

Contents

Preface

xv

1 A Brief Introduction

1

- 1.1 A Note on Exploratory Data Analysis, 3
- 1.2 Computing Considerations and Software, 4
- 1.3 A Brief Outline of the Book, 5
- 1.4 Data Sets and Case Studies, 6
 - 1.4.1 The Golub Data, 6
 - 1.4.2 The Mouse5 Data, 7
 - 1.4.3 The Khan Data, 7
 - 1.4.4 The Sialin Data, 7
 - 1.4.5 The Behavioral Study Data, 8
 - 1.4.6 The Spiked-In Data, 8
 - 1.4.7 The APOAI Study, 8
 - 1.4.8 The Breast Cancer Data, 8
 - 1.4.9 Platinum Spike Data Set, 9
 - 1.4.10 Human Epidermal Squamous Carcinoma Cell Line A431 Experiment, 9
 - 1.4.11 Note: Public Repositories of Microarray Data, 9

2 Genomics Basics

10

- 2.1 Genes, 10
- 2.2 Deoxyribonucleic Acid, 11
- 2.3 Gene Expression, 12
- 2.4 Hybridization Assays and Other Laboratory Techniques, 14

2.5	The Human Genome, 16	
2.6	Genome Variations and Their Consequences, 17	
2.7	Genomics, 19	
2.8	The Role of Genomics in Pharmaceutical Research and Clinical Practice, 19	
2.9	Proteins, 22	
2.10	Bioinformatics, 23	
	Supplementary Reading, 24	
3	Microarrays	25
3.1	Types of Microarray Experiments, 26	
3.1.1	Experiment Type 1: Tissue-Specific Gene Expression, 26	
3.1.2	Experiment Type 2: Developmental Genetics, 26	
3.1.3	Experiment Type 3: Genetic Diseases, 27	
3.1.4	Experiment Type 4: Complex Diseases, 28	
3.1.5	Experiment Type 5: Pharmacological Agents, 28	
3.1.6	Experiment Type 6: Plant Breeding, 29	
3.1.7	Experiment Type 7: Environmental Monitoring, 29	
3.2	A Very Simple Hypothetical Microarray Experiment, 30	
3.3	A Typical Microarray Experiment, 31	
3.3.1	Microarray Preparation, 32	
3.3.2	Sample Preparation, 33	
3.3.3	The Hybridization Step, 33	
3.3.4	Scanning the Microarray, 34	
3.3.5	Interpreting the Scanned Image, 34	
3.4	Multichannel cDNA Microarrays, 35	
3.5	Oligonucleotide Microarrays, 36	
3.6	Bead-Based Arrays, 37	
3.7	Confirmation of Microarray Results, 37	
	Supplementary Reading and Electronic References, 37	
4	Processing the Scanned Image	39
4.1	Converting the Scanned Image to the Spotted Image, 39	
4.1.1	Gridding, 40	
4.1.2	Segmentation, 40	
4.1.3	Quantification, 41	
4.2	Quality Assessment, 42	
4.2.1	Visualizing the Spotted Image, 43	

- 4.2.2 Numerical Evaluation of Array Quality, 44
- 4.2.3 Spatial Problems, 45
- 4.2.4 Spatial Randomness, 46
- 4.2.5 Quality Control of Arrays, 47
- 4.2.6 Assessment of Spot Quality, 48
- 4.3 Adjusting for Background, 49
 - 4.3.1 Estimating the Background, 49
 - 4.3.2 Adjusting for the Estimated Background, 52
- 4.4 Expression-Level Calculation for Two-Channel cDNA Microarrays, 53
- 4.5 Expression-Level Calculation for Oligonucleotide Microarrays, 53
 - 4.5.1 The Average Difference, 54
 - 4.5.2 A Weighted Average Difference, 54
 - 4.5.3 Perfect Matches Only, 55
 - 4.5.4 Background Adjustment Approach, 56
 - 4.5.5 Model-Based Approach, 56
 - 4.5.6 Absent-Present Calls, 57
- Supplementary Reading, 58
- Software Notes, 58

5 Preprocessing Microarray Data

60

- 5.1 Logarithmic Transformation, 60
- 5.2 Variance Stabilizing Transformations, 62
- 5.3 Sources of Bias, 63
- 5.4 Normalization, 63
- 5.5 Intensity-Dependent Normalization, 65
 - 5.5.1 Smooth Function Normalization, 68
 - 5.5.2 Quantile Normalization, 68
 - 5.5.3 Stagewise Normalization, 71
 - 5.5.4 Normalization of Two-Channel Arrays, 71
 - 5.5.5 Spatial Normalization, 76
 - 5.5.6 Normalization of Oligonucleotide Arrays, 76
- 5.6 Judging the Success of a Normalization, 77
- 5.7 Outlier Identification, 80
- 5.8 Nonresistant Rules for Outlier Identification, 80
- 5.9 Resistant Rules for Outlier Identification, 81
- 5.10 Assessing Replicate Array Quality, 83
- Software Notes, 83

6	Summarization	86
6.1	Replication, 86	
6.2	Technical Replicates, 87	
6.3	Biological Replicates, 91	
6.4	Experiments with Both Technical and Biological Replicates, 91	
6.5	Multiple Oligonucleotide Arrays, 95	
6.6	Estimating Fold Change in Two-Channel Experiments, 96	
6.7	Bayes Estimation of Fold Change, 97	
6.8	Estimating Fold Change Affymetrix Data, 98	
6.9	RMA Summarization of Multiple Oligonucleotide Arrays Revisited, 100	
6.10	Factor Analysis for Robust Microarray Summarization, FARMS, 105 Software Notes, 106	
7	Two-Group Comparative Experiments	109
7.1	Basics of Statistical Hypothesis Testing, 111	
7.2	Fold Changes, 112	
7.3	The Two-Sample t -Test, 114	
7.4	Diagnostic Checks, 116	
7.5	Robust t -Tests, 118	
7.6	The Mann-Whitney-Wilcoxon Rank Sum Test, 119	
7.7	Multiplicity Adjustment: The Familywise Error Rate, 121	
7.7.1	A Pragmatic Approach to the Issue of Multiplicity, 122	
7.7.2	Simple Multiplicity Adjustments, 123	
7.7.3	Sequential Multiplicity Adjustments, 123	
7.8	Multiplicity Adjustment: The False Discovery Rate, 125	
7.8.1	Benjamini and Hochberg Procedure, 125	
7.8.2	The Positive False Discovery Rate, 129	
7.9	Resampling-Based Multiple Testing Procedures, 130	
7.10	Small-Variance-Adjusted t -Tests and SAM, 132	
7.10.1	Modifying the t -Statistic, 135	
7.10.2	Assessing Significance with the SAM t Statistic, 135	
7.10.3	Strategies for Using SAM, 138	
7.10.4	An Empirical Bayes Framework, 139	
7.10.5	Understanding the SAM Adjustment, 139	
7.11	Conditional t , 141	
7.12	Borrowing Strength Across Genes, 145	
7.12.1	Simple Methods, 147	

7.12.2	A Bayesian Model, 148	
7.13	Two-Channel Experiments, 149	
7.13.1	The Paired Sample t Test and SAM, 150	
7.13.2	Borrowing Strength Via Hierarchical Modeling, 150	
7.14	Filtering, 152	
7.14.1	Filtering Based on Summarized Data, 152	
7.14.2	Filtering Based on Probe-Level Data, 156	
	Supplementary Reading, 160	
	Software Notes, 161	
8	Model-Based Inference and Experimental Design Considerations	166
8.1	The F -Test, 167	
8.2	The Basic Linear Model, 168	
8.3	Fitting the Model in Two Stages, 171	
8.4	Multichannel Experiments, 171	
8.5	Experimental Design Considerations, 172	
8.5.1	Comparing Two Varieties with Two-Channel Microarrays, 172	
8.5.2	Comparing Multiple Varieties with Two-Channel Microarrays, 174	
8.5.3	Single-Channel Microarray Experiments, 176	
8.6	Miscellaneous Issues, 177	
8.7	Model-Based Analysis of Affymetrix Arrays, 177	
8.7.1	One-Way ANOVA, 177	
8.7.2	Linear Models for Microarray Data (Limma), 179	
8.7.3	A Joint Model for Gene Expression and Response, 184	
8.7.4	Analysis of Dose-Response Microarray Experiments, 187	
8.7.5	Analysis of Time Course Data, 191	
	Supplementary Reading, 195	
	Software Notes, 196	
9	Analysis of Gene Sets	200
9.1	Methods for Identifying Enriched Gene Sets, 201	
9.1.1	MLP and Fisher's Test, 202	
9.1.2	GSEA and the Kolmogorov-Smirnov Test, 205	
9.2	ORA and Fisher's Exact Test, 205	
9.3	Interpretation of Results, 206	
9.4	Example, 206	
	Software Notes, 206	

10	Pattern Discovery	210
10.1	Initial Considerations, 210	
10.2	Cluster Analysis, 212	
10.2.1	Dissimilarity Measures and Similarity Measures, 213	
10.2.2	Guilt by Association, 215	
10.2.3	Hierarchical Clustering, 216	
10.2.4	Partitioning Methods, 222	
10.2.5	Model-Based Clustering, 226	
10.2.6	Chinese Restaurant Clustering, 227	
10.2.7	Ensemble Methods for Clustering Samples, 228	
10.2.8	Discussion, 229	
10.3	Seeking Patterns Visually, 229	
10.3.1	Principal Components Analysis, 230	
10.3.2	Factor Analysis, 234	
10.3.3	Biplots, 237	
10.3.4	Spectral Map Analysis, 237	
10.3.5	Multidimensional Scaling, 239	
10.3.6	Projection Pursuit, 239	
10.3.7	Data Visualization with the Grand Tour and Projection Pursuit, 241	
10.4	Biclustering, 242	
10.4.1	Block Clustering, 243	
10.4.2	Gene Shaving, 244	
10.4.3	The Plaid Model, 244	
	Supplementary Reading, 247	
	Software Notes, 247	
11	Class Prediction	250
11.1	Initial Considerations, 251	
11.1.1	Misclassification Rates, 251	
11.1.2	Reducing the Number of Classifiers, 252	
11.2	Linear Discriminant Analysis, 258	
11.3	Extensions of Fisher's LDA, 261	
11.4	Penalized Methods, 263	
11.5	Nearest Neighbors, 264	
11.6	Recursive Partitioning, 265	
11.6.1	Classification Trees, 267	
11.6.2	Activity Region Finding, 269	

11.7	Ensemble Methods, 270	
11.7.1	Random Forest, 271	
11.7.2	Enriched Random Forest, 272	
11.8	Enriched Ensemble Classifiers, 273	
11.9	Neural Networks, 273	
11.10	Support Vector Machines, 276	
11.11	Generalized Enriched Methods, 277	
11.11.1	Enriched Principal Components Analysis and Biplots, 279	
11.11.2	Enriched Penalized Methods: Lasso, SVM, P-SVM, 281	
11.11.3	Enriched Partial Least Squares (PLS), 283	
11.12	Integration of Genome Information, 285	
11.12.1	Integration of Gene Expression Data and Molecular Structure Data, 285	
11.12.2	Pathway Inference, 285	
	Supplementary Reading, 286	
	Software Notes, 286	
12	Protein Arrays	290
12.1	Introduction, 290	
12.2	Protein Array Experiments, 291	
12.3	Special Issues with Protein Arrays, 292	
12.4	Analysis, 293	
12.5	Using Antibody Antigen Arrays to Measure Protein Concentrations, 294	
	References	298
	Index	313