
Contents

1 Introduction to Genomic and Proteomic Data Analysis	
<i>Daniel Berrar, Martin Granzow, and Werner Dubitzky</i>	1
1.1 Introduction	1
1.2 A Short Overview of Wet Lab Techniques	3
1.2.1 Transcriptomics Techniques in a Nutshell	3
1.2.2 Proteomics Techniques in a Nutshell	5
1.3 A Few Words on Terminology	6
1.4 Study Design	7
1.5 Data Mining	8
1.5.1 Mapping Scientific Questions to Analytical Tasks	9
1.5.2 Visual Inspection	11
1.5.3 Data Pre-Processing	13
1.5.3.1 Handling of Missing Values	13
1.5.3.2 Data Transformations	14
1.5.4 The Problem of Dimensionality	15
1.5.4.1 Mapping to Lower Dimensions	15
1.5.4.2 Feature Selection and Significance Analysis	16
1.5.4.3 Test Statistics for Discriminatory Features	17
1.5.4.4 Multiple Hypotheses Testing	19
1.5.4.5 Random Permutation Tests	21
1.5.5 Predictive Model Construction	22
1.5.5.1 Basic Measures of Performance	24
1.5.5.2 Training, Validating, and Testing	25
1.5.5.3 Data Resampling Strategies	27
1.5.6 Statistical Significance Tests for Comparing Models	29
1.6 Result Post-Processing	31
1.6.1 Statistical Validation	31
1.6.2 Epistemological Validation	32
1.6.3 Biological Validation	32
1.7 Conclusions	32
References	33

2 Design Principles for Microarray Investigations

<i>Kathleen F. Kerr</i>	39
2.1 Introduction	39
2.2 The "Pre-Planning" Stage	39
2.2.1 Goal 1: Unsupervised Learning	40
2.2.2 Goal 2: Supervised Learning	41
2.2.3 Goal 3: Class Comparison	41
2.3 Statistical Design Principles, Applied to Microarrays	42
2.3.1 Replication	42
2.3.2 Blocking	43
2.3.3 Randomization	46
2.4 Case Study	47
2.5 Conclusions	47
References	48

3 Pre-Processing DNA Microarray Data

<i>Benjamin M. Bolstad</i>	51
3.1 Introduction	51
3.1.1 Affymetrix GeneChips	53
3.1.2 Two-Color Microarrays	55
3.2 Basic Concepts	55
3.2.1 Pre-Processing Affymetrix GeneChip Data	56
3.2.2 Pre-Processing Two-Color Microarray Data	59
3.3 Advantages and Disadvantages	62
3.3.1 Affymetrix GeneChip Data	62
3.3.1.1 Advantages	62
3.3.1.2 Disadvantages	62
3.3.2 Two-Color Microarrays	62
3.3.2.1 Advantages	62
3.3.2.2 Disadvantages	63
3.4 Caveats and Pitfalls	63
3.5 Alternatives	63
3.5.1 Affymetrix GeneChip Data	63
3.5.2 Two-Color Microarrays	64
3.6 Case Study	64
3.6.1 Pre-Processing an Affymetrix GeneChip Data Set	64
3.6.2 Pre-Processing a Two-Channel Microarray Data Set	69
3.7 Lessons Learned	73
3.8 List of Tools and Resources	74
3.9 Conclusions	74
3.10 Mathematical Details	74
3.10.1 RMA Background Correction Equation	74
3.10.2 Quantile Normalization	75
3.10.3 RMA Model	75
3.10.4 Quality Assessment Statistics	75

3.10.5 Computation of M and A Values for Two-Channel Microarray Data	76
3.10.6 Print-Tip Loess Normalization	76
References	76
4 Pre-Processing Mass Spectrometry Data	
<i>Kevin R. Coombes, Keith A. Baggerly, and Jeffrey S. Morris</i>	79
4.1 Introduction	79
4.2 Basic Concepts	82
4.3 Advantages and Disadvantages	83
4.4 Caveats and Pitfalls	87
4.5 Alternatives	89
4.6 Case Study: Experimental and Simulated Data Sets for Comparing Pre-Processing Methods	92
4.7 Lessons Learned	98
4.8 List of Tools and Resources	98
4.9 Conclusions	99
References	99
5 Visualization in Genomics and Proteomics	
<i>Xiaochun Li and Jaroslaw Harezlak</i>	103
5.1 Introduction	103
5.2 Basic Concepts	105
5.2.1 Metric Scaling	107
5.2.2 Nonmetric Scaling	109
5.3 Advantages and Disadvantages	109
5.4 Caveats and Pitfalls	110
5.5 Alternatives	112
5.6 Case Study: MDS on Mass Spectrometry Data	113
5.7 Lessons Learned	118
5.8 List of Tools and Resources	119
5.9 Conclusions	120
References	121
6 Clustering – Class Discovery in the Post-Genomic Era	
<i>Joaquín Dopazo</i>	123
6.1 Introduction	123
6.2 Basic Concepts	126
6.2.1 Distance Metrics	126
6.2.2 Clustering Methods	127
6.2.2.1 Aggregative Hierarchical Clustering	128
6.2.2.2 <i>k</i> -Means	129
6.2.2.3 Self-Organizing Maps	130
6.2.2.4 Self-Organizing Tree Algorithm	130
6.2.2.5 Model-Based Clustering	130
6.2.3 Biclustering	131

6.2.4 Validation Methods	131
6.2.5 Functional Annotation	132
6.3 Advantages and Disadvantages	132
6.4 Caveats and Pitfalls	134
6.4.1 On Distances	135
6.4.2 On Clustering Methods	135
6.5 Alternatives	136
6.6 Case Study	137
6.7 Lessons Learned	139
6.8 List of Tools and Resources	140
6.8.1 General Resources	140
6.8.1.1 Multiple Purpose Tools (Including Clustering)	140
6.8.2 Clustering Tools	141
6.8.3 Biclustering Tools	141
6.8.4 Time Series	141
6.8.5 Public-Domain Statistical Packages and Other Tools	141
6.8.6 Functional Analysis Tools	142
6.9 Conclusions	142
References	143
7 Feature Selection and Dimensionality Reduction in Genomics and Proteomics	
<i>Milos Hauskrecht, Richard Pelikan, Michal Valko, and James Lyons-Weiler</i>	149
7.1 Introduction	149
7.2 Basic Concepts	151
7.2.1 Filter Methods	151
7.2.1.1 Criteria Based on Hypothesis Testing	151
7.2.1.2 Permutation Tests	152
7.2.1.3 Choosing Features Based on the Score	153
7.2.1.4 Feature Set Selection and Controlling False Positives ..	153
7.2.1.5 Correlation Filtering	154
7.2.2 Wrapper Methods	155
7.2.3 Embedded Methods	155
7.2.3.1 Regularization/Shrinkage Methods	155
7.2.3.2 Support Vector Machines	156
7.2.4 Feature Construction	156
7.2.4.1 Clustering	156
7.2.4.2 Clustering Algorithms	158
7.2.4.3 Probabilistic (Soft) Clustering	158
7.2.4.4 Clustering Features	158
7.2.4.5 Principal Component Analysis	159
7.2.4.6 Discriminative Projections	159
7.3 Advantages and Disadvantages	160
7.4 Case Study: Pancreatic Cancer	161

7.4.1 Data and Pre-Processing.....	161
7.4.2 Filter Methods	162
7.4.2.1 Basic Filter Methods	162
7.4.2.2 Controlling False Positive Selections	162
7.4.2.3 Correlation Filters	164
7.4.3 Wrapper Methods	165
7.4.4 Embedded Methods	166
7.4.5 Feature Construction Methods	167
7.4.6 Summary of Analysis Results and Recommendations	168
7.5 Conclusions.....	169
7.6 Mathematical Details	169
References	170
 8 Resampling Strategies for Model Assessment and Selection	
<i>Richard Simon</i>	173
8.1 Introduction	173
8.2 Basic Concepts.....	174
8.2.1 Resubstitution Estimate of Prediction Error.....	174
8.2.2 Split-Sample Estimate of Prediction Error	175
8.3 Resampling Methods.....	176
8.3.1 Leave-One-Out Cross-Validation	177
8.3.2 k -fold Cross-Validation	178
8.3.3 Monte Carlo Cross-Validation	178
8.3.4 Bootstrap Resampling	179
8.3.4.1 The .632 Bootstrap.....	179
8.3.4.2 The .632+ Bootstrap	180
8.4 Resampling for Model Selection and Optimizing Tuning Parameters	181
8.4.1 Estimating Statistical Significance of Classification Error Rates	183
8.4.2 Comparison to Classifiers Based on Standard Prognostic Variables	183
8.5 Comparison of Resampling Strategies	184
8.6 Tools and Resources	184
8.7 Conclusions.....	185
References	186
 9 Classification of Genomic and Proteomic Data Using Support Vector Machines	
<i>Peter Johansson and Markus Ringnér</i>	187
9.1 Introduction	187
9.2 Basic Concepts.....	187
9.2.1 Support Vector Machines	188
9.2.2 Feature Selection	190
9.2.3 Evaluating Predictive Performance	191
9.3 Advantages and Disadvantages	192
9.3.1 Advantages	192

9.3.2 Disadvantages	192
9.4 Caveats and Pitfalls	192
9.5 Alternatives	193
9.6 Case Study: Classification of Mass Spectral Serum Profiles Using Support Vector Machines	193
9.6.1 Data Set	193
9.6.2 Analysis Strategies	194
9.6.2.1 Strategy A: SVM without Feature Selection	196
9.6.2.2 Strategy B: SVM with Feature Selection	196
9.6.2.3 Strategy C: SVM Optimized Using Test Samples Performance	196
9.6.2.4 Strategy D: SVM with Feature Selection Using Test Samples	196
9.6.3 Results	196
9.7 Lessons Learned	197
9.8 List of Tools and Resources	197
9.9 Conclusions	198
9.10 Mathematical Details	198
References	200
10 Networks in Cell Biology	
<i>Carlos Rodríguez-Caso and Ricard V. Solé</i>	203
10.1 Introduction	203
10.1.1 Protein Networks	204
10.1.2 Metabolic Networks	205
10.1.3 Transcriptional Regulation Maps	205
10.1.4 Signal Transduction Pathways	206
10.2 Basic Concepts	206
10.2.1 Graph Definition	206
10.2.2 Node Attributes	207
10.2.3 Graph Attributes	208
10.3 Caveats and Pitfalls	212
10.4 Case Study: Topological Analysis of the Human Transcription Factor Interaction Network	213
10.5 Lessons Learned	218
10.6 List of Tools and Resources	219
10.7 Conclusions	220
10.8 Mathematical Details	220
References	221
11 Identifying Important Explanatory Variables for Time-Varying Outcomes	
<i>Oliver Bembom, Maya L. Petersen, and Mark J. van der Laan</i>	227
11.1 Introduction	227
11.2 Basic Concepts	229

11.3 Advantages and Disadvantages	233
11.3.1 Advantages	233
11.3.2 Disadvantages	234
11.4 Caveats and Pitfalls	235
11.5 Alternatives	237
11.6 Case Study: HIV Drug Resistance Mutations	239
11.7 Lessons Learned	245
11.8 List of Tools and Resources	246
11.9 Conclusions	247
References	248
12 Text Mining in Genomics and Proteomics	
<i>Robert Hoffmann</i>	251
12.1 Introduction	251
12.1.1 Text Mining	251
12.1.2 Interactive Literature Exploration	253
12.2 Basic Concepts	253
12.2.1 Information Retrieval	253
12.2.2 Entity Recognition	254
12.2.3 Information Extraction	254
12.2.4 Biomedical Text Resources	255
12.2.5 Assessment and Comparison of Text Mining Methods	256
12.3 Caveats and Pitfalls	256
12.3.1 Entity Recognition	256
12.3.2 Full Text	257
12.3.3 Distribution of Information	257
12.3.4 The Impossible	258
12.3.5 Overall Performance	258
12.4 Alternatives	259
12.4.1 Functional Coherence Analysis of Gene Groups	259
12.4.2 Co-Occurrence Networks	260
12.4.3 Superimposition of Experimental Data to the Literature Network	260
12.4.4 Gene Ontologies	261
12.5 Case Study	261
12.6 Lessons Learned	265
12.7 List of Tools and Resources	266
12.8 Conclusion	266
12.9 Mathematical Details	270
References	270
Index	275