Methods in Molecular Biology™

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

For other titles published in this series, go to
www.springer.com/series/7651

Statistical Methods in Molecular Biology

Edited by

Heejung Bang, Xi Kathy Zhou, and Madhu Mazumdar

Weill Medical College of Cornell University, New York, NY, USA

Heather L. Van Epps

Rockefeller University Press, New York, NY, USA

Humana Press

Editors
Heejung Bang
Division of Biostatistics and Epidemiology
Department of Public Health
Weill Cornell Medical College
402 East 67th St.
New York, NY 10065
USA
heb2013@med.cornell.edu

Heather L. Van Epps
Rockefeller University Press
Journal of Experimental Medicine
1114 First Ave.
New York, NY 10021
3rd Floor
USA
hvanepps@rockefeller.edu

Xi Kathy Zhou
Division of Biostatistics and Epidemiology
Department of Public Health
Weill Cornell Medical College
402 East 67th St.
New York, NY 10065
USA
kaz2004@med.cornell.edu

Madhu Mazumdar
Division of Biostatistics and Epidemiology
Department of Public Health
Weill Cornell Medical College
402 East 67th St.
New York, NY 10065
USA
mam2073@med.cornell.edu

ISSN 1064-3745 e-ISSN 1940-6029
ISBN 978-1-60761-578-1 e-ISBN 978-1-60761-580-4
DOI 10.1007/978-1-60761-580-4
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2009942427

Cover illustration: Adapted from Figure 19 of Chapter 2. **Traditional MDS map showing genes clustered by coregulation (background) and significance of the uni-variate *p*-values (size of red circles)**. The overlayed network indicates the "most significant" gene pairs (green) and trios (blue) (right). Font size indicating the smallest of the uni-, bi- and tri-variate *p*-values.

Printed on acid-free paper

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

Knowing is not enough; we must apply. Willing is not enough; we must do.

-Johann Wolfgang Von Goethe

Preface

This book is intended for molecular biologists who perform quantitative analyses on data emanating from their field and for the statisticians who work with molecular biologists and other biomedical researchers. There are many excellent textbooks that provide fundamental components for statistical training curricula. There are also many "by experts for experts" books in statistics and molecular biology which require in-depth knowledge in both subjects to be taken full advantage of. So far, no book in statistics has been published that provides the basic principles of proper statistical analyses and progresses to a more advanced statistics in response to rapidly developing technologies and methodologies in the field of molecular biology.

Responding to this situation, our book aims at bridging the gap between these two extremes. Molecular biologists will benefit from the progressive style of the book where basic statistical methods are introduced and gradually elevated to an intermediate level. Similarly, statisticians will benefit from learning the various biological data generated from the field of molecular biology, the types of questions of interest to molecular biologists, and the statistical approaches to analyzing the data. The statistical concepts and methods relevant to studies in molecular biology are presented in a simple and practical manner. Specifically, the book covers basic and intermediate statistics that are useful for classical and molecular biology settings and advanced statistical techniques that can be used to help solve problems commonly encountered in modern molecular biology studies, such as supervised and unsupervised learning, hidden Markov models, manipulation and analysis of data from high-throughput microarray and proteomic platform, and synthesis of these evidences. A tutorial-type format is used to maximize learning in some chapters. Advice from journal editors on peer-reviewed publication and some useful information on software implementation are also provided.

This book is recommended for use as supplementary material both inside and outside classrooms or as a self-learning guide for students, scientists, and researchers who deal with numeric data in molecular biology and related fields. Those who start as beginners, but desire to be at an intermediate level, will find this book especially useful in their learning pathway.

We want to thank John Walker (series editor), Patrick Marton, David Casey, and Anne Meagher, (editors at Springer and Humana) and Shanthy Jaganathan (Integra-India). The following persons provided useful advice and comments on selection of topics, referral to experts in each topic, and/or chapter reviews that we truly appreciate: Stephen Looney (a former editor of this book), Stan Young, Dmitri Zaykin, Douglas Hawkins, Wei Pan, Alexandre Almeida, John Ho, Rebecca Doerge, Paula Trushin, Kevin Morgan, Jason Osborne, Peter Westfall, Jenny Xiang, Ya-lin Chiu, Yolanda Barron, Huibo Shao, Alvin Mushlin, and Ronald Fanta. Drs. Bang, Zhou, and Mazumdar were partially supported by Clinical Translational Science Center (CTSC) grant (UL1-RR024996).

Heejung Bang

Contents

Contributors

CHUL AHN • *Division of Biostatistics, Department of Clinical Sciences, The Harold C. Simmons Comprehensive Cancer Center, University of Texas Southwestern Medical Center, Dallas, TX, USA*

SAMPRIT BANERJEE • *Division of Biostatistics and Epidemiology, Department of Public Health, Weill Cornell Medical College, New York, NY, USA*

HEEJUNG BANG • *Division of Biostatistics and Epidemiology, Weill Cornell Medical College, New York, NY, USA*

MARTINA BREMER • *Department of Mathematics, San Jose State University, San Jose, CA, USA*

JENNIFER SOUSA BRENNAN • *Department of Biostatistics, Yale University, New Haven, CT, USA*

BRADLEY M. BROOM • *Department of Bioinformatics and Computational Biology, University of Texas M. D. Anderson Cancer Center, Houston, TX, USA*

FABIEN CAMPAGNE • *HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Department of Physiology and Biophysics, Weill Cornell Medical College, New York, NY, USA*

FENG CHENG • *Department of Biophysics, University of Virginia, Charlottesville, VA, USA*

SANG-HOON CHO • *Department of Public Health Sciences, University of Virginia, Charlottesville, VA, USA*

YUEHUA CUI • *Department of Statistics and Probability, Michigan State University, East Lansing, MI, USA*

MARIE DAVIDIAN • *Department of Statistics, North Carolina State University, Raleigh, NC, USA*

XUTAO DENG • *HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Cornell Medical College, New York, NY, USA*

KIM-ANH DO • *Department of Biostatistics, University of Texas M. D. Anderson Cancer Center, Houston, TX, USA*

MELISSA J. FAZZARI • *Division of Biostatistics, Department of Epidemiology and Population Health, Department of Genetics, Albert Einstein College of Medicine, Bronx, NY, USA*

SUJIT K. GHOSH • *Department of Statistics, North Carolina State University, Raleigh, NC, USA*

MITHAT GÖNEN • *Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, NY, USA*

JOHN M. GREALLY • *Department of Genetics, Department of Medicine, Albert Einstein College of Medicine, Bronx, NY, USA*

EDWARD HIMELBLAU • *Department of Biological Science, California Polytechnic State University, San Luis Obispo, CA, USA*

KLAUS JUNG • *Department of Medical Statistics, Georg-August-University Göttingen, Göttingen, Germany*

SIN-HO JUNG • *Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA*
TRECIA A. KIPPOLA • *Department of Statistics, Oklahoma State University, Stillwater, OK, USA*
CHRISTOPH LANGE • *Center for Genomic Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA; Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA*
JAE K. LEE • *Department of Public Health Sciences, University of Virginia, Charlottesville, VA, USA*
YOONKYUNG LEE • *Department of Statistics, The Ohio State University, Columbus, OH, USA*
GENGXIN LI • *Department of Statistics and Probability, Michigan State University, East Lansing, MI, USA*
LEXIN LI • *Department of Statistics, North Carolina State University, Raleigh, NC, USA*
SHAOYU LI • *Department of Statistics and Probability, Michigan State University, East Lansing, MI, USA*
FEI LIU • *Department of Statistics, University of Missouri, Columbia, MO, USA*
ANDREAS MADLUNG • *Department of Biology, University of Puget Sound, Tacoma, WA, USA*
MADHU MAZUMDAR • *Division of Biostatistics and Epidemiology, Department of Public Health, Weill Cornell Medical College, New York, NY, USA*
MARINA MEILA • *Department of Statistics, University of Washington, Seattle, WA, USA*
AMY MURPHY • *Channing Laboratory, Center for Genomic Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA*
REBECCA NUGENT • *Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, USA*
LAJOS PUSZTAI • *Department of Breast Medical Oncology, University of Texas M. D. Anderson Cancer Center, Houston, TX, USA*
WAREE RINSURONGKAWONG • *Department of Biostatistics, University of Texas M. D. Anderson Cancer Center, Houston, TX, USA*
STEPHANIE A. SANTORICO • *Department of Mathematical and Statistical Sciences, University of Colorado, Denver, CO, USA*
TINGTING SONG • *Center for Clinical and Translational Science, The Rockefeller University, New York, NY, USA*
HEATHER L. VAN EPPS • *Journal of Experimental Medicine, Rockefeller University Press, New York, NY, USA*
SCOTT T. WEISS • *Channing Laboratory, Center for Genomic Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA*
KNUT M. WITTKOWSKI • *Center for Clinical and Translational Science, The Rockefeller University, New York, NY, USA*
JING WU • *Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, USA*
RONGLING WU • *Departments of Public Health Sciences and Statistics, Pennsylvania State University, Hershey, PA, USA*
ZHIJIN WU • *Center for Statistical Sciences, Brown University, Providence, RI, USA*
ZHIQIANG WU • *Department of Electrical Engineering, Wright State University, Dayton, OH, USA*
JUN XIE • *Department of Statistics, Purdue University, West Lafayette, IN, USA*

YANG XIE • *Division of Biostatistics, Department of Clinical Sciences, The Harold C. Simmons Comprehensive Cancer Center, University of Texas Southwestern Medical Center, Dallas, TX, USA*