TUP-Springer Project

Rui Jiang
Xuegong Zhang
Michael Q. Zhang *Editors*

生物信息学课程导引

——生物信息学研究生暑期学校讲义

Basics of Bioinformatics

Lecture Notes of the Graduate Summer School on Bioinformatics of China

清华大学出版社 Springer

TUP-Springer Project

Rui Jiang
Xuegong Zhang
Michael Q. Zhang *Editors*

生物信息学课程导引

——生物信息学研究生暑期学校讲义

Basics of Bioinformatics

Lecture Notes of the Graduate Summer School on Bioinformatics of China

常州大学图书馆
藏书章

清华大学出版社
北京

Springer

内 容 简 介

本书根据清华大学承办的全国生物信息学暑期学校课程，高度概括地介绍了与生物信息学研究紧密相关的11门基础课程和15个前沿专题报告。全书分12章，包括：生物信息学引论、生物信息学中的基础统计、计算基因组学专题、生物信息学中的高级统计、计算生物学算法基础、生物信息学中的多元统计、人类疾病关联研究方法与实例、生物信息学中的数据挖掘与知识发现、生物信息学应用工具、蛋白质结构与功能基础、中医药研究的计算系统生物学方法、生物信息学与计算系统生物学前沿等。本书不仅可以作为生物信息学初学者的入门读物，还可作为生物信息学领域专业研究人员高度概括而又不失系统性的参考书籍。

本书封面贴有清华大学出版社防伪标签，无标签者不得销售。
版权所有，侵权必究。侵权举报电话：010-62782989　13701121933

图书在版编目(CIP)数据

生物信息学课程导引＝Basics of Bioinformatics：英文/江瑞，张学工，张奇伟主编. --北京：清华大学出版社，2014

ISBN 978-7-302-32359-4

Ⅰ. ①生… Ⅱ. ①江… ②张… ③张… Ⅲ. ①生物信息论－高等学校－教学参考资料－英文 Ⅳ. ①Q811.4

中国版本图书馆CIP数据核字(2013)第094210号

责任编辑：王一玲
封面设计：何凤霞
责任校对：白　蕾
责任印制：刘海龙

出版发行：清华大学出版社
网　址：http://www.tup.com.cn，http://www.wqbook.com
地　址：北京清华大学学研大厦A座　　邮　编：100084
社总机：010-62770175　　邮　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质　量　反　馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：三河市春园印刷有限公司
经　　销：全国新华书店
开　　本：155mm×235mm　　**印　张**：26　　**字　数**：462千字
版　　次：2014年5月第1版　　**印　次**：2014年5月第1次印刷
印　　数：1～1000
定　　价：99.00元

产品编号：036335-01

Editors
Rui Jiang
Xuegong Zhang
Department of Automation
Tsinghua University
Beijing
China, People's Republic

Michael Q. Zhang
Department of Molecular and Cell Biology
The University of Texas at Dallas
Richardson, TX, USA

Tsinghua National Laboratory
for Information Science and Technology
Tsinghua University
Beijing, China, People's Republic

ISBN 978-3-642-38950-4 ISBN 978-3-642-38951-1 (eBook)
DOI 10.1007/978-3-642-38951-1
Springer Heidelberg New York Dordrecht London

Jointly published with Tsinghua University Press, Beijing
ISBN: 978-7-302-32359-4 Tsinghua University Press, Beijing

Library of Congress Control Number: 2013950934

© Tsinghua University Press, Beijing and Springer-Verlag Berlin Heidelberg 2013
This work is subject to copyright. All rights are reserved by the Publishers, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publishers' locations, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publishers can accept any legal responsibility for any errors or omissions that may be made. The publishers make no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Foreword

This ambitious volume is the result of the successful 2007 Graduate Summer School on Bioinformatics of China held at Tsinghua University. It is remarkable for its range of topics as well as the depth of coverage. Bioinformatics draws on many subjects for analysis of the data generated by the biological sciences and biotechnology. This foreword will describe briefly each of the 12 chapters and close with additional general comments about the field. Many of the chapters overlap and include useful introductions to concepts such as gene or Bayesian methods. This is a valuable aspect of the volume allowing a student various angles of approach to a new topic.

Chapter 1, "Basics for Bioinformatics," defines bioinformatics as "the storage, manipulation and interpretation of biological data especially data of nucleic acids and amino acids, and studies molecular rules and systems that govern or affect the structure, function and evolution of various forms of life from computational approaches." Thus, the first subject they turn to is molecular biology, a subject that has had an enormous development in the last decades and shows no signs of slowing down. Without a basic knowledge of biology, the bioinformatics student is greatly handicapped. From basic biology the authors turn to biotechnology, in particular, methods for DNA sequencing, microarrays, and proteomics. DNA sequencing is undergoing a revolution. The mass of data collected in a decade of the Human Genome Project from 1990 to 2001 can be generated in 1 day in 2010. This is changing the science of biology at the same time. A 1,000 genome project became a 10,000 genome project 2 years later, and one expects another zero any time now. Chromatin Immunoprecipitation or ChIP allows access to DNA bound by proteins and thus to a large number of important biological processes. Another topic under the umbrella of biological sciences is genetics, the study of heredity and inherited characteristics (phenotypes). Heredity is encoded in DNA and thus is closely related to the goals of bioinformatics. This whole area of genetics beginning with Mendel's laws deserves careful attention, and genetics is a key aspect of the so-called genetic mapping and other techniques where the chromosomal locations of disease genes are sought.

此为试读,需要完整PDF请访问: www.ertongbook.com

Chapter 2, "Basic Statistics for Bioinformatics," presents important material for the understanding and analysis of data. Probability and statistics are basic to bioinformatics, and this chapter begins with the fundamentals including many classical distributions (including the binomial, Poisson, and normal). Usually the observation of complete populations such as "all people in China over 35 years old" is not practical to obtain. Instead random samples of the population of interest are obtained and then inferences about parameters of the population are made. Statistics guides us in making those inferences and gaining information about the quality of the estimates. The chapter describes techniques such as method of moments, maximum likelihood, and Bayesian methods. Bayesian methods have become indispensable in the era of powerful computing machines. The chapter treats hypothesis testing which is less used than parameter estimation, but hypothesis testing provides understanding of p-values which are ubiquitous in bioinformatics and data analysis. Classical testing situations reveal useful statistics such as the t-statistic. Analysis of variance and regression analysis are crucial for testing and fitting large data sets. All of these methods and many more are included in the free open-source package called *R*.

Chapter 3, "Topics in Computational Genomics," takes us on a tour of important topics that arise when complete genome information is available. The subject did not begin until nearly 2000 when complete genome sequences became a possibility. The authors present us with a list of questions, some of which are listed next. What are the genes of an organism? How are they turned off and on? How do they interact with each other? How are introns and exons organized and expressed in RNA transcripts? What are the gene products, both structure and function? How has a genome evolved? This last question has to be asked with other genomes and with members of the population comprising the species. Then the authors treat some of the questions in detail. They describe "finding protein coding genes," "identifying promoters," "genomic arrays and a CGH/CNP analysis," "modeling regulatory elements," "predicting transcription factor binding sites," and motif enrichment and analysis. Within this last topic, for example, various word counting methods are employed including the Bayesian methods of expectation maximization and Gibbs sampling.

An alert reader will have noticed the prominence of Bayesian methods in the preceding paragraphs. Chapter 4, "Statistical Methods in Bioinformatics," in this collection focuses on this subject. There is a nice discussion of statistical modeling and then Bayesian inference. Dynamic programming, a recursive method of optimization, is introduced and then employed in the development of Hidden Markov Models (HMMs). Of course the basics of Markov chains must also be covered. The Metropolis-Hastings algorithm, Monte Carlo Markov chains (MCMC), and Gibbs sampling are carefully presented. Then these ideas find application in the analysis of microarray data. Here the challenging aspects of multiple hypothesis testing appear, and false discovery rate analysis is described. Hierarchical clustering and bi-clustering appear naturally in the context of microarray analysis. Then the issues of sequence analysis (especially multiple sequence analysis) are approached using these HHM and Bayesian methods along with pattern discovery in the sequences.

Discovering regulatory sequence patterns is an especially important topic in this section. The topics of this chapter appear in computer science as "machine learning" or under "data mining"; here the subject is called statistical or Bayesian methods. Whatever it is named, this is an essential area for bioinformatics.

The next chapter (Chap. 5), "Algorithms in Computational Biology," takes up the formal computational approach to our biological problems. It should be pointed out that the previous chapters contained algorithmic content, but there it was less acknowledged. It is my belief that the statistical and algorithmic approaches go hand in hand. Even with the Euclid's algorithm example of the present chapter, there are statistical issues nearby. For example, the three descriptions of Euclid's algorithm are analyzed for time complexity. It is easy to ask how efficient the algorithms are on randomly chosen pairs of integers. What is the expected running time of the algorithms? What is the variance? Amazingly these questions have answers which are rather deep. The authors soon turn to dynamic programming (DP), and once again they present clear illustrative examples, in this case Fibonacci numbers. Designing DP algorithms for sequence alignment is covered. Then a more recently developed area of genome rearrangements is described along with some of the impressive (and deep) results from the area. This topic is relevant to whole genome analysis as chromosomes evolve on a larger scale than just alterations of individual letters as covered by sequence alignment.

In Chap. 6, "Multivariate Statistical Methods in Bioinformatics Research," we have a thorough excursion into multivariate statistics. This can be viewed as the third statistical chapter in this volume. Here the multivariate normal distribution is studied in its many rich incarnations. This is justified by the ubiquitous nature of the normal distribution. Just as with the bell-shaped curve which appears in one dimension due to the central limit theorem (add up enough independent random variables and suitably normalized, one gets the normal under quite general conditions), there is also a multivariate central limit theorem. Here detailed properties are described as well as related distributions such as the Wishart distribution (the analog of the chi-square). Estimation is relevant as is a multivariate t-test. Principal component analysis, factor analysis, and linear discriminant analysis are all covered with some nice examples to illustrate the power of approaches. Then classification problems and variable selection both give platforms to further illustrate and develop the methods on important bioinformatics application areas.

Chapter 7, "Association Analysis for Human Diseases: Methods and Examples," gives us the opportunity to look more deeply into aspects of genetics. While this chapter emphasizes statistics, be aware that computational issues also drive much of the research and cannot be ignored. Population genetics is introduced and then the important subjects of genetic linkage analysis and association studies. Genomic information such as single-nucleotide polymorphisms (SNPs) provide voluminous data for many of these studies, where multiple hypothesis testing is a critical issue.

Chapter 8, "Data Mining and Knowledge Discovery Methods with Case Examples," deals with the area of knowledge discovery and data mining. To quote the authors, this area "has emerged as an important research direction for extracting useful information from vast repositories of data of various types. The basic

concepts, problems and challenges deals with the area of knowledge discovery and data mining that has emerged as an important research direction for extracting useful information from vast repositories of data of various types. The basic concepts, problems and challenges are first briefly discussed. Some of the major data mining tasks like classification, clustering and association rule mining are then described in some detail. This is followed by a description of some tools that are frequently used for data mining. Two case examples of supervised and unsupervised classification for satellite image analysis are presented. Finally an extensive bibliography is provided."

The valuable chapter on Applied Bioinformatics Tools (Chap. 9) provides a step-by-step description of the application tools used in the course and data sources as well as a list of the problems. It should be strongly emphasized that no one learns this material without actually having hands-on experience with the derivations and the applications. This is not a subject for contemplation only!

Protein structure and function is a vast and critically important topic. In this collection it is covered by Chap. 10, "Foundations for the Study of Structure and Function of Proteins." There the detailed structure of amino acids is presented with their role in the various levels of protein structure (including amino acid sequence, secondary structure, tertiary structure, and spatial arrangements of the subunits). The geometry of the polypeptide chain is key to these studies as are the forces causing the three-dimensional structures (including electrostatic and van der Waals forces). Secondary structural units are classified into α-helix, β-sheets, and β-turns. Structural motifs and folds are described. Protein structure prediction is an active field, and various approaches are described including homology modeling and machine learning.

Systems biology is a recently described approach to combining system-wide data of biology in order to gain a global understanding of a biological system, such as a bacterial cell. The science is far from succeeding in this endeavor in general, let alone having powerful techniques to understand the biology of multicellular organisms. It is a grand challenge goal at this time. The fascinating chapter on Computational Systems Biology Approaches for Deciphering Traditional Chinese Medicine (Chap. 11) seeks to apply the computational systems biology (CSB) approach to traditional Chinese medicine (TCM). The chapter sets up parallel concepts between CSB and CTM. In Sect. 11.3.2 the main focus is "on a CSB-based case study for TCM ZHENG—a systems biology approach with the combination of computational analysis and animal experiment to investigate Cold ZHENG and Hot ZHENG in the context of the neuro-endocrine-immune (NEI) system." With increasing emphasis on the so-called nontraditional medicine, these studies have great potential to unlock new understandings for both CSB and TCM.

Finally I close with a few remarks about this general area. Biology is a major science for our new century; perhaps it will be the major science of the twenty-first century. However, if someone is not excited by biology, then they should find a subject that does excite them. I have almost continuously found the new discoveries such as introns or microRNA absolutely amazing. It is such a young science when such profound wonders keep showing up. Clearly no one analysis subject can

solve all the problems arising in modern computational molecular biology. Statistics alone, computer science alone, experimental molecular biology alone, none of these are sufficient in isolation. Protein structure studies require an entire additional set of tools such as classical mechanics. And as systems biology comes into play, systems of differential equations and scientific computing will surely be important. None of us can learn everything, but everyone working in this area needs a set of well-understood tools. We all learn new techniques as we proceed, learning things required to solve the problems. This requires people who evolve with the subject. This is exciting, but I admit it is hard work too. Bioinformatics will evolve as it confronts new data created by the latest biotechnology and biological sciences.

University of Southern California
Los Angeles, USA
March 2, 2013

Michael S. Waterman

Contents

此为试读,需要完整PDF请访问: www.ertongbook.com

此为试读,需要完整PDF请访问: www.ertongbook.com