Preface

Bioinformatics is the science of managing, analyzing, extracting, and interpreting information from biological sequences and molecules. It has been an active research area since late 1980's. After the human genome project was completed in April 2003, this area has drawn even more attention. With more genome sequencing projects undertaken, data in the field such as DNA sequences, protein sequences, and protein structures are exponentially growing. Facing this huge amount of data, the biologist cannot simply use the traditional techniques in biology to analyze the data. In order to understand the mystery of life, instead, information technologies are needed.

There are a lot of online databases and analysis tools for bioinformatics on the World Wide Web. Information technologies have made a tremendous contribution to the field. The database technology helps collect and annotate the data. And the retrieval of the data is made easy. The networking and Web technology facilitates data and information sharing and distribution. The visualization technology allows us to investigate the RNA and protein structures more easily. As for the analysis of the data, online sequence alignment tools are quite mature and ready for use all the time. But to conquer more complicated tasks such as microarray data analysis, protein-protein interaction, gene mapping, biochemical pathways, and systems biology, sophisticated techniques are needed.

Data mining is defined as uncovering meaningful, previously unknown information from a mass of data. It is an emerging field since mid 1990's boosted by the flood of data on the Internet. It combines traditional databases, statistics, and machine learning technologies under the same goal. Computer algorithms are also important in speeding up the process while dealing with a large amount of data. State-of-the-art techniques in data mining are, for example, information retrieval, data warehousing, Bayesian learning, hidden Markov model, neural networks, fuzzy logic, genetic algorithms, and support vector machines. Generally, data mining techniques deal with three major problems, i.e., classification, clustering, and association. In analyzing biological data, these three kinds of problems can be seen quite often. The technologies in data mining have been applied to bioinformatics research in the past few years with quite a success. But more research in this field is necessary since a lot of tasks are still undergoing. Furthermore, while tremendous progress has been made over the years, many of the fundamental problems in bioinformatics are still open. Data mining will play a fundamental role in understanding the emerging problems in genomics and proteomics. This book wishes to cover advanced data mining technologies in solving such problems.

The audiences of this book are senior or graduate students majoring in computer science, computer engineering, or management information system (MIS) with interests in data mining and applications to bioinformatics. Professional instructors and researchers will also find that the book is very helpful. Readers can benefit from this book in understanding basics and problems of bioinformatics, as well as the applications of data mining technologies in tackling the problems and the essential research topics in the field.

The uniqueness of this book is that it covers important bioinformatics research topics with applications of data mining technologies on them. It includes a few advanced data mining technologies. Actually, in order to solve bioinformatics problems, there is plenty of room for improvement in data mining technologies. This book covers basic concepts of data mining and technologies from data preprocessing like hierarchical profiling, information fusion, sequence visualization, and data management, to data mining algorithms for a variety of bioinformatics problems like phylogenetics, protein threading, gene discovery, protein sequence clustering, protein-protein interaction, protein interaction networks, and gene annotations. The summaries of all chapters of the book are as follows.

Chapter I introduces the concept and the process of data mining, plus its relationship with bioinformatics. Tasks and techniques of data mining are also presented. At the end, selected bioinformatics problems related to data mining are discussed. It provides an overview on data mining in bioinformatics.

Chapter II reviews the recent developments related to hierarchical profiling where the attributes are not independent, but rather are correlated in a hierarchy. It discusses in detail several clustering and classification methods where hierarchical correlations are tackled with effective and efficient ways, by incorporation of domain specific knowledge. Relations to other statistical learning methods and more potential applications are also discussed.

Chapter III presents a method, called Combinatorial Fusion Analysis (CFA), for analyzing combination and fusion of multiple scoring systems. Both rank combination and score combination are explored as to their combinatorial complexity and computational efficiency. Information derived from the scoring characteristics of each scoring system is used to perform system selection and to decide method combination. Various applications of the framework are illustrated using examples in information retrieval and biomedical informatics.

Chapter IV introduces various visualization (i.e., graphical representation) schemes of symbolic DNA sequences, which are basically represented by character strings in conventional sequence databases. Further potential applications based on the visualized sequences are also discussed. By understanding the visualization process, the researchers will be able to analyze DNA sequences by designing signal processing algorithms for specific purposes such as sequence alignment, feature extraction, and sequence clustering.

Chapter V provides a rudimentary review of the field of proteomics as it applies to mass spectrometry, data handling and analysis. It points out the potential significance of the field suggesting that the study of nuclei acids has its limitations and that the progressive field of proteomics with spectrometry in tandem with transcription studies could potentially elucidate the link between RNA transcription and concomitant protein expression. Furthermore, the chapter describes the fundamentals of proteomics with mass spectrometry and expounds the methodology necessary to manage the vast amounts of data generated in order to facilitate statistical analysis.

Chapter VI considers the prominent problem of reconstructing the basal phylogenetic tree topology when several subclades have already been identified or are well-known by other means, such as morphological characteristics. Whereas most available tools attempt to estimate a fully resolved tree from scratch, the profile neighbor-joining (PNJ) method focuses directly on the mentioned problem and has proven a robust and efficient method for large-scale data sets, especially when used in an iterative way. The chapter also describes an implementation of this idea, the ProfDist software package, and applies the method to estimate the phylogeny of the eukaryotes.

Chapter VII provides an overview of computational problems and techniques for protein threading. Protein threading can be modeled as an optimization problem. This chapter explains the ideas employed in various algorithms developed for finding optimal or near optimal solutions. It also gives brief explanations of related problems: protein threading with constraints, comparison of RNA secondary structures, and protein structure alignment.

Chapter VIII introduces hybrid methods to tackle the major challenges of power and reproducibility of the dynamic differential gene temporal patterns. Hybrid clustering methods are developed based on resulting profiles from several clustering methods. The developed hybrid analysis is demonstrated through an application to a time course gene expression data from interferon-β-1a treated multiple sclerosis patients. The resulting integrated-condensed clusters and overrepresented gene lists demonstrate that the hybrid methods can successfully be applied.

Chapter IX discusses the issue of parameterless clustering technique for gene expression analysis. Two novel, parameterless and efficient clustering methods that fit for analysis of gene expression data are introduced. The unique feature of the methods is that they incorporate the validation techniques into the clustering process so that high quality results can be obtained. Through experimental evaluation, these methods are shown to outperform other clustering methods greatly in terms of clustering quality, efficiency, and automation on both of synthetic and real data sets.

Chapter X introduces gene selection approaches in microarray data analysis for two purposes: cancer classification and tissue heterogeneity correction. In the first part, jointly discriminatory genes which are most responsible to classification of tissue samples for diagnosis are searched for. In the second part, tissue heterogeneity correction techniques are studied. Also, non-negative matrix factorization (NMF) is employed to computationally decompose molecular signatures based on the fact that the expression values in microarray profiling are non-negative. Throughout the chapter, a real world gene expression profile data was used for experiments.

Chapter XI introduces computational methods for detecting complex disease loci with haplotype analysis. It argues that the haplotype analysis, which plays a major role in the study of population genetics, can be computationally modeled and systematically implemented as a means for detecting causative genes of complex diseases. The explanation of the system and some real examples of the haplotype analysis not only provide researchers with better understanding of current theory and practice of genetic association studies, but also present a computational perspective on the gene discovery research for the common diseases.

Chapter XII presents a Bayesian framework for improving clustering accuracy of protein sequences based on association rules. Most of the existing protein-clustering algorithms compute the similarity between proteins based on one-to-one pairwise sequence alignment instead of multiple sequences alignment. Furthermore, the traditional clustering methods are ad-hoc and the resulting clustering often converges to local optima. The experimental results manifest that the introduced framework can significantly improve the performance of traditional clustering methods.

Chapter XIII reviews high-throughput experimental methods for identification of protein-protein interactions, existing databases of protein-protein interactions, computational approaches to predicting protein-protein interactions at both residue and protein levels, various statistical and machine learning techniques to model protein-protein interactions, and applications of protein-protein interactions in predicting protein functions. Intrinsic drawbacks of the existing approaches and future research directions are also discussed.

Chapter XIV discusses the use of differential association rules to study the annotations of proteins in one or more interaction networks. Using this technique, the differences in the annotations of interacting proteins in a network can be found. The concept is extended to compare annotations of interacting proteins across different definitions of interaction networks. Both cases reveal instances of rules that explain known and unknown characteristics of the network(s). By taking advantage of such data mining techniques, a large number of interesting patterns can be effectively explored that otherwise would not be.

Chapter XV introduces the use of Text Mining in scientific literature for biological research, with a special focus on automatic gene and protein annotation. The chapter describes the main approaches adopted and analyzes systems that have been developed for automatically annotating genes or proteins. To illustrate how text-mining tools fit in biological databases curation processes, the chapter also presents a tool that assists protein annotation. At last, it presents the main open problems in using text-mining tools for automatic annotation of genes and proteins.

Chapter XVI surveys systems that can be used for annotating genomes by comparing multiple genomes and discusses important issues in designing genome comparison systems such as extensibility, scalability, reconfigurability, flexibility, usability, and data mining functionality. Further issues in developing genome comparison systems where users can perform genome comparison flexibly on the sequence analysis level are also discussed.