
Contents

Acknowledgments	v
1 Introduction	1
2 DNA Microarray Technology	5
2.1 Overview	5
2.2 Measuring Label Intensity	5
2.3 Labeling Methods	6
2.4 Printed Microarrays	7
2.5 Affymetrix GeneChip TM Arrays	9
2.6 Other Microarray Platforms	10
3 Design of DNA Microarray Experiments	11
3.1 Introduction	11
3.2 Study Objectives	12
3.2.1 Class Comparison	12
3.2.2 Class Prediction	13
3.2.3 Class Discovery	13
3.2.4 Pathway Analysis	13
3.3 Comparing Two RNA Samples	13
3.4 Sources of Variation and Levels of Replication	14
3.5 Pooling of Samples	16
3.6 Pairing Samples on Dual-Label Microarrays	17
3.6.1 The Reference Design	17
3.6.2 The Balanced Block Design	19
3.6.3 The Loop Design	20
3.7 Reverse Labeling (Dye Swap)	21
3.8 Number of Biological Replicates Needed	23

4	Image Analysis	29
4.1	Image Generation	29
4.2	Image Analysis for cDNA Microarrays	30
4.2.1	Image Display	30
4.2.2	Gridding	30
4.2.3	Segmentation	31
4.2.4	Foreground Intensity Extraction.....	32
4.2.5	Background Correction.....	33
4.2.6	Image Output File.....	34
4.3	Image Analysis for Affymetrix GeneChip TM	35
5	Quality Control	39
5.1	Introduction	39
5.2	Probe-Level Quality Control for Two-Color Arrays.....	40
5.2.1	Visual Inspection of the Image File	40
5.2.2	Spots Flagged at Image Analysis	40
5.2.3	Spot Size	41
5.2.4	Weak Signal	42
5.2.5	Large Relative Background Intensity.....	43
5.3	Gene Level Quality Control for Two-Color Arrays	44
5.3.1	Poor Hybridization and Printing	45
5.3.2	Probe Quality Control Based on Duplicate Spots	45
5.3.3	Low Variance Genes	46
5.4	Array-Level Quality Control for Two-Color Arrays.....	47
5.5	Quality Control for GeneChip TM Arrays.....	48
5.6	Data Imputation	50
6	Array Normalization	53
6.1	Introduction	53
6.2	Choice of Genes for Normalization	53
6.2.1	Biologically Defined Housekeeping Genes	53
6.2.2	Spiked Controls	54
6.2.3	Normalize Using All Genes	55
6.2.4	Identification of Housekeeping Genes Based on Observed Data	55
6.3	Normalization Methods for Two-Color Arrays	55
6.3.1	Linear or Global Normalization	56
6.3.2	Intensity-Based Normalization	57
6.3.3	Location-Based Normalization	59
6.3.4	Combination Location and Intensity Normalization	61
6.4	Normalization of GeneChip TM Arrays.....	61
6.4.1	Linear or Global Normalization	61
6.4.2	Intensity-Based Normalization	62

- 7 Class Comparison** 65
 - 7.1 Introduction 65
 - 7.2 Examining Whether a Single Gene is Differentially Expressed
Between Classes 66
 - 7.2.1 *t*-Test 67
 - 7.2.2 Permutation Tests 68
 - 7.2.3 More Than Two Classes 71
 - 7.2.4 Paired-Specimen Data 73
 - 7.3 Identifying Which Genes Are Differentially Expressed
Between Classes 75
 - 7.3.1 Controlling for No False Positives 76
 - 7.3.2 Controlling the Number of False Positives 80
 - 7.3.3 Controlling the False Discovery Proportion 81
 - 7.4 Experiments with Very Few Specimens from Each Class 84
 - 7.5 Global Tests of Gene Expression Differences Between Classes . 86
 - 7.6 Experiments with a Single Specimen from Each Class 88
 - 7.7 Regression Model Analysis; Generalizations of Class
Comparison 90
 - 7.8 Evaluating Associations of Gene Expression to Survival 91
 - 7.9 Models for Nonreference Designs on Dual-Label Arrays 92

- 8 Class Prediction** 95
 - 8.1 Introduction 95
 - 8.2 Feature Selection 97
 - 8.3 Class Prediction Methods 98
 - 8.3.1 Nomenclature 98
 - 8.3.2 Discriminant Analysis 98
 - 8.3.3 Variants of Diagonal Linear Discriminant Analysis 101
 - 8.3.4 Nearest Neighbor Classification 103
 - 8.3.5 Classification Trees 104
 - 8.3.6 Support Vector Machines 106
 - 8.3.7 Comparison of Methods 107
 - 8.4 Estimating the Error Rate of the Predictor 108
 - 8.4.1 Bias of the Re-Substitution Estimate 108
 - 8.4.2 Cross-Validation and Bootstrap Estimates of Error Rate 110
 - 8.4.3 Reporting Error Rates 112
 - 8.4.4 Statistical Significance of the Error Rate 113
 - 8.4.5 Validation Dataset 113
 - 8.5 Example 114
 - 8.6 Prognostic Prediction 118

- 9 Class Discovery** 121
 - 9.1 Introduction 121
 - 9.2 Similarity and Distance Metrics 122
 - 9.3 Graphical Displays 125

9.3.1	Classical Multidimensional Scaling	125
9.3.2	Nonmetric Multidimensional Scaling	131
9.4	Clustering Algorithms	131
9.4.1	Hierarchical Clustering	131
9.4.2	<i>k</i> -Means Clustering	138
9.4.3	Self-Organizing Maps	142
9.4.4	Other Clustering Procedures	145
9.5	Assessing the Validity of Clusters	146
9.5.1	Global Tests of Clustering	148
9.5.2	Estimating the Number of Clusters	150
9.5.3	Assessing Reproducibility of Individual Clusters	152
A	Basic Biology of Gene Expression	157
A.1	Introduction	157
B	Description of Gene Expression Datasets Used as Examples 165	
B.1	Introduction	165
B.2	Bittner Melanoma Data	165
B.3	Luo Prostate Data	166
B.4	Perou Breast Data	166
B.5	Tamayo HL-60 Data	167
B.6	Hedenfalk Breast Cancer Data	168
C	BRB-ArrayTools	169
C.1	Software Description	169
C.2	Analysis of Bittner Melanoma Data	171
C.3	Analysis of Perou Breast Cancer Chemotherapy Data	178
C.4	Analysis of Hedenfalk Breast Cancer Data	182
	References	185
	Index	195