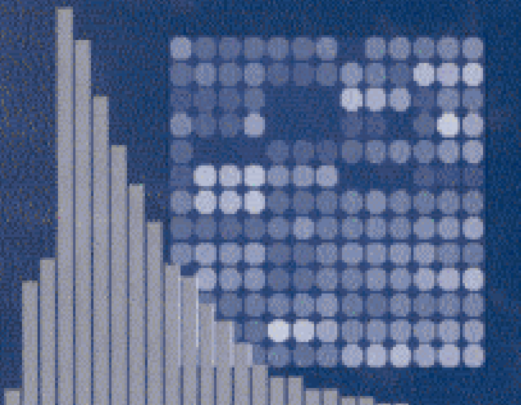
WILEY

EXPLORATION AND ANALYSIS OF DNA MICROARRAY AND PROTEIN ARRAY DATA

DHAMMIKA AMARATUNGA
JAVIER CABRERA

Wiley Series in Probability and Statistics



Contents

Preface **xiii**

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

2 Genomics Basics **8**

2.1 Genes, 8

2.2 DNA, 9

2.3 Gene Expression, 10

2.4 Hybridization Assays and Other Laboratory Techniques, 12

2.5 The Human Genome, 14

2.6 Genome Variations and Their Consequences, 15

2.7 Genomics, 17

2.8 The Role of Genomics in Pharmaceutical Research, 18

2.9 Proteins, 20

2.10 Bioinformatics, 21

Supplementary Reading, 22

Exercises, 22

3 Microarrays **23**

3.1 Types of Microarray Experiments, 24

3.1.1 Experiment Type 1: Tissue-Specific Gene Expression, 24

3.1.2 Experiment Type 2: Developmental Genetics, 24

3.1.3 Experiment Type 3: Genetic Diseases, 25
3.1.4 Experiment Type 4: Complex Diseases, 26
3.1.5 Experiment Type 5: Pharmacological Agents, 26
3.1.6 Experiment Type 6: Plant Breeding, 27
3.1.7 Experiment Type 7: Environmental Monitoring, 27
3.2 A Very Simple Hypothetical Microarray Experiment, 28
3.3 A Typical Microarray Experiment, 30
3.3.1 Microarray Preparation, 30
3.3.2 Sample Preparation, 32
3.3.3 The Hybridization Step, 32
3.3.4 Scanning the Microarray, 33
3.3.5 Interpreting the Scanned Image, 33
3.4 Multichannel cDNA Microarrays, 34
3.5 Oligonucleotide Arrays, 35
3.6 Bead-Based Arrays, 36
3.7 Confirmation of Microarray Results, 37
Supplementary Reading and Electronic References, 37
Exercises, 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, 50
4.3.2 Adjusting for the Estimated Background, 53
4.4 Expression Level Calculation for Two-Channel cDNA Microarrays, 53
4.5 Expression Level Calculation for Oligonucleotide Arrays, 54

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, 56
Supplementary Reading, 58
Exercises, 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 Normalization of Oligonucleotide Arrays, 70
5.5.4 Normalization of Two-Channel Arrays, 70
5.5.5 Spatial Normalization, 71
5.5.6 Stagewise Normalization, 72
5.6 Judging the Success of a Normalization, 73
5.7 Outlier Identification, 77
5.7.1 Nonresistant Rules for Outlier Identification, 77
5.7.2 Resistant Rules for Outlier Identification, 78
5.8 Assessing Replicate Array Quality, 79
Exercises, 80

6 Summarization 82

6.1 Replication, 82
6.2 Technical Replicates, 83
6.3 Biological Replicates, 86
6.4 Experiments with Both Technical and Biological Replicates, 87
6.5 Multiple Oligonucleotide Arrays, 90
6.6 Estimating Fold Change in Two-Channel Experiments, 92
6.7 Bayes Estimation of Fold Change, 93
Exercises, 94

7 Two-Group Comparative Experiments 95

7.1 Basics of Statistical Hypothesis Testing, 96
7.2 Fold Changes, 99
7.3 The Two-Sample *t* Test, 100
7.4 Diagnostic Checks, 103
7.5 Robust *t* Tests, 104
7.6 Randomization Tests, 105
7.7 The Mann–Whitney–Wilcoxon Rank Sum Test, 108
7.8 Multiplicity, 109
7.8.1 A Pragmatic Approach to the Issue of Multiplicity, 109
7.8.2 Simple Multiplicity Adjustments, 110
7.8.3 Sequential Multiplicity Adjustments, 111
7.9 The False Discovery Rate, 113
7.9.1 The Positive False Discovery Rate, 114
7.10 Small Variance-Adjusted *t* Tests and SAM, 115
7.10.1 Modifying the *t* Statistic, 117
7.10.2 Assesing Significance with the SAM *t* Statistic, 117
7.10.3 Strategies for Using SAM, 120
7.10.4 An Empirical Bayes Framework, 120
7.10.5 Understanding the SAM Adjustment, 121
7.11 Conditional *t*, 123
7.12 Borrowing Strength across Genes, 126
7.12.1 Simple Methods, 127
7.12.2 A Bayesian Model, 129
7.13 Two-Channel Experiments, 130
7.13.1 The Paired Sample *t* Test and SAM, 131
7.13.2 Borrowing Strength via Hierarchical Modeling, 131
Supplementary Reading, 133
Exercises, 133

8 Model-Based Inference and Experimental Design Considerations 135

8.1 The *F* Test, 136
8.2 The Basic Linear Model, 138
8.3 Fitting the Model in Two Stages, 140
8.4 Multichannel Experiments, 141
8.5 Experimental Design Considerations, 141

8.5.1 Comparing Two Varieties with Two-Channel Microarrays, 141
8.5.2 Comparing Multiple Varieties with Two-Channel Microarrays, 143
8.5.3 Single-Channel Microarray Experiments, 145
8.6 Miscellaneous Issues, 146
Supplementary Reading, 147
Exercises, 147

9 Pattern Discovery **149**

9.1 Initial Considerations, 149
9.2 Cluster Analysis, 151
9.2.1 Dissimilarity Measures and Similarity Measures, 152
9.2.2 Guilt by Association, 155
9.2.3 Hierarchical Clustering, 155
9.2.4 Partitioning Methods, 160
9.2.5 Model-Based Clustering, 166
9.2.6 Chinese Restaurant Clustering, 167
9.2.7 Discussion, 168
9.3 Seeking Patterns Visually, 168
9.3.1 Principal Components Analysis, 169
9.3.2 Factor Analysis, 174
9.3.3 Biplots, 176
9.3.4 Spectral Map Analysis, 177
9.3.5 Multidimensional Scaling, 179
9.3.6 Projection Pursuit, 179
9.3.7 Data Visualization with the Grand Tour and Projection Pursuit, 181
9.4 Two-Way Clustering, 182
9.4.1 Block Clustering, 182
9.4.2 Gene Shaving, 182
9.4.3 The Plaid Model, 183
Software Notes, 184
Supplementary Reading, 184
Exercises, 185

10 Class Prediction **186**

10.1 Initial Considerations, 187

10.1.1 Misclassification Rates, 188

10.1.2 Reducing the Number of Classifiers, 189

10.2 Linear Discriminant Analysis, 193

10.3 Extensions of Fisher's LDA, 197

10.4 Nearest Neighbors, 200

10.5 Recursive Partitioning, 201

10.5.1 Classification Trees, 201

10.5.2 Activity Region Finding, 206

10.6 Neural Networks, 206

10.7 Support Vector Machines, 208

10.8 Integration of Genomic Information, 210

10.8.1 Integration of Gene Expression Data and Molecular Structure Data, 210

10.8.2 Pathway Inference, 211

Software Notes, 211

Supplementary Reading, 212

Exercises, 212

11 Protein Arrays **214**

11.1 Introduction, 214

11.2 Protein Array Experiments, 215

11.3 Special Issues with Protein Arrays, 216

11.4 Analysis, 217

11.5 Using Antibody Antigen Arrays to Measure Protein Concentrations, 218

Exercises, 221

References **222**

Author Index **237**

Subject Index **241**