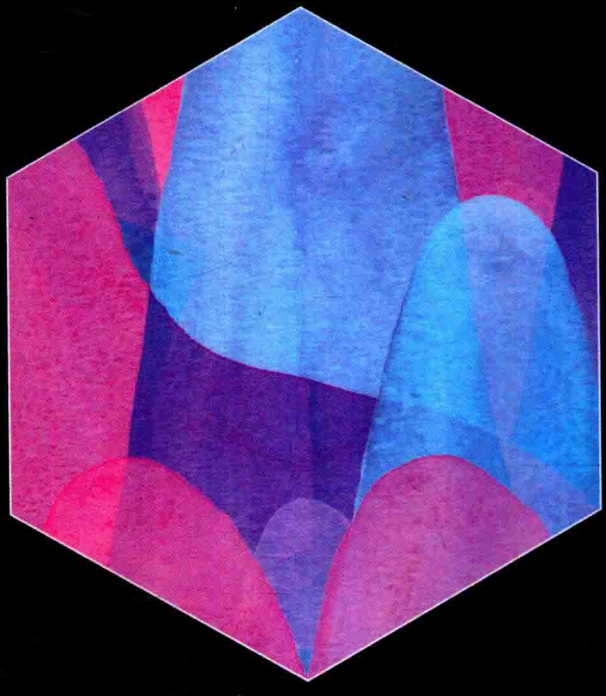
Chapman & Hall/CRC
Mathematical and Computational Biology Series

Statistical Modeling and Machine Learning for Molecular Biology



Alan M. Moses

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Chapman & Hall/CRC Mathematical and Computational Biology Series

Statistical Modeling and Machine Learning for Molecular Biology

Alan M. Moses
University of Toronto, Canada

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed by CPI on sustainably sourced paper
Version Date: 20160930

International Standard Book Number-13: 978-1-4822-5859-2 (Paperback)

Library of Congress Cataloging-in-Publication Data

Names: Moses, Alan M., author.
Title: Statistical modeling and machine learning for molecular biology / Alan M. Moses.
Description: Boca Raton : CRC Press, 2016. | Includes bibliographical references and index.
Identifiers: LCCN 2016028358| ISBN 9781482258592 (hardback : alk. paper) | ISBN 9781482258615 (e-book) | ISBN 9781482258622 (e-book) | ISBN 9781482258608 (e-book)
Subjects: LCSH: Molecular biology–Statistical methods. | Molecular biology–Data processing.
Classification: LCC QH506 .M74 2016 | DDC 572.8–dc23
LC record available at https://lccn.loc.gov/2016028358

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Statistical Modeling and Machine Learning for Molecular Biology

CHAPMAN & HALL/CRC
Mathematical and Computational Biology Series

Aims and scope:
This series aims to capture new developments and summarize what is known over the entire spectrum of mathematical and computational biology and medicine. It seeks to encourage the integration of mathematical, statistical, and computational methods into biology by publishing a broad range of textbooks, reference works, and handbooks. The titles included in the series are meant to appeal to students, researchers, and professionals in the mathematical, statistical and computational sciences, fundamental biology and bioengineering, as well as interdisciplinary researchers involved in the field. The inclusion of concrete examples and applications, and programming techniques and examples, is highly encouraged.

Series Editors

Proposals for the series should be submitted to one of the series editors above or directly to:
CRC Press, Taylor & Francis Group
3 Park Square, Milton Park
Abingdon, Oxfordshire OX14 4RN
UK

Published Titles

An Introduction to Systems Biology: Design Principles of Biological Circuits
Uri Alon

Glycome Informatics: Methods and Applications
Kiyoko F. Aoki-Kinoshita

Computational Systems Biology of Cancer
Emmanuel Barillot, Laurence Calzone, Philippe Hupé, Jean-Philippe Vert, and Andrei Zinovyev

Python for Bioinformatics
Sebastian Bassi

Quantitative Biology: From Molecular to Cellular Systems
Sebastian Bassi

Methods in Medical Informatics: Fundamentals of Healthcare Programming in Perl, Python, and Ruby
Jules J. Berman

Computational Biology: A Statistical Mechanics Perspective
Ralf Blossey

Game-Theoretical Models in Biology
Mark Broom and Jan Rychtář

Computational and Visualization Techniques for Structural Bioinformatics Using Chimera
Forbes J. Burkowski

Structural Bioinformatics: An Algorithmic Approach
Forbes J. Burkowski

Spatial Ecology
Stephen Cantrell, Chris Cosner, and Shigui Ruan

Cell Mechanics: From Single Scale-Based Models to Multiscale Modeling
Arnaud Chauvière, Luigi Preziosi, and Claude Verdier

Bayesian Phylogenetics: Methods, Algorithms, and Applications
Ming-Hui Chen, Lynn Kuo, and Paul O. Lewis

Statistical Methods for QTL Mapping
Zehua Chen

Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems
Qiang Cui and Ivet Bahar

Kinetic Modelling in Systems Biology
Oleg Demin and Igor Goryanin

Data Analysis Tools for DNA Microarrays
Sorin Draghici

Statistics and Data Analysis for Microarrays Using R and Bioconductor, Second Edition
Sorin Drăghici

Computational Neuroscience: A Comprehensive Approach
Jianfeng Feng

Biological Sequence Analysis Using the SeqAn C++ Library
Andreas Gogol-Döring and Knut Reinert

Gene Expression Studies Using Affymetrix Microarrays
Hinrich Göhlmann and Willem Talloen

Handbook of Hidden Markov Models in Bioinformatics
Martin Gollery

Meta-analysis and Combining Information in Genetics and Genomics
Rudy Guerra and Darlene R. Goldstein

Differential Equations and Mathematical Biology, Second Edition
D.S. Jones, M.J. Plank, and B.D. Sleeman

Knowledge Discovery in Proteomics
Igor Jurisica and Dennis Wigle

Introduction to Proteins: Structure, Function, and Motion
Amit Kessel and Nir Ben-Tal

RNA-seq Data Analysis: A Practical Approach
Eija Korpelainen, Jarno Tuimala, Panu Somervuo, Mikael Huss, and Garry Wong

Introduction to Mathematical Oncology
Yang Kuang, John D. Nagy, and Steffen E. Eikenberry

Biological Computation
Ehud Lamm and Ron Unger

Published Titles (continued)

Optimal Control Applied to Biological Models
Suzanne Lenhart and John T. Workman

Clustering in Bioinformatics and Drug Discovery
John D. MacCuish and Norah E. MacCuish

Spatiotemporal Patterns in Ecology and Epidemiology: Theory, Models, and Simulation
Horst Malchow, Sergei V. Petrovskii, and Ezio Venturino

Stochastic Dynamics for Systems Biology
Christian Mazza and Michel Benaïm

Statistical Modeling and Machine Learning for Molecular Biology
Alan M. Moses

Engineering Genetic Circuits
Chris J. Myers

Pattern Discovery in Bioinformatics: Theory & Algorithms
Laxmi Parida

Exactly Solvable Models of Biological Invasion
Sergei V. Petrovskii and Bai-Lian Li

Computational Hydrodynamics of Capsules and Biological Cells
C. Pozrikidis

Modeling and Simulation of Capsules and Biological Cells
C. Pozrikidis

Cancer Modelling and Simulation
Luigi Preziosi

Introduction to Bio-Ontologies
Peter N. Robinson and Sebastian Bauer

Dynamics of Biological Systems
Michael Small

Genome Annotation
Jung Soh, Paul M.K. Gordon, and Christoph W. Sensen

Niche Modeling: Predictions from Statistical Distributions
David Stockwell

Algorithms in Bioinformatics: A Practical Introduction
Wing-Kin Sung

Introduction to Bioinformatics
Anna Tramontano

The Ten Most Wanted Solutions in Protein Bioinformatics
Anna Tramontano

Combinatorial Pattern Matching Algorithms in Computational Biology Using Perl and R
Gabriel Valiente

Managing Your Biological Data with Python
Allegra Via, Kristian Rother, and Anna Tramontano

Cancer Systems Biology
Edwin Wang

Stochastic Modelling for Systems Biology, Second Edition
Darren J. Wilkinson

Big Data Analysis for Bioinformatics and Biomedical Discoveries
Shui Qing Ye

Bioinformatics: A Practical Approach
Shui Qing Ye

Introduction to Computational Proteomics
Golan Yona

For my parents

Acknowledgments

First, I'd like to acknowledge the people who taught me statistics and computers. As with most of the people that will read this book, I took the required semester of statistics as an undergraduate. Little of what I learned proved useful for my scientific career. I came to statistics and computers late, although I learned some html during a high-school job at PCI Geomatics and tried (and failed) to write my first computer program as an undergraduate hoping to volunteer in John Reinitz's lab (then at Mount Sinai in New York). I finally did manage to write some programs as an undergraduate summer student, thanks to Tim Gardner (then a grad student in Marcelo Magnasco's lab), who first showed me PERL codes.

Most of what I learned was during my PhD with Michael Eisen (who reintroduced cluster analysis to molecular biologists with his classic paper in 1998) and postdoctoral work with Richard Durbin (who introduced probabilistic models from computational linguistics to molecular biologists, leading to such universal resources as Pfam, and wrote a classic bioinformatics textbook, to which I am greatly indebted). During my PhD and postdoctoral work, I learned a lot of what is found in this book from Derek Chiang, Audrey Gasch, Justin Fay, Hunter Fraser, Dan Pollard, David Carter, and Avril Coughlan. I was also very fortunate to take courses with Terry Speed, Mark van der Laan, and Michael Jordan while at UC Berkeley and to have sat in on Geoff Hinton's advanced machine learning lectures in Toronto in 2012 before he was whisked off to Google. Most recently, I've been learning from Quaid Morris, with whom I cotaught the course that inspired this book.

I'm also indebted to everyone who read this book and gave me feedback while I was working on it: Miranda Calderon, Drs. Gelila Tilahun, Muluye Liku, and Derek Chiang, my graduate students Mitchell Li Cheong Man, Gavin Douglas, and Alex Lu, as well as an anonymous reviewer.

Much of this book was written while I was on sabbatical in 2014–2015 at Michael Elowitz's lab at Caltech, so I need to acknowledge Michael's generosity to host me and also the University of Toronto for continuing the tradition of academic leave. Michael and Joe Markson introduced me to the ImmGen and single-cell sequence datasets that I used for many of the examples in this book.

Finally, to actually make this book (and the graduate course that inspired it) possible, I took advantage of countless freely available software, R packages, Octave, PERL, bioinformatics databases, *Wikipedia* articles and open-access publications, and supplementary data sets, many of which I have likely neglected to cite. I hereby acknowledge all of the people who make this material available and enable the progress of pedagogy.

Contents

I

Overview

The first four chapters give necessary background. The first chapter is background to the book: what it covers and why I wrote it. The next three chapters are background material needed for the statistical modeling and machine learning methods covered in the later chapters. However, although I've presented that material as background, I believe that the review of modeling and statistics (in Chapters 2, 3 and 4) might be valuable to readers, whether or not they intend to go on to the later chapters.