

Preface

This book arose primarily out of a compelling need for a comprehensive reference in bioinformatics that will cater to students, research, and industry. We strongly believe that this new field evolved from the active interaction of two fast-developing disciplines: biology and information technology. Solving modern biological problems requires advanced computational methods. Key techniques include database management, data modeling, pattern recognition, data mining, query processing, and visualization of biological data. Until very recently, virtually all public databases were based on large flat files stored in simple formats. Navigation among databases required expert knowledge and considerable patience. The huge quantities of biological data and escalating demands of modern biological research increasingly require the sophistication and computing power of information technology (IT) tools. More specifically, optimal use of these tools requires proximal information – knowing which data points are in the surrounding area of others. In this book, we will present methodologies and data structures for arriving at high quality biological information, which can then be used as foundation to develop practical tools for clustering and visualization in biological data mining and database management.

Throughout the book, we will demonstrate the application of well established concepts and techniques of information technology to the management and analysis of biological data. Biological analysis requires the integration of software tools used in data mining, such as clustering, classification, decision trees and decision tables, and sequence and structural modeling such as data modeling. A distinctive feature of our book is the integration of advanced database technologies with visualization techniques such as query-interactive user interfaces, visual descriptions, and advanced 3-D visual modeling.

Biological data continue to grow exponentially in size and complexity. As a result, they introduce new data types not previously seen even in molecular biology. It is vital and urgent that advanced information technologies, in particular, database technologies and visual analysis, be applied to support biological research and innovation based on biological data. Specific IT-motivated activities are taking root in some parts of the biological

research community, and we foresee that they will benefit information technology.

"Bioinformatics technologies" is a comprehensive book that covers these two important areas, viz., IT and biology, which have become interwoven in recent years. Many international experts have made contributions to this book. Each article is written in a way that a practitioner of bioinformatics can easily understand and then apply the knowledge gained to extract useful information from biological data. Each article covers one topic, and can be read independently of each other. The book provides both a general survey of the topic and an in-depth exposition of the state-of-the-art. Practitioners will certainly find this book very resourceful and handy when looking for solutions to practical problems in bioinformatics. Researchers can use this book as a source for obtaining background information, current trends and developments; this provides them also with the most important references on these topics.

The book covers the basic principles and applications of bioinformatics technologies. It also contains many articles that specifically address bioinformatics databases and emerging topics in bioinformatics technologies such as patterns discovery, data mining, simulation and visualization. The central issue in bioinformatics is how to transform biological data into meaningful and valuable information. It implies that the biological knowledge related to the problem domain is incorporated into the requirements analysis phase of the bioinformatics.

However, it has been recently recognized that in the twenty-first century bioinformatics will play an increasingly important role. For this reason, the international conference series on Asia-Pacific Bioinformatics (first bioinformatics conference in the IT domain) was founded in 2002. The underlying goal behind this conference series is to recognize the interdisciplinary nature of bioinformatics in the interplay between biology and IT and how information technology can be applied to biology.

Even though a great deal of attention is paid to this area in terms of research and investment, the theoretical understanding needs further refinement to bring the outcome of the biological analysis effectively to the service of mankind. In editing this book, this viewpoint has been carefully taken into consideration to conceptually organize the recent progress in bioinformatics.

The book is organized into twelve chapters that cover twelve important technologies in bioinformatics.

Chapter 1, Introduction to Bioinformatics, provides an overview of bioinformatics technology, and different techniques within bioinformatics. Further, it introduces the relationships between the other chapters.

Chapter 2, Overview of Structural Bioinformatics, presents an overview of structural bioinformatics. The chapter describes organization of structural bioinformatics, the Protein Data Bank, secondary resources and applications, and using structural bioinformatics approaches in drug design. It also includes structural classification, structure prediction, functional assignments in structural genomics, protein-protein interactions and protein-ligand interactions. The role of structural bioinformatics in systems biology is also briefly discussed.

Chapter 3, Database Warehousing in Bioinformatics, deals with the basics in database warehousing, transforming biological data into knowledge, data warehouse architectures and data quality in bioinformatics.

Chapter 4, Data Mining for Bioinformatics, discusses the basics of data mining applicable to bioinformatics. The main types of data analysis, namely, biomedical data analysis, DNA data analysis, protein data analysis and microarray data analysis, are elaborated upon. Biomedical data analysis includes a major nucleotide sequence database, a protein sequence database, a gene expression database, and software tools for bioinformatics research. DNA data analysis covers DNA sequence and DNA data analysis. Protein data analysis encompasses protein and amino acid sequence and protein data analysis.

Chapter 5, Machine Learning in Bioinformatics, dwells on the theory behind machine learning applied to bioinformatics. It includes neural network architectures and applications. We also describe other machine learning techniques, such as genetic algorithms and fuzzy systems.

Chapter 6, Systems Biotechnology: a New Paradigm in Biotechnology Development, describes a new paradigm in biotechnology development called system biotechnology. It covers integrative approaches and *in silico* modeling and simulation of cellular processes.

Chapter 7, Computational Modeling of Biological Processes with Petri Net-Based Architecture, describes computational modeling of biological processes with a Petri net-based architecture, a hybrid Petri net and a hybrid dynamic net, and a hybrid functional Petri net. The chapter also covers the implementation of a HFPNe in a genomic object net and the modeling of biological processes with a HFPNe and a genomic object net and its visualizer.

Chapter 8, Biological Sequence Assembly and Alignment, illustrates biological sequence assembly and alignment. It covers large-scale sequence assembly, Euler sequence assembly, PESA sequence assembly, large-scale pairwise sequence alignment, large-scale multiple sequence alignment, and load balancing and communication overheads.

Chapter 9, Modeling for Bioinformatics, covers the basics of modeling techniques related to bioinformatics. It includes the major modeling tech-

niques, namely, hidden Markov modeling for biological data analysis, comparative modeling and molecular modeling. An elaborate discussion is made to apply hidden Markov modeling on biological data to have sequence identification, sequence classification, and multiple alignment generation. Comparative modeling comprises protein comparative modeling, comparative genomic modeling, and probabilistic modeling. The probabilistic modeling encompasses Bayesian networks, stochastic context-free grammars, and probabilistic Boolean networks. Finally, we describe molecular modeling, which deals with molecular and related visualization applications, molecular mechanics, and modern computer programs used in molecular modeling.

Chapter 10, Pattern Matching for Motifs, addresses the issues in pattern matching for discovering motifs. Topics include gene regulation and promoter organization. We include motif recognition and motif detection strategies. The chapter also includes two different approaches, namely, the single gene multi-species approach and the multi-gene multi-species approach.

Chapter 11, Visualization and Fractal Analysis of Biological Sequences, deals with visualization and fractal analysis of biological sequences. It elaborates on the fractal analysis, the recurrent iterated function system model, the moment method to estimate the parameters of the IFS (RIFS) model, multifractal analysis, the DNA walk model, and chaos game representation of biological sequences. Two-dimensional portrait representation of DNA sequences and one-dimensional measure representations of biological sequences are also introduced.

Chapter 12, Microarray Data Analysis, discusses the techniques used to analyze microarray data and microarray technology used for genome expression study, image analysis for data extraction, and data analysis for pattern discovery.

In a rapidly expanding area such as bioinformatics, no book can claim to cover the topics that suit the interests of everyone. However, it is hoped that this book is comprehensive enough to serve as a useful and handy guide for both practitioners and researchers. This book will help both IT professionals and biologists to understand the bioinformatics world.

We would like to thank all authors who contributed the chapters in this book, without whom the mission would have been impossible. Special thanks to the reviewers for their professional inputs. We thank Ricky Chen and Chinnu Subramaniam for helping us check parts of the manuscript at short notice. We have taken care to cite referenced work. If we have missed any citation, we apologize for the lapse. We thank all researchers for their permission to use their figures in this book. We also wish to thank the Springer publisher Ralf Gerstner for his final step of checking and

timely help before publication. Finally, we wish to thank our families and friends for their support.

We are sure that some errors may stay behind in the book. Your input for improvement will be helpful for future reprints and editions. Comments, corrections, and constructive suggestions should be sent to Springer or by electronic mail to phoebe@deakin.edu.au

January 2005

Yi-Ping Phoebe Chen

1	1.1 Introduction	1
2	1.2 Needs of Bioinformatics Technologies	2
3	1.3 An Overview of Bioinformatics Technologies	3
4	1.4 A Brief Discussion on the Chapters	4
5	2 Overview of Structural Bioinformatics	5
6	2.1 Introduction	6
7	2.2 Organization of Structural Bioinformatics	7
8	2.3 Primary Resource: Protein Data Bank	8
9	2.3.1 Data Format	8
10	2.3.2 Growth of Data	10
11	2.3.3 Data Processing and Quality Control	20
12	2.3.4 The Future of the PDB	21
13	2.3.5 Visualization	21
14	2.4 Secondary Resources and Applications	22
15	2.4.1 Structural Classification	22
16	2.4.2 Structure Prediction	28
17	2.4.3 Functional Assignments in Structural Genomics	30
18	2.4.4 Protein-Protein Interactions	32
19	2.4.5 Protein-Ligand Interactions	34
20	2.5 Using Structural Bioinformatics Approaches in Drug Design	37
21	2.6 The Future	39
22	2.6.1 Integration over Multiple Resources	39
23	2.6.2 The Impact of Structural Genomics	39
24	2.6.3 The Role of Structural Bioinformatics in Systems Biology	39
25	References	40
26	3 Database Warehousing in Bioinformatics	45
27	3.1 Introduction	45
28	3.2 Bioinformatics Data	48
29	3.3 Transforming Data to Knowledge	51
30	3.4 Data Warehousing	54
31	3.5 Data Warehouse Architecture	56
32	3.6 Data Quality	58
33	3.7 Concluding Remarks	60