PROBLEMS AND SOLUTIONS IN BIOLOGICAL SEQUENCE ANALYSIS

This book is the first of its kind to provide a large collection of bioinformatics problems with accompanying solutions. Notably, the problem set includes all of the problems offered in *Biological Sequence Analysis* (*BSA*), by Durbin *et al.*, widely adopted as a required text for bioinformatics courses at leading universities worldwide. Although many of the problems included in *BSA* as exercises for its readers have been repeatedly used for homework and tests, no detailed solutions for the problems were available. Bioinformatics instructors had therefore frequently expressed a need for fully worked solutions and a larger set of problems for use in courses.

This book provides just that: following the same structure as *BSA*, and significantly extending the set of workable problems, it will facilitate a better understanding of the contents of the chapters in *BSA* and will help its readers develop problem solving skills that are vitally important for conducting successful research in the growing field of bioinformatics. All of the material has been class-tested by the authors at Georgia Tech, where the first ever M.Sc. degree program in Bioinformatics was held.

MARK BORODOVSKY is the Regents' Professor of Biology and Biomedical Engineering and Director of the Center for Bioinformatics and Computational Biology at Georgia Institute of Technology in Atlanta. He is the founder of the Georgia Tech M.Sc. and Ph.D. degree programs in Bioinformatics. His research interests are in bioinformatics and systems biology. He has taught Bioinformatics courses since 1994.

SVETLANA EKISHEVA is a research scientist at the School of Biology, Georgia Institute of Technology, Atlanta. Her research interests are in bioinformatics, applied statistics, and stochastic processes. Her expertise includes teaching probability theory and statistics at universities in Russia and in the USA.

PROBLEMS AND SOLUTIONS IN BIOLOGICAL SEQUENCE ANALYSIS

MARK BORODOVSKY AND SVETLANA EKISHEVA

CAMBRIDGE
UNIVERSITY PRESS

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

Cambridge University Press
The Edinburgh Building, Cambridge CB2 2RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521847544

First published 2006

Printed in the United Kingdom at the University Press, Cambridge

A catalogue record for this publication is available from the British Library

ISBN-13 978-0-521-84754-4 hardback
ISBN-10 0-521-84754-0 hardback

ISBN-13 978-0-521-61230-2 paperback
ISBN-10 0-521-61230-6 paperback

M. B.:
To Richard and Judy Lincoff

S. E.:
To Sergey and Natasha

Contents

Preface *page* xi

1 Introduction 1
- 1.1 Original problems 2
- 1.2 Additional problems 5
- 1.3 Further reading 23

2 Pairwise alignment 24
- 2.1 Original problems 24
- 2.2 Additional problems and theory 43
  - 2.2.1 Derivation of the amino acid substitution matrices (PAM series) 46
  - 2.2.2 Distributions of similarity scores 57
  - 2.2.3 Distribution of the length of the longest common word among several unrelated sequences 62
- 2.3 Further reading 65

3 Markov chains and hidden Markov models 67
- 3.1 Original problems 68
- 3.2 Additional problems and theory 77
  - 3.2.1 Probabilistic models for sequences of symbols: selection of the model and parameter estimation 86
  - 3.2.2 Bayesian approach to sequence composition analysis: the segmentation model by Liu and Lawrence 95
- 3.3 Further reading 102

4 Pairwise alignment using HMMs 104
- 4.1 Original problems 105
- 4.2 Additional problems 113
- 4.3 Further reading 125

5 Profile HMMs for sequence families 126
5.1 Original problems 127
5.2 Additional problems and theory 137
5.2.1 Discrimination function and maximum discrimination weights 150
5.3 Further reading 161

6 Multiple sequence alignment methods 162
6.1 Original problem 163
6.2 Additional problems and theory 163
6.2.1 Carrillo–Lipman multiple alignment algorithm 164
6.2.2 Progressive alignments: the Feng–Doolittle algorithm 171
6.2.3 Gibbs sampling algorithm for local multiple alignment 179
6.3 Further reading 181

7 Building phylogenetic trees 183
7.1 Original problems 183
7.2 Additional problems 211
7.3 Further reading 215

8 Probabilistic approaches to phylogeny 218
8.1 Original problems 219
8.1.1 Bayesian approach to finding the optimal tree and the Mau–Newton–Larget algorithm 235
8.2 Additional problems and theory 259
8.2.1 Relationship between sequence evolution models described by the Markov and the Poisson processes 264
8.2.2 Thorne–Kishino–Felsenstein model of sequence evolution with substitutions, insertions, and deletions 270
8.2.3 More on the rates of substitution 275
8.3 Further reading 277

9 Transformational grammars 279
9.1 Original problems 280
9.2 Further reading 290

10 RNA structure analysis 291
10.1 Original problems 292
10.2 Further reading 308

11 Background on probability 311
11.1 Original problems 311
11.2 Additional problem 326
11.3 Further reading 327

References 328

Index 343

Preface

Bioinformatics, an integral part of post-genomic biology, creates principles and ideas for computational analysis of biological sequences. These ideas facilitate the conversion of the flood of sequence data unleashed by the recent information explosion in biology into a continuous stream of discoveries. Not surprisingly, the new biology of the twenty-first century has attracted the interest of many talented university graduates with various backgrounds. Teaching bioinformatics to such a diverse audience presents a well-known challenge. The approach requiring students to advance their knowledge of computer programming and statistics prior to taking a comprehensive core course in bioinformatics has been accepted by many universities, including the Georgia Institute of Technology, Atlanta, USA.

In 1998, at the start of our graduate program, we selected the then recently published book *Biological Sequence Analysis* (*BSA*) by Richard Durbin, Anders Krogh, Sean R. Eddy, and Graeme Mitchison as a text for the core course in bioinformatics. Through the years, *BSA*, which describes the ideas of the major bioinformatic algorithms in a remarkably concise and consistent manner, has been widely adopted as a required text for bioinformatics courses at leading universities around the globe. Many problems included in *BSA* as exercises for its readers have been repeatedly used for homeworks and tests. However, the detailed solutions to these problems have not been available. The absence of such a resource was noticed by students and teachers alike.

The goal of this book, *Problems and Solutions in Biological Sequence Analysis* is to close this gap, extend the set of workable problems, and help its readers develop problem-solving skills that are vitally important for conducting successful research in the growing field of bioinformatics. We hope that this book will facilitate understanding of the content of the *BSA* chapters and also will provide an additional perspective for in-depth *BSA* reading by those who might not be able to take a formal bioinformatics course. We have augmented the set of original *BSA* problems with many new problems, primarily those that were offered to the Georgia Tech graduate students.

Probabilistic modeling and statistical analysis are frequently used in bioinformatics research. The mainstream bioinformatics algorithms, those for pairwise and multiple sequence alignment, gene finding, detecting orthologs, and building phylogenetic trees, would not work without rational model selection, parameter estimation, properly justified scoring systems, and assessment of statistical significance. These and many other elements of efficient bioinformatic tools require one to take into account the random nature of DNA and protein sequences.

As it has been illustrated by the *BSA* authors, probabilistic modeling laid the foundation for the development of powerful methods and algorithms for biological sequence interpretation and the revelation of its functional meaning and evolutionary connections. Notably, probabilistic modeling is a generalization of strictly deterministic modeling, which has a remarkable tradition in natural science. This tradition could be traced back to the explanation of astronomic observations on the motion of solar system planets by Isaac Newton, who suggested a concise model combining the newly discovered law of gravity and the laws of dynamics.

The maximum likelihood principle of statistics, notwithstanding the fashion of its traditional application, also has its roots in "deterministic" science that suggests that the chosen structure and parameters of a theoretical model should provide the best match of predictions to experimental observations. For instance, one could recognize the maximum likelihood approach in Francis Crick and James Watson's inference of the DNA double helix model, chosen from the combinatorial number of biochemically viable alternatives as the best fit to the X-ray data on DNA three-dimensional structure and other experimental data available.

In studying the processes of inheritance and molecular evolution, where random factors play important roles, fully fledged probabilistic models enter the picture. A classic cycle of experiments, data analysis, and modeling with search for a best fit of the models to data was designed and implemented by Gregor Mendel. His remarkable long term research endeavor provided proof of the existence of discrete units of inheritance, the genes.

When we deal with data coming from a less controllable environment, such as data on natural biological evolution spanning time periods on a scale of millions of years, the problem is even more challenging. Still, the situation is hopeful. The models of molecular evolution proposed by Dayhoff and co-authors, Jukes and Cantor, and Kimura, are classical examples of fundamental advances in modeling of the complex processes of DNA and protein evolution. Notably these models focus on only a single site of a molecular sequence and require the further simplifying assumption that evolution of sequence sites occurs independently from each other. Nevertheless, such models are useful starting points for understanding the

function and evolution of biological sequences as well as for designing algorithms elucidating these functional and evolutionary connections.

For instance, amino acid substitution scores are critically important parameters of the optimal global (Needleman and Wunsch) and local (Smith and Waterman) sequence alignment algorithms. Biologically sensible derivation of the substitution scores is impossible without models of protein evolution.

In the mid 1990s the notion of the hidden Markov model (HMM), having been of great practical use in speech recognition, was introduced to bioinformatics and quickly entered the mainstream of the modeling techniques in biological sequence analysis.

Theoretical advances that have occurred since the mid 1990s have shown that the sequence alignment problem has a natural probabilistic interpretation in terms of hidden Markov models. In particular, the dynamic programming (DP) algorithm for pairwise and multiple sequence alignment has the HMM-based algorithmic equivalent, the Viterbi algorithm. If the type of probabilistic model for a biological sequence has been chosen, parameters of the model could be inferred by statistical (machine learning) methods. Two competitive models could be compared to identify the one with the best fit.

The events and selective forces of the past, moving the evolution of biological species, have to be reconstructed from the current biological sequence data containing significant noise caused by all the changes that have occurred in the lifetime of disappeared generations. This difficulty can be overcome to some extent by the use of the general concept of self-consistent models with parameters adjusted iteratively to fit the growing collection of sequence data. Subsequently, implementation of this concept requires the expectation–maximization type algorithms able to estimate the model parameters simultaneously with rearranging data to produce the data structure (such as a multiple alignment) that fits the model better. *BSA* describes several algorithms of expectation–maximization type, including the self-training algorithm for a profile HMM and the self-training algorithm for a phylogenetic HMM. Given that the practice with many algorithms described in *BSA* requires significant computer programming, one may expect that describing the solutions would lead us into heavy computer codes, thus moving far away from the initial concepts and ideas. However, the majority of the *BSA* exercises have analytical solutions. On several occasions we have illustrated the implementations of the algorithms by "toy" examples. The computer codes written in C++ and Perl languages for such examples are available at opal.biology.gatech.edu/PSBSA. Note, that in the "Further reading" sections we include mostly papers that were published later than 1998, the year of *BSA* publication. Finally, we should mention that the references in the text to the pages in the BSA book cite the 2006 edition.

Acknowledgements

We thank Sergey Latkin, Svetlana's husband, for the remarkable help with preparation of LaTex figures and tables. We are grateful to Alexandre Lomsadze, Ryan Mills, Yuan Tian, Burcu Bakir, Jittima Piriyapongsa, Vardges Ter-Hovhannisyan, Wenhan Zhu, Jeffrey Yunes, and Matthew Berginski for invaluable technical assistance in preparation of the book materials; to Soojin Yi, and Galina Glazko for useful references on molecular evolution; to Michael Roytberg for helpful discussions on transformational grammars and finite automata. We cordially thank our editor Katrina Halliday for tremendous patience and constant support, without which this book would never have come to fruition. We are especially grateful to Richard Durbin, Anders Krogh, Sean R. Eddy, and Graeme Mitchison, for encouragement, helpful criticism and suggestions. Further, it is our pleasure to acknowledge firm support from the Georgia Tech School of Biology and the Wallace H. Coulter Department of Biomedical Engineering at Georgia Tech and Emory University. Finally, we wish to express our particular gratitude to our families for great patience and constant understanding.

M.B. and S.E.