CONTENTS

PREFACE xi

ACKNOWLEDGMENTS xvii

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
CONTENTS

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 (Taxonomy-Related Analysis) 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
4.3 The Method 202
  4.3.1 Identification of the Informative Set of Genes 203
  4.3.2 Primary Expression Patterns of the Informative Set of Genes 208
  4.3.3 The Most Frequently Used Genes of the Primary Expression Patterns 211
4.4 Using the Informative Set of Genes 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 (UniProt) 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 *Informative Set of Genes* (Exercises 4.2–4.6) 265  
6.3.1 The *Informative Set of Genes* 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 *Informative Set of Genes* 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 *Informative Set of Genes* 276  
6.6.5 Primary Expression Patterns of the *Informative Set of Genes* 280  
6.6.6 The Most Frequently Used Genes of the Primary Expression Patterns 281  
6.6.7 Using the *Informative Set of Genes* to Identify Robust Multivariate Markers 285  
6.6.8 Validating Biomarkers on an Independent Test Data Set 287

*REFERENCES* 289

*INDEX* 307