
CONTENTS

<i>PREFACE</i>	<i>xiii</i>
----------------	-------------

<i>ACKNOWLEDGMENTS</i>	<i>xvii</i>
------------------------	-------------

1 INTRODUCTION	1
-----------------------	----------

1.1 Basic Terminology	2
1.1.1 The Central Dogma of Molecular Biology	2
1.1.2 Genome	3
1.1.3 Proteome	4
1.1.4 DNA (Deoxyribonucleic Acid)	5
1.1.5 RNA (Ribonucleic Acid)	6
1.1.6 mRNA (messenger RNA)	7
1.1.7 Genetic Code	7
1.1.8 Gene	9
1.1.9 Gene Expression and the Gene Expression Level	12
1.1.10 Protein	13
1.2 Overlapping Areas of Research	14
1.2.1 Genomics	14
1.2.2 Proteomics	14
1.2.3 Bioinformatics	14
1.2.4 Transcriptomics and Other -omics ...	14
1.2.5 Data Mining	15

2 BASIC ANALYSIS OF GENE EXPRESSION MICROARRAY DATA	17
--	-----------

2.1 Introduction	17
2.2 Microarray Technology	18
2.2.1 Spotted Microarrays	19
2.2.2 Affymetrix GeneChip® Microarrays	20
2.2.3 Bead-Based Microarrays	24
2.3 Low-Level Preprocessing of Affymetrix Microarrays	25
2.3.1 MAS5	27
2.3.2 RMA	31
2.3.3 GCRMA	33
2.3.4 PLIER	34
2.4 Public Repositories of Microarray Data	34
2.4.1 Microarray Gene Expression Data Society (MGED) Standards	34
2.4.2 Public Databases	37
2.4.2.1 Gene Expression Omnibus (GEO)	37
2.4.2.2 ArrayExpress	38

2.5	Gene Expression Matrix	38
2.5.1	Elements of Gene Expression Microarray Data Analysis	42
2.6	Additional Preprocessing, Quality Assessment, and Filtering	43
2.6.1	Quality Assessment	45
2.6.2	Filtering	50
2.7	Basic Exploratory Data Analysis	52
2.7.1	t Test	54
2.7.1.1	t Test for Equal Variances	55
2.7.1.2	t Test for Unequal Variances	55
2.7.2	ANOVA F Test	56
2.7.3	SAM t Statistic	57
2.7.4	Limma	59
2.7.5	Adjustment for Multiple Comparisons	59
2.7.5.1	Single-Step Bonferroni Procedure	61
2.7.5.2	Single-Step Sidak Procedure	61
2.7.5.3	Step-Down Holm Procedure	61
2.7.5.4	Step-Up Benjamini and Hochberg Procedure	62
2.7.5.5	Permutation Based Multiplicity Adjustment	63
2.8	Unsupervised Learning (<i>Taxonomy-Related Analysis</i>)	64
2.8.1	Cluster Analysis	65
2.8.1.1	Measures of Similarity or Distance	67
2.8.1.2	K -Means Clustering	70
2.8.1.3	Hierarchical Clustering	71
2.8.1.4	Two-Way Clustering and Related Methods	78
2.8.2	Principal Component Analysis	80
2.8.3	Self-Organizing Maps	85
	Exercises	90
3	BIOMARKER DISCOVERY AND CLASSIFICATION	95
3.1	Overview	95
3.1.1	Gene Expression Matrix ... Again	98
3.1.2	Biomarker Discovery	100
3.1.3	Classification Systems	105
3.1.3.1	Parametric and Nonparametric Learning Algorithms	106
3.1.3.2	Terms Associated with Common Assumptions Underlying Parametric Learning Algorithms	106
3.1.3.3	Visualization of Classification Results	110
3.1.4	Validation of the Classification Model	111
3.1.4.1	Reclassification	111
3.1.4.2	Leave-One-Out and K -Fold Cross-Validation	111
3.1.4.3	External and Internal Cross-Validation	112
3.1.4.4	Holdout Method of Validation	113
3.1.4.5	Ensemble-Based Validation (Using Out-of-Bag Samples)	113
3.1.4.6	Validation on an Independent Data Set	114
3.1.5	Reporting Validation Results	114
3.1.5.1	Binary Classifiers	115
3.1.5.2	Multiclass Classifiers	117
3.1.6	Identifying Biological Processes Underlying the Class Differentiation	119
3.2	Feature Selection	119
3.2.1	Introduction	119

3.2.2	Univariate Versus Multivariate Approaches	121
3.2.3	Supervised Versus Unsupervised Methods	123
3.2.4	Taxonomy of Feature Selection Methods	126
3.2.4.1	Filters, Wrappers, Hybrid, and Embedded Models	126
3.2.4.2	Strategy: Exhaustive, Complete, Sequential, Random, and Hybrid Searches	131
3.2.4.3	Subset Evaluation Criteria	133
3.2.4.4	Search-Stopping Criteria	133
3.2.5	Feature Selection for Multiclass Discrimination	133
3.2.6	Regularization and Feature Selection	134
3.2.7	Stability of Biomarkers	135
3.3	Discriminant Analysis	136
3.3.1	Introduction	136
3.3.2	Learning Algorithm	139
3.3.3	A Stepwise Hybrid Feature Selection with T^2	147
3.4	Support Vector Machines	149
3.4.1	Hard-Margin Support Vector Machines	150
3.4.2	Soft-Margin Support Vector Machines	157
3.4.3	Kernels	160
3.4.4	SVMs and Multiclass Discrimination	165
3.4.4.1	One-Versus-the-Rest Approach	165
3.4.4.2	Pairwise Approach	165
3.4.4.3	All-Classes-Simultaneously Approach	166
3.4.5	SVMs and Feature Selection: Recursive Feature Elimination	166
3.4.6	Summary	167
3.5	Random Forests	168
3.5.1	Introduction	168
3.5.2	Random Forests Learning Algorithm	172
3.5.3	Random Forests and Feature Selection	174
3.5.4	Summary	176
3.6	Ensemble Classifiers, Bootstrap Methods, and The <i>Modified Bagging</i> Schema	177
3.6.1	Ensemble Classifiers	177
3.6.1.1	Parallel Approach	177
3.6.1.2	Serial Approach	177
3.6.1.3	Ensemble Classifiers and Biomarker Discovery	177
3.6.2	Bootstrap Methods	178
3.6.3	Bootstrap and Linear Discriminant Analysis	179
3.6.4	The <i>Modified Bagging</i> Schema	180
3.7	Other Learning Algorithms	182
3.7.1	k -Nearest Neighbor Classifiers	183
3.7.2	Artificial Neural Networks	185
3.7.2.1	Perceptron	186
3.7.2.2	Multilayer Feedforward Neural Networks	187
3.7.2.3	Training the Network (Supervised Learning)	192
3.8	Eight Commandments of Gene Expression Analysis (for Biomarker Discovery)	197
	Exercises	198

4	THE INFORMATIVE SET OF GENES	201
4.1	Introduction	201
4.2	Definitions	202

4.3	The Method	202
4.3.1	Identification of the <i>Informative Set of Genes</i>	203
4.3.2	Primary Expression Patterns of the <i>Informative Set of Genes</i>	208
4.3.3	The Most Frequently Used Genes of the Primary Expression Patterns	211
4.4	Using the <i>Informative Set of Genes</i> to Identify Robust Multivariate Biomarkers	211
4.5	Summary	212
	Exercises	215
5	ANALYSIS OF PROTEIN EXPRESSION DATA	219
5.1	Introduction	219
5.2	Protein Chip Technology	222
5.2.1	Antibody Microarrays	223
5.2.2	Peptide Microarrays	225
5.2.3	Protein Microarrays	225
5.2.4	Reverse Phase Microarrays	226
5.3	Two-Dimensional Gel Electrophoresis	226
5.4	MALDI-TOF and SELDI-TOF Mass Spectrometry	228
5.4.1	MALDI-TOF Mass Spectrometry	229
5.4.2	SELDI-TOF Mass Spectrometry	230
5.5	Preprocessing of Mass Spectrometry Data	232
5.5.1	Introduction	232
5.5.2	Elements of Preprocessing of SELDI-TOF Mass Spectrometry Data	234
5.5.2.1	Quality Assessment	234
5.5.2.2	Calibration	235
5.5.2.3	Baseline Correction	235
5.5.2.4	Noise Reduction and Smoothing	235
5.5.2.5	Peak Detection	235
5.5.2.6	Intensity Normalization	236
5.5.2.7	Peak Alignment Across Spectra	237
5.6	Analysis of Protein Expression Data	237
5.6.1	Additional Preprocessing	239
5.6.2	Basic Exploratory Data Analysis	239
5.6.3	Unsupervised Learning	240
5.6.4	Supervised Learning—Feature Selection and Biomarker Discovery	242
5.6.5	Supervised Learning—Classification Systems	243
5.7	Associating Biomarker Peaks with Proteins	244
5.7.1	Introduction	244
5.7.2	The Universal Protein Resource (<i>UniProt</i>)	246
5.7.3	Search Programs	247
5.7.4	Tandem Mass Spectrometry	249
5.8	Summary	251
6	SKETCHES FOR SELECTED EXERCISES	253
6.1	Introduction	253
6.2	Multiclass Discrimination (Exercise 3.2)	254
6.2.1	Data Set Selection, Downloading, and Consolidation	254
6.2.2	Filtering Probe Sets	256
6.2.3	Designing a Multistage Classification Schema	257

6.3	Identifying the <i>Informative Set of Genes</i> (Exercises 4.2–4.6)	265
6.3.1	The <i>Informative Set of Genes</i>	266
6.3.2	Primary Expression Patterns of the Informative Set	267
6.3.3	The Most Frequently Used Genes of the Primary Expression Patterns	270
6.4	Using the <i>Informative Set of Genes</i> to Identify Robust Multivariate Markers (Exercise 4.8)	271
6.5	Validating Biomarkers on an Independent Test Data Set (Exercise 4.8)	272
6.6	Using a Training Set that Combines More than One Data Set (Exercises 3.5 and 4.1–4.8)	274
6.6.1	Combining the Two Data Sets into a Single Training Set	275
6.6.2	Filtering Probe Sets of the Combined Data	276
6.6.3	Assessing the Discriminatory Power of the Biomarkers and Their Generalization	276
6.6.4	Identifying the <i>Informative Set of Genes</i>	276
6.6.5	Primary Expression Patterns of the <i>Informative Set of Genes</i>	280
6.6.6	The Most Frequently Used Genes of the Primary Expression Patterns	281
6.6.7	Using the <i>Informative Set of Genes</i> to Identify Robust Multivariate Markers	285
6.6.8	Validating Biomarkers on an Independent Test Data Set	287

REFERENCES	289
------------	-----

INDEX	307
-------	-----