BAYESIAN INFERENCE FOR GENE EXPRESSION AND PROTEOMICS

The interdisciplinary nature of bioinformatics presents a research challenge in integrating concepts, methods, software, and multiplatform data. Although there have been rapid developments in new technology and an inundation of statistical methodology and software for the analysis of microarray gene expression arrays, there exist few rigorous statistical methods for addressing other types of high-throughput data, such as proteomic profiles that arise from mass spectrometry experiments. This book discusses the development and application of Bayesian methods in the analysis of high-throughput bioinformatics data that arise from medical, in particular cancer, research, as well as molecular and structural biology. The Bayesian approach has the advantage that evidence can be easily and flexibly incorporated into statistical models.

A basic overview of the biological and technical principles behind multiplatform high-throughput experimentation is followed by expert reviews of Bayesian methodology, tools, and software for single group inference, group comparisons, classification and clustering, motif discovery and regulatory networks, and Bayesian networks and gene interactions.

Kim-Anh Do is a professor in the Department of Biostatistics and Applied Mathematics and the University of Texas M.D. Anderson Cancer Center. Her research interests are in computer-intensive statistical methods with recent focus in the development of methodology and software to analyze data produced from high-throughput technologies.

Peter Müller is also a professor in the Department of Biostatistics and Applied Mathematics and the University of Texas M.D. Anderson Cancer Center. His research interests and contributions are in the areas of Markov chain Monte Carlo posterior simulation, nonparametric Bayesian inference, hierarchical models, mixture models, and Bayesian decision problems.

Marina Vannucci is a professor of Statistics at Texas A&M University. Her research focuses on the theory and practice of Bayesian variable selection techniques and on the development of wavelet-based statistical models and their applications. Her work is often motivated by real problems that need to be addressed with suitable statistical methods.

BAYESIAN INFERENCE FOR GENE EXPRESSION AND PROTEOMICS

Edited by

KIM-ANH DO
University of Texas M.D. Anderson Cancer Center

PETER MÜLLER
University of Texas M.D. Anderson Cancer Center

MARINA VANNUCCI
Texas A&M University

CAMBRIDGE UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314-321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi - 110025, India

103 Penang Road, #05-06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107636989

First published 2006

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

Library of Congress Cataloging in Publication data
Bayesian inference for gene expression and proteomics /
edited by Kim-Anh Do, Peter Müller, Marina Vannucci.
p. cm.
Includes bibliographical references.
ISBN-13: 978-0-521-86092-5 (hardback)
ISBN-10: 0-521-86092-X (hardback)
1. Gene expression – Statistical methods. 2. Proteomics –
Statistical methods. I. Do, Kim-Anh, 1960– II. Müller, Peter, 1963–
III. Vannucci, Marina, 1966– IV. Title.
QH450.B39 2006
572.8´6501519542 – dc22 2006005635

ISBN 978-0-521-86092-5 Hardback
ISBN 978-1-107-63698-9 Paperback

Contents

Contributors

Keith A. Baggerly, *Department of Biostatistics & Applied Mathematics, University of Texas M.D. Anderson Cancer Center, 1515 Holcombe Blvd., Houston, TX 77030-4075*

Veerabhadran Baladandayuthapani, *Department of Biostatistics & Applied Mathematics, University of Texas M.D. Anderson Cancer Center, 1515 Holcombe Blvd., Houston, TX 77030-4075*

Ghislain Bidaut, *Department of Genetics, The University of Pennsylvania School of Medicine, 1423 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104-6021*

Andrea Bild, *Institute for Genome Sciences and Policy, Duke University, Durham, NC 27710*

Linda L. Breeden, *Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N., Mailstop A2-168, P.O. Box 19024, Seattle, WA 98109-1024*

Bradley M. Broom, *Department of Biostatistics & Applied Mathematics, University of Texas M.D. Anderson Cancer Center, 1515 Holcombe Blvd., Houston, TX 77030-4075*

Philip J. Brown, *Institute of Mathematics, Statistics and Actuarial Science, Room E218, Cornwallis Building, University of Kent, Canterbury, Kent CT2 7NF, UK*

Raymond J. Carroll, *Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143*

Carlos Carvalho, *Institute of Statistics and Decision Sciences, Box 90251, Duke University, Durham, NC 27708-0251*

Meng Chen, *Department of Statistics, The University of Wisconsin-Madison, 1220 Medical Sciences Center, 1300 University Ave., Madison, WI 53703*

Merlise A. Clyde, *Institute of Statistics and Decision Sciences, Box 90251, Duke University, Durham, NC 27708-0251*

Kevin R. Coombes, *Department of Biostatistics & Applied Mathematics, University of Texas M.D. Anderson Cancer Center, 1515 Holcombe Blvd., Houston, TX 77030-4075*

David B. Dahl, *Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143*

Kim-Anh Do, *Department of Biostatistics & Applied Mathematics, University of Texas M.D. Anderson Cancer Center, 1515 Holcombe Blvd., Houston, TX 77030-4075*

Arnoldo Frigessi, *Department of Biostatistics, University of Oslo, P.O. Box 1122, Blindern, 0317 Oslo, Norway*

Elizabeth Garrett-Mayer, *Johns Hopkins Kimmel Cancer Center, Johns Hopkins University, Suite 1103, 550 N. Broadway, Baltimore, MD 21205*

Ingrid K. Glad, *Department of Mathematics, University of Oslo, P.O. Box 1053, Blindern, 0316 Oslo, Norway*

Michele Guindani, *Department of Biostatistics & Applied Mathematics, University of Texas M.D. Anderson Cancer Center, 1515 Holcombe Blvd., Houston, TX 77030-4075*

Mayetri Gupta, *Department of Biostatistics, University of North Carolina at Chapel Hill, McGavran, Greenberg Hall B CB#7420, Chapel Hill, NC 27599-7420*

Alexander J. Hartemink, *Department of Computer Science, Duke University, Box 90129, Durham, NC 27708-0129*

Anne-Mette K. Hein, *Department of Epidemiology and Public Health, Imperial School of Medicine, St. Mary's Campus, Norfolk Place, London W2 1PG, UK*

Marit Holden, *Norwegian Computing Center, P.O. Box 114, Blindern, 0314 Oslo, Norway*

Chris C. Holmes, *Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK*

Leanna L. House, *Institute of Statistics and Decision Sciences, Box 90251, Duke University, Durham, NC 27708-0251*

Christina Kendziorski, *Department of Biostatistics and Medical Informatics, 6785 Medical Sciences Center, 1300 University Ave., Madison, WI 53703*

Sinae Kim, *Department of Biostatistics, School of Public Health, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029*

Andrew V. Kossenkov, *Division of Population Science, Fox Chase Cancer Center, 333 Cottman Ave., Philadelphia, PA 19111-2497*

Deukwoo Kwon, *Radiation Epidemiology Branch, National Cancer Institute, 6120 Executive Blvd, MSC 7238, Executive Plaza South, Room 7045, Bethesda, MD 20892-7238*

Alex Lewin, *Department of Epidemiology and Public Health, Imperial School of Medicine, St. Mary's Campus, Norfolk Place, London W2 1PG, UK*

Jun S. Liu, *Statistics Department, Harvard University, Science Center, One Oxford Street, Cambridge, MA 02138-2901*

Joseph Lucas, *Institute of Statistics and Decision Sciences, Box 90251, Duke University, Durham, NC 27708-0251*

Heidi Lyng, *Department of Biophysics, The Norwegian Radium Hospital, Montebello, 0310 Oslo, Norway*

Bani K. Mallick, *Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143*

Jeffrey S. Morris, *Department of Biostatistics & Applied Mathematics, University of Texas M.D. Anderson Cancer Center, 1515 Holcombe Blvd., Houston, TX 77030-4075*

Peter Müller, *Department of Biostatistics & Applied Mathematics, University of Texas M.D. Anderson Cancer Center, 1515 Holcombe Blvd., Houston, TX 77030-4075*

Michael A. Newton, *Department of Statistics, University of Wisconsin-Madison, Medical Sciences Center, RM 1245A, 1300 University Ave., Madison, WI 53706-1532*

Joseph R. Nevins, *Institute for Genome Sciences and Policy, Duke University, Durham, NC 27710*

Michael F. Ochs, *Division of Population Science, Fox Chase Cancer Center, 333 Cottman Ave., Philadelphia, PA 19111-2497*

Sylvia Richardson, *Department of Epidemiology and Public Health, Imperial School of Medicine, St. Mary's Campus, Norfolk Place, London W2 1PG, UK*

Christian Robert, *Ceremade – Université Paris-Dauphine, Bureau C638, Place du Maréchal de Lattre de Tassigny, 75775 Paris Cedex 16, France*

Judith Rousseau, *Ceremade – Université Paris-Dauphine Bureau B638, Place du Maréchal de Lattre de Tassigny, 75775 Paris Cedex 16, France*

Robert Scharpf, *Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, E3034 615 N. Wolfe Street, Baltimore, MD 21205-2179*

Naijun Sha, *Department of Mathematical Science, University of Texas at El Paso, Bell Hall 203, El Paso, TX 79968-0514*

Devika Subramanian, *Department of Computer Science, Rice University, 6100 Main Street, Houston, TX 77005*

Ning Sun, *Division of Biostatistics, Department of Epidemiology and Public Health, Yale University School of Medicine, 60 College Street, P.O. Box 208034, New Haven, CT 06520-8034*

Michael Swartz, *Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143* and *Department of Biostatistics & Applied Mathematics, University of Texas M.D. Anderson Cancer Center, 1515 Holcombe Blvd., Houston, TX 77030-4075*

Mahlet G. Tadesse, *Department of Epidemiology & Biostatistics, University of Pennsylvania School of Medicine, 918 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104-6021*

Mark A. van de Wiel, *Department of Mathematics, Vrije Universiteit, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands*

Marina Vannucci, *Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143*

Jon C. Wakefield, *University of Washington, Department of Biostatistics, Box 357232, Seattle, WA 98195-7232* and *Department of Statistics, Box 354322, Seattle, WA 98195-4322*

Ping Wang, *Department of Statistics, The University of Wisconsin-Madison, B248 Medical Sciences Center, 1300 University Ave., Madison, WI 53703*

Quanli Wang, *Institute for Genome Sciences & Policy, Duke University, Durham, NC 27710* and *Institute of Statistics & Decision Sciences, Duke University, Durham, NC 27708-0251*

Mike West, *Institute of Statistics and Decision Sciences, Box 90251, Duke University, Durham, NC 27708-0251*

Robert L. Wolpert, *Institute of Statistics and Decision Sciences, Box 90251, Duke University, Durham, NC 27708-0251*

Hongyu Zhao, *Division of Biostatistics, Department of Epidemiology and Public Health, Yale University School of Medicine, 60 College Street, P.O. Box 208034, New Haven, CT 06520-8034*

Chuan Zhou, *Department of Biostatistics, Vanderbilt University, S-2323 Medical Center North, Nashville, TN 37232-2158*

Preface

Recent rapid technical advances in genome sequencing (genomics) and protein identification (proteomics) have given rise to research problems that require combined expertise from statistics, biology, computer science, and other fields. The interdisciplinary nature of bioinformatics presents many research challenges related to integrating concepts, methods, software, and multiplatform data. In addition to new tools for investigating biological systems via high-throughput genomic and proteomic measurements, statisticians face many novel methodological research questions generated by such data. The work in this book is dedicated to the development and application of Bayesian statistical methods in the analysis of high-throughput bioinformatics data that arise from problems in medical research, in particular cancer research, and molecular and structural biology. This book does not aim to be comprehensive in all areas of bioinformatics. Rather, it presents a broad overview of statistical inference problems related to three main high-throughput platforms: microarray gene expression, serial analysis gene expression (SAGE), and mass spectrometry proteomic profiles. The book's main focus is on the design, statistical inference, and data analysis, from a Bayesian perspective, of data sets arising from such high-throughput experiments.

Chapter 1 provides a detailed introduction to the three main data platforms and sets the scene for subsequent methodology chapters. This chapter is mainly aimed at nonbiologists and covers elementary biological concepts, details the unique measurement technology with associated idiosyncrasies for the different platforms, and generates an overall outline of issues that statistical methodology can address.

Subsequent chapters focus on specific methodology developments and are grouped approximately by the main bioinformatics platform, with several chapters discussing the integration of at least two platforms. The central statistical topics addressed include experimental design, single group inference, group comparisons, classification and clustering, motif discovery and regulatory networks, and Bayesian networks and gene interactions. The general theme of each

chapter is to review existing methods, followed by a specific novel method developed by the author(s). Results are often demonstrated on simulated data and/or a real application data set. Additionally, relevant software may be discussed.

Chapters 2 through 11 are concerned with Bayesian inference for gene expression, focusing on microarray data. Chapter 2 discusses inference about differential expression based on hierarchical mixture models, including a discussion of more than two patterns of differential expression. Inference is based on conjugate parametric models, with empirical Bayes estimation of hyperparameters. Chapter 3 explores the use of Bayesian hierarchical models for an integrated approach to the analysis of microarray data that includes flexible model-based normalization. The chapter defines a Bayesian gene expression (BGE) index as a gene-specific mean parameter in a hierarchical model. In Chapter 4, a model that mimics the detailed experimental process, from gene preparation to image analysis, is developed. The detailed process-based model allows to estimate absolute and relative mRNA concentrations. The use of Bayesian variable selection methods for biomarker selection in classification and clustering problems is reviewed in Chapter 5. The model for classification includes a mixture prior that specifies a positive probability of a given gene not being included in the model. In the clustering setting, the group structure of the data is uncovered by specifying mixture models where the random inclusion of genes can be interpreted as an attribute selection. Chapter 6 applies multivariate adaptive regression splines (MARS) to define a flexible model for the relationship between gene expression and disease status. For a binary classification problem the MARS model is defined on the logistic transformation of the class probability. The resulting classification boundaries are highly nonlinear and account for gene interactions. Chapter 7 reviews the popular probability of expression (POE) model for differential gene expression. The model postulates a mixture of a uniform submodel for under- and overexpression, and a central normal for typical expression. Posterior probabilities of mixture indicators in this model define the POE scale. Chapter 8 explores the use of sparsity priors in multivariate regression and latent factor regression, applied to inference for gene expression. The sparsity prior is a variation of a mixture prior on regression coefficients, with a positive point mass at zero- and a second-level mixture allowing for gene- and covariate-specific relative weights in the mixture. Chapter 9 develops a model for cell cycle gene expression, using a first-order Fourier model. The set of all genes is partitioned into subsets of different frequency and time-dependent amplitude, including a zero class of not-cycle-dependent genes. The prior on the random partitioning is defined as a Dirichlet distribution for cluster membership indicators. Chapter 10 defines a semiparametric Bayesian model for gene expression. The model exploits the clustering that is implicitly defined by the Dirichlet process prior to define subsets of genes based on gene-specific mean and sampling precision. Chapter 11 reviews the

expression quantitative trait (eQTL) mapping problem and commonly used approaches. A new method is proposed to facilitate eQTL interval mapping that can account for multiplicities across transcripts. The problem is to match microarray gene expression (phenotype) with a set of genetic markers (genetic map). The mixture over markers (MOM) model defines a mixture model for gene expression, with the mixture being defined over submodels corresponding to the transcript mapping to one of the considered markers. An extension to interval mapping is discussed.

Chapters 12 through 15 discuss statistical inference for protein spectrometry. Chapter 12 reviews the use of semiparametric mixture models for inference on differential gene expression, for protein mass/charge spectra, and for SAGE data. The underlying models are Dirichlet process mixtures of normals, mixtures of beta kernels, and Dirichlet process mixtures of Poissons, respectively. Specific focus on SAGE data alone is detailed in Chapter 13, highlighting the two main characteristics of such data: skewness in the distribution of relative abundances, and small sample size relative to the dimension. A new Bayesian procedure based on the mixture Dirichlet prior is reviewed and specific properties depicted in terms of efficiency advantages over existing methods. Chapters 14 and 15 present two different Bayesian approaches of analyzing MALDI-TOF mass spectrometry. Chapter 14 generalizes the linear mixed model to the case of functional data by using a wavelet-based functional mixed model. In contrast, Chapter 15 presents model-based inference by focusing on nonparametric Bayesian models; a sum of kernel functions is chosen as basis functions for modeling spectral peaks.

Chapters 16 through 21 review motif discovery and regulatory networks. During the process of gene transcription, proteins (transcription factors) interact with control points of DNA sequences known as cis-acting regulatory sequences, called motifs. Chapter 16 focuses on the problem of locating these short sequence patterns in the DNA. The chapter provides an extensive explanation of the biological background and then describes a Bayesian framework for gene regulatory binding site discovery. Alternative models are also reviewed. Recent efforts in the development of methods that aid the detection of transcription factor binding sites have attempted to integrate sequence data with gene expression. In Chapter 17 the authors describe a regression model and a Bayesian variable selection formulation that helps refine the search for candidate motifs by selecting those that correlate with the gene expression. The authors also propose a model extension that includes gene regulators. Data integration is also the focus of Chapter 18, where gene expression and protein–DNA binding information, obtained from ChIP-Chip data, are integrated via a linear classification model to aid the reconstruction of gene regulatory networks. Chapter 19 deals with models that exploit the links between transcription factors and signaling pathways via gene ontologies and annotation databases. Bayesian decomposition models allow the extraction of overlapping

transcriptional signatures. Popular approaches to infer gene regulatory networks are those that utilize probabilistic network models. Bayesian networks, in particular, have received much attention. Chapter 20 provides an extensive survey of computational techniques for learning Bayesian networks from gene expression data. State-of-the art and open questions are also addressed. Chapter 21 discusses the developments of Bayesian networks and dynamic networks. Additional data are used to derive informative prior structures, in particular combining gene expression data with protein–DNA binding information.

The final chapter, Chapter 22, addresses the choice of the sample size for microarray experiments. The authors take a decision theory point of view that attempts to minimize a conditional expected loss. The method is exemplified using a mixture Gamma/Gamma model.

We thank our friends and collaborators for contributing their ideas and insights toward this collection. We are excited by the continuing opportunities for statistical challenges in the area of high-throughput bioinformatics data. We hope our readers will join us in being engaged with changing technologies and statistical development.

Kim-Anh Do
Peter Müller
Marina Vannucci