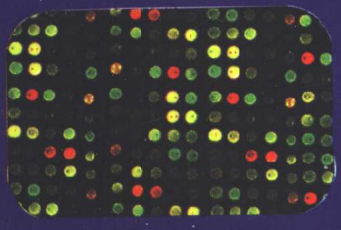
Current Topics in Computational Molecular Biology

计算分子生物学前沿课题

Edited by Tao Jiang Ying Xu Michael Q. Zhang



清华大学出版社
http://www.tup.tsinghua.edu.cn

The MIT Press
http://mitpress.mit.edu

Current Topics in Computational Molecular Biology

计算分子生物学前沿课题

Edited by

Tao Jiang

Ying Xu

Michael Q. Zhang

Tsinghua University Press The MIT Press

内 容 简 介

本书概观综述了计算生物学的传统课题，如蛋白质结构模拟计算和基因序列比较，以及目前的发展热点，如基因表达数据分析和比较基因学，并阐述了算法，统计，数据库和一些以人工智能为基础的技术在解决生物问题上的应用。本书还包括了对计算生物学基础知识介绍，具体实用统计模型和分子生物学的计算方法实例。本书每章节围绕一具体议题阐述，自成一体，以便阅读。

Published in association with Tsinghua University Press, Beijing, China, as part of TUP's Frontiers of Science and Technology for the 21st Century Series.

This book was set in Times New Roman on 3B2 by Asco Typesetters, Hong Kong and was printed and bound in the United States of America.

（以下内容为在中国印制部分）

书　名：计算分子生物学前沿课题

作　者：Tao Jiang　Ying Xu　Michael Q. Zhang

出版者：清华大学出版社（北京清华大学学研大厦，邮编 100084）

http://www.tup.tsinghua.edu.cn

印刷者：清华大学印刷厂

发行者：新华书店总店北京发行所

开　本：787×960　1/16　印张：35

版　次：2002 年 4 月第 1 版　2003 年 4 月第 2 次印刷

书　号：ISBN 7-302-05378-2

ISBN 0-262-10092-4

印　数：801～1300（中国印制部分）

定　价：70.00 元

21世纪科技前沿丛书
Frontiers of Science and Technology for the 21st Century

《21 世纪科技前沿》

丛书序言

由清华大学出版社出版的这套丛书是基础科学和应用科学领域内的专门著作。除了可作为研究生教材外，也可作为科研和工程技术人员的参考书。在丛书的题材选择中，着重考虑目前比较活跃而且具有发展前景的新兴学科。因此，这套丛书大都涉及交叉和新兴学科的内容。编写的方式大多由主编策划并组织本学科有影响的专家共同执笔完成，从而使每一本书的系统性和各章节内容的连贯性得到了充分的兼顾。丛书涵盖学科的最新学术进展，兼顾到基本理论和新技术、新方法的介绍，并引入必要的导论和充分的参考文献以适应具有不同学术背景的读者。编撰一套容纳多学科的科技丛书是一项浩繁的工作，我们希望通过主编和作者的集体努力和精诚协作，使整套丛书的学术水准能够保持在较高的水平上。

编辑《21 世纪科技前沿》丛书是由“旅美中国科学家工程师协会”发起的一项国际科技界的合作。传递信息，加强交流，促进新世纪的科技繁荣是编著者们参与此项工作的共同信念。此外，这套丛书还具有特别的纪念意义。20 年前，历史的进程使成千上万的中国学生、学者有机会走出国门，到世界各地学习和从事科学研究。今天，活跃在世界科技前沿领域的中华学子们没有忘记振兴祖国科技教育事业的责任和推动国际学术交流与合作的义务。正是基于这一共同的心愿，大家积极参与这套系列丛书的撰写、组稿和编辑工作。为此，我们愿以这套丛书来纪念中国改革开放 20 周年。

编 委 会

1999. 6

《Frontiers of Science and Technology for the 21st Century》

FOREWORD

Over the next several years, Tsinghua University Press will publish a series of books addressing progress in basic sciences and innovations in technology. We have made no attempt to pursue a comprehensive coverage of all disciplines of science and technology. Rather, topics for this series were selected with an emphasis on the currently active forefront of science and technology that will be contemporary in the next century. Most books in this series will deal with subjects of cross disciplines and newly emerging fields. Each book will be completed by individual authors or in a collaborative effort managed by an editor(s), and will be self-consistent, with contents systematically focused on review of the most recent advances and description of current progresses in the field. Sufficient introduction and references will be provided for readers with varying backgrounds. We have realize clearly the challenge of encompassing the diverse subjects of science and technology in one series. However, we hope that, through intensive collaboration between the authors and editors, high standards in editorial quality and scientific merit will be maintained for the entire series.

The international collaboration on this series has been coordinated by the Association of Chinese Scientists and Engineers-USA(ACSE). In the science community, authors voluntarily publish their results and discoveries in the full

conviction that science should serve human society. The editors and authors of this series share this academic tradition, and many of them are fulfilling a spiritual commitment as well. For our editors and authors who were graduated from universities in China and further educated abroad in science and engineering, this is an opportunity to dedicate their work to the international education community and to commemorate the historical open-door movement that began in China two decades ago. When the human society enters the information age, there is no geographic boundary for science. The Editorial committee hopes that this series will promote further international collaboration in scientific research and education at the dawn of the new century.

The Editorial Committee
1999. 6

Current Topics in Computational Molecular Biology

Science is advanced by new observations and technologies. The Human Genome Project has led to a massive outpouring of genomic data, which has in turn fueled the rapid developments of high-throughput biotechnologies. We are witnessing a revolution driven by the high-throughput biotechnologies and data, a revolution that is transforming the entire biomedical research field into a new systems level of genomics, transcriptomics, and proteomics, fundamentally changing how biological science and medical research are done. This revolution would not have been possible if there had not been a parallel emergence of the new field of computational molecular biology, or bioinformatics, as many people would call it. Computational molecular biology/bioinformatics is interdisciplinary by nature and calls upon expertise in many different disciplines—biology, mathematics, statistics, physics, chemistry, computer science, and engineering; and is ubiquitous at the heart of all large-scale and high-throughput biotechnologies. Though, like many emerging interdisciplinary fields, it has not yet found its own natural home department within traditional university settings, it has been identified as one of the top strategic growing areas throughout academic as well as industrial institutions because of its vital role in genomics and proteomics, and its profound impact on health and medicine.

At the eve of the completion of the human genome sequencing and annotation, we believe it would be very useful and timely to bring out this up-to-date survey of current topics in computational molecular biology. Because this is a rapidly developing field and covers a very wide range of topics, it is extremely difficult for any individual to write a comprehensive book. We are fortunate to be able to pull together a team of renowned experts who have been actively working at the forefront of each major area of the field. This book covers most of the important topics in computational molecular biology, ranging from traditional ones such as protein structure modeling and sequence alignment, to the recently emerged ones such as expression data analysis and comparative genomics. It also contains a general introduction to the field, as well as a chapter on general statistical modeling and computational techniques in molecular biology. Although there are already several books on computational molecular biology/bioinformatics, we believe that this book is unique as it covers a wide spectrum of topics (including a number of new ones not covered in existing books, such as gene expression analysis and pathway databases) and it combines algorithmic, statistical, database, and AI-based methods for biological problems.

Although we have tried to organize the chapters in a logical order, each chapter is a self-contained review of a specific subject. It typically starts with a brief overview of a particular subject, then describes in detail the computational techniques used and the computational results generated, and ends with open challenges. Hence the reader need not read the chapters sequentially. We have selected the topics carefully so that

the book would be useful to a broad readership, including students, nonprofessionals, and bioinformatic experts who want to brush up topics related to their own research areas.

The 19 chapters are grouped into four sections. The introductory section is a chapter by Temple Smith, who attempts to set bioinformatics into a useful historical context. For over half a century, mathematics and even computer-based analyses have played a fundamental role in bringing our biological understanding to its current level. To a very large extent, what is new is the type and sheer volume of new data. The birth of bioinformatics was a direct result of this new data explosion. As this interdisciplinary area matures, it is providing the data and computational support for functional genomics, which is defined as the research domain focused on linking the behavior of cells, organisms, and populations to the information encoded in the genomes.

The second of the four sections consists of six chapters on computational methods for comparative sequence and genome analyses.

Liu's chapter presents a systematic development of the basic Bayesian methods alongside contrasting classical statistics procedures, emphasizing the conceptual importance of statistical modeling and the coherent nature of the Bayesian methodology. The missing data formulation is singled out as a constructive framework to help one build comprehensive Bayesian models and design efficient computational strategies. Liu describes the powerful computational techniques needed in Bayesian analysis, including the expectation-maximization algorithm for finding the marginal mode, Markov chain Monte Carlo algorithms for simulating from complex posterior distributions, and dynamic programming-like recursive procedures for marginalizing out uninteresting parameters or missing data. Liu shows that the popular motif sampler used for finding gene regulatory binding motifs and for aligning subtle protein motifs can be derived easily from a Bayesian missing data formulation.

Huang's chapter focuses on methods for comparing two sequences and their applications in the analysis of DNA and protein sequences. He presents a global alignment algorithm for comparing two sequences that are entirely similar. He also describes a local alignment algorithm for comparing sequences that contain locally similar regions. The chapter gives efficient computational techniques for comparing two long sequences and comparing two sets of sequences, and it provides real applications to illustrate the usefulness of sequence alignment programs in the analysis of DNA and protein sequences.

The chapter by Jiang and Wang provides a survey on computational methods for multiple sequence alignment, which is a fundamental and challenging problem in computational molecular biology. Algorithms for multiple sequence alignment are routinely used to find conserved regions in biomolecular sequences, to construct

family and superfamily representations of sequences, and to reveal evolutionary histories of species (or genes). The authors discuss some of the most popular mathematical models for multiple sequence alignment and efficient approximation algorithms for computing optimal multiple alignment under these models. The main focus of the chapter is on recent advances in combinatorial (as opposed to stochastic) algorithms.

Kearney's chapter illustrates the basic concepts in phylogenetics, the design and development of computational tools for evolutionary analyses, using the quartet method as an example. Quartet methods have recently received much attention in the research community. This chapter begins by examining the mathematical, computational, and biological foundations of the quartet method. A survey of the major contributions to the method reveals an excess of diverse and interesting concepts indicative of a ripening research topic. These contributions are examined critically with strengths, weakness, and open problems.

Sankoff and El-Mabrouk's chapter describes the basic concepts of genome rearrangement and applications. Genome structure evolves through a number of nonlocal rearrangement processes that may involve an arbitrarily large proportion of a chromosome. The formal analysis of rearrangements differs greatly from DNA and protein comparison algorithms. In this chapter, the authors formalize the notion of a genome in terms of a set of chromosomes, each consisting of an ordered set of genes. The chapter surveys genomic distance problems, including the Hannenhalli-Pevzner theory for reversals and translocations, and covers the progress to date on phylogenetic extensions of rearrangement analysis. Recent work focuses on problems of gene and genome duplication and their implications for genomic distance and genome-based phylogeny.

The chapter by Li describes the author's work on compressing DNA sequences and applications. The chapter concentrates on two programs the author has developed: a lossless compression algorithm, GenCompress, which achieves the best compression ratios for benchmark sequences; and an entropy estimation program, GTAC, which achieves the lowest entropy estimation for benchmark DNA sequences. The author then discusses a new information-based distance measure between two sequences and shows how to use the compression programs as heuristics to realize such distance measures. Some experiments are described to demonstrate how such a theory can be used to compare genomes.

The third section covers computational methods for mining biological data and discovering patterns hidden in the data.

The chapter by Xu presents an overview of the major statistical techniques for quantitative trait analysis. Quantitative traits are defined as traits that have a con-

tinuous phenotypic distribution. Variances of these traits are often controlled by the segregation of multiple loci plus an environmental variance. Localization of these quantitative trait loci (QTL) on the chromosomes and estimation of their effects using molecular markers are called QTL linkage analysis or QTL mapping. Results of QTL mapping can help molecular biologists target particular chromosomal regions and eventually clone genes of functional importance.

The chapter by Solovyev describes statistically based methods for the recognition of eukaryotic genes. Computational gene identification is an issue of vital importance as a tool of identifying biologically relevant features (protein coding sequences), which often cannot be found by the traditional sequence database searching technique. Solovyev reviews the structure and significant characteristics of gene components, and discusses recent advances and open problems in gene-finding methodology and its application to sequence annotation of long genomic sequences.

Zhang's chapter gives an overview of computational methods currently used for identifying eukaryotic PolII promoter elements and the transcriptional start sites. Promoters are very important genetic elements. A PolII promoter generally resides in the upstream region of each gene; it controls and regulates the transcription of the downstream gene.

In their chapter, Shamir and Sharan describe some of the main algorithmic approaches to clustering gene expression data, and briefly discuss some of their properties. DNA chip technologies allow for the first time a global, simultaneous view of the transcription levels of many thousands of genes, under various cellular conditions. This opens great opportunities in medical, agricultural, and basic scientific research. A key step in the analysis of gene expression data is the identification of groups of genes that manifest similar expression patterns. This translates to the algorithmic problem of clustering gene expression data. The authors also discuss methods for evaluating the quality of clustering solutions in various situations, and demonstrate the performance of the algorithms on yeast cell cycle data.

The chapter by Kanehisa and Goto dsecribes the latest developments of the KEGG database. A key objective of the KEGG project is to computerize data and knowledge on molecular pathways and complexes that are involved in various cellular processes. Currently KEGG consists of (1) a pathway database, (2) a genes database, (3) a genome database, (4) a gene expression database, (5) a database of binary relations between proteins and other biological molecules, and (6) a ligand database, plus various classification information. It is well known that the analysis of individual molecules would not be sufficient for understanding higher order functions of cells and organisms. KEGG provides a computational resource for analyzing biological networks.

The chapter by Wong presents an introduction to what has come to be known as datamining and knowledge discovery in the biomedical context. The major reason that datamining has attracted increasing attention in the biomedical industry in recent years is due to the increased availability of huge amount of biomedical data and the imminent need to turn such data into useful information and knowledge. The knowledge gained can lead to improved drug targets, improved diagnostics, and improved treatment plans.

The last section of the book, which consists of six chapters, covers computational approaches for structure prediction and modeling of macromolecules.

Wang and Zhang's chapter presents an overview of predictions of RNA secondary structures. The secondary structure of an RNA is a set of base-pairs (nucleotide pairs) that form bonds between A-U and C-G. These bonds have been traditionally assumed to be noncrossing in a secondary structure. Two major prediction approaches considered are thermodynamic energy minimization methods and phylogenetic comparative methods. Thermodynamic energy minimization methods have been used to predict secondary structures from a single RNA sequence. Phylogenetic comparative methods have been used to determine secondary structures from a set of homologous RNAs whose sequences can be reliably aligned.

The chapter by Solovyev and Shindyalov provides a survey of computational methods for protein secondary structure predictions. Secondary structures describe regular features of the main chain of a protein molecule. Experimental investigation of polypeptides and small proteins suggest that a secondary structure can form in isolation, implying the possibility of identifying rules for its computational prediction. Predicting the secondary structure from an amino acid sequence alone is an important step toward our understanding of protein structures and functions. It may provide a starting point for tertiary structure modeling, especially in the absence of a suitable homologous template structure, reducing the search space in the simulation of protein folding.

The chapter by Chan et al. surveys currently available physics-based computational approaches to protein folding. A spectrum of methods—ranging from all-atom molecular dynamics to highly coarse-grained lattice modeling—have been employed to address physicochemical aspects of protein folding at various levels of structural and energetic resolution. The chapter discusses the strengths and limitations of some of these methods. In particular, the authors emphasize the primacy of self-contained chain models and how they differ logically from non-self-contained constructs with ad hoc conformational distributions. The important role of a protein's aqueous environment and the general non-additivity of solvent-mediated protein interactions are illustrated by examples in continuum electrostatics and atomic treatments of hydro-

phobic interactions. Several recent applications of simple lattice protein models are discussed in some detail.

In their chapter, Peitsch et al. discuss how protein models can be applied to functional analysis, as well as some of the current issues and limitations inherent to these methods. Functional analysis of the proteins discovered in fully sequenced genomes represents the next major challenge of life science research, and computational methods play an increasingly important part. Among them, comparative protein modeling will play a major role in this challenge, especially in light of the Structural Genomics programs about to be started around the world.

Xu and Xu's chapter presents a survey on protein threading as a computational technique for protein structure calculation. The fundamental reason for protein threading to be generally applicable is that the number of unique folds in nature is quite small, compared to the number of protein sequences, and a significant portion of these unique folds are already solved. A new trend in the development of computational modeling methods for protein structures, particularly in threading, is to incorporate partial structural information into the modeling process as constraints. This trend will become more clear as a great amount of structural data will be generated by the high-throughput structural genomics centers funded by the NIH Structural Genonics Initiative. The authors outline their recent work along this direction.

The chapter by Nussinov, Ma, and Wolson describes highly efficient, computer-vision and robotics based algorithms for docking and for the generation and matching of epitopes on molecular surfaces. The goal of frequently used approaches, both in searches for molecular similarity and for docking, that is, molecular complementarity, is to obtain highly accurate matching of respective molecular surfaces. Yet, owing to the variability of molecular surfaces in solution, to flexibility, to mutational events, and to the need to use modeled structures in addition to high resolution ones, utilization of epitopes may ultimately prove a more judicious approach to follow.

This book would not have been possible without the timely cooperation from all the authors and the patience of the publisher. Many friends and colleagues who have served as chapter reviewers have contributed tremendously to the quality and readability of the book. We would like to take this opportunity to thank them individually. They are: Nick Alexandrov, Vincent Berry, Mathieu Blanchette, David Bryant, Alberto Caprara, Kun-Mao Chao, Jean-Michel Claverie, Hui-Hsien Chou, Bhaskar DasGupta, Ramana Davuluri, Jim Fickett, Damian Gessler, Dan Gusfield, Loren Hauser, Xiaoqiu Huang, Larry Hunter, Shuyun Le, Sonia Leach, Hong Liu, Satoru Miyano, Ruth Nussinov, Victor Olman, Jose N. Onuchic, Larry Ruzzo, Gavin Sherlock, Jay Snoddy, Chao Tang, Ronald Taylor, John Tromp, Ilya A. Vakser, Martin Vingron, Natascha Vukasinovic, Mike Waterman, Liping Wei, Dong Xu, Zhenyu

Xuan, Lisa Yan, Louxin Zhang, and Zheng Zhang. We would also like to thank Ray Zhang for the artistic design of the cover page. Finally, we would like to thank Katherine Almeida, Katherine Innis, Ann Rae Jonas, Robert V. Prior, and Michael P. Rutter from The MIT Press for their great support and assistance throughout the process, and Dr. Guokui Liu for connecting us with the Tsinghua University Press (TUP) of China and facilitating copublication of this book by TUP in China.

Preface vii

I INTRODUCTION 1

1 **The Challenges Facing Genomic Informatics** 3
Temple F. Smith

II COMPARATIVE SEQUENCE AND GENOME ANALYSIS 9

2 **Bayesian Modeling and Computation in Bioinformatics Research** 11
Jun S. Liu

3 **Bio-Sequence Comparison and Applications** 45
Xiaoqiu Huang

4 **Algorithmic Methods for Multiple Sequence Alignment** 71
Tao Jiang and Lusheng Wang

5 **Phylogenetics and the Quartet Method** 111
Paul Kearney

6 **Genome Rearrangement** 135
David Sankoff and Nadia El-Mabrouk

7 **Compressing DNA Sequences** 157
Ming Li

III DATA MINING AND PATTERN DISCOVERY 173

8 **Linkage Analysis of Quantitative Traits** 175
Shizhong Xu

9 **Finding Genes by Computer: Probabilistic and Discriminative Approaches** 201
Victor V. Solovyev

10 **Computational Methods for Promoter Recognition** 249
Michael Q. Zhang

11 **Algorithmic Approaches to Clustering Gene Expression Data** 269
Ron Shamir and Roded Sharan

12 **KEGG for Computational Genomics** 301
Minoru Kanehisa and Susumu Goto

13	**Datamining: Discovering Information from Bio-Data** Limsoon Wong	317
IV	**COMPUTATIONAL STRUCTURAL BIOLOGY**	343
14	**RNA Secondary Structure Prediction** Zhuozhi Wang and Kaizhong Zhang	345
15	**Properties and Prediction of Protein Secondary Structure** Victor V. Solovyev and Ilya N. Shindyalov	365
16	**Computational Methods for Protein Folding: Scaling a Hierarchy of Complexities** Hue Sun Chan, Hüseyin Kaya, and Seishi Shimizu	403
17	**Protein Structure Prediction by Comparison: Homology-Based Modeling** Manuel C. Peitsch, Torsten Schwede, Alexander Diemand, and Nicolas Guex	449
18	**Protein Structure Prediction by Protein Threading and Partial Experimental Data** Ying Xu and Dong Xu	467
19	**Computational Methods for Docking and Applications to Drug Design: Functional Epitopes and Combinatorial Libraries** Ruth Nussinov, Buyong Ma, and Haim J. Wolfson	503
	Contributors	525
	Index	527