**COMPUTATIONAL AND
STATISTICAL APPROACHES TO
GENOMICS, SECOND EDITION**

COMPUTATIONAL AND STATISTICAL APPROACHES TO GENOMICS, SECOND EDITION

edited by

Wei Zhang

University of Texas, M.D. Anderson Cancer Center

and

Ilya Shmulevich

Institute for Systems Biology, Seattle, WA

Library of Congress Cataloging-in-Publication Data

Computational and statistical approaches to genomics / edited by Wei Zhang and Ilya Shmulevich. – 2nd ed.

p. cm.

Includes bibliographical references and index.

ISBN-13: 978-0-387-26287-1

ISBN-10: 0-387-26287-3 (alk. paper)

ISBN-13: 978-0-387-26288-8 (e-book)

ISBN-10: 0-387-26288-1 (e-book)

1. Genomics—Mathematical models. 2. Genomics—Statistical methods. 3 Genomics—Data processing. 4 DNA microarrays. I. Zhang, Wei, 1963- II. Shmulevich, Ilya, 1969-

QH438.4.M3C65 2005

572.8'6'015118—dc22

2005049761

© 2006 Springer Science+Business Media, 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 Science+Business Media, Inc., 233 Spring Street, New York, NY 10013, 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 in this publication of trade names, trademarks, service marks and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed in the United States of America.

9 8 7 6 5 4 3 2 1

SPIN 11419631

springeronline.com

This book is dedicated
to our families and
colleagues

Contents

Preface	ix
1	
Microarray Image Analysis and Gene Expression Ratio Statistics	1
<i>Yidong Chen, Edward R. Dougherty, Michael L. Bittner, Paul Meltzer, and Jeffery Trent</i>	
2	
Statistical Considerations in the Assessment of cDNA Microarray Data Obtained Using Amplification	21
<i>Jing Wang, Kevin R. Coombes, Keith Baggerly, Limei Hu, Stanley R. Hamilton, and Wei Zhang</i>	
3	
Sources of Variation in Microarray Experiments	37
<i>Kathleen F. Kerr, Edward H. Leiter, Laurent Picard, and Gary A. Churchill</i>	
4	
Studentizing Microarray Data	49
<i>Keith A. Baggerly, Kevin R. Coombes, Kenneth R. Hess, David N. Stivers, Lynne V. Abruzzo, and Wei Zhang</i>	
5	
Exploratory Clustering of Gene Expression Profiles of Mutated Yeast Strains	61
<i>Merja Oja, Janne Nikkilä, Petri Törönen, Garry Wong, Eero Castrén, and Samuel Kaski</i>	
6	
Selecting Informative Genes for Cancer Classification Using Gene Expression Data	75
<i>Tatsuya Akutsu and Satoru Miyano</i>	
7	
Finding Functional Structures in Glioma Gene-Expressions Using Gene Shaving Clustering and MDL Principle	89
<i>Ciprian D. Giurcaneanu, Cristian Mircean, Gregory N. Fuller, and Ioan Tabus</i>	

8	Design Issues and Comparison of Methods for Microarray-Based Classification	119
	<i>Edward R. Dougherty and Sanju N. Attoor</i>	
9	Analyzing Protein Sequences Using Signal Analysis Techniques	137
	<i>Karen M. Bloch and Gonzalo R. Arce</i>	
10	Scale-Dependent Statistics of the Numbers of Transcripts and Protein Sequences Encoded in the Genome	163
	<i>Vladimir A. Kuznetsov</i>	
11	Statistical Methods in Serial Analysis of Gene Expression (SAGE)	209
	<i>Ricardo Z. N. Vêncio and Helena Brentani</i>	
12	Normalized Maximum Likelihood Models for Boolean Regression with Application to Prediction and Classification in Genomics	235
	<i>Ioan Tabus, Jorma Rissanen, and Jaakko Astola</i>	
13	Inference of Genetic Regulatory Networks via Best-Fit Extensions	259
	<i>Harri Lähdesmäki, Ilya Shmulevich, Olli Yli-Harja, and Jaakko Astola</i>	
14	Regularization and Noise Injection for Improving Genetic Network Models	279
	<i>Eugene van Someren, Lodewyk Wessels, Marcel Reinders, and Eric Backer</i>	
15	Parallel Computation and Visualization Tools for Codetermination Analysis of Multivariate Gene Expression Relations	297
	<i>Edward B. Suh, Edward R. Dougherty, Seungchan Kim, Michael L. Bittner, Yidong Chen, Daniel E. Russ, and Robert L. Martino</i>	
16	Single Nucleotide Polymorphisms and Their Applications	311
	<i>Rudy Guerra and Zhaoxia Yu</i>	
17	The Contribution of Alternative Transcription and Alternative Splicing to the Complexity of Mammalian Transcriptomes	351
	<i>Mihaela Zavolan and Christian Schönbach</i>	
18	Computational Imaging, and Statistical Analysis of Tissue Microarrays: Quantitative Automated Analysis of Tissue Microarrays	381
	<i>Aaron J. Berger, Robert L. Camp, and David L. Rimm</i>	
	Index	405

Preface

During the three years after the publication of the first edition of this book, the computational and statistical research in genomics have become increasingly more important and indispensable for understanding cellular behavior under a variety of environmental conditions and for tackling challenging clinical problems. In the first edition, the organizational structure was: *data* $\rightarrow$ *analysis* $\rightarrow$ *synthesis* $\rightarrow$ *application*. In the second edition, we have kept the same structure but have decided to eliminate chapters that primarily focused on applications.

Our decision was motivated by several factors. Firstly, the main focus of this book is computational and statistical approaches in genomics research. Thus, the main emphasis is on methods rather than on applications. Secondly, many of the chapters already include numerous examples of applications of the discussed methods to current problems in biology.

We have tried to further broaden the range of topics to which end we have included newly contributed chapters on topics such as alternative splicing, tissue microarray image and data analysis, single nucleotide polymorphisms, serial analysis of gene expression, and gene shaving. Additionally, a number of chapters have been updated or revised. We thank all the contributing authors for their contributions and hope that you enjoy reading this book.

Wei Zhang
Houston, TX

Ilya Shmulevich
Seattle, WA