

Contents

Preface	xv
1 Microarrays in Gene Expression Studies	1
1.1 Introduction	1
1.2 Background Biology	2
1.2.1 Genome, Genotype, and Gene Expression	2
1.2.2 Of Wild-Types and Other Alleles	3
1.2.3 Aspects of Underlying Biology and Physiochemistry	4
1.3 Polymerase Chain Reaction	5
1.4 cDNA	6
1.4.1 Expressed Sequence Tag	6
1.5 Microarray Technology and Application	7
1.5.1 History of Microarray Development	8
1.5.2 Tools of Microarray Technology	10
1.5.3 Limitations of Microarray Technology	18
1.5.4 Oligonucleotides versus cDNA Arrays	20
1.5.5 SAGE: Another Method for Detecting and Measuring Gene Expression Levels	23
1.5.6 Emerging Technologies	24
1.6 Sampling of Relevant Research Entities and Public Resources	24

2	Cleaning and Normalization	31
2.1	Introduction	31
2.2	Cleaning Procedures	32
2.2.1	Image Processing to Extract Information	32
2.2.2	Missing Value Estimation	36
2.2.3	Sources of Nonlinearity	38
2.3	Normalization and Plotting Procedures for Oligonucleotide Arrays	38
2.3.1	Global Approaches for Oligonucleotide Array Data	38
2.3.2	Spiked Standard Approaches	39
2.3.3	Geometric Mean and Linear Regression Normalization for Multiple Arrays	41
2.3.4	Nonlinear Normalization for Multiple Arrays Using Smooth Curves	42
2.4	Normalization Methods for cDNA Microarray Data	44
2.4.1	Single-Array Normalization	46
2.4.2	Multiple Slides Normalization	48
2.4.3	ANOVA and Related Methods for Normalization	49
2.4.4	Mixed-Model Method for Normalization	50
2.4.5	SNOMAD	51
2.5	Transformations and Replication	52
2.5.1	Importance of Replication	52
2.5.2	Transformations	53
2.6	Analysis of the Alon Data Set	56
2.7	Comparison of Normalization Strategies and Discussion	56
3	Some Cluster Analysis Methods	61
3.1	Introduction	61
3.2	Reduction in the Dimension of the Feature Space	62
3.3	Cluster Analysis	63
3.4	Some Hierarchical Agglomerative Techniques	64
3.5	k -Means Clustering	68
3.6	Cluster Analysis with No <i>A Priori</i> Metric	69
3.7	Clustering via Finite Mixture Models	69
3.7.1	Definition	69
3.7.2	Advantages of Model-Based Clustering	71
3.8	Fitting Mixture Models Via the EM Algorithm	72
3.8.1	E-Step	73
3.8.2	M-Step	74

4.5	The EMMIX-GENE Clustering Procedure	103
4.6	Step 1: Screening of Genes	104
4.7	Step 2: Clustering of Genes: Formation of Metagenes	105
4.8	Step 3: Clustering of Tissues	107
4.9	EMMIX-GENE Software	108
4.10	Example: Clustering of Alon Data	108
4.10.1	Clustering on Basis of 446 Genes	108
4.10.2	Clustering on Basis of Gene Groups	109
4.10.3	Clustering on Basis of Metagenes	112
4.11	Example: Clustering of van 't Veer Data	112
4.11.1	Screening and Clustering of Genes	113
4.11.2	Usefulness of the Selected Genes	115
4.11.3	Clustering of Tissues	121
4.11.4	Use of Underlying Signatures with Clinical Data	123
4.12	Choosing the Number of Clusters in Microarray Data	124
4.12.1	Some Previous Attempts	124
4.13	Likelihood Ratio Test Applied to Microarray Data	125
4.13.1	Golub Data	125
4.13.2	Alizadeh Data	126
4.13.3	Bittner Data	127
4.13.4	van 't Veer Data	127
4.14	Effect of Selection Bias on the Number of Clusters	128
4.15	Clustering on Microarray and Clinical Data	128
4.16	Discussion	130
5	Screening and Clustering of Genes	133
5.1	Detection of Differentially Expressed Genes	133
5.1.1	Introduction	133
5.1.2	Fold Change	134
5.1.3	Multiplicity Problem	134
5.1.4	Overview of Literature	135
5.2	Test of a Single Hypothesis	137
5.3	Gene Statistics	138
5.3.1	Calculation of Interactions via ANOVA Models	138
5.3.2	Two-Sample t -Statistics	139
5.4	Multiple Hypothesis Testing	139
5.4.1	Outcomes with Multiple Hypotheses	140
5.4.2	Controlling the FWER	140

5.4.3	False Discovery Rate (FDR)	141
5.4.4	Benjamini-Hochberg Procedure	142
5.4.5	False Nondiscovery Rate (FNR)	143
5.4.6	Positive FDR	143
5.4.7	Positive FNR	143
5.4.8	Linking False Rates with Posterior Probabilities	143
5.5	Null Distribution of Test Statistic	144
5.5.1	Permutation Method	144
5.5.2	Null Replications of the Test Statistic	145
5.5.3	The SAM Method	146
5.5.4	Application of SAM Method to Alon Data	146
5.6	Recent Approaches for Strong Control of the FDR	148
5.6.1	The q -Value	148
5.6.2	Technical Definition of q -Value	149
5.6.3	Controlling FDR Strongly	150
5.6.4	Selecting Genes via the q -Value	151
5.6.5	Application to Hedenfalk Data	152
5.7	Two-Component Mixture Model Framework	154
5.7.1	Definition of Model	154
5.7.2	Bayes Rule	155
5.7.3	Estimated FDR	155
5.7.4	Bayes Risk in terms of Estimated FDR and FNR	156
5.8	Nonparametric Empirical Bayes Approach	158
5.8.1	Method of Efron et al. (2001)	158
5.8.2	Mixture Model Method (MMM)	158
5.8.3	Nonparametric Bayesian Approach	159
5.8.4	Application of Empirical Bayes Methods to Alon Data	159
5.9	Parametric Mixture Models for Differential Gene Expression	160
5.9.1	Parametric Empirical Bayes Methods	160
5.9.2	Finding Clusters of Differentially Expressed Genes	164
5.9.3	Example: Fitting Normal Mixtures to t -Statistic Values	165
5.10	Use of the P -Value as a Summary Statistic	166
5.10.1	Beta Mixture for Distribution of P -Values	168
5.10.2	Example: Fitting Beta Mixtures to P -Values	169
5.11	Clustering of Genes	171
5.12	Finding Correlated Genes	173

5.13	Clustering of Genes via Full Expression Profiles	173
5.14	Clustering of Genes via PCA of Expression Profiles	174
5.15	Clustering of Genes with Repeated Measurements	175
5.15.1	A Mixture Model for Technical Replicates	175
5.15.2	Application of EM Algorithm	176
5.15.3	M-Step	176
5.16	Gene Shaving	177
5.16.1	Introduction	177
5.16.2	Methodology and implementation	177
5.16.3	Optimal cluster size via the Gap statistic	178
5.16.4	Supervised Gene Shaving	179
5.16.5	Real Data Example	179
5.16.6	Computer Software	180
6	Discriminant Analysis	185
6.1	Introduction	185
6.2	Basic Notation	185
6.3	Error Rates	187
6.4	Decision-Theoretic Approach	187
6.5	Training Data	189
6.6	Different Types of Error Rates	190
6.7	Sample-Based Discriminant Rules	191
6.8	Parametric Discriminant Rules	192
6.9	Discrimination via Normal Models	193
6.9.1	Heteroscedastic Normal Model	193
6.9.2	Plug-in Sample NQDR	194
6.9.3	Homoscedastic Normal Model	195
6.9.4	Optimal Error Rates	197
6.9.5	Plug-in Sample NLDR	197
6.9.6	Normal Mixture Model	198
6.10	Fisher's Linear Discriminant Function	199
6.10.1	Separation Approach	199
6.10.2	Regression Approach	199
6.11	Logistic Discrimination	201
6.12	Nearest-Centroid Rule	202
6.13	Support Vector Machines	203
6.13.1	Two Classes	203
6.13.2	Selection of Feature Variables	204

6.13.3	Multiple Classes	205
6.13.4	Computer Software	206
6.14	Variants of Support Vector Machines	207
6.15	Neural Networks	207
6.16	Nearest-Neighbor Rules	208
6.16.1	Introduction	208
6.16.2	Definition of a k -NN Rule	209
6.17	Classification Trees	210
6.18	Error-Rate Estimation	211
6.18.1	Apparent Error Rate	211
6.18.2	Bias Correction of the Apparent Error Rate	213
6.19	Cross-Validation	213
6.19.1	Leave-One-Out(LOO) Estimator	213
6.19.2	q -Fold Cross-Validation	214
6.20	Error-Rate Estimation via the Bootstrap	214
6.20.1	The 0.632 Estimator	214
6.20.2	Mean Squared Error of the Estimated Error Rate	215
6.21	Selection of Feature Variables	216
6.22	Error-Rate Estimation with Selection Bias	218
6.22.1	Selection Bias	218
6.22.2	External Cross-Validation	218
6.22.3	The 0.632+ Estimator	219
7	Supervised Classification of Tissue Samples	221
7.1	Introduction	221
7.2	Reducing the Dimension of the Feature Space of Genes	222
7.2.1	Principal Components	223
7.2.2	Partial Least Squares	223
7.2.3	Ranking of Genes	223
7.2.4	Grouping of Genes	224
7.3	SVM with Recursive Feature Elimination (RFE)	224
7.4	Selection Bias: SVM with RFE	226
7.5	Selection Bias: Fisher's Rule with Forward Selection	228
7.6	Selection Bias: Noninformative Data	230
7.7	Discussion of Selection Bias	232
7.8	Selection of Marker Genes with SVM	233
7.8.1	Description of van de Vijver Breast Cancer Data	233
7.8.2	Application of SVM with RFE	234