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

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.

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

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 Mazzarelli, and Eugene Buehler. We also thank Warren Gish for his help on BLAST, and for letting us reproduce his BLAST printout examples. We thank Chris Burge, Sandrine Dudoit, Terry Speed, Matt Werner, Alessandra Gallinari, Sam Sokolovski, Helen Murphy, Ethan Fingerman, 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