression data and provide some answers to a challenging question: Is there any relationship between gene expression profiles and patient survival?

In recent years, RNA interference (RNAi) has surged into the spotlight of pharmacy, genomics, and system biology. In Chapter 7, S. Qiu and T. Lane describe the biological mechanisms of RNAi, covering both secondary RNA and microRNA as interfering initiators, followed by in-depth discussions of a few computational topics, including RNAi specificity, microRNA gene and target prediction, and siRNA silencing efficacy estimation. The authors also highlight some open questions related to RNAi research.

Currently, our ability to produce sequence information far outpaces the rate at which we can produce structural and functional information. Consequently, researchers rely increasingly on computational techniques to extract useful information from known structures contained in large databases, although such approaches remain incomplete. Unraveling the relationship between pure sequence information and three-dimensional structure thus remains one of the great fundamental problems in molecular biology. In Chapter 8, H. Rangwala, K. DeRonne, and G. Karypis show several ways in which researchers try to characterize the structural, functional, and evolutionary nature of proteins.

Public genomic databases involve three main aspects of data use: data representation, data storage, and data access. In Chapter 9, A. Robinson, W. Rahayu, and D. Taniar present a comprehensive study of genomic databases covering these three aspects. Data representation in the biology domain consists of three principal structures: sequence-centric, annotation-centric, and XML formatting. Data storage is usually achieved by maintaining flat files or a management system such as a relational database management system. Genetic data access is provided by a varied range of user interfaces; the main groups are single-database access points, cross-referencing, multidatabase access points, and tool-based interfaces.

A vast number of unstructured data become a difficult challenge in text mining. To tackle this issue, M. Song and I.-Y. Song propose a novel technique that integrates information retrieval techniques with text mining. In Chapter 10, they present a new unsupervised query expansion technique that utilizes keyphrases and part-of-speech (POS) phrase categorization. The keyphrases are extracted from the documents retrieved and are weighted with an algorithm based on information gain and cooccurrence of phrases. The keyphrases selected are translated into disjunctive normal form based on the POS phrase categorization technique for better query refomulation. Additionally, the authors discuss whether ontologies such as WordNet and MeSH improve retrieval performance in conjunction with the keyphrases.

Understanding the molecular basis of complex formations between proteins as well as between modules and domains in multimodular protein systems is central to the development of strategies for human intervention in biological processes. In Chapter 11, L. S. Swapna, B. Offmann, and N. Srinivasan trace the drift of protein–protein interaction surfaces between families related at superfamily level or between

subfamilies related at family level when their interacting surfaces are not equivalent. They have investigated such evolutionary drifts in protein–protein interfaces in the class I glutamine amidotransferase–like superfamily of proteins and its constituent families.

Comparing and aligning RNA secondary structures is fundamental to knowledge discovery in biomolecular informatics. In recent years, much progress has been made in RNA structure alignment and comparison. However, existing tools either require a large number of prealigned structures or suffer from high time complexities. This makes it difficult for the tools to process RNAs whose prealigned structures are unavailable or to process very large RNA structure databases. In Chapter 12, J. T. L. Wang, D. Wen, and J. Liu present an efficient method, called RSmatch, for comparing and aligning RNA secondary structures. RSmatch can find the optimal global or local alignment between two RNA secondary structures. Also presented is a visualization tool, called RSview, which is capable of displaying the output of RSmatch in a colorful and graphic manner.

Despite an extensive effort to use computational methods in deciphering transcriptional regulatory networks, research on translation regulatory networks has attracted little attention in the bioinformatics and computational biology community, due probably to the nature of data available and to a bias in the conventional wisdom, In Chapter 13, D. D. Wu and X. Hu present a global network analysis of protein translation networks in yeast, a first step in attempting to facilitate elucidation of the structures and properties of translation networks. They extract the translation proteome using the MIPS functional category and analyze it in the context of the full protein–protein interaction network and derive individual translation networks from a full interaction network using the proteome extracted. They show that in contrast to a full network, protein translation networks do not exhibit power-law degree distributions. These results have potential implications for understanding mechanisms of translational control from a systems perspective.

Membrane proteins account for roughly one-third of all proteins and play a crucial role in processes such as cell-to-cell signaling, transport of ions across membranes, and energy metabolism, and are a prime target for therapeutic drugs. In Chapter 14, M. Q. Yang, J. Y. Yang, and C. W. Codrington emphasize the prediction of α-helical transmembrane regions in proteins using variants of the self-organizing feature map algorithm. To identify features that are useful for this task, the authors have conducted a detailed analysis of the physiochemical properties of transmembrane and intrinsically unstructured proteins.

In Chapter 15, L. Zhao and M. J. Zaki present a novel, efficient, deterministic, triclustering method called triCluster that addresses the challenge issues in three-dimensional microarray data analysis.

Clustering in the PPI network context groups together proteins that share a higher number of interactions. The results of this process can illuminate the structure of the PPI network and suggest possible functions for members of the cluster that were previously uncharacterized. In Chapter 16, C. Lin, Y.-R. Cho, W.-C. Hwang, P. Pei, and A. Zhang begin with a brief introduction to the properties of protein–protein interaction networks, including a review of data generated by both experimental and

computational approaches. A variety of methods that have been employed to cluster these networks are also presented. These approaches are broadly characterized as either distance- or graph-based clustering methods. Techniques for validating the results of these approaches are also discussed.

We would like to express our sincere thanks to all authors for their important contributions. We would also like to thank the referees for their efforts in reviewing the chapters and providing valuable comments and suggestions. We would like to extend our deepest gratitude to Paul Petralia (senior editor) and Whitney A. Lesch (editorial assistant) at Wiley for their guidance and help in finalizing the book. Xiaohua Hu would like to thank his parents, Zhikun Hu and Zhuanhui Wang; his wife, Shuetyue Tsang; and his son, Michael Hu, for their love and encouragement. Yi Pan would like to thank Sherry for friendship, help, and encouragement during preparation of the book.

XIAOHUA HU
YI PAN