CAMBRIDGE

Cambridge University Press
0521600820 - Essential Bioinformatics
Jin Xiong
Frontmatter
More information

ESSENTIAL BIOINFORMATICS

Essential Bioinformatics is a concise yet comprehensive textbook of bioinformatics that provides a broad introduction to the entire field. Written specifically for a life science audience, the basics of bioinformatics are explained, followed by discussions of the state-of-the-art computational tools available to solve biological research problems. All key areas of bioinformatics are covered including biological databases, sequence alignment, gene and promoter prediction, molecular phylogenetics, structural bioinformatics, genomics, and proteomics. The book emphasizes how computational methods work and compares the strengths and weaknesses of different methods. This balanced yet easily accessible text will be invaluable to students who do not have sophisticated computational backgrounds. Technical details of computational algorithms are explained with a minimum use of mathematical formulas; graphical illustrations are used in their place to aid understanding. The effective synthesis of existing literature as well as in-depth and up-to-date coverage of all key topics in bioinformatics make this an ideal textbook for all bioinformatics courses taken by life science students and for researchers wishing to develop their knowledge of bioinformatics to facilitate their own research.

Jin Xiong is an assistant professor of biology at Texas A&M University, where he has taught bioinformatics to graduate and undergraduate students for several years. His main research interest is in the experimental and bioinformatics analysis of photosystems.

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
0521600820 - Essential Bioinformatics
Jin Xiong
Frontmatter
More information

Essential Bioinformatics

JIN XIONG
Texas A&M University

CAMBRIDGE UNIVERSITY PRESS

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
0521600820 - Essential Bioinformatics
Jin Xiong
Frontmatter
More information

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
40 West 20th Street, New York, NY 10011-4211, USA

www.cambridge.org
Information on this title: www.cambridge.org/9780521840989

© Jin Xiong 2006

This publication is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without
the written permission of Cambridge University Press.

First published 2006

Printed in the United States of America

A catalog record for this publication is available from the British Library.

Library of Congress Cataloging in Publication Data

Xiong, Jin. Essential bioinformatics / Jin Xiong.
p. ; cm.
Includes bibliographical references and index.
ISBN-13: 978-0-521-84098-9 (hardback)
ISBN-10: 0-521-84098-8 (hardback)
ISBN-13: 978-0-521-60082-8 (pbk.)
ISBN-10: 0-521-60082-0 (pbk.)
1. Bioinformatics. I. Title.
[DNLM: 1. Computational Biology. QU 26.5 X6e 2006]
QH324.2.X56 2006
572.80285 – dc22 2005029453

ISBN-13 978-0-521-84098-9 hardback
ISBN-10 0-521-84098-8 hardback

ISBN-13 978-0-521-60082-8 paperback
ISBN-10 0-521-60082-0 paperback

Cambridge University Press has no responsibility for
the persistence or accuracy of URLs for external or
third-party Internet Web sites referred to in this publication
and does not guarantee that any content on such
Web sites is, or will remain, accurate or appropriate.

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
0521600820 - Essential Bioinformatics
Jin Xiong
Frontmatter
More information

Contents

Preface ■ ix

SECTION I INTRODUCTION AND BIOLOGICAL DATABASES

1 Introduction ■ 3
- What Is Bioinformatics? ■ 4
- Goal ■ 5
- Scope ■ 5
- Applications ■ 6
- Limitations ■ 7
- New Themes ■ 8
- Further Reading ■ 8

2 Introduction to Biological Databases ■ 10
- What Is a Database? ■ 10
- Types of Databases ■ 10
- Biological Databases ■ 13
- Pitfalls of Biological Databases ■ 17
- Information Retrieval from Biological Databases ■ 18
- Summary ■ 27
- Further Reading ■ 27

SECTION II SEQUENCE ALIGNMENT

3 Pairwise Sequence Alignment ■ 31
- Evolutionary Basis ■ 31
- Sequence Homology versus Sequence Similarity ■ 32
- Sequence Similarity versus Sequence Identity ■ 33
- Methods ■ 34
- Scoring Matrices ■ 41
- Statistical Significance of Sequence Alignment ■ 47
- Summary ■ 48
- Further Reading ■ 49

4 Database Similarity Searching ■ 51
- Unique Requirements of Database Searching ■ 51
- Heuristic Database Searching ■ 52
- Basic Local Alignment Search Tool (BLAST) ■ 52
- FASTA ■ 57
- Comparison of FASTA and BLAST ■ 60
- Database Searching with the Smith–Waterman Method ■ 61

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
0521600820 - Essential Bioinformatics
Jin Xiong
Frontmatter
More information

Summary ■ 61
Further Reading ■ 62

5 Multiple Sequence Alignment ■ 63
Scoring Function ■ 63
Exhaustive Algorithms ■ 64
Heuristic Algorithms ■ 65
Practical Issues ■ 71
Summary ■ 73
Further Reading ■ 74

6 Profiles and Hidden Markov Models ■ 75
Position-Specific Scoring Matrices ■ 75
Profiles ■ 77
Markov Model and Hidden Markov Model ■ 79
Summary ■ 84
Further Reading ■ 84

7 Protein Motifs and Domain Prediction ■ 85
Identification of Motifs and Domains in Multiple Sequence Alignment ■ 86
Motif and Domain Databases Using Regular Expressions ■ 86
Motif and Domain Databases Using Statistical Models ■ 87
Protein Family Databases ■ 90
Motif Discovery in Unaligned Sequences ■ 91
Sequence Logos ■ 92
Summary ■ 93
Further Reading ■ 94

SECTION III GENE AND PROMOTER PREDICTION

8 Gene Prediction ■ 97
Categories of Gene Prediction Programs ■ 97
Gene Prediction in Prokaryotes ■ 98
Gene Prediction in Eukaryotes ■ 103
Summary ■ 111
Further Reading ■ 111

9 Promoter and Regulatory Element Prediction ■ 113
Promoter and Regulatory Elements in Prokaryotes ■ 113
Promoter and Regulatory Elements in Eukaryotes ■ 114
Prediction Algorithms ■ 115
Summary ■ 123
Further Reading ■ 124

SECTION IV MOLECULAR PHYLOGENETICS

10 Phylogenetics Basics ■ 127
Molecular Evolution and Molecular Phylogenetics ■ 127
Terminology ■ 128
Gene Phylogeny versus Species Phylogeny ■ 130

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
0521600820 - Essential Bioinformatics
Jin Xiong
Frontmatter
More information

Forms of Tree Representation ■ 131
Why Finding a True Tree Is Difficult ■ 132
Procedure ■ 133
Summary ■ 140
Further Reading ■ 141

11 Phylogenetic Tree Construction Methods and Programs ■ 142
Distance-Based Methods ■ 142
Character-Based Methods ■ 150
Phylogenetic Tree Evaluation ■ 163
Phylogenetic Programs ■ 167
Summary ■ 168
Further Reading ■ 169

SECTION V STRUCTURAL BIOINFORMATICS

12 Protein Structure Basics ■ 173
Amino Acids ■ 173
Peptide Formation ■ 174
Dihedral Angles ■ 175
Hierarchy ■ 176
Secondary Structures ■ 178
Tertiary Structures ■ 180
Determination of Protein Three-Dimensional Structure ■ 181
Protein Structure Database ■ 182
Summary ■ 185
Further Reading ■ 186

13 Protein Structure Visualization, Comparison, and Classification ■ 187
Protein Structural Visualization ■ 187
Protein Structure Comparison ■ 190
Protein Structure Classification ■ 195
Summary ■ 199
Further Reading ■ 199

14 Protein Secondary Structure Prediction ■ 200
Secondary Structure Prediction for Globular Proteins ■ 201
Secondary Structure Prediction for Transmembrane Proteins ■ 208
Coiled Coil Prediction ■ 211
Summary ■ 212
Further Reading ■ 213

15 Protein Tertiary Structure Prediction ■ 214
Methods ■ 215
Homology Modeling ■ 215
Threading and Fold Recognition ■ 223
Ab Initio Protein Structural Prediction ■ 227
CASP ■ 228
Summary ■ 229
Further Reading ■ 230

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
0521600820 - Essential Bioinformatics
Jin Xiong
Frontmatter
More information

16 RNA Structure Prediction ■ 231
Introduction ■ 231
Types of RNA Structures ■ 233
RNA Secondary Structure Prediction Methods ■ 234
Ab Initio Approach ■ 234
Comparative Approach ■ 237
Performance Evaluation ■ 239
Summary ■ 239
Further Reading ■ 240

SECTION VI GENOMICS AND PROTEOMICS

17 Genome Mapping, Assembly, and Comparison ■ 243
Genome Mapping ■ 243
Genome Sequencing ■ 245
Genome Sequence Assembly ■ 246
Genome Annotation ■ 250
Comparative Genomics ■ 255
Summary ■ 259
Further Reading ■ 259

18 Functional Genomics ■ 261
Sequence-Based Approaches ■ 261
Microarray-Based Approaches ■ 267
Comparison of SAGE and DNA Microarrays ■ 278
Summary ■ 279
Further Reading ■ 280

19 Proteomics ■ 281
Technology of Protein Expression Analysis ■ 281
Posttranslational Modification ■ 287
Protein Sorting ■ 289
Protein–Protein Interactions ■ 291
Summary ■ 296
Further Reading ■ 296

APPENDIX

Appendix 1. Practical Exercises ■ 301

Appendix 2. Glossary ■ 318

Index ■ 331

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
0521600820 - Essential Bioinformatics
Jin Xiong
Frontmatter
More information

Preface

With a large number of prokaryotic and eukaryotic genomes completely sequenced and more forthcoming, access to the genomic information and synthesizing it for the discovery of new knowledge have become central themes of modern biological research. Mining the genomic information requires the use of sophisticated computational tools. It therefore becomes imperative for the new generation of biologists to be familiar with many bioinformatics programs and databases to tackle the new challenges in the genomic era. To meet this goal, institutions in the United States and around the world are now offering graduate and undergraduate students bioinformatics-related courses to introduce them to relevant computational tools necessary for the genomic research. To support this important task, this text was written to provide comprehensive coverage on the state-of-the-art of bioinformatics in a clear and concise manner.

The idea of writing a bioinformatics textbook originated from my experience of teaching bioinformatics at Texas A&M University. I needed a text that was comprehensive enough to cover all major aspects in the field, technical enough for a college-level course, and sufficiently up to date to include most current algorithms while at the same time being logical and easy to understand. The lack of such a comprehensive text at that time motivated me to write extensive lecture notes that attempted to alleviate the problem. The notes turned out to be very popular among the students and were in great demand from those who did not even take the class. To benefit a larger audience, I decided to assemble my lecture notes, as well as my experience and interpretation of bioinformatics, into a book.

This book is aimed at graduate and undergraduate students in biology, or any practicing molecular biologist, who has no background in computer algorithms but wishes to understand the fundamental principles of bioinformatics and use this knowledge to tackle his or her own research problems. It covers major databases and software programs for genomic data analysis, with an emphasis on the theoretical basis and practical applications of these computational tools. By reading this book, the reader will become familiar with various computational possibilities for modern molecular biological research and also become aware of the strengths and weaknesses of each of the software tools.

The reader is assumed to have a basic understanding of molecular biology and biochemistry. Therefore, many biological terms, such as *nucleic acids*, *amino acids*, *genes*, *transcription*, and *translation*, are used without further explanation. One exception is *protein structure*, for which a chapter about fundamental concepts is included so that

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
0521600820 - Essential Bioinformatics
Jin Xiong
Frontmatter
More information

algorithms and rationales for protein structural bioinformatics can be better understood. Prior knowledge of advanced statistics, probability theories, and calculus is of course preferable but not essential.

This book is organized into six sections: biological databases, sequence alignment, genes and promoter prediction, molecular phylogenetics, structural bioinformatics, and genomics and proteomics. There are nineteen chapters in total, each of which is relatively independent. When information from one chapter is needed for understanding another, cross-references are provided. Each chapter includes definitions and key concepts as well as solutions to related computational problems. Occasionally there are boxes that show worked examples for certain types of calculations. Since this book is primarily for molecular biologists, very few mathematical formulas are used. A small number of carefully chosen formulas are used where they are absolutely necessary to understand a particular concept. The background discussion of a computational problem is often followed by an introduction to related computer programs that are available online. A summary is also provided at the end of each chapter.

Most of the programs described in this book are online tools that are freely available and do not require special expertise to use them. Most of them are rather straightforward to use in that the user only needs to supply sequences or structures as input, and the results are returned automatically. In many cases, knowing which programs are available for which purposes is sufficient, though occasionally skills of interpreting the results are needed. However, in a number of instances, knowing the names of the programs and their applications is only half the journey. The user also has to make special efforts to learn the intricacies of using the programs. These programs are considered to be on the other extreme of user-friendliness. However, it would be impractical for this book to try to be a computer manual for every available software program. That is not my goal in writing the book. Nonetheless, having realized the difficulties of beginners who are often unaware of or, more precisely, intimidated by the numerous software programs available, I have designed a number of practical Web exercises with detailed step-by-step procedures that aim to serve as examples of the correct use of a combined set of bioinformatics tools for solving a particular problem. The exercises were originally written for use on a UNIX workstation. However, they can be used, with slight modifications, on any operating systems with Internet access.

In the course of preparing this book, I consulted numerous original articles and books related to certain topics of bioinformatics. I apologize for not being able to acknowledge all of these sources because of space limitations in such an introductory text. However, a small number of articles (mainly recent review articles) and books related to the topics of each chapter are listed as "Further Reading" for those who wish to seek more specialized information on the topics. Regarding the inclusion of computational programs, there are often a large number of programs available for a particular task. I apologize for any personal bias in the selection of the software programs in the book.

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
0521600820 - Essential Bioinformatics
Jin Xiong
Frontmatter
More information

One of the challenges in writing this text was to cover sufficient technical background of computational methods without extensive display of mathematical formulas. I strived to maintain a balance between explaining algorithms and not getting into too much mathematical detail, which may be intimidating for beginning students and nonexperts in computational biology. This sometimes proved to be a tough balance for me because I risk either sacrificing some of the original content or losing the reader. To alleviate this problem, I chose in many instances to use graphics instead of formulas to illustrate a concept and to aid understanding.

I would like to thank the Department of Biology at Texas A&M University for the opportunity of letting me teach a bioinformatics class, which is what made this book possible. I thank all my friends and colleagues in the Department of Biology and the Department of Biochemistry for their friendship. Some of my colleagues were kind enough to let me participate in their research projects, which provided me with diverse research problems with which I could hone my bioinformatics analysis skills. I am especially grateful to Lisa Peres of the Molecular Simulation Laboratory at Texas A&M, who was instrumental in helping me set up and run the laboratory section of my bioinformatics course. I am also indebted to my former postdoctoral mentor, Carl Bauer of Indiana University, who gave me the wonderful opportunity to learn evolution and phylogenetics in great depth, which essentially launched my career in bioinformatics. Also importantly, I would like to thank Katrina Halliday, my editor at Cambridge University Press, for accepting the manuscript and providing numerous suggestions for polishing the early draft. It was a great pleasure working with her. Thanks also go to Cindy Fullerton and Marielle Poss for their diligent efforts in overseeing the copyediting of the book to ensure a quality final product.

Jin Xiong

© Cambridge University Press www.cambridge.org