Algebraic Statistics for Computational Biology

"If you can't stand algebra, keep out of evolutionary biology"

– John Maynard Smith
[Smith, 1998, page ix]

Algebraic Statistics for Computational Biology

Edited by
Lior Pachter and Bernd Sturmfels
University of California at Berkeley

CAMBRIDGE
UNIVERSITY PRESS

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/9780521857000

First published 2005

Printed in the United States of America

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

Library of Congress Cataloging-in-Publication Data

Algebraic statistics for computational biology / edited by Lior Pachter, Bernd Sturmfels.
p. cm.
Includes bibliographical references and index.
ISBN 0-521-85700-7
1. Biometry. 2. Algebra. I. Pachter, Lior, 1973– II. Sturmfels, Bernd, 1962–
QH323.5.A43 2005
572.8′6 – dc22 2005050070

ISBN-13 978-0-521-85700-0 hardback
ISBN-10 0-521-85700-7 hardback

Contents

Preface *page* ix
Guide to the chapters xi
Acknowledgment of support xii

Part I Introduction to the four themes 1

1 Statistics ***L. Pachter and B. Sturmfels*** 3
1.1 Statistical models for discrete data 4
1.2 Linear models and toric models 9
1.3 Expectation Maximization 17
1.4 Markov models 24
1.5 Graphical models 33

2 Computation ***L. Pachter and B. Sturmfels*** 43
2.1 Tropical arithmetic and dynamic programming 44
2.2 Sequence alignment 49
2.3 Polytopes 59
2.4 Trees and metrics 67
2.5 Software 75

3 Algebra ***L. Pachter and B. Sturmfels*** 85
3.1 Varieties and Gröbner bases 86
3.2 Implicitization 94
3.3 Maximum likelihood estimation 102
3.4 Tropical geometry 109
3.5 The tree of life and other tropical varieties 117

4 Biology ***L. Pachter and B. Sturmfels*** 125
4.1 Genomes 126
4.2 The data 132
4.3 The problems 137
4.4 Statistical models for a biological sequence 141
4.5 Statistical models of mutation 147

Part II Studies on the four themes 161

5 Parametric Inference ***R. Mihaescu*** 165
5.1 Tropical sum-product decompositions 166
5.2 The polytope propagation algorithm 169
5.3 Algorithm complexity 173
5.4 Specialization of parameters 177

6 Polytope Propagation on Graphs ***M. Joswig*** 181
6.1 Polytopes from directed acyclic graphs 181
6.2 Specialization to hidden Markov models 185
6.3 An implementation in `polymake` 186
6.4 Returning to our example 191

7 Parametric Sequence Alignment
C. Dewey and K. Woods 193
7.1 Few alignments are optimal 193
7.2 Polytope propagation for alignments 195
7.3 Retrieving alignments from polytope vertices 199
7.4 Biologically correct alignments 202

8 Bounds for Optimal Sequence Alignment
S. Elizalde and F. Lam 206
8.1 Alignments and optimality 206
8.2 Geometric interpretation 208
8.3 Known bounds 211
8.4 The square root conjecture 212

9 Inference Functions ***S. Elizalde*** 215
9.1 What is an inference function? 215
9.2 The few inference functions theorem 217
9.3 Inference functions for sequence alignment 220

10 Geometry of Markov Chains ***E. Kuo*** 226
10.1 Viterbi sequences 226
10.2 Two- and three-state Markov chains 229
10.3 Markov chains with many states 231
10.4 Fully observed Markov models 233

11 Equations Defining Hidden Markov Models
N. Bray and J. Morton 237
11.1 The hidden Markov model 237
11.2 Gröbner bases 238
11.3 Linear algebra 240
11.4 Combinatorially described invariants 247

12 The EM Algorithm for Hidden Markov Models
I. B. Hallgrímsdóttir, R. A. Milowski and J. Yu 250
12.1 The EM algorithm for hidden Markov models 250
12.2 An implementation of the Baum–Welch algorithm 254
12.3 Plots of the likelihood surface 257
12.4 The EM algorithm and the gradient of the likelihood 261

13 Homology Mapping with Markov Random Fields ***A. Caspi*** 264
13.1 Genome mapping 264
13.2 Markov random fields 267
13.3 MRFs in homology assignment 270
13.4 Tractable MAP inference in a subclass of MRFs 273
13.5 The Cystic Fibrosis Transmembrane Regulator 276

14 Mutagenetic Tree Models ***N. Beerenwinkel and M. Drton*** 278
14.1 Accumulative evolutionary processes 278
14.2 Mutagenetic trees 279
14.3 Algebraic invariants 282
14.4 Mixture models 287

15 Catalog of Small Trees
M. Casanellas, L. D. Garcia, and S. Sullivant 291
15.1 Notation and conventions 291
15.2 Fourier coordinates 295
15.3 Description of website features 297
15.4 Example 298
15.5 Using the invariants 303

16 The Strand Symmetric Model
M. Casanellas and S. Sullivant 305
16.1 Matrix-valued Fourier transform 306
16.2 Invariants for the 3-taxa tree 310
16.3 G-tensors 314
16.4 Extending invariants 318
16.5 Reduction to $K_{1,3}$ 319

17 Extending Tree Models to Splits Networks ***D. Bryant*** 322
17.1 Trees, splits and splits networks 322
17.2 Distance-based models for trees and splits graphs 325
17.3 A graphical model on a splits network 326
17.4 Group-based mutation models 327
17.5 Group-based models for trees and splits 330
17.6 A Fourier calculus for splits networks 332

18 Small Trees and Generalized Neighbor-Joining 335
M. Contois and D. Levy 335
18.1 From alignments to dissimilarity 335
18.2 From dissimilarity to trees 337
18.3 The need for exact solutions 342
18.4 Jukes–Cantor triples 344

19 Tree Construction using Singular Value Decomposition 347
N. Eriksson 347
19.1 The general Markov model 347
19.2 Flattenings and rank conditions 348
19.3 Singular Value Decomposition 351
19.4 Tree-construction algorithm 352
19.5 Performance analysis 355

20 Applications of Interval Methods to Phylogenetics 359
R. Sainudiin and R. Yoshida 359
20.1 Brief introduction to interval analysis 360
20.2 Enclosing the likelihood of a compact set of trees 366
20.3 Global optimization 366
20.4 Applications to phylogenetics 371

21 Analysis of Point Mutations in Vertebrate Genomes 375
J. Al-Aidroos and S. Snir 375
21.1 Estimating mutation rates 375
21.2 The ENCODE data 378
21.3 Synonymous substitutions 379
21.4 The rodent problem 381

22 Ultra-Conserved Elements in Vertebrate and Fly Genomes 387
M. Drton, N. Eriksson and G. Leung 387
22.1 The data 387
22.2 Ultra-conserved elements 390
22.3 Biology of ultra-conserved elements 392
22.4 Statistical significance of ultra-conservation 400
References 403
Index 418

Preface

The title of this book reflects who we are: a computational biologist and an algebraist who share a common interest in statistics. Our collaboration sprang from the desire to find a mathematical language for discussing biological sequence analysis, with the initial impetus being provided by the introductory workshop on *Discrete and Computational Geometry* at the Mathematical Sciences Research Institute (MSRI) held at Berkeley in August 2003. At that workshop we began exploring the similarities between tropical matrix multiplication and the Viterbi algorithm for hidden Markov models. Our discussions ultimately led to two articles [Pachter and Sturmfels, 2004a,b] which are explained and further developed in various chapters of this book.

In the fall of 2003 we held a graduate seminar on *The Mathematics of Phylogenetic Trees*. About half of the authors of the second part of this book participated in that seminar. It was based on topics from the books [Felsenstein, 2003, Semple and Steel, 2003] but we also discussed other projects, such as Michael Joswig's polytope propagation on graphs (now Chapter 6). That seminar got us up to speed on research topics in phylogenetics, and led us to participate in the conference on *Phylogenetic Combinatorics* which was held in July 2004 in Uppsala, Sweden. In Uppsala we were introduced to David Bryant and his statistical models for split systems (now Chapter 17).

Another milestone was the workshop on *Computational Algebraic Statistics*, held at the American Institute for Mathematics (AIM) at Palo Alto in December 2003. That workshop was built on the algebraic statistics paradigm, which is that statistical models for discrete data can be regarded as solutions to systems of polynomial equations. Our current understanding of algebraic statistical models, maximum likelihood estimation and expectation maximization was shaped by the excellent discussions and lectures at AIM.

These developments led us to offer a mathematics graduate course titled *Algebraic Statistics for Computational Biology* in the fall of 2004. The course was attended mostly by mathematics students curious about computational biology, but also by computer scientists, statisticians, and bioengineering students interested in understanding the mathematical foundations of bioinformatics. Participants ranged from postdocs to first-year graduate students and even one undergraduate. The format consisted of lectures by us on basic principles

of algebraic statistics and computational biology, as well as student participation in the form of group projects and presentations. The class was divided into four sections, reflecting the four themes of algebra, statistics, computation and biology. Each group was assigned a handful of projects to pursue, with the goal of completing a written report by the end of the semester. In some cases the groups worked on the problems we suggested, but, more often than not, original ideas by group members led to independent research plans.

Halfway through the semester, it became clear that the groups were making fantastic progress, and that their written reports would contain many novel ideas and results. At that point, we thought about preparing a book. The first half of the book would be based on our own lectures, and the second half would consist of chapters based on the final term papers. A tight schedule was seen as essential for the success of such an undertaking, given that many participants would be leaving Berkeley and the momentum would be lost. It was decided that the book should be written by March 2005, or not at all.

We were fortunate to find a partner in Cambridge University Press, which agreed to work with us on our concept. We are especially grateful to our editor, David Tranah, for his strong encouragement, and his trust that our half-baked ideas could actually turn into a readable book. After all, we were proposing to write to a book with twenty-nine authors during a period of three months.

The project did become reality and the result is in your hands. It offers an accurate snapshot of what happened during our seminars at UC Berkeley in 2003 and 2004. Nothing more and nothing less. The choice of topics is certainly biased, and the presentation is undoubtedly very far from perfect. But we hope that it may serve as an invitation to biology for mathematicians, and as an invitation to algebra for biologists, statisticians and computer scientists. Following this preface, we have included a guide to the chapters and suggested entry points for readers with different backgrounds and interests. Additional information and supplementary material may be found on the book website at
`http://bio.math.berkeley.edu/ascb/`

Many friends and colleagues provided helpful comments and inspiration during the project. We especially thank Elizabeth Allman, Ruchira Datta, Manolis Dermitzakis, Serkan Hoşten, Ross Lippert, John Rhodes and Amelia Taylor. Serkan Hoşten was also instrumental in developing and guiding research which is described in Chapters 15 and 18.

Most of all, we are grateful to our wonderful students and postdocs from whom we learned so much. Their enthusiasm and hard work have been truly amazing. You will enjoy meeting them in Part II.

Lior Pachter and Bernd Sturmfels
Berkeley, California, May 2005

Guide to the chapters

The introductory Chapters 1–4 can be studied as a unit or read in parts with specific topics in mind. Although there are some dependencies and shared examples, the individual chapters are largely independent of each other. Suggested introductory sequences of study for specific topics are:

- Algebraic statistics: 1.1, 1.2, 1.4, 1.5.
- Maximum likelihood estimation: 1.1, 1.2, 1.3, 3.3.
- Tropical geometry: 2.1, 3.4, 3.5.
- Gröbner bases: 3.1, 3.2, 2.5.
- Comparative genomics: 4.1, 4.2, 4.3, 4.4, 4.5, 2.5.
- Sequence alignment: 1.1, 1.2, 1.4, 2.1, 2.2, 2.3.
- Phylogenetics: 1.1, 1.2, 1.4, 2.4, 3.4, 3.5, 4.5.

Dependencies of the Part II chapters on Part I are summarized in the table below. This should help readers interested in reading a specific chapter or section to find the location of background material. Pointers are also provided to related chapters that may be of interest.

Chapter	Prerequisites	Further reading
5	1.4, 2.2, 2.3	6, 7, 8, 9
6	1.1, 1.2, 1.4, 2.3	5, 7, 8, 9
7	2.2, 2.3	8
8	1.1, 1.2, 1.4, 2.1, 2.2, 2.3	5, 7, 9
9	1.5, 2.2, 2.3, 4.4	5, 8
10	1.1, 1.2, 1.4	9, 11
11	1.1, 1.2, 1.3, 3.1, 3.2	12
12	1.3, 1.4	4.4, 11
13	1.1, 1.2, 1.4, 1.5	22
14	1.1, 1.2, 1.4, 1.5, 3.1	11, 16
15	1.4, 3.1, 3.2, 3.3, 4.5	16, 17, 18, 19, 20
16	1.4, 3.1, 3.2, 4.5	15, 19
17	1.1, 1.2, 1.4, 1.5, 2.4 4.5	15, 18, 19
18	2.4, 4.5	20
19	2.4, 3.1, 4.5	15, 18
20	2.4, 4.5	17
21	1.4, 2.5, 4.5	17, 19
22	1.4, 4	7, 13, 21

Acknowledgment of support

We were fortunate to receive support from many agencies and institutions while working on the book. The following list is an acknowledgment of support for the many research activities that formed part of the *Algebraic Statistics for Computational Biology* book project.

Niko Beerenwinkel was funded by Deutsche Forschungsgemeinschaft (DFG) under Grant No. BE 3217/1-1. David Bryant was supported by NSERC grant number 238975-01 and FQRNT grant number 2003-NC-81840. Marta Casanellas was partially supported by RyC program of "Ministerio de Ciencia y Tecnologia", BFM2003-06001 and BIO2000-1352-C02-02 of "Plan Nacional I+D" of Spain. Anat Caspi was funded through the Genomics Training Grant at UC Berkeley: NIH 5-T32-HG00047. Mark Contois was partially supported by NSF grant DEB-0207090. Mathias Drton was support by NIH grant R01-HG02362-03. Dan Levy was supported by NIH grant GM 68423 and NSF grant DMS 9971169. Radu Mihaescu was supported by the Hertz foundation. Raaz Sainudiin was partly supported by a joint DMS/NIGMS grant 0201037. Sagi Snir was supported by NIH grant R01-HG02362-03. Kevin Woods was supported by NSF Grant DMS 0402148. Eric Kuo, Seth Sullivant and Josephine Yu were supported by NSF graduate research fellowships.

Lior Pachter was supported by NSF CAREER award CCF 03-47992, NIH grant R01-HG02362-03 and a Sloan Research Fellowship. He also acknowledges support from the Programs for Genomic Application (NHLBI). Bernd Sturmfels was supported by NSF grant DMS 0200729 and the Clay Mathematics Institute (July 2004). He was the Hewlett–Packard Research Fellow at the Mathematical Sciences Research Institute (MSRI) Berkeley during the year 2003–2004 which allowed him to study computational biology.

Finally, we thank staff at the University of California at Berkeley, Universitat de Barcelona (2001SGR-00071), the Massachusetts Institute of Technology and MSRI for extending hospitality to visitors at various times during which the book was being written.