

Contents

Foreword	V
Preface	XIX
List of Contributors	XXIII

Part One General Biological and Statistical Basics 1

1	The Biology of MYC in Health and Disease: A High Altitude View	3
	<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
2	Cancer Stem Cells – Finding and Capping the Roots of Cancer	25
	<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
3	Multiple Testing Methods	45
	<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

4	Making Mountains Out of Molehills: Moving from Single Gene to Pathway Based Models of Colon Cancer Progression	75
	<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
5	Gene-Set Expression Analysis: Challenges and Tools	89
	<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
8	Multivariate Analysis of Microarray Data: Application of MANOVA	151
	<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
9	Testing Significance of a Class of Genes	167
	<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

10	Differential Dependency Network Analysis to Identify Topological Changes in Biological Networks	185
	<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 ℓ_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
11	An Introduction to Time-Varying Connectivity Estimation for Gene Regulatory Networks	205
	<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
12	A Systems Biology Approach to Construct A Cancer-Perturbed Protein-Protein Interaction Network for Apoptosis by Means of Microarray and Database Mining	231
	<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
13	A New Gene Expression Meta-Analysis Technique and Its Application to Co-Analyze Three Independent Lung Cancer Datasets	253
	<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
14	Kernel Classification Methods for Cancer Microarray Data	279
	<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	Metritization 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
15	Predicting Cancer Survival Using Expression Patterns	305
	<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
16	Integration of Microarray Datasets	325
	<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
17	Model Averaging for Biological Networks with Prior Information	347
	<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
	Index	373