

*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).

**Limit of Liability/Disclaimer of Warranty:** While the publisher and authors have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty can be created or extended by sales representatives or written sales materials. The Advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

**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>.

© 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany

Wiley-Blackwell is an imprint of John Wiley & Sons, formed by the merger of Wiley's global Scientific, Technical, and Medical business with Blackwell Publishing.

All rights reserved (including those of translation into other languages). No part of this book may be reproduced in any form – by photoprinting, microfilm, or any other means – nor transmitted or translated into a machine language without written permission from the publishers. Registered names, trademarks, etc. used in this book, even when not specifically marked as such, are not to be considered unprotected by law.

**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.

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

<b>Foreword</b>	V
<b>Preface</b>	XIX
<b>List of Contributors</b>	XXIII

### Part One General Biological and Statistical Basics 1

<b>1</b>	<b>The Biology of MYC in Health and Disease: A High Altitude View</b>	<b>3</b>
	<i>Brian C. Turner, Gregory A. Bird, and Yosef Refaeli</i>	
1.1	Introduction	3
1.2	MYC and Normal Physiology	4
1.3	Regulation of Transcription and Gene Expression	4
1.4	Metabolism	6
1.5	Cell-Cycle Regulation and Differentiation	7
1.6	Protein Synthesis	7
1.7	Cell Adhesion	7
1.8	Apoptosis	8
1.9	MicroRNAs	9
1.10	Physiological Effects of Loss and Gain of <i>c-myc</i> Function in Mice	9
1.10.1	Loss of Function	9
1.10.2	Gain of Function: Inducible Transgenic Animals	10
1.11	Contributions of MYC to Tumor Biology	11
1.12	Introduction of Hematopoietic Malignancies	12
1.13	Mechanisms of MYC Dysregulation in Hematological Malignancies	13
1.14	Mutation(s) in the MYC Gene in Hematological Cancers	14
1.15	Role of MYC in Cell Cycle Regulation and Differentiation in Hematological Cancers	14
1.16	Role of BCR Signaling in Conjunction with MYC Overexpression in Lymphoid Malignancies	15
1.17	Deregulation of Auxiliary Proteins in Addition to MYC in Hematological Cancers	16

1.18	Conclusion	17
	References	18
<b>2</b>	<b>Cancer Stem Cells – Finding and Capping the Roots of Cancer</b>	<b>25</b>
	<i>Eike C. Buss and Anthony D. Ho</i>	
2.1	Introduction – Stem Cells and Cancer Stem Cells	25
2.1.1	What are Stem Cells?	25
2.1.2	Concept of Cancer Stem Cells (CSCs)	25
2.2	Hematopoietic Stem Cells as a Paradigm	28
2.2.1	Leukemia as a Paradigmatic Disease for Cancer Research	28
2.2.2	CFUs	29
2.2.3	LTC-ICs	29
2.2.4	<i>In Vivo</i> Repopulation	30
2.2.5	Importance of the Bone Marrow Niche	30
2.2.6	Leukemic Stem Cells	31
2.2.6.1	Leukemic Stem Cells in the Bone Marrow Niche	31
2.2.7	CML as a Paradigmatic Entity	32
2.3	Current Technical Approach to the Isolation and Characterization of Cancer Stem Cells	33
2.3.1	Tools for the Detection of Cancer Stem Cells	33
2.3.2	Phenotype of Cancer Stem Cells	34
2.4	Cancer Stem Cells in Solid Tumors	35
2.4.1	Breast Cancer	36
2.4.2	Prostate Cancer	36
2.4.3	Colon Cancer	37
2.4.4	Other Cancers	37
2.5	Open Questions of the Cancer Stem Cell Hypothesis	37
2.6	Clinical Relevance of Cancer Stem Cells	38
2.6.1	Diagnostic Relevance of Cancer Stem Cells	38
2.6.2	Therapeutic Relevance – New Drugs Directed Against Cancer Stem Cells	39
2.7	Outlook	40
	References	40
<b>3</b>	<b>Multiple Testing Methods</b>	<b>45</b>
	<i>Alessio Farcomeni</i>	
3.1	Introduction	45
3.1.1	A Brief More Focused Introduction	46
3.1.2	Historic Development of the Field	47
3.2	Statistical Background	48
3.2.1	Tests	48
3.2.2	Test Statistics and <i>p</i> -Values	49
3.2.3	Resampling Based Testing	49
3.3	Type I Error Rates	50
3.4	Introduction to Multiple Testing Procedures	52

3.4.1	Adjusted $p$ -values	52
3.4.2	Categories of Multiple Testing Procedures	52
3.4.3	Estimation of the Proportion of False Nulls	53
3.5	Multiple Testing Procedures	55
3.5.1	Procedures Controlling the FWER	55
3.5.2	Procedures Controlling the FDR	56
3.5.3	Procedures Controlling the FDX	59
3.6	Type I Error Rates Control Under Dependence	61
3.6.1	FWER Control	62
3.6.2	FDR and FDX Control	62
3.7	Multiple Testing Procedures Applied to Gene Discovery in DNA Microarray Cancer Studies	63
3.7.1	Gene identification in Colon Cancer	64
3.7.1.1	Classification of Lymphoblastic and Myeloid Leukemia	64
3.8	Conclusions	67
	References	69

## Part Two Statistical and Computational Analysis Methods 73

<b>4</b>	<b>Making Mountains Out of Molehills: Moving from Single Gene to Pathway Based Models of Colon Cancer Progression</b>	<b>75</b>
	<i>Elena Edelman, Katherine Garman, Anil Potti, and Sayan Mukherjee</i>	
4.1	Introduction	75
4.2	Methods	76
4.2.1	Data Collection and Standardization	76
4.2.2	Stratification and Mapping to Gene Sets	77
4.2.3	Regularized Multi-task Learning	78
4.2.4	Validation via Mann–Whitney Test	79
4.2.5	Leave-One-Out Error	79
4.3	Results	80
4.3.1	Development and Validation of Model Statistics	82
4.3.2	Comparison of Single Gene and Gene Set Models	83
4.3.3	Novel Pathway Findings and Therapeutic Implications	84
4.4	Discussion	85
	References	86
<b>5</b>	<b>Gene-Set Expression Analysis: Challenges and Tools</b>	<b>89</b>
	<i>Assaf P. Oron</i>	
5.1	The Challenge	89
5.2	Survey of Gene-Set Analysis Methods	91
5.2.1	Motivation for GS Analysis	91
5.2.2	Some Notable GS Analysis Methods	92
5.2.3	Correlations and Permutation Tests	95
5.3	Demonstration with the “ALL” Dataset	97
5.3.1	The Dataset	97

5.3.2	The Gene-Filtering Dilemma	97
5.3.3	Basic Diagnostics: Testing Normalization and Model Fit	99
5.3.4	Pinpointing Aneuploidies via Outlier Identification	102
5.3.5	Signal-to-Noise Evaluation: The Sex Variable	103
5.3.6	Confounding, and Back to Basics: The Age Variable	106
5.3.7	How it all Reflects on the Bottom Line: Inference	107
5.4	Summary and Future Directions	108
	References	111

**6 Multivariate Analysis of Microarray Data Using Hotelling's  $T^2$  Test** 113

*Yan Lu, Peng-Yuan Liu, and Hong-Wen Deng*

6.1	Introduction	113
6.2	Methods	114
6.2.1	Wishart Distribution	114
6.2.2	Hotelling's $T^2$ Statistic	115
6.2.3	Two-Sample $T^2$ Statistic	115
6.2.4	Multiple Forward Search (MFS) Algorithm	116
6.2.5	Resampling	117
6.3	Validation of Hotelling's $T^2$ Statistic	118
6.3.1	Human Genome U95 Spike-In Dataset	118
6.3.2	Identification of DEGs	118
6.4	Application Examples	118
6.4.1	Human Liver Cancers	118
6.4.1.1	Dataset	118
6.4.1.2	Identification of DEGs	120
6.4.1.3	Classification of Human Liver Tissues	122
6.4.2	Human Breast Cancers	124
6.4.2.1	Dataset	124
6.4.2.2	Cluster Analysis	124
6.5	Discussion	124
	References	128

**7 Interpreting Differential Coexpression of Gene Sets** 131

*Ju Han Kim, Sung Bum Cho, and Jihun Kim*

7.1	Coexpression and Differential Expression Analyses	131
7.2	Gene Set-Wise Differential Expression Analysis	133
7.3	Differential Coexpression Analysis	134
7.4	Differential Coexpression Analysis of Paired Gene Sets	135
7.5	Measuring Coexpression of Gene Sets	136
7.6	Measuring Differential Coexpression of Gene Sets	137
7.7	Gene Pair-Wise Differential Coexpression	138
7.8	Datasets and Gene Sets	139
7.8.1	Datasets	139
7.8.2	Gene Sets	139

7.9	Simulation Study	139
7.10	Lung Cancer Data Analysis Results	140
7.11	Duchenne's Muscular Dystrophy Data Analysis Results	142
7.12	Discussion	145
	References	150
<b>8</b>	<b>Multivariate Analysis of Microarray Data: Application of MANOVA</b>	<b>151</b>
	<i>Taeyoung Hwang and Taesung Park</i>	
8.1	Introduction	151
8.2	Importance of Correlation in Multiple Gene Approach	152
8.2.1	Small Effects Coordinate to Make a Big Difference	154
8.2.2	Significance of the Correlation	155
8.3	Multivariate ANalysis of VAriance (MANOVA)	155
8.3.1	ANOVA	156
8.3.2	MANOVA	157
8.4	Applying MANOVA to Microarray Data Analysis	159
8.5	Application of MANOVA: Case Studies	160
8.5.1	Identifying Disease Specific Genes	160
8.5.2	Identifying Significant Pathways from Public Pathway Databases	161
8.5.3	Identification of Subnetworks from Protein–Protein Interaction Data	162
8.6	Conclusions	163
	References	165
<b>9</b>	<b>Testing Significance of a Class of Genes</b>	<b>167</b>
	<i>James J. Chen and Chen-An Tsai</i>	
9.1	Introduction	167
9.2	Competitive versus Self-Contained Tests	169
9.3	One-Sided and Two-Sided Hypotheses	171
9.4	Over-Representation Analysis (ORA)	171
9.5	GCT Statistics	172
9.5.1	One-Sided Test	174
9.5.1.1	OLS Global Test	174
9.5.1.2	GSEA Test	175
9.5.2	Two-Sided Test	175
9.5.2.1	MANOVA Test	175
9.5.2.2	SAM-GS Test	176
9.5.2.3	ANCOVA Test	177
9.6	Applications	177
9.6.1	Diabetes Dataset	177
9.6.2	p53 Dataset	180
9.7	Discussion	181
	References	182

<b>10</b>	<b>Differential Dependency Network Analysis to Identify Topological Changes in Biological Networks</b>	<b>185</b>
	<i>Bai Zhang, Huai Li, Robert Clarke, Leena Hilakivi-Clarke, and Yue Wang</i>	
10.1	Introduction	185
10.2	Preliminaries	187
10.2.1	Probabilistic Graphical Models and Dependency Networks	187
10.2.2	Graph Structure Learning and $\ell_1$ -Regularization	188
10.3	Method	188
10.3.1	Local Dependency Model in DDN	188
10.3.2	Local Structure Learning	189
10.3.3	Detection of Statistically Significant Topological Changes	191
10.3.4	Identification of “Hot Spots” in the Network and Extraction of the DDN	192
10.4	Experiments and Results	192
10.4.1	A Simulation Experiment	192
10.4.1.1	Experiment Data	193
10.4.1.2	Application of DDN Analysis	193
10.4.1.3	Algorithm Analysis	195
10.4.2	Breast Cancer Dataset Analysis	196
10.4.2.1	Experiment Background and Data	196
10.4.2.2	Application of DDN Analysis	197
10.4.3	<i>In Utero</i> Excess E2 Exposed Adult Mammary Glands Analysis	198
10.4.3.1	Experiment Background and Data	198
10.4.4	Application of DDN Analysis	198
10.5	Closing Remarks	199
	References	200
<b>11</b>	<b>An Introduction to Time-Varying Connectivity Estimation for Gene Regulatory Networks</b>	<b>205</b>
	<i>André Fujita, João Ricardo Sato, Marcos Angelo Almeida Demasi, Satoru Miyano, Mari Cleide Sogayar, and Carlos Eduardo Ferreira</i>	
11.1	Regulatory Networks and Cancer	205
11.2	Statistical Approaches	207
11.2.1	Causality and Granger Causality	207
11.2.2	Vector Autoregressive Model – VAR	209
11.2.2.1	Estimation Procedure	210
11.2.2.2	Hypothesis Testing	211
11.2.3	Dynamic Vector Autoregressive Model – DVAR	211
11.2.3.1	Estimation Procedure	214
11.2.3.2	Covariance Matrix Estimation	215
11.2.3.3	Hypothesis Testing	215
11.3	Simulations	216
11.4	Application of the DVAR Method to Actual Data	218
11.5	Final Considerations	222
11.6	Conclusions	224

11.A	Appendix	225
	References	227
<b>12</b>	<b>A Systems Biology Approach to Construct A Cancer-Perturbed Protein–Protein Interaction Network for Apoptosis by Means of Microarray and Database Mining</b>	<b>231</b>
	<i>Liang-Hui Chu and Bor-Sen Chen</i>	
12.1	Introduction	231
12.2	Methods	233
12.2.1	Microarray Experimental Data	233
12.2.2	Construction of Initial Protein–Protein Interaction (PPI) Networks	233
12.2.3	Nonlinear Stochastic Interaction Model	233
12.2.4	Identification of Interactions in the Initial Protein–Protein Interaction Network	236
12.2.5	Modification of Initial PPI Networks	238
12.3	Results	239
12.3.1	Construction of a Cancer-Perturbed PPI Network for Apoptosis	239
12.3.2	Prediction of Apoptosis Drug Targets by Means of Cancer-Perturbed PPI Networks for Apoptosis	241
12.3.2.1	Common Pathway: CASP3	244
12.3.2.2	Extrinsic Pathway and Cross-Talk: TNF	244
12.3.2.3	Intrinsic Pathway: BCL2, BAX, and BCL2L1	244
12.3.2.4	Apoptosis Regulators: TP53, MYC, and EGFR	245
12.3.2.5	Stress-Induced Signaling: MAPK1 and MAPK3	245
12.3.2.6	Others: CDKN1A	245
12.3.3	Prediction of More Apoptosis Drug Targets by Decreasing the Degree of Perturbation	246
12.3.3.1	Prediction of More Cancer Drug Targets by Decreasing the Degree of Perturbation	246
12.3.3.2	Prediction of New GO Annotations of the Four Proteins: CDKN1A, CCND, PRKCD, and PCNA	246
12.4	Apoptosis Mechanism at the Systems Level	247
12.4.1	Caspase Family and Caspase Regulators	247
12.4.2	Extrinsic Pathway, Intrinsic Pathway, and Cross-Talk	248
12.4.3	Regulation of Apoptosis at the Systems Level	248
12.5	Conclusions	248
	References	249
<b>13</b>	<b>A New Gene Expression Meta-Analysis Technique and Its Application to Co-Analyze Three Independent Lung Cancer Datasets</b>	<b>253</b>
	<i>Irit Fishel, Alon Kaufman, and Eytan Ruppin</i>	
13.1	Background	253
13.1.1	DNA Microarray Technology	253
13.1.1.1	cDNA Microarray	253

13.1.1.2	Oligonucleotide Microarray	255
13.1.2	Machine Learning Background	255
13.1.2.1	Basic Definitions and Terms in Machine Learning	255
13.1.2.2	Supervised Learning in the Context of Gene Expression Data	256
13.1.3	Support Vector Machines	256
13.1.4	Support Vector Machine Recursive Feature Elimination	258
13.2	Introduction	259
13.3	Methods	260
13.3.1	Overview and Definitions	260
13.3.2	A Toy Example	261
13.3.3	Datasets	263
13.3.4	Data Pre-processing	263
13.3.5	Probe Set Reduction	264
13.3.6	Constructing a Predictive Model	264
13.3.7	Constructing Predictive Gene Sets	264
13.3.8	Estimating the Predictive Performance	266
13.3.9	Constructing a Repeatability-Based Gene List	266
13.3.10	Ranking the Joint Core Genes	267
13.4	Results	267
13.4.1	Unstable Ranked Gene Lists in a Tumor Versus Normal Binary Classification Task	267
13.4.2	Constructing a Consistent Repeatability-Based Gene List	268
13.4.3	Repeatability-Based Gene Lists are Stable	269
13.4.4	Comparing Gene Rankings between Datasets	269
13.4.5	Joint Core Magnitude	270
13.4.6	The Joint Core is Transferable	271
13.4.7	Biological Significance of the Joint Core Genes	272
13.5	Discussion	273
	References	275
<b>14</b>	<b>Kernel Classification Methods for Cancer Microarray Data</b>	<b>279</b>
	<i>Tsuyoshi Kato and Wataru Fujibuchi</i>	
14.1	Introduction	279
14.1.1	Notation	280
14.2	Support Vector Machines and Kernels	281
14.2.1	Support Vector Machines	281
14.2.2	Kernel Matrix	284
14.2.3	Polynomial Kernel and RBF Kernel	285
14.2.4	Pre-process of Kernels	286
14.2.4.1	Normalization	286
14.2.4.2	SVD Denoising	287
14.3	Metritzation Kernels: Kernels for Microarray Data	288
14.3.1	Partial Distance (or kNND)	288
14.3.2	Maximum Entropy Kernel	289
14.3.3	Other Distance-Based Kernels	290



14.4	Applications to Cancer Data	290
14.4.1	Leave-One-Out Cross Validation	291
14.4.2	Data Normalization and Classification Analysis	291
14.4.3	Parameter Selection	292
14.4.4	Heterogeneous Kidney Carcinoma Data	292
14.4.5	Problems in Training Multiple Support Vector Machines for All Sub-data	293
14.4.6	Effects of Partial Distance Denoising in Homogeneous Leukemia Data	293
14.4.7	Heterogeneous Squamous Cell Carcinoma Metastasis Data	295
14.4.8	Advantages of ME Kernel	296
14.5	Conclusion	296
14.A	Appendix	298
	References	300
<b>15</b>	<b>Predicting Cancer Survival Using Expression Patterns</b>	<b>305</b>
	<i>Anupama Reddy, Louis-Philippe Kronek, A. Rose Brannon, Michael Seiler, Shridar Ganesan, W. Kimryn Rathmell, and Gyan Bhanot</i>	
15.1	Introduction	305
15.2	Molecular Subtypes of ccRCC	307
15.3	Logical Analysis of Survival Data	308
15.4	Bagging LASD Models	311
15.5	Results	312
15.5.1	Prediction Results are More Accurate after Stratifying Data into Subtypes	313
15.5.2	LASD Performs Significantly Better than Cox Regression	313
15.5.3	Bagging Improves Robustness of LASD Predictions	314
15.5.4	LASD Patterns have Distinct Survival Profiles	314
15.5.5	Importance Scores for Patterns and an Optimized Risk Score	314
15.5.6	Risk Scores could be used to Classify Patients into Distinct Risk Groups	316
15.5.7	LASD Survival Prediction is Highly Predictive When Compared with Clinical Parameters (Stage, Grade, and Performance)	318
15.6	Conclusion and Discussion	318
	References	322
<b>16</b>	<b>Integration of Microarray Datasets</b>	<b>325</b>
	<i>Ki-Yeol Kim and Sun Young Rha</i>	
16.1	Introduction	325
16.2	Integration Methods	325
16.2.1	Existing Methods for Adjusting Batch Effects	326
16.2.1.1	Singular Value Decomposition (SVD) and Distance Weighted Discrimination (DWD)	326
16.2.1.2	ANOVA (Analysis of Variance) Model	327
16.2.1.3	Empirical Bayesian Method for Adjusting Batch Effect	327

16.2.2	Transformation Method	329
16.2.2.1	Standardization of Expression Data	329
16.2.2.2	Transformation of Datasets Using a Reference Dataset	330
16.2.3	Discretization Methods	332
16.2.3.1	Equal Width and Equal Frequency Discretizations	332
16.2.3.2	ChiMerge Method	333
16.2.3.3	Discretization Based on Recursive Minimal Entropy	333
16.2.3.4	Nonparametric Scoring Method for Microarray Data	333
16.2.3.5	Discretization by Rank of Gene Expression in Microarray Dataset: Proposed Method	335
16.3	Statistical Method for Significant Gene Selection and Classification	336
16.3.1	Chi-Squared Test for Significant Gene Selection	336
16.3.2	Random Forest for Calculating Prediction Accuracy	337
16.4	Example	337
16.4.1	Dataset	338
16.4.2	Prediction Accuracies Using the Combined Dataset	339
16.4.2.1	Data Preprocessing	339
16.4.2.2	Improvement of Prediction Accuracy Using Combined Datasets by the Proposed Method	339
16.4.2.3	Description of Significant Genes Selected from a Combined Dataset by the Proposed Method	340
16.4.2.4	Improvement of Prediction Accuracies by Combining Datasets Performed using Different Platforms	340
16.4.3	Conclusions	341
16.5	Summary	342
	References	342
<b>17</b>	<b>Model Averaging for Biological Networks with Prior Information</b>	<b>347</b>
	<i>Sach Mukherjee, Terence P. Speed, and Steven M. Hill</i>	
17.1	Introduction	347
17.2	Background	349
17.2.1	Bayesian Networks	349
17.2.2	Model Scoring	350
17.2.3	Model Selection and Model Averaging	351
17.2.4	Markov Chain Monte Carlo on Graphs	354
17.3	Network Priors	356
17.3.1	A Motivating Example	356
17.3.2	General Framework	357
17.3.2.1	Specific Edges	357
17.3.2.2	Classes of Vertices	358
17.3.2.3	Higher-Level Network Features	358
17.3.2.4	Network Sparsity	358
17.3.2.5	Degree Distributions	359
17.3.2.6	Constructing a Prior	359

17.3.3	Prior-Based Proposals	359
17.4	Some Results	360
17.4.1	Simulated Data	360
17.4.1.1	Priors	361
17.4.1.2	MCMC	362
17.4.1.3	ROC Analysis	362
17.4.2	Prior Sensitivity	362
17.4.3	A Biological Network	362
17.4.3.1	Data	363
17.4.3.2	Priors	364
17.4.3.3	MCMC	365
17.4.3.4	Single Best Graph	365
17.4.3.5	Network Features	365
17.4.3.6	Prior Sensitivity	365
17.5	Conclusions and Future Prospects	366
17.6	Appendix	369
	References	370
	<b>Index</b>	<b>373</b>



## 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, Danaïl 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