

Contents

Preface	xi
1 Preliminaries	1
1.1 Using the R Computing Environment	1
1.1.1 Installing smida	2
1.1.2 Loading smida	3
1.2 Data Sets from Biological Experiments	3
1.2.1 Arabidopsis experiment: Anna Amtmann	4
1.2.2 Skin cancer experiment: Nighean Barr	6
1.2.3 Breast cancer experiment: John Bartlett	7
1.2.4 Mammary gland experiment: Gusterson group	9
1.2.5 Tuberculosis experiment: B μ G@S group	10
I Getting Good Data	13
2 Set-up of a Microarray Experiment	15
2.1 Nucleic Acids: DNA and RNA	15
2.2 Simple cDNA Spotted Microarray Experiment	16
2.2.1 Growing experimental material	17
2.2.2 Obtaining RNA	17
2.2.3 Adding spiking RNA and poly-T primer	18
2.2.4 Preparing the enzyme environment	19
2.2.5 Obtaining labelled cDNA	19
2.2.6 Preparing cDNA mixture for hybridization	19
2.2.7 Slide hybridization	20
3 Statistical Design of Microarrays	23
3.1 Sources of Variation	24
3.2 Replication	26
3.2.1 Biological and technical replication	27
3.2.2 How many replicates?	29
3.2.3 Pooling samples	30

- 3.3 Design Principles 36
 - 3.3.1 Blocking, crossing and randomization 37
 - 3.3.2 Design and normalization 39
- 3.4 Single-channel Microarray Design 40
 - 3.4.1 Design issues 41
 - 3.4.2 Design layout 42
 - 3.4.3 Dealing with technical replicates 42
- 3.5 Two-channel Microarray Designs 44
 - 3.5.1 Optimal design of dual-channel arrays 44
 - 3.5.2 Several practical two-channel designs 50
- 4 Normalization 57**
 - 4.1 Image Analysis 57
 - 4.1.1 Filtering 58
 - 4.1.2 Gridding 60
 - 4.1.3 Segmentation 61
 - 4.1.4 Quantification 62
 - 4.2 Introduction to Normalization 62
 - 4.2.1 Scale of gene expression data 63
 - 4.2.2 Using control spots for normalization 65
 - 4.2.3 Missing data 65
 - 4.3 Normalization for Dual-channel Arrays 69
 - 4.3.1 Order for the normalizations 70
 - 4.3.2 Spatial correction 71
 - 4.3.3 Background correction 76
 - 4.3.4 Dye effect normalization 80
 - 4.3.5 Normalization within and across conditions 84
 - 4.4 Normalization of Single-channel Arrays 93
 - 4.4.1 Affymetrix data structure 93
 - 4.4.2 Normalization of Affymetrix data 94
- 5 Quality Assessment 103**
 - 5.1 Using MIAME in Quality Assessment 104
 - 5.1.1 Components of MIAME 104
 - 5.2 Comparing Multivariate Data 105
 - 5.2.1 Measurement scale 105
 - 5.2.2 Dissimilarity and distance measures 106
 - 5.2.3 Representing multivariate data 111
 - 5.3 Detecting Data Problems 113
 - 5.3.1 Clerical errors 114
 - 5.3.2 Normalization problems 117
 - 5.3.3 Hybridization problems 119
 - 5.3.4 Array mishandling 121
 - 5.4 Consequences of Quality Assessment Checks 123

6	Microarray Myths: Data	125
6.1	Design	125
6.1.1	Single-versus dual-channel designs?	125
6.1.2	Dye-swap experiments	129
6.2	Normalization	129
6.2.1	Myth: ‘microarray data is Gaussian’	129
6.2.2	Myth: ‘microarray data is not Gaussian’	131
6.2.3	Confounding spatial and dye effect	132
6.2.4	Myth: ‘non-negative background subtraction’	133
II	Getting Good Answers	135
7	Microarray Discoveries	137
7.1	Discovering Sample Classes	137
7.1.1	Why cluster samples?	138
7.1.2	Sample dissimilarity measures	139
7.1.3	Clustering methods for samples	144
7.2	Exploratory Supervised Learning	155
7.2.1	Labelled dendrograms	156
7.2.2	Labelled PAM-type clusterings	157
7.3	Discovering Gene Clusters	160
7.3.1	Similarity measures for expression profiles	160
7.3.2	Gene clustering methods	163
8	Differential Expression	177
8.1	Introduction	177
8.1.1	Classical versus Bayesian hypothesis testing	177
8.1.2	Multiple testing ‘problem’	179
8.2	Classical Hypothesis Testing	179
8.2.1	What is a hypothesis test?	180
8.2.2	Hypothesis tests for two conditions	183
8.2.3	Decision rules	192
8.2.4	Results from skin cancer experiment	195
8.3	Bayesian Hypothesis Testing	196
8.3.1	A general testing procedure	200
8.3.2	Bayesian <i>t</i> -test	203
9	Predicting Outcomes with Gene Expression Profiles	211
9.1	Introduction	211
9.1.1	Probabilistic classification theory	212
9.1.2	Modelling and predicting continuous variables	217
9.2	Curse of Dimensionality: Gene Filtering	218
9.2.1	Use only significantly expressed genes	218
9.2.2	PCA and gene clustering	220

- 9.2.3 Penalized methods 222
- 9.2.4 Biological selection 222
- 9.3 Predicting Class Memberships 223
 - 9.3.1 Variance-bias trade-off in prediction 223
 - 9.3.2 Linear discriminant analysis 227
 - 9.3.3 k -nearest neighbour classification 231
- 9.4 Predicting Continuous Responses 235
 - 9.4.1 Penalized regression: LASSO 235
 - 9.4.2 k -nearest neighbour regression 243

- 10 Microarray Myths: Inference 247**
 - 10.1 Differential Expression 247
 - 10.1.1 Myth: ‘Bonferroni is too conservative’ 247
 - 10.1.2 FPR and collective multiple testing 248
 - 10.1.3 Misinterpreting FDR 248
 - 10.2 Prediction and Learning 249
 - 10.2.1 Cross-validation 249

- Bibliography 251**

- Index 259**