# METHODS IN MOLECULAR BIOLOGY™

*Series Editor*
**John M. Walker**
**School of Life Sciences**
**University of Hertfordshire**
**Hatfield, Hertfordshire, AL10 9AB, UK**

For further volumes:
http://www.springer.com/series/7651

# Statistical Methods for Microarray Data Analysis

## Methods and Protocols

Edited by

**Andrei Y. Yakovlev**

*School of Medicine and Dentistry, Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY, USA*

**Lev Klebanov**

*Department of Probability and Statistics, Charles University, Prague, Czech Republic*

**Daniel Gaile**

*State University of New York at Buffalo, Buffalo, NY, USA*

Humana Press

*Editors*
Andrei Y. Yakovlev
School of Medicine and Dentistry
Department of Biostatistics and Computational Biology
University of Rochester
Rochester, NY, USA

Daniel Gaile
State University of New York at Buffalo
Buffalo, NY, USA

Lev Klebanov
Department of Probability and Statistics
Charles University
Prague, Czech Republic

ISSN 1064-3745 ISSN 1940-6029 (electronic)
ISBN 978-1-60327-336-7 ISBN 978-1-60327-337-4 (eBook)
DOI 10.1007/978-1-60327-337-4
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2012956545



Printed on acid-free paper

Humana Press is a brand of Springer
Springer is part of Springer Science+Business Media (www.springer.com)

# Preface

Microarrays for simultaneous measurement of redundancy of RNA species are widely used for fundamental biology research. They are also tested for their use in personalized medicine in disease diagnosis and prognosis. From the point of mathematical statistics, the invention of microarray technology in the mid-1990s allowed the simultaneous monitoring of the expression levels of thousands of genes. Microarrays for simultaneous measurement of redundancy of RNA species are used in fundamental biology as well as in medical research. Microarray may be considered as an observation of very high dimensionality equal to the number of expression levels. There arise some needs to develop new statistical methods to handle the data of such large dimensionality, especially connected to the fact of small number of observations (which is the number of arrays). Because of the small number of observations the standard asymptotic methods of multivariate statistical analysis appear to be inapplicable.

The aim of the book is to familiarize the readers with statistical methods used nowadays in microarray analysis. It is addressed to everybody who is involved or is planning to be involved in statistical data analysis of microarrays, mostly to statisticians but also to biological researchers.

It was impossible to come over all statistical methods published to date in a book. The selection was made on a basis of mathematical correctness of corresponding methods. Some approaches based on intuitive impressions and having no mathematical support were excluded from our consideration. However, the selection criteria were sometimes not scientific. In many cases they reflect personal taste of the editors. Nevertheless, the editors took pains that other valuable methods should be described or mentioned. The editors invite the interested readers to continue their study of the material beyond this book. We are very grateful to all authors for their writing.

Each chapter can be read as a separate entry. The style of the individual writing is essential to show the knowledge and experience of each team that contributed. However, the reader will be guided from microarray technology to statistical problems of corresponding data analysis. Chapters 1 and 2 provide such prolegomena.

In Chapter 3, an introduction to current multiple testing methodology are presented, with the objective of clarifying the methodological issues involved, and hopefully providing the reader with some basis with which to compare and select methods.

Chapter 4 discusses a method of selecting differentially expressed genes based on a newly discovered structure termed as the δ-sequence. Together with the nonparametric empirical Bayes methodology, it leads to dramatic gains in terms of the mean numbers of true and false discoveries, and in the stability of the results of testing. The results of this chapter can be viewed also as a new method of normalization of microarray data.

Chapter 5 is in some sense connected to Chapter 4. It studies different normalization procedures of gene expression levels. Normalization procedures are used for removing systematical variation which affects the measure of expression levels.

Chapter 6 is dedicated to constructing of multivariate prognostic gene signatures with censored survival data. Modern high-throughput technologies allow us to simultaneously measure the expressions of a huge number of candidate predictors, some of which are likely to be associated with survival. One difficult task is to search among an enormous number of potential predictors and to correctly identify most of the important ones, without mistakenly identifying too many spurious associations. Mere variable selection is insufficient, however, for the information from the multiple predictors must be intelligently combined and calibrated to form the final composite predictor. Many commonly used procedures overfit the training data, miss many important predictors, or both. Author proposes a method that offers a middle ground where some partial multivariate adjustments can be made in an adaptive fashion, regardless of the number of candidate predictors. He demonstrates the performance of our proposed procedure in a simulation study within the Cox proportional hazards regression framework, and applies this new method to a publicly available data set to construct a novel prognostic gene signature for breast cancer survival.

Chapter 7 considers clustering problems for gene-expression data.

There are two distinct but related clustering problems with microarray data. One problem concerns the clustering of the tissue samples (gene signatures) on the basis of the genes; the other concerns the clustering of the genes on the basis of the tissues (gene profiles). The clusters of tissues so obtained in the first problem can play a useful role in the discovery and understanding of new subclasses of diseases. The clusters of genes obtained in the second problem can be used to search for genetic pathways or groups of genes that might be regulated together. Authors focus here on mixtures of normals to provide a model-based clustering of tissue samples (gene signatures) and of gene profiles.

Network-based analysis of multivariate gene expression data is given in Chapter 8. Such data are collected to study genomic responses under special conditions for which the expression levels of given genes are expected to be dependent. One important question from such multivariate gene expression experiments is to identify genes that show different expression patterns over treatment dosages or over time; these genes can also point to the pathways that are perturbed during a given biological process. Several empirical Bayes approaches have been developed for identifying the differentially expressed genes in order to account for the parallel structure of the data and to borrow information across all the genes.

In Chapter 9, author discusses the statistical problem, termed oncogene outlier detection, and discusses a variety of proposals to this problem. A statistical model in the multiclass situation is described; links with multiple testing concepts are established. Some new nonparametric procedures are described and compared to existing methods using simulation studies.

Data quality is intrinsically influenced by design, technical, and analytical parameters. Quality parameters have not yet been well defined for gene expression analysis by microarrays, though ad interim, following recommended good experimental practice guidelines should ensure generation of reliable and reproducible data. In Chapter 10 author summarizes essential practical recommendations for experimental design, technical considerations, feature annotation issues, and standardization efforts.

Inferring gene regulatory networks from microarray data has become a popular activity in recent years, resulting in an ever increasing volume of publications. There are many pitfalls in network analysis that remain either unnoticed or scantily understood. A critical discussion of such pitfalls is long overdue. Chapter 11 discuss one feature of microarray data the investigators need to be aware of when embarking on a study of putative associations between elements of networks and pathways.

Finally, Chapter 12 considers the problem of normality of logs of gene expression levels. In the literature there is no unique point on the fact of normality (or nonnormality) of the distribution of gene expression levels. This chapter discusses different approaches to testing of normality in this situation.

The editors and the contributors assume from the reader a basic knowledge of biological concepts of gene expression and statistical methods of gene expression analysis.

Editors are grateful to all contributors.

*Prague, Czech Republic* *Lev Klebanov*

# Contents

# Contributors


Anthony Almudevar • *Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY, USA*

Garrett Brodeur • *Department of Pediatrics, Children's Hospital of Philadelphia, Philadelphia, PA, USA*

Peter Bubelíny • *Department of Probability and Statistics, Charles University, Prague, Czech Republic*

Linlin Chen • *School of Mathematical Sciences, Rochester Institute of Technology, Rochester, NY, USA*

Manhong Dai • *Psychiatry Department and Molecular and Behavioral Neuroscience Institute, University of Michigan, University of Michigan, Ann Arbor, MI, USA*

L.K. Flack • *Department of Rheumatology, University of NSW, Ryde, NSW, Australia*

Debashis Ghosh • *Departments of Statistics and Public Health Sciences, Penn State University, DuBios, PA, USA*

Diana M. Kelmansky • *Instituto de Cálculo, Ciudad Universitaria, Buenos Aires, Argentina*

Lev Klebanov • *Department of Probability and Statistics Charles University Prague, Czech Republic*

Hongzhe Li • *Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA, USA*

*Department of Pediatrics, Children's Hospital of Philadelphia, Philadelphia, PA, USA*

G.J. McLachlan • *Department of Mathematics, University of Queensland, Brisbane, Australia*

Fan Meng • *Psychiatry Department and Molecular and Behavioral Neuroscience Institute, University of Michigan, University of Michigan, Ann Arbour, MI, USA*

Jane Minturn • *Department of Pediatrics, Children's Hospital of Philadelphia, Philadelphia, PA, USA*

S.K. Ng • *School of Medicine, Griffith University, Meadowbrook, Australia*

Derick R. Peterson • *Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY, USA*

Xing Qiu • *Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY, USA*

Eric Rappaport • *Children's Hospital of Philadelphia, Philadelphia, PA, USA*

Andreas Scherer • *Genomics, Biomarker Development, Spheromics, Kontiolahti, Joensuu, Finland*

Bobosharif Shokirov • *Department of Probability and Statistics, MFF, Charles University, Prague, Czech Republic*

K. Wang • *University of Queensland, Melbourne, Australia*

Stephen Welle • *Functional Genomics Center, University of Rochester, Rochester, NY, USA*

Wei Zhi • *Department of Biostatistics and Epidemiology, New Jersey Institute of Technology, Newark, NJ, USA*