



Contents

Preface	ix
Acknowledgements	xi
Chapter 1 Introduction	
1.1 The central dogma of molecular biology	3
1.2 What are microarrays and how do they work?	4
1.3 Gene function and drug discovery	5
1.4 Data generation, processing and analysis: an overview	7
1.5 Data management	10
References	12
Chapter 2 Experimental design	
2.1 Experimental objectives and features of microarray data	14
2.2 General principles of experimental design	16
2.2.1 Reducing the number of variables	17
2.2.2 Time courses vs. independent data points	18
2.2.3 Replicates and repeated measurements	18
2.2.4 Reference samples	22
2.2.5 Exogenous ('spiked-in') controls	23
2.2.6 Dual labelling/dye swapping	24
2.2.7 Validation of results	25
2.3 Choice and preparation of samples	26
2.3.1 Obtaining the appropriate sample	26
2.3.1.1 Strain background	26
2.3.1.2 Mutants	26
2.3.1.3 Reagents	27
2.3.1.4 Sample and sample composition	27
2.3.2 Small amounts of sample: RNA amplification and pooling	28
2.3.3 Preparation of the labelled extract	28
2.3.4 Assessing the quality of the labelled extract	29

2.4 Choice and design of arrays	30
2.4.1 Choice of array platform	30
2.4.2 Oligonucleotides vs. PCR products	31
2.4.3 Replicate, guide and control features	32
2.4.4 Cross-hybridisation	32
2.5 Hybridisation, scanning and quality control	33
2.6 Long-term considerations	35
2.6.1 Record keeping	35
2.6.2 Standardisation	36
References	37

Chapter 3 Image processing, normalisation and data transformation

3.1 Introduction	40
3.2 Preliminary processing of the data	41
3.2.1 Image analysis	41
3.2.2 Measuring and reporting expression	42
3.2.2.1 Saturated pixels	44
3.2.2.2 The appropriate number of pixels	44
3.2.2.3 Estimating background	46
3.2.2.4 Reporting expression with Affymetrix GeneChips TM	47
3.2.2.5 Expression ratios: the starting point for sample comparison	48
3.2.2.6 Transformations of the expression ratio	49
3.2.2.7 Situations where expression does not correlate with spot intensity	51
3.3 Normalisation	51
3.3.1 Total intensity normalisation	54
3.3.2 Mean log centring	55
3.3.3 Linear regression	56
3.3.4 Chen's ratio statistics	57
3.3.5 Lowess normalisation	57
3.3.6 Global vs. local normalisation	59
3.4 Data filtering	60
3.4.1 Filtering low intensity data	61
3.4.2 Setting floors and ceilings	61
3.4.3 Use of replicate data	62
3.4.4 Experimental design strategies	62
3.4.5 Replicate filtering	63
3.4.6 Averaging replicate data	64
3.5 Identification of differentially expressed genes	66
3.5.1 Intensity-dependent estimation of differential expression	66
3.5.2 Analysis of variance	67
References	69

Chapter 4 Analysis of gene expression data matrices

4.1	Introduction	71
4.2	Gene expression data matrices: their features and representations	75
4.2.1	Gene expression matrices	75
4.2.2	Representation of expression data as vector space – sample space and gene space	79
4.2.3	Distance and similarity measures in expression space	81
4.2.3.1	Euclidean, Minkowski, Manhattan, angle and chord distances	82
4.2.3.2	Pearson correlation distance, adjusting the mean and variance, correlation matrices, and the relationship between Euclidean and correlation distances	85
4.2.3.3	Spearman's rank correlation	89
4.2.3.4	Distances in discretised space, and mutual information	90
4.2.4	Principal component analysis, eigen-vectors and eigen-genes	92
4.2.5	Dealing with missing values	94
4.2.6	Representation of gene expression data by graphs (networks)	95
4.2.7	Gene expression matrix annotation	95
4.3	Clustering	98
4.3.1	Types of clustering	99
4.3.2	Hierarchical agglomerative clustering	100
4.3.3	Hierarchical divisive clustering	103
4.3.4	Non-hierarchical clustering – K -means	104
4.3.5	Self-organising maps and trees	105
4.3.6	Relationship between clustering and PCA	106
4.3.7	'Gene shaving'	107
4.3.8	Clustering in discretised space	108
4.3.9	Graph-based clustering	108
4.3.10	Bayesian or model-based clustering and fuzzy clustering	109
4.3.11	Clustering genes and samples – applications of clustering	110
4.3.12	Cluster scoring and validation	112
4.4	Classification algorithms and class prediction	113
4.4.1	Definition of the problem	114
4.4.2	Linear discriminants	115
4.4.3	Support vector machines	117
4.4.4	K -nearest neighbour method	118
4.4.5	Neural networks, decision trees and applications of classification	119
4.4.6	Partially supervised analysis	120
4.4.7	Class discovery	120

4.5 Time series analysis	121
4.6 Visualisation	124
4.7 Downstream from expression profile analysis	126
4.7.1 Identification of regulatory signals	128
References	130
Appendix: Non-commercial software	134
A.1 Statistical analysis	134
A.2 Normalisation, clustering and classification	136
A.3 Visualisation	139
A.4 Multifunctional software	139
References	141
Glossary	143
Index	154

Colour plates fall between pp. 88 and 89.