

Statistics for Biology and Health

**Warren J. Ewens
Gregory R. Grant**

**Statistical
Methods in
Bioinformatics**

An Introduction



Springer

Warren J. Ewens

Gregory R. Grant

Statistical Methods in Bioinformatics: An Introduction

With 30 Illustrations



Springer

Warren J. Ewens
Department of Biology
University of Pennsylvania
Philadelphia, PA 19104-6018
USA
wewens@sas.upenn.edu

Gregory R. Grant
Penn Center for Computational Biology
University of Pennsylvania
Philadelphia, PA 19104
USA
ggrant@pcbi.upenn.edu

Series Editors

K. Dietz
Institut für Medizinische Biometrie
Universität Tübingen
Westbahnhofstrasse 55
D-72070 Tübingen
Germany

M. Gail
National Center Institute
Rockville, MD 20892
USA

K. Krickeberg
Le Chatelet
F-63270 Manglieu
France

J. Samet
School of Public Health
Department of Epidemiology
Johns Hopkins University
615 Wolfe Street
Baltimore, MD 21205-2103
USA

A. Tsiatis
Department of Statistics
North Carolina State University
Raleigh, NC 27695
USA

Library of Congress Cataloging-in-Publication Data

Ewens, W.J. (Warren John), 1937–

Statistical methods in bioinformatics : an introduction / Warren Ewens, Gregory Grant.

p. cm. — (Statistics for biology and health)

Includes bibliographical references and index.

ISBN 0-387-95229-2 (alk. paper)

1. Bioinformatics. 2. Statistical Methods. I. Grant, Gregory, 1965– II. Title. III. Series.
R858 .E986 2001

570'.1'5195—dc21

00-067923

Printed on acid-free paper.

© 2001 Springer-Verlag New York, Inc.

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer-Verlag New York, Inc., 175 Fifth Avenue, New York, NY 10010, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden. The use of general descriptive names, trade names, trademarks, etc., in this publication, even if the former are not especially identified, is not to be taken as a sign that such names, as understood by the Trade Marks and Merchandise Marks Act, may accordingly be used freely by anyone.

Production managed by MaryAnn Brickner; manufacturing supervised by Jerome Basma.

Camera-ready copy prepared from the authors' LaTeX files.

Printed and bound by Edwards Brothers, Inc., Ann Arbor, MI.

Printed in the United States of America.

9 8 7 6 5 4 3 (Corrected third printing, 2002)

ISBN 0-387-95229-2

SPIN 10868159

Springer-Verlag New York Berlin Heidelberg

A member of BertelsmannSpringer Science+Business Media GmbH

Statistics for Biology and Health

Series Editors

K. Dietz, M. Gail, K. Krickeberg, J. Samet, A. Tsiatis

Springer

New York

Berlin

Heidelberg

Barcelona

Hong Kong

London

Milan

Paris

Singapore

Tokyo

Statistics for Biology and Health

Everitt/Rabe-Hesketh: Analyzing Medical Data Using S-PLUS

Ewens/Grant: Statistical Methods in Bioinformatics: An Introduction.

Hougaard: Analysis of Multivariate Survival Data.

Klein/Moeschberger: Survival Analysis: Techniques for Censored and Truncated Data.

Kleinbaum: Logistic Regression: A Self-Learning Text.

Kleinbaum: Survival Analysis: A Self-Learning Text.

Lange: Mathematical and Statistical Methods for Genetic Analysis.

Manton/Singer/Suzman: Forecasting the Health of Elderly Populations.

Salsburg: The Use of Restricted Significance Tests in Clinical Trials.

Therneau/Grambsch: Modeling Survival Data: Extending the Cox Model.

Zhang/Singer: Recursive Partitioning in the Health Sciences.

For Kathy and Elisabetta

Preface

We take *bioinformatics* to mean the emerging field of science growing from the application of mathematics, statistics, and information technology, including computers and the theory surrounding them, to the study and analysis of very large biological, and particularly genetic, data sets. The field has been fueled by the increase in DNA data generation leading to the massive data sets already generated, and yet to be generated, in particular the data from the human genome project, as well as other genome projects.

Bioinformatics does not aim to lay down fundamental mathematical laws that govern biological systems parallel to those laid down in physics. Such laws, if they exist, are a long way from being determined for biological systems. Instead, at this stage the main utility of mathematics in the field is in the creation of tools that investigators can use to analyze data. For example, biologists need tools for the statistical assessment of the similarity between two or more DNA or protein sequences, for finding genes in genomic DNA, and for estimating differences in how genes are expressed in different tissues. Such tools involve statistical modeling of biological systems, and it is our belief that there is a need for a book that introduces probability, statistics, and stochastic processes in the context of bioinformatics. We hope to fill that need here.

The material in this text assumes little or no background in biology. The basic notions of biology that one needs in order to understand the material are outlined in Appendix A. Some further details that are necessary to understand particular applications are given in the context of those applications. The necessary background in mathematics is introductory courses in calculus and linear algebra. In order to be clear about notation and

terminology, as well as to organize several results that are needed in the text, a review of basic notions in mathematics is given in Appendix B. No computer science knowledge is assumed and no programming is necessary to understand the material.

Why are probability and statistics so important in bioinformatics? Bioinformatics involves the analysis of biological data. Many chance mechanisms are involved in the creation of these data, most importantly the many random processes inherent in biological evolution and the randomness inherent in any sampling process. Stochastic process theory involves the description of the evolution of random processes occurring over time or space. Biological evolution over eons has provided the outcome of one of the most complex stochastic processes imaginable, and one that requires complex stochastic process theory for its description and analysis.

Our aim is to give an introductory account of some of the probability theory, statistics, and stochastic process theory appropriate to computational biology and bioinformatics. This is not a “how-to” book, of which there are several in the literature, but it aims to fill a gap in the literature in the statistical and probabilistic aspects of bioinformatics. The earlier chapters in this book contain standard introductory material suitable for any statistics course. Even here, however, we have departed somewhat from well-trodden paths, beginning to focus on material of interest in bioinformatics, for example the theory of the maximum of several random variables, moment-generating functions, geometric random variables and their various generalizations, together with information theory and entropy. We have also provided, in Appendix B, some standard mathematical results that are needed as background for this and other material.

This text is by no means comprehensive. There are several books that cover some of the topics we consider at a more advanced level. The reader should approach this text in part as an introduction and a means of assessing his/her interest in and ability to pursue this field further. Thus we have not tried to cover a comprehensive list of topics, nor to cover those topics discussed in complete detail. No book can ever fulfill the task of providing a complete introduction to this subject, since bioinformatics is evolving too quickly. To learn this subject as it evolves one must ultimately turn to the literature of published articles in journals. We hope that this book will provide a first stepping stone leading the reader in this direction.

We also wish to appeal to trained statisticians and to give them an introduction to bioinformatics, since their contribution will be vital to the analysis of the biological data already at hand and, more important, to developing analyses for new forms of data to arrive in the future. Such readers should be able to read the latter chapters directly.

The statistical procedures currently used in this subject are often ad hoc, with different methods being used in different parts of the subject. We have tried to provide as many threads running through the book as possible in order to overcome this problem and to integrate the material.

One such thread is provided by aspects of the material on stochastic processes. BLAST is one of the most frequently used algorithms in applied statistics, one BLAST search being made every few seconds on average by bioinformatics researchers around the world. However, the stochastic process theory behind the statistical calculations used in this algorithm is not widely understood. We approach this theory by starting with random walks, and through these to sequential analysis theory and to Markov chains, and ultimately to BLAST. This sequence also leads to the theory of hidden Markov models and to evolutionary analyses.

We have chosen this thread for three reasons. The first is that BLAST theory is intrinsically important. The second, as just mentioned, is that this provides a coherent thread to the often unconnected aspects of stochastic process theory used in various areas of bioinformatics. The final reason is that, with the human genome data and the genomes of other important species complete at least in first draft, we wish to emphasize procedures that lead to the analysis of these data. The analysis of these data will require new and currently unpredictable statistical analyses, and in particular, the theory for the most recent and sophisticated versions of BLAST, and for its further developments, will require new advanced theory.

So far as more practical matters are concerned, we are well aware of the need for precision in presenting any mathematically based topic. However, we are also aware, for an applied field, of the perils of a too mathematically precise approach to probability, perhaps through measure theory. Our approach has tended to be less rather than more pedantic, and detailed qualifications that interrupt the flow and might annoy the reader have been omitted. As one example, we assume throughout that all random variables we consider have finite moments of all orders. This assumption enables us to avoid many minor (and in practice unimportant) qualifications to the analysis we present.

So far as statistical theory is concerned, the focus in this book is on *discrete* as opposed to *continuous* random variables, since (especially with DNA and protein sequences) discrete random variables are more relevant to bioinformatics. However, some aspects of the theory of discrete random variables are difficult, with no limiting distribution theory available for the maximum of these random variables. In this case progress is made by using theory from continuous random variables to provide bounds and approximations. Thus continuous random variables are also discussed in some detail in the early chapters.

The focus in this book is, as stated above, on probability, statistics, and stochastic processes. We do, however, discuss aspects of the important algorithmic side of bioinformatics, especially when relevant to these probabilistic topics. In particular, the dynamic programming algorithm is introduced because of its use in various probability applications, especially in hidden Markov models. Several books are already available that are devoted to algorithmic aspects of the subject.

In a broad interpretation of the word “bioinformatics” there are several areas of the application of statistics to bioinformatics that we do not develop. Thus we do not cover aspects of the statistical theory in genetics associated with disease finding and linkage analysis. This subject deserves an entire book on its own. Nor do we discuss the increasingly important applications of bioinformatics in the stochastic theory of evolutionary population genetics. Again, all these topics deserve a complete treatment of their own.

This book is based on lectures given to students in the two-semester course in bioinformatics and computational biology at the University of Pennsylvania given each year during the period 1995–2000. We are most indebted to Elisabetta Manduchi, from PCBI/CBIL, who helped at every stage in revising the material. We are also grateful to the late Christian Overton for guidance, inspiration, and friendship. We thank all other members of PCBI/CBIL who supported us in this task and patiently answered many questions, in particular Brian Brunk, Jonathan Schug, Chris Stoeckert, Jonathan Crabtree, Angel Pizarro, Deborah Pinney, Shannon McWeeney, Joan Mazzairelli, and Eugene Buehler. We also thank Warren Gish for his help on BLAST, and for letting us reproduce his BLAST print-out examples. We thank Chris Burge, Sandrine Dudoit, Terry Speed, Matt Werner, Alessandra Gallinari, Sam Sokolovski, Helen Murphy, Ethan Fingergerman, Aaron Shaver, and Sue Wilson for their help. Finally, we thank students in the computational biology courses we have taught for their comments on the material, which we have often incorporated into this book. Any errors or omissions are, of course, our own responsibility. An archive of errata will be maintained at <http://www.textbook-errata.org>.

Warren J. Ewens
Gregory R. Grant
Philadelphia, Pennsylvania
February, 2001

Contents

Preface	vii
1 Probability Theory (i): One Random Variable	1
1.1 Introduction	1
1.2 Discrete Random Variables, Definitions	3
1.2.1 Probability Distributions and Parameters	3
1.2.2 Independence	6
1.3 Six Important Discrete Probability Distributions	7
1.3.1 One Bernoulli Trial	7
1.3.2 The Binomial Distribution	7
1.3.3 The Uniform Distribution	9
1.3.4 The Geometric Distribution	10
1.3.5 The Negative Binomial and the Generalized Geometric Distributions	12
1.3.6 The Poisson Distribution	13
1.4 The Mean of a Discrete Random Variable	13
1.5 The Variance of a Discrete Random Variable	16
1.6 General Moments of a Probability Distribution	18
1.7 The Probability-Generating Function	19
1.8 Continuous Random Variables	21
1.9 The Mean, Variance, and Median of a Continuous Random Variable	22
1.9.1 Definitions	22
1.9.2 Chebyshev's Inequality	23

1.10	Five Important Continuous Distributions	24
1.10.1	The Uniform Distribution	24
1.10.2	The Normal Distribution	24
1.10.3	The Normal Approximation to a Discrete Distribution	27
1.10.4	The Exponential Distribution	28
1.10.5	The Gamma Distribution	31
1.10.6	The Beta Distribution	32
1.11	The Moment-Generating Function	32
1.12	Events	36
1.12.1	What Are Events?	36
1.12.2	Complements, Unions, and Intersections	36
1.12.3	Probabilities of Events	37
1.12.4	Conditional Probabilities	39
1.12.5	Independence of Events	41
1.13	The Memoryless Property of the Geometric and the Exponential Distributions	42
1.14	Entropy and Related Concepts	43
1.14.1	Entropy	43
1.14.2	Relative Entropy	44
1.14.3	Scores and Support	45
1.15	Transformations	46
1.16	Empirical Methods	47
2	Probability Theory (ii): Many Random Variables	55
2.1	Many Discrete Random Variables	55
2.1.1	The Independent Case	56
2.1.2	The Dependent Case and the Multinomial Distribution	57
2.1.3	Properties of Random Variables from pgfs	57
2.2	Many Continuous Random Variables	59
2.2.1	The Independent Case	59
2.2.2	The Dependent Case	60
2.3	Moment-Generating Functions	60
2.4	Indicator Random Variables	62
2.4.1	Definitions	62
2.4.2	Example: Sequencing EST Libraries	63
2.5	Marginal Distributions	66
2.6	Conditional Distributions	68
2.7	Expected Values	71
2.8	Covariance and Correlation	73
2.9	Asymptotic Distributions	75
2.10	Derived Random Variables (i): Sums, Averages, and Minima	75
2.10.1	Sums and Averages	76
2.10.2	The Minimum of n Random Variables	79
2.11	Derived Random Variables (ii): The Maximum of n Random Variables	81

2.11.1	Distributional Properties: Continuous Random Variables	81
2.11.2	Distributional Properties: Discrete Random Variables	84
2.11.3	An Asymptotic Formula for the Distribution of X_{\max}	87
2.11.4	Bounds for the Distributions of the Maxima of Geometric and Geometric-Like Random Variables	89
2.12	Introduction to P -Values for Y_{\max}	91
2.12.1	Introduction	91
2.12.2	P -Values for Geometric-Like Random Variables	91
2.12.3	P -Values for the Generalized Geometric Distribution	92
2.13	Large Observations and Rare Events	94
2.14	Order Statistics	94
2.14.1	Definition	94
2.14.2	Example: The Uniform Distribution	96
2.14.3	The Sample Median	97
2.15	Transformations	97
2.15.1	Theory	97
2.15.2	Example: The Uniform Distribution (Continued)	98
3	Statistics (i): An Introduction to Statistical Inference	105
3.1	Introduction	105
3.2	Classical and Bayesian Methods	105
3.3	Classical Estimation Methods	107
3.3.1	Unbiased Estimation	107
3.3.2	Biased Estimators	109
3.4	Classical Hypothesis Testing	110
3.4.1	General Principles	110
3.4.2	P -Values	113
3.4.3	Hypothesis Testing Examples	113
3.4.4	Likelihood Ratios, Information, and Support	117
3.4.5	Hypothesis Testing Using X_{\max} as Test Statistic	118
3.5	Nonparametric Alternatives to the Two-Sample t -Test	119
3.6	The Bayesian Approach to Hypothesis Testing	122
3.7	The Bayesian Approach to Estimation	124
3.8	Multiple Testing	125
4	Stochastic Processes (i): Poisson Processes and Markov Chains	129
4.1	The Homogeneous Poisson Process and the Poisson Distribution	129
4.2	The Poisson and the Binomial Distributions	132
4.3	The Poisson and the Gamma Distributions	133
4.4	Introduction to Finite Markov Chains	133
4.5	Transition Probabilities and the Transition Probability Matrix	135
4.6	Markov Chains with Absorbing States	137

4.7	Markov Chains with No Absorbing States	137
4.7.1	Stationary Distributions	138
4.7.2	Example	139
4.8	The Graphical Representation of a Markov Chain	140
4.9	Modeling	141
5	The Analysis of One DNA Sequence	147
5.1	Shotgun Sequencing	147
5.2	Modeling DNA	155
5.3	Modeling Signals in DNA	156
5.3.1	Introduction	156
5.3.2	Weight Matrices: Independence	157
5.3.3	Markov Dependencies	158
5.3.4	Maximal Dependence Decomposition	158
5.4	Long Repeats	160
5.5	r -Scans	162
5.6	The Analysis of Patterns	165
5.7	Overlaps Counted	166
5.7.1	General Comments	166
5.7.2	Number of Occurrences	166
5.7.3	Distance Between Occurrences	169
5.8	Overlaps Not Counted	173
5.8.1	General Comments	173
5.8.2	Distance Between Recurrences	173
5.8.3	Number of Recurrences	174
5.9	Generalizations	175
5.9.1	Arbitrary Nucleotide Probabilities	175
5.9.2	Arbitrary Word Length	176
6	The Analysis of Multiple DNA or Protein Sequences	181
6.1	Two Sequences: Frequency Comparisons	181
6.2	Alignments	183
6.3	Simple Tests for Significant Similarity in an Alignment . . .	184
6.4	Alignment Algorithms for Two Sequences	189
6.4.1	Introduction	189
6.4.2	Gapped Global Comparisons and Dynamic Programming Algorithms	191
6.4.3	Fitting One Sequence into Another Using a Linear Gap Model	194
6.4.4	Local Alignments with a Linear Gap Model	195
6.4.5	Other Gap Models	196
6.4.6	Limitations of the Dynamic Programming Alignment Algorithms	198
6.5	Protein Sequences and Substitution Matrices	199
6.5.1	Introduction	199

6.5.2	BLOSUM Substitution Matrices	200
6.5.3	PAM Substitution Matrices	205
6.5.4	A Simple Symmetric PAM Matrix	210
6.6	Multiple Sequences	211
7	Stochastic Processes (ii): Random Walks	219
7.1	Introduction	219
7.2	The Simple Random Walk	221
7.2.1	Introduction	221
7.3	The Difference Equation Approach	222
7.3.1	Absorption Probabilities	222
7.3.2	Mean Number of Steps Taken Until the Walk Stops	223
7.4	The Moment-Generating Function Approach	224
7.4.1	Absorption Probabilities	224
7.4.2	Mean Number of Steps Until the Walk Stops	226
7.4.3	An Asymptotic Case	227
7.5	General Walks	228
7.6	General Walks: Asymptotic Theory	230
7.6.1	Introduction	230
7.6.2	The Renewal Theorem	231
7.6.3	Unrestricted Walks	231
7.6.4	Restricted Walks	234
8	Statistics (ii): Classical Estimation and Hypothesis Testing	237
8.1	Introduction	237
8.2	Classical Estimation Theory	237
8.2.1	Criteria for "Good" Estimators	237
8.2.2	Maximum Likelihood Estimation	239
8.3	Classical Hypothesis Testing: Simple Fixed Sample Size Tests	246
8.3.1	The Likelihood Ratio	246
8.4	Classical Hypothesis Testing: Composite Fixed Sample Size Tests	249
8.4.1	Introduction	249
8.4.2	The $-2 \log \lambda$ Approximation	254
8.5	Sequential Analysis	258
8.5.1	The Sequential Probability Ratio Test	258
8.5.2	The Power Curve for a Sequential Test	261
8.5.3	The Mean Sample Size	262
8.5.4	The Effect of Boundary Overlap	264
9	BLAST	269
9.1	Introduction	269
9.2	The Comparison of Two Aligned Sequences	270
9.2.1	Introduction	270

9.2.2	Ladder Points and Excursions	270
9.2.3	Parameter Calculations	271
9.2.4	The Choice of a Score	273
9.2.5	Bounds for the P -Value	275
9.2.6	The Normalized and the Bit Scores	276
9.2.7	The Number of High-Scoring Excursions	278
9.2.8	The Karlin–Altschul Sum Statistic	279
9.3	The Comparison of Two Unaligned Sequences	281
9.3.1	Introduction	281
9.3.2	Theoretical and Empirical Background	281
9.3.3	Edge Effects	283
9.3.4	Multiple Testing	285
9.4	The Comparison of a Query Sequence Against a Database	286
9.5	Printouts	288
9.5.1	Example	288
9.5.2	A More Complicated Example	291
9.6	Minimum Significance Lengths	294
9.6.1	A Correct Choice of n	294
9.6.2	An Incorrect Choice of n	296
9.7	BLAST: A Parametric or a Nonparametric Test?	298
9.8	Relation to Sequential Analysis	298
9.9	Further Developments of BLAST and Further Theory	300
10	Stochastic Processes (iii): Markov Chains	303
10.1	Introduction	303
10.2	Markov Chains with No Absorbing States	304
10.2.1	Introduction	304
10.2.2	Convergence to the Stationary Distribution	304
10.2.3	Stationary Distributions: A Numerical Example	305
10.2.4	Reversibility and Detailed Balance	306
10.3	Higher-Order Markov Dependence	308
10.3.1	Testing for Higher-Order Markov Dependence	308
10.3.2	Testing for a Uniform Stationary Distribution	308
10.4	Patterns in Sequences with First-Order Markov Dependence	309
10.5	Markov Chain Monte Carlo	310
10.5.1	The Hastings–Metropolis Algorithm	311
10.5.2	Gibbs Sampling	311
10.5.3	Simulated Annealing	314
10.6	Markov Chains with Absorbing States	317
10.6.1	Theory	317
10.6.2	Examples	318
10.7	Continuous-Time Markov Chains	321
10.7.1	Definitions	321
10.7.2	Time-Dependent Solutions	322
10.7.3	The Stationary Distribution	323

10.7.4 Detailed Balance	323
10.7.5 Exponential Holding Times	323
10.7.6 The Embedded Chain	324
11 Hidden Markov Models	327
11.1 What is a Hidden Markov Model?	327
11.2 Three Algorithms	329
11.2.1 The Forward and Backward Algorithms	329
11.2.2 The Viterbi Algorithm	331
11.2.3 The Estimation Algorithms	332
11.3 Applications	335
11.3.1 Modeling Protein Families	335
11.3.2 Multiple Sequence Alignments	337
11.3.3 Pfam	339
11.3.4 Gene Finding	340
12 Computationally Intensive Methods	349
12.1 Introduction	349
12.2 Estimation	350
12.2.1 Classical Estimation Methods	350
12.2.2 Bootstrap Estimation and Confidence Intervals	351
12.3 Hypothesis Testing: The Bootstrap Alternative to the Two-Sample t -Test	354
12.4 Multiple Testing Revisited	356
12.4.1 Introduction	356
12.4.2 Step-Down Methods: t -Tests	356
12.4.3 Step-Down Methods: A Permutation Test	359
12.4.4 Step-Down Methods Applied to Expression Arrays	361
13 Evolutionary Models	365
13.1 Models of Nucleotide Substitution	365
13.2 Discrete-Time Models	366
13.2.1 The Jukes–Cantor Model	366
13.2.2 The Kimura Models	368
13.2.3 Further Generalizations of the Kimura Models	369
13.2.4 The Felsenstein Model	370
13.2.5 The HKY Model	371
13.2.6 Other Models	372
13.2.7 The Reversibility Criterion	372
13.2.8 The Simple Symmetric PAM Model	373
13.3 Continuous-Time Models	374
13.3.1 The Continuous-Time Jukes–Cantor Model	374
13.3.2 The Continuous-Time Kimura Model	378
13.3.3 The Continuous-Time Felsenstein Model	381
13.3.4 The Continuous-Time HKY Model	382