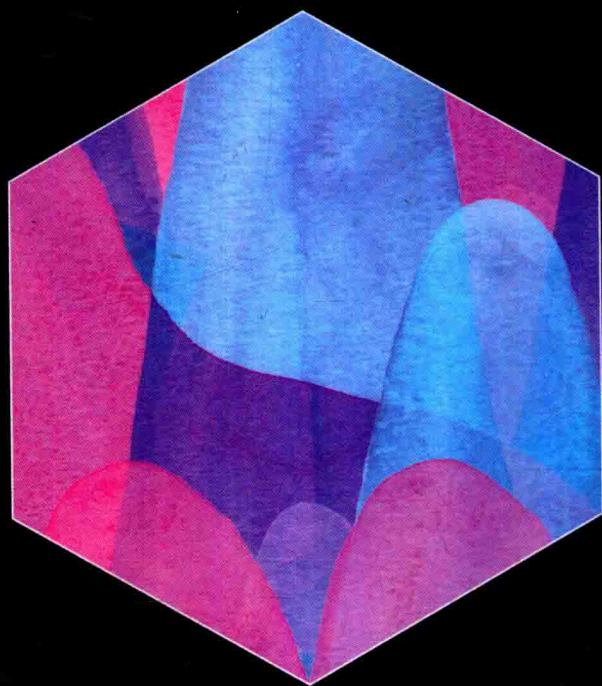


Chapman & Hall/CRC  
Mathematical and Computational Biology Series

# 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

© 2017 by Taylor & Francis Group, LLC  
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works

Printed by CPI on sustainably sourced paper  
Version Date: 20160930

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

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access [www.copyright.com](http://www.copyright.com) (<http://www.copyright.com/>) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

**Trademark Notice:** Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

---

#### 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://lcn.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**

N. F. Britton

*Department of Mathematical Sciences  
University of Bath*

Xihong Lin

*Department of Biostatistics  
Harvard University*

Nicola Mulder

*University of Cape Town  
South Africa*

Maria Victoria Schneider

*European Bioinformatics Institute*

Mona Singh

*Department of Computer Science  
Princeton University*

Anna Tramontano

*Department of Physics  
University of Rome La Sapienza*

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 Draghici*

### **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

---

Acknowledgments, xv

SECTION I   **Overview**

CHAPTER 1   ■ Across Statistical Modeling and Machine Learning on a Shoestring	3
1.1   ABOUT THIS BOOK	3
1.2   WHAT WILL THIS BOOK COVER?	4
1.2.1   Clustering	4
1.2.2   Regression	5
1.2.3   Classification	6
1.3   ORGANIZATION OF THIS BOOK	6
1.4   WHY ARE THERE MATHEMATICAL CALCULATIONS IN THE BOOK?	8
1.5   WHAT WON'T THIS BOOK COVER?	11
1.6   WHY IS THIS A BOOK?	12
REFERENCES AND FURTHER READING	14
CHAPTER 2   ■ Statistical Modeling	15
2.1   WHAT IS STATISTICAL MODELING?	15
2.2   PROBABILITY DISTRIBUTIONS ARE THE MODELS	18
2.3   AXIOMS OF PROBABILITY AND THEIR CONSEQUENCES: "RULES OF PROBABILITY"	23
2.4   HYPOTHESIS TESTING: WHAT YOU PROBABLY ALREADY KNOW ABOUT STATISTICS	26

2.5	TESTS WITH FEWER ASSUMPTIONS	30
2.5.1	Wilcoxon Rank-Sum Test, Also Known As the Mann–Whitney $U$ Test (or Simply the WMW Test)	30
2.5.2	Kolmogorov–Smirnov Test (KS-Test)	31
2.6	CENTRAL LIMIT THEOREM	33
2.7	EXACT TESTS AND GENE SET ENRICHMENT ANALYSIS	33
2.8	PERMUTATION TESTS	36
2.9	SOME POPULAR DISTRIBUTIONS	38
2.9.1	The Uniform Distribution	38
2.9.2	The $T$ -Distribution	39
2.9.3	The Exponential Distribution	39
2.9.4	The Chi-Squared Distribution	39
2.9.5	The Poisson Distribution	39
2.9.6	The Bernoulli Distribution	40
2.9.7	The Binomial Distribution	40
	EXERCISES	40
	REFERENCES AND FURTHER READING	41
CHAPTER 3	Multiple Testing	43
3.1	THE BONFERRONI CORRECTION AND GENE SET ENRICHMENT ANALYSIS	43
3.2	MULTIPLE TESTING IN DIFFERENTIAL EXPRESSION ANALYSIS	46
3.3	FALSE DISCOVERY RATE	48
3.4	eQTLs: A VERY DIFFICULT MULTIPLE-TESTING PROBLEM	49
	EXERCISES	51
	REFERENCES AND FURTHER READING	52
CHAPTER 4	Parameter Estimation and Multivariate Statistics	53
4.1	FITTING A MODEL TO DATA: OBJECTIVE FUNCTIONS AND PARAMETER ESTIMATION	53
4.2	MAXIMUM LIKELIHOOD ESTIMATION	54
4.3	LIKELIHOOD FOR GAUSSIAN DATA	55

4.4	HOW TO MAXIMIZE THE LIKELIHOOD ANALYTICALLY	56
4.5	OTHER OBJECTIVE FUNCTIONS	60
4.6	MULTIVARIATE STATISTICS	64
4.7	MLEs FOR MULTIVARIATE DISTRIBUTIONS	69
4.8	HYPOTHESIS TESTING REVISITED: THE PROBLEMS WITH HIGH DIMENSIONS	77
4.9	EXAMPLE OF LRT FOR THE MULTINOMIAL: GC CONTENT IN GENOMES	80
	EXERCISES	83
	REFERENCES AND FURTHER READING	83

## SECTION II   **Clustering**

CHAPTER 5 ■	Distance-Based Clustering	87
5.1	MULTIVARIATE DISTANCES FOR CLUSTERING	87
5.2	AGGLOMERATIVE CLUSTERING	91
5.2	CLUSTERING DNA AND PROTEIN SEQUENCES	95
5.4	IS THE CLUSTERING RIGHT?	98
5.5	K-MEANS CLUSTERING	100
5.6	SO WHAT IS LEARNING ANYWAY?	106
5.7	CHOOSING THE NUMBER OF CLUSTERS FOR K-MEANS	107
5.8	K-MEDOIDS AND EXEMPLAR-BASED CLUSTERING	109
5.9	GRAPH-BASED CLUSTERING: "DISTANCES" VERSUS "INTERACTIONS" OR "CONNECTIONS"	110
5.10	CLUSTERING AS DIMENSIONALITY REDUCTION	113
	EXERCISES	113
	REFERENCES AND FURTHER READING	115
CHAPTER 6 ■	Mixture Models and Hidden Variables for Clustering and Beyond	117
6.1	THE GAUSSIAN MIXTURE MODEL	118
6.2	E-M UPDATES FOR THE MIXTURE OF GAUSSIANS	123

6.3	DERIVING THE E-M ALGORITHM FOR THE MIXTURE OF GAUSSIANS	127
6.4	GAUSSIAN MIXTURES IN PRACTICE AND THE CURSE OF DIMENSIONALITY	131
6.5	CHOOSING THE NUMBER OF CLUSTERS USING THE AIC	131
6.6	APPLICATIONS OF MIXTURE MODELS IN BIOINFORMATICS	133
	EXERCISES	141
	REFERENCES AND FURTHER READING	142

### SECTION III Regression

CHAPTER 7 ■ Univariate Regression	145
7.1 SIMPLE LINEAR REGRESSION AS A PROBABILISTIC MODEL	145
7.2 DERIVING THE MLEs FOR LINEAR REGRESSION	146
7.3 HYPOTHESIS TESTING IN LINEAR REGRESSION	149
7.4 LEAST SQUARES INTERPRETATION OF LINEAR REGRESSION	154
7.5 APPLICATION OF LINEAR REGRESSION TO eQTLs	155
7.6 FROM HYPOTHESIS TESTING TO STATISTICAL MODELING: PREDICTING PROTEIN LEVEL BASED ON mRNA LEVEL	157
7.7 REGRESSION IS NOT JUST “LINEAR”—POLYNOMIAL AND LOCAL REGRESSIONS	161
7.8 GENERALIZED LINEAR MODELS	165
EXERCISES	167
REFERENCES AND FURTHER READING	167
CHAPTER 8 ■ Multiple Regression	169
8.1 PREDICTING Y USING MULTIPLE Xs	169
8.2 HYPOTHESIS TESTING IN MULTIPLE DIMENSIONS: PARTIAL CORRELATIONS	171

8.3	EXAMPLE OF A HIGH-DIMENSIONAL MULTIPLE REGRESSION: REGRESSING GENE EXPRESSION LEVELS ON TRANSCRIPTION FACTOR BINDING SITES	174
8.4	AIC AND FEATURE SELECTION AND OVERFITTING IN MULTIPLE REGRESSION	179
	EXERCISES	182
	REFERENCES AND FURTHER READING	183
CHAPTER 9 ■ Regularization in Multiple Regression and Beyond		185
9.1	REGULARIZATION AND PENALIZED LIKELIHOOD	186
9.2	DIFFERENCES BETWEEN THE EFFECTS OF $L_1$ AND $L_2$ PENALTIES ON CORRELATED FEATURES	189
9.3	REGULARIZATION BEYOND SPARSITY: ENCOURAGING YOUR OWN MODEL STRUCTURE	190
9.4	PENALIZED LIKELIHOOD AS MAXIMUM A POSTERIORI (MAP) ESTIMATION	192
9.5	CHOOSING PRIOR DISTRIBUTIONS FOR PARAMETERS: HEAVY-TAILS IF YOU CAN	193
	EXERCISES	197
	REFERENCES AND FURTHER READING	199
SECTION IV <b>Classification</b>		
CHAPTER 10 ■ Linear Classification		203
10.1	CLASSIFICATION BOUNDARIES AND LINEAR CLASSIFICATION	205
10.2	PROBABILISTIC CLASSIFICATION MODELS	206
10.3	LOGISTIC REGRESSION	208
10.4	LINEAR DISCRIMINANT ANALYSIS (LDA) AND THE LOG LIKELIHOOD RATIO	210
10.5	GENERATIVE AND DISCRIMINATIVE MODELS FOR CLASSIFICATION	214
10.6	NAÏVE BAYES: GENERATIVE MAP CLASSIFICATION	215

10.7 TRAINING NAÏVE BAYES CLASSIFIERS	221
10.8 NAÏVE BAYES AND DATA INTEGRATION	222
EXERCISES	223
REFERENCES AND FURTHER READING	223
<b>CHAPTER 11 ■ Nonlinear Classification</b>	<b>225</b>
11.1 TWO APPROACHES TO CHOOSE NONLINEAR BOUNDARIES: DATA-GUIDED AND MULTIPLE SIMPLE UNITS	226
11.2 DISTANCE-BASED CLASSIFICATION WITH <i>k</i> -NEAREST NEIGHBORS	228
11.3 SVMs FOR NONLINEAR CLASSIFICATION	230
11.4 DECISION TREES	234
11.5 RANDOM FORESTS AND ENSEMBLE CLASSIFIERS: THE WISDOM OF THE CROWD	236
11.6 MULTICLASS CLASSIFICATION	237
EXERCISES	238
REFERENCES AND FURTHER READING	239
<b>CHAPTER 12 ■ Evaluating Classifiers</b>	<b>241</b>
12.1 CLASSIFICATION PERFORMANCE STATISTICS IN THE IDEAL CLASSIFICATION SETUP	241
12.2 MEASURES OF CLASSIFICATION PERFORMANCE	242
12.3 ROC CURVES AND PRECISION-RECALL PLOTS	245
12.4 EVALUATING CLASSIFIERS WHEN YOU DON'T HAVE ENOUGH DATA	248
12.5 LEAVE-ONE-OUT CROSS-VALIDATION	251
12.6 BETTER CLASSIFICATION METHODS VERSUS BETTER FEATURES	253
EXERCISES	254
REFERENCES AND FURTHER READING	255



# 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.