Edited by
Frank Emmert-Streib and Matthias Dehmer

Medical Biostatistics for Complex Diseases

WILEY-BLACKWELL

Edited by
Frank Emmert-Streib and
Matthias Dehmer

Medical Biostatistics
for Complex Diseases

Related Titles

Emmert-Streib, F., Dehmer, M. (eds.)

Analysis of Microarray Data

A Network-Based Approach

2008

ISBN: 978-3-527-31822-3

Dehmer, M., Emmert-Streib, F. (eds.)

Analysis of Complex Networks

From Biology to Linguistics

2009

ISBN: 978-3-527-32345-6

Biswas, A., Datta, S., Fine, J. P., Segal, M. R. (eds.)

Statistical Advances in the Biomedical Sciences

Clinical Trials, Epidemiology, Survival Analysis, and Bioinformatics

2008

ISBN: 978-0-471-94753-0

Knudsen, S.

Cancer Diagnostics with DNA Microarrays

2006

ISBN: 978-0-471-78407-4

Nagl, S. (ed.)

Cancer Bioinformatics: From Therapy Design to Treatment

2006

ISBN: 978-0-470-86304-6

Edited by
Frank Emmert-Streib and Matthias Dehmer

Medical Biostatistics for Complex Diseases

WILEY-BLACKWELL

The Editors

Prof. Dr. Frank Emmert-Streib
Queen's University
Cancer Research
School of Medicine & Biomedical Sciences
97, Lisburn Road
Belfast BT9 7BL
United Kingdom

Prof. Dr. Matthias Dehmer
Universität für Gesundheitswissenschaften UMIT
Institut für Bioinformatik
Eduard Wallnöfer Zentrum 1
6060 Thaur
Austria

Cover
A heatmap of residuals diagnosing model fit in gene-set expression analysis, as described in chapter 5 by A.P. Oron. It was produced using the R open-source statistical language, via the 'heatmap' function. The reader can produce a similar heatmap by running the tutorial script for the 'GSEAlm' package, available at the Bioconductor repository: http://www.bioconductor.org/packages/2.6/bioc/vignettes/GSEAlm/inst/doc/GSEAlm.R (Copyright 2008 by Oxford University Press).

Library of Congress Card No.: applied for

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.d-nb.de.

Composition Thomson Digital, Noida, India
Printing and Binding Strauss GmbH, Mörlenbach
Cover Design Grafik-Design Schulz, Fußgönheim

Printed in the Federal Republic of Germany
Printed on acid-free paper

ISBN: 978-3-527-32585-6

Foreword

The evolution of disease in cancer, metabolic disorders, and immunological disorders is still poorly understood. During the past few decades research has revealed that, in most instances, a complex interaction network of micro-environmental factors including cytokines and cytokine receptors as well as a complex network of signaling pathways and metabolic events contribute to disease evolution and disease progression. In addition, the genetic background, somatic mutations, and epigenetic mechanisms are involved in disease manifestation and disease progression. The heterogeneity of disease points to the complexity of events and factors that may all act together to lead to a frank disorder in the individual patient. Based on this assumption, the evaluation of such complex diseases with respect to the affected cells and cell systems by appropriate biostatistical analysis, including high capacity assays and highly developed multi-parameter evaluation-assays, is a clear medical need.

This book, *Medical Biostatistics for Complex Diseases*, reviews statistical and computational methods for the analysis of high-throughput data and their interpretation with special emphasis on the applicability in biomedical and clinical science. One major aim is to discuss methodologies and assays in order to analyze pathway-specific patterns in various disorders and disease-categories. Such approaches are especially desired because they avoid many problems of methods that focus solely on a single-gene level. For example, detecting differentially expressed genes among experimental conditions or disease stages has received tremendous interest since the introduction of DNA microarrays. However, the inherent problem of a causal connection between a genetic characteristic and a phenotypic trait becomes especially problematic in the context of complex diseases because such diseases involve many factors, externally and internally, and their collective processing. For this reason pathway-approaches form an important step towards a full integration of multilevel factors and interactions to establish a systems biology perspective of physiological processes. From an educational point of view this point cannot be stressed enough because the gene-centric view is still prevalent and dominant in genetics, molecular biology, and medicine. That is why this book can serve as a basis to train a new generation of scientists and to forge their way of thinking.

Medical Biostatistics for Complex Diseases. Edited by Frank Emmert-Streib and Matthias Dehmer

ISBN: 978-3-527-32585-6

Deciphering complex diseases like cancer is a collaborative endeavor requiring the coordinated effort of an interdisciplinary team and highly developed multivariate methods through which the complexity of disorders can be addressed appropriately. For this reason it is notable that the present book also provides a brief introduction to the molecular biological mechanisms of cancer and cancer stem cells. This will be very helpful for biostatisticians and computational biologists, guiding their interpretations with related projects.

It will be very interesting to observe the development of this field during the next few years and to witness, hopefully, many exciting results that blossom from the methods and concepts presented in this book.

Vienna, February 2010 *Peter Valent*

Acknowledgments

I would like to thank Matthias Dehmer and Frank Emmert-Streib for fruitful discussions.

Contents

Preface

This book, *Medical Biostatistics for Complex Diseases*, presents novel approaches for the statistical and computational analysis of high-throughput data from complex diseases. A complex disease is characterized by an intertwined interplay between several genes that are responsible for the pathological phenotype instead of a single gene. This interplay among genes and their products leads to a bio-complexity that makes a characterization and description of such a disease intricate. For this reason, it has been realized that single-gene-specific methods are less insightful than methods based on groups of genes [1]. A possible explanation for this is that the orchestral behavior of genes in terms of their molecular interactions form gene networks [2, 3] that are composed of functional units (subnetworks) that are called pathways. In this respect, analysis methods based on groups of genes may resemble biological pathways and, hence, functional units of the biological system. This is in the spirit of systems theory [4, 5], which requires that a functional part of a system under investigation has to be studied to gain information about its functioning. The transfer of this conceptual framework to biological problems has been manifested in systems biology [6–8]. For this reason, the methods presented in this book emphasize pathway-based approaches. In contrast to network-based approaches for the analysis of high-throughput data [9] a pathway has a less stringent definition than a network [10] which may correspond to the causal molecular interactions or merely to a set of genes constituting it while neglecting their relational structure. Hence, the methodological analysis methods for both types of approaches vary considerably. Further, the present book emphasizes statistical methods because, for example, the need to test for significance or classify robustly is omnipresent in the context of high-throughput data from complex diseases. In a nutshell, the book focuses on a certain perspective of systems biology for the analysis of high-throughput data to help elucidating aspects of complex diseases that may otherwise remain covered.

The book is organized in the following way. The first part consists of three introductory chapters about basic cancer biology, cancer stem cells, and multiple correction methods for hypotheses testing. These chapters cover topics that recur during the book at various degrees and for this reason should be read first. The provided biological knowledge and the statistical methods are indispensable for a systematic design, analysis, and interpretation of high-throughput data from cancer but also other complex diseases. Despite the fact that the present book has a

methodological focus on statistical analysis methods we consider it essential to include also some chapters that provide information about basic biological mechanisms that may be crucial to understand aspects of complex diseases.

The second part of the book presents statistical and computational analysis methods and their application to high-throughput data sets from various complex diseases. Specifically, biological data sets studied are from acute myeloid leukemia (AML), acute lymphoblastic leukemia (ALL), breast cancer, cervical cancer, conventional renal cell carcinoma (cRCC), colorectal cancer, liver cancer, and lung cancer. In addition to these data sets from cancer, also microarray data from diabetes and Duchenne muscular dystrophy (DMD) are used. These biological datasets are complemented by simulated data to study methods theoretically. This part of the book presents chapters that apply and develop methods for identifying differentially expressed genes, integration of data sets, inference of regulatory network, gene set analysis, predicting disease stages or survival times, and pathway analysis. From a methodological point of view the chapters in the second part comprise, for example, analysis of covariance (ANCOVA), bagging, Bayesian networks, dynamic vector autoregressive model, empirical Bayes, false discovery rate (FDR), Granger causality, Hotelling's T^2, kernel methods, least angle regression (LARS), least absolute shrinkage and selection operator (Lasso), Markov chain Monte Carlo (MCMC), model averaging, multiple hypotheses testing, multivariate analysis of variance (MANOVA), random forest, resampling methods, singular-value decomposition (SVD), and support vector machine (SVM).

Regarding the organization of each chapter we decided that the chapters should be presented comprehensively accessible not only to researchers from this field but also to researchers from related fields or even students that have passed already introductory courses. For this reason each chapter presents not only some novel results but also provides some background knowledge necessary to understand, for example, the mathematical method or the biological problem under consideration. In research articles this background information is either completely omitted or the reader is referred to an original article. Hence, this book could also serve as textbook for, e.g., an interdisciplinary seminar for advanced students, not only because of the comprehensiveness of the chapters but also because of its size, which allowing it to fill a complete semester.

The present book is intended for researchers in the interdisciplinary fields of computational biology, biostatistics, bioinformatics, and systems biology studying problems in biomedical sciences. Despite the fact that these fields emerged from traditional disciplines like biology, biochemistry, computer science, electrical engineering, mathematics, medicine, statistics, or physics we want to emphasize that they are now becoming independent. The reasons for this are at least three-fold. First, these fields study problems that cannot be assigned to one of the traditional fields alone, neither biologically nor methodologically. Second, the studied problems are considered of general importance, not only for science itself but society because of their immediate impact on public health. Third, biomedical problems *demand* the development of novel statistical and computational methodology for their problem-oriented and efficient investigation. This implies that none of the traditional

quantitative fields provide ready-to-use solutions to many of the urgent problems we are currently facing when studying the basic molecular mechanisms of complex diseases. This explains the eruption of methodological papers that appeared during the last two decades. Triggered by continuing technological developments leading to new or improved high-throughput measurement devices it is expected that this process will continue. The quest for a systematic understanding of complex diseases is intriguing not only because we acquire a precise molecular and cellular "picture" of organizational processes within and among cells but especially because of consequences that may result from this. For example, insights from such studies may translate directly into rational drug design and stem cell research.

Many colleagues, whether consciously or unconsciously, have provided us with input, help, and support before and during the formation of the present book. In particular we would like to thank Andreas Albrecht, Gökmen Altay, Gökhan Bakır, Igor Bass, David Bialy, Danail Bonchev, Ulrike Brandt, Stefan Borgert, Mieczysław Borowiecki, Andrey A. Dobrynin, Michael Drmota, Maria Duca, Dean Fennell, Isabella Fritz, Maria Fonoberova, Boris Furtula, Bernhard Gittenberger, Galina Glazko, Armin Graber, Martin Grabner, Earl Glynn, Ivan Gutman, Arndt von Haeseler, Peter Hamilton, Bernd Haas, Des Higgins, Dirk Husmeier, Wilfried Imrich, Puthen Jithesh, Patrick Johnston, Frank Kee, Jürgen Kilian, Elena Konstantinova, Terry Lappin, D. D. Lozovanu, Dennis McCance, Alexander Mehler, Abbe Mowshowitz, Ken Mills, Arcady Mushegian, Klaus Pawelzik, Andrei Perjan, Marina Popovscaia, William Reeves, Bert Rima, Armindo Salvador, Heinz Georg Schuster, Helmut Schwegler, Chris Seidel, Andre Ribeiro, Ricardo de Matos Simoes, Francesca Shearer, Brigitte Senn-Kircher, Fred Sobik, Doru Stefanescu, John Storey, Robert Tibshirani, Shailesh Tripathi, Kurt Varmuza, Suzanne D. Vernon, Robert Waterston, Bruce Weir, Olaf Wolkenhauer, Bohdan Zelinka, Shu-Dong Zhang, and Dongxiao Zhu, and apologize to all who have not been named mistakenly. We would like also to thank our editors Andreas Sendtko and Gregor Cicchetti from Wiley-VCH who have been always available and helpful. Last but not least we would like to thank our families for support and encouragement during all that time.

Finally, we hope this book helps to spread our enthusiasm and joy we have for this field and inspires people regarding their own practical or theoretical research problems.

Belfast and Hall/Tyrol
January 2010

F. Emmert-Streib and
M. Dehmer

References

1 Emmert-Streib, F. (2007) The chronic fatigue syndrome: a comparative pathway analysis. *J. Comput. Biol.*, **14** (7), 961–972.

2 Barabási, A.L. and Oltvai, Z.N. (2004) Network biology: Understanding the cell's functional organization. *Nat. Rev. Genet.*, **5**, 101–113.

3 Kauffman, S.A. (1969) Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.*, **22**, 437–467.
4 Bertalanffy, L.v. (1950) An outline of general systems theory. *Br. J. Philos. Sci.*, **1** (2), 134–165.
5 Bertalanffy, L.v. (1976) *General System Theory: Foundations, Development, Applications*, revised edn, George Braziller, New York.
6 Alon, U. (2006) *An Introduction to Systems Biology: Design Principles of Biological Circuits*, Chapman & Hall/CRC.
7 Kitano, H. (ed.) (2001) *Foundations of Systems Biology*, MIT Press.
8 Palsson, B.O. (2006) *Systems Biology: Properties of Reconstructed Networks*, Cambridge University Press, New York.
9 Emmert-Streib, F. and Dehmer, M. (eds) (2008) *Analysis of Microarray Data: A Network-Based Approach*, Wiley-VCH Verlag, Weinheim.
10 Dehmer, M. and Emmert-Streib, F. (eds) (2009) *Analysis of Complex Networks: From Biology to Linguistics*, Wiley-VCH Verlag, Weinheim.

List of Contributors

Gyan Bhanot
Rutgers University
BioMaPs Institute for Quantitative Biology
610 Taylor Road
Piscataway, NJ 08854
USA

Gregory A. Bird
University of Colorado Denver
Department of Dermatology
Charles C. Gates Program in Regenerative Medicine and Stem Cell Biology
12800 E. 19th Avenue
Aurora, CO 80045
USA

A. Rose Brannon
University of North Carolina
Lineberger Comprehensive Cancer Center
CB 7295
Chapel Hill, NC 27599-7295
USA

Eike C. Buss
Heidelberg University
Department of Internal Medicine V
Im Neuenheimer Feld 410
69120 Heidelberg
Germany

Bor-Sen Chen
National Tsing Hua University
Department of Electrical Engineering
101, Section 2, Kuang-Fu Road
Hsinchu 30013
Taiwan

James J. Chen
FDA/National Center for Toxicological Research
Division of Personalized Nutrition and Medicine
3900 NCTR Road
Jefferson, AR 72079
USA

Sung Bum Cho
Seoul National University
College of Medicine
Division of Biomedical and Healthcare Informatics
28 Yongon-dong, Chongno-gu
Seoul 110-799
Korea

Liang-Hui Chu
National Tsing Hua University
Department of Electrical Engineering
101, Section 2, Kuang-Fu Road
Hsinchu 30013
Taiwan

Robert Clarke
Georgetown University Medical Center
Department of Oncology
3970 Reservoir Rd NW
Washington DC, 20057
USA

Marcos Angelo Almeida Demasi
University of Sao Paulo
Institute of Mathematics and Statistics
Av. Prof. Lineu Prestes, 748
Butanta
05508-900 Sao Paulo
Brazil

Hong-Wen Deng
University of Missouri-Kansas City
Departments of Orthopedic Surgery and
Basic Medical Sciences
2411 Holmes Street
Kansas City, MO 64108
USA

Elena Edelman
Harvard Medical School
Department of Medicine
185 Cambridge Street
CPZN 4200
Boston, MA 02114
USA

Alessio Farcomeni
Sapienza - University of Rome
Piazzale Aldo Moro, 5
00186 Rome
Italy

Carlos Eduardo Ferreira
University of Sao Paulo
Institute of Mathematics and Statistics
Av. Prof. Lineu Prestes, 748
Butanta
05508-900 Sao Paulo
Brazil

Irit Fishel
Tel Aviv University
School of Computer Sciences and
School of Medicine
Schreiber Building Ramat Aviv
69978 Tel Aviv
Israel

Wataru Fujibuchi
National Institute of Advanced
Industrial Science and Technology
(ASIT)
Computational Biology Research Centre
2-42 Aomi, Koto-ku
Tokyo 135-0064
Japan

André Fujita
University of Sao Paulo
Institute of Mathematics and Statistics
Av. Prof. Lineu Prestes, 748
Butanta
05508-900 Sao Paulo
Brazil

Shridar Ganesan
Cancer Institute of New Jersey
195 Little Albany Street
New Brunswick, NJ 08903
USA

Katherine Garman
Duke University
Institute for Genome Sciences & Policy
Department of Medicine
101 Science Drive
Durham, NC 27708
USA

Leena Hilakivi-Clarke
Georgetown University Medical Center
Department of Oncology
3970 Reservoir Rd NW
Washington, DC 20057
USA

Steven M. Hill
University of Warwick
Centre for Complexity Science
Zeeman Building
Coventry CV4 7AL
UK

Anthony D. Ho
University of Heidelberg
Department of Internal Medicine V
Im Neuenheimer Feld 410
69120 Heidelberg
Germany

Taeyoung Hwang
Seoul National University
Department of Statistics
56-1 Shillim-Dong, Kwang-Gu
Seoul 151-747
Korea

Tsuyoshi Kato
National Institute of Advanced
Industrial Science and Technology
(ASIT)
Computational Biology Research Centre
2-42 Aomi, Koto-ku
Tokyo 135-0064
Japan

Jihun Kim
Seoul National University
College of Medicine
Division of Biomedical and Healthcare
Informatics
28 Yongon-dong, Chongno-gu
Seoul 110-799
Korea

Ju Han Kim
Seoul National University
College of Medicine
Division of Biomedical and Healthcare
Informatics
28 Yongon-dong, Chongno-gu
Seoul 110-799
Korea

Ki-Yeol Kim
Yonsei University College of Dentistry
Oral Cancer Research Institute
250 Seongsanno Seodaemun-gu
Seoul 120-752
Korea

Louis-Philippe Kronek
G-SCOP, Grenoble-Science Conception
Organization and Production
46 Avenue Viallet
38031 Grenoble
France

Huai Li
Bioinformatics Unit
National Institute on Aging
National Institutes of Health
Baltimore, MD 21224
USA

Peng-Yuan Liu
Washington University School of Medicine
Department of Surgery and
the Alvin J. Siteman Cancer Center
660 South Euclid Avenue
Campus Box 8109
St. Louis, MO 63110
USA

Yan Lu
Washington University School of Medicine
Department of Surgery and
the Alvin J. Siteman Cancer Center
660 South Euclid Avenue
Campus Box 8109
St. Louis, MO 63110
USA

Satoru Miyano
University of Sao Paulo
Institute of Mathematics and Statistics
Av. Prof. Lineu Prestes, 748
Butanta
05508-900 Sao Paulo
Brazil

Sach Mukherjee
University of Warwick
Centre for Complexity Science
Zeeman Building
Coventry CV4 7AL
UK

Sayan Mukherjee
Duke University
Institute for Genome Sciences & Policy
Departments of Statistical Science,
Computer Science, and Mathematics
214 Old Chemistry Building
Durham, NC 27708
USA

Assaf P. Oron
University of Washington
Department of Statistics
Box 354322
Seattle, WA 98195
USA

Taesung Park
Seoul National University
Department of Statistics
56-1 Shillim-Dong, Kwang-Gu
Seoul 151-747
Korea

Anil Potti
Duke University
Institute for Genome Sciences & Policy
Department of Medicine
101 Science Drive
Durham, NC 27708
USA

W. Kimryn Rathmell
University of North Carolina
Lineberger Comprehensive Cancer Center
CB 7295
Chapel Hill, NC 27599
USA

Anupama Reddy
Rutgers University
RUTCOR - Rutgers Center for Operations Research
640 Bartholomew Rd.
Piscataway, NJ 08854
USA

Yosef Refaeli
University of Colorado Denver
Department of Dermatology
Charles C. Gates Program in
Regenerative Medicine and Stem Cell
Biology
12800 E. 19th Avenue
Aurora, CO 80045
USA

Sun Young Rha
Yonsei University College of Medicine
Yonsei Cancer Center
134 Shinchon-Dong Seodaemun-Ku
Seoul 120-752
Korea

Eytan Ruppin
Tel Aviv University
School of Computer Sciences and
School of Medicine
Schreiber Building Ramat Aviv
69978 Tel Aviv
Israel

João Ricardo Sato
University of Sao Paulo
Institute of Mathematics and Statistics
Av. Prof. Lineu Prestes, 748
Butanta
05508-900 Sao Paulo
Brazil

Michael Seiler
Rutgers University
BioMaPs Institute for Quantitative
Biology
610 Taylor Road
Piscataway, NJ 08854
USA

Mari Cleide Sogayar
University of Sao Paulo
Institute of Mathematics and Statistics
Av. Prof. Lineu Prestes, 748
Butanta
05508-900 Sao Paulo
Brazil

Terence P. Speed
Walter and Eliza Hall Institute of
Medical Research
1G Royal Parade
Parkville Victoria 3052
Australia

Chen-An Tsai
China Medical University
Graduate Institute of Biostatistics &
Biostatistics Center
Taichung
91 Hsueh-Shih Road
Taiwan 40402
R.O.C.

Brian C. Turner
University of Colorado Denver
Department of Dermatology
Charles C. Gates Program in
Regenerative Medicine and Stem Cell
Biology
12800 E. 19th Avenue
Aurora, CO 80045
USA

Yue Wang
Virginia Polytechnic Institute and State
University
Department of Electrical and Computer
Engineering
4300 Wilson Blvd.
Arlington, VA 22203
USA

Bai Zhang
Virginia Polytechnic Institute and State University
Department of Electrical and Computer Engineering
4300 Wilson Blvd., Suite 750
Arlington, VA 22203
USA