PREFACE

Over the past decade, computer scientists have increasingly been enlisted as "bioinformaticians" to assist molecular biologists in their research. This book is conceived as a practical introduction to bioinformatics for these computer scientists. While it is not possible to be exhaustive in coverage, the chapters are in-depth discussions by expert bioinformaticians on both general techniques and specific approaches to a range of selected bioinformatics problems. Let us provide a brief overview of these chapters here.

A practical bioinformatician must learn to speak the language of biology. We thus begin with Chapter 1 by Ng and Chapter 2 by Schönbach that form an overview of modern molecular biology and of the planning and execution of bioinformatics experiments. At the same time, a practical bioinformatician must also be conversant in a broad spectrum of topics in computer science—data mining, machine learning, mathematical modeling, sequence alignment, data integration, database management, workflow development, and so on. These diversed topics are surveyed in three separate chapters, *viz.* Chapter 3 by Li *et al.* which provides an in-depth review of data mining techniques that are amongst the key computing technologies for the analysis of biological data; Chapter 10 by Brown *et al.* which discusses the advances through the past thirty years in both global and local alignment, and present methods for general purpose homology that are widely adopted; and Chapter 17 by Wong which reviews some of the requirements and technologies relevant to data integration and warehousing.

DNA sequences contain a number of functional and regulatory sites, such as the site where the transcription of a gene begins, the site where the translation of a gene begins, the site where an exon of a gene ends and an intron begins, and so on. The next several chapters of the book deal with the computational recognition of these sites, *viz.* Chapter 4 by Li *et al.* which is an in-depth survey spanning two decades of research of methods for computational recognization of translation initiation sites from mRNA, cDNA, and DNA sequences; and Chapters 5–7 by Bajić *et al.* which discuss general frameworks, conceptual issues, and performance tuning related to the use of statistical and neural network modeling for computational recognition of promoters and regulatory sites. The recognition of translation initiation sites is among the simplest problems in the recognition of functional sites from DNA sequences. On the other hand, among the toughest problems in the recognition of functional sites from DNA sequences is the determination of locations of promoters and related regulatory elements and functional sites. Thus we hope these four chapters together can bring out clearly the whole range of approaches to this group of problems.

We next move to analysis of RNA sequences. We consider the problem of predicting the secondary structure of RNAs, which is relevant to applications such as function classification, evolution study, and pseudogene detection. Chapter 8 by Sung provides a detailed review of computational methods for predicting secondary structure from RNA sequences. Unlike the prediction methods introduced in earlier chapters for recognition of functional sites from DNA sequences, which are mostly data mining and machine learning methods, the methods described in this chapter come from the realm of mathematical modeling.

After looking at RNA sequences, we move on to protein sequences, and look at aspects relating to protein function prediction. For example, each compartment in a cell has a unique set of functions, and it is thus reasonable to assume that the compartment or membrane in which a protein resides is a determinant of its function. So we have Chapter 9 by Horton *et al.* to discuss various aspects of protein subcellular localization in the context of bioinformatics and review the twenty years of progress in predicting protein subcellular localization. As another example, the homology relationship between a protein and another protein is also suggestive of the function of that protein. So we have Chapter 12 by Kaplan *et al.* to describe two bioinformatics tools, ProtoNet and PANDORA. ProtoNet uses an approach of protein sequence hierarchical clustering to detect remote protein relatives. PANDORA uses a graph-based method to interpret complex protein groups through their annotations. As a third example, motifs are commonly used to classify protein sequences and to provide functional clues on binding sites, catalytic sites, and active sites, or structure/functions relations. So we have Chapter 16 by Schönbach and Matsuda to present a case study of a workflow for mining new motifs from the FANTOM1 mouse cDNA clone collection by a linkage-clustering method, with an all-to-all sequence comparison, followed by visual inspection, sequence, topological, and literature analysis of the motif candidates.

Next we sample the fascinating topic of phylogenetics—the study of the origin, development, and death of a taxon—based on sequence and other information. Chapter 11 by Meng is an introduction to phylogenetics using a case study on Saururaceae. In contrast to earlier chapters, which emphasize the computational aspect, this chapter is written from the perspective of a plant molecular biologist, and emphasizes instead the care that must be exercised in the use of computational tools and the analysis that must be performed on the results produced by computational tools.

The genomics and proteomics efforts have helped identify many new genes and proteins in living organisms. However, simply knowing the existence of genes and proteins does not tell us much about the biological processes in which they participate. Many major biological processes are controlled by protein interaction networks and gene regulation networks. Thus we have Chapter 13 by Tan and Ng to give an overview of the various current methods for discovering protein-protein interactions experimentally and computationally.

The development of microarray technology in the last decade has made possible the simultaneous monitoring of the expression of thousands of genes. This development offers great opportunities in advancing the diagnosis of dieases, the treatment of diseases, and the understanding of gene functions. Chapter 14 by Li and Wong is an in-depth survey of several approaches to some of the gene expression analysis challenges that accompany these opportunities. On the other hand, Chapter 15 by Lin *et al.* presents a method for selecting probes in the design of a microarray to profile genome-wide gene expression of a given genome.

Biological data is being created at ever-increasing rates as different high-throughput technologies are implemented for a wide variety of discovery platforms. It is crucial for researchers to be able to not only access this information but also to integrate it well and synthesize new holistic ideas about various topics. So it is appropriate that we devote the remaining chapters of this book to the issues of integrating databases, cleansing databases, and large-scale experimental and computational analysis workflows as follows. Chapter 18 by Kolatkar and Lin demonstrates the construction of a purpose-built integrated database PPDB using the powerful general data integration engine Kleisli. Chapter 19 by Wu and Barker presents the classification-driven rule-based approach in PIR database to the functional annotation of proteins. Wu and Barker also provide two case studies: the first looks at error propagation to secondary databases using the example of IMP Dehydrogenase; the second looks at transitive identification error using the example of His-I bifunctional proteins. Finally, Chapter 20 by Scheetz and Casavant describes the sophisticated informatics tools and workflow underlying the large-scale effort in EST-based gene discovery in Rat, Human, Mouse, and other species being conducted at the University of Iowa.

Limsoon Wong
11 December 2003

CONTENTS