

advanced analysis of
gene expression
microarray data

aidong zhang

Contents

<i>Preface</i>	vii
1. Introduction	1
1.1 The Microarray: Key to Functional Genomics and Systems Biology	1
1.2 Applications of Microarray	2
1.2.1 Gene Expression Profiles in Different Tissues	3
1.2.2 Developmental Genetics	3
1.2.3 Gene Expression Patterns in Model Systems	3
1.2.4 Differential Gene Expression Patterns in Diseases	4
1.2.5 Gene Expression Patterns in Pathogens	5
1.2.6 Gene Expression in Response to Drug Treatments	6
1.2.7 Genotypic Analysis	7
1.2.8 Mutation Screening of Disease Genes	7
1.3 Framework of Microarray Data Analysis	8
1.4 Summary	11
2. Basic Concepts of Molecular Biology	13
2.1 Introduction	13
2.2 Cells	13
2.3 Proteins	15
2.4 Nucleic Acids	19
2.4.1 DNA	19
2.4.2 RNA	22
2.5 Central Dogma of Molecular Biology	22
2.5.1 Genes and the Genetic Code	23
2.5.2 Transcription and Gene Expression	25

2.5.3	Translation and Protein Synthesis	26
2.6	Genotype and Phenotype	27
2.7	Summary	30
3.	Overview of Microarray Experiments	31
3.1	Introduction	31
3.2	Microarray Chip Manufacture	32
3.2.1	Deposition-Based Manufacture	33
3.2.2	<i>In Situ</i> Manufacture	34
3.2.2.1	The Affymetrix GeneChip	35
3.3	Steps of Microarray Experiments	36
3.3.1	Sample Preparation and Labeling	36
3.3.2	Hybridization	39
3.3.3	Image Scanning	39
3.4	Image Processing	40
3.5	Microarray Data Cleaning and Preprocessing	42
3.5.1	Data Transformation	42
3.5.2	Missing Value Estimation	43
3.6	Data Normalization	45
3.6.1	Global Normalization Approaches	46
3.6.1.1	Standardization	46
3.6.1.2	Iterative linear regression	46
3.6.2	Intensity-Dependent Normalization	47
3.6.2.1	LOWESS: Locally weighted linear regression	47
3.6.2.2	Distribution normalization	49
3.7	Summary	49
4.	Analysis of Differentially-Expressed Genes	51
4.1	Introduction	51
4.2	Basic Concepts in Statistics	53
4.2.1	Statistical Inference	53
4.2.2	Hypothesis Test	54
4.3	Fold Change Methods	56
4.3.1	k -fold Change	56
4.3.2	Unusual Ratios	57
4.3.3	Model-Based Methods	60
4.4	Parametric Tests	62
4.4.1	Paired t -Test	62

4.4.2	Unpaired t -Test	63
4.4.3	Variants of t -Test	64
4.5	Non-Parametric Tests	65
4.5.1	Classical Non-Parametric Statistics	65
4.5.2	Other Non-Parametric Statistics	66
4.5.3	Bootstrap Analysis	67
4.6	Multiple Testing	69
4.6.1	Family-Wise Error Rate	70
4.6.1.1	Šidák correction and Bonferroni correction	70
4.6.1.2	Holm's step-wise correction	71
4.6.2	False Discovery Rate	71
4.6.3	Permutation Correction	72
4.6.4	SAM: Significance Analysis of Microarrays	73
4.7	ANOVA: Analysis of Variance	77
4.7.1	One-Way ANOVA	79
4.7.2	Two-Way ANOVA	80
4.8	Summary	82
5.	Gene-Based Analysis	83
5.1	Introduction	83
5.2	Proximity Measurement for Gene Expression Data	85
5.2.1	Euclidean Distance	85
5.2.2	Correlation Coefficient	86
5.2.2.1	Pearson's correlation coefficient	86
5.2.2.2	Jackknife correlation	88
5.2.2.3	Spearman's rank-order correlation	88
5.2.3	Kullback-Leibler Divergence	88
5.3	Partition-Based Approaches	90
5.3.1	K-means and its Variations	90
5.3.2	SOM and its Extensions	92
5.3.3	Graph-Theoretical Approaches	94
5.3.3.1	HCS and CLICK	94
5.3.3.2	CAST: Cluster affinity search technique	96
5.3.4	Model-Based Clustering	98
5.4	Hierarchical Approaches	99
5.4.1	Agglomerative Algorithms	99
5.4.2	Divisive Algorithms	102
5.4.2.1	DAA: Deterministic annealing algorithm	102
5.4.2.2	SPC: Super-paramagnetic clustering	103

5.5	Density-Based Approaches	104
5.5.1	DBSCAN	105
5.5.2	OPTICS	106
5.5.3	DENCLUE	107
5.6	GPX: Gene Pattern eXplorer	110
5.6.1	The Attraction Tree	115
5.6.1.1	The distance measure	115
5.6.1.2	The density definition	116
5.6.1.3	The attraction tree	118
5.6.1.4	An example of attraction tree	120
5.6.2	Interactive Exploration of Coherent Patterns	122
5.6.2.1	Generating the index list	123
5.6.2.2	The coherent pattern index and its graph	125
5.6.2.3	Drilling down to subgroups	126
5.6.3	Experimental Results	128
5.6.3.1	Interactive exploration of Iyer's data and Spellman's data	129
5.6.3.2	Comparison with other algorithms	129
5.6.4	Efficiency and Scalability	134
5.7	Cluster Validation	135
5.7.1	Homogeneity and Separation	136
5.7.2	Agreement with Reference Partition	137
5.7.3	Reliability of Clusters	138
5.7.3.1	P -value of a cluster	138
5.7.3.2	Prediction strength	139
5.8	Summary	139
6.	Sample-Based Analysis	141
6.1	Introduction	141
6.2	Selection of Informative Genes	144
6.2.1	Supervised Approaches	145
6.2.1.1	Differentially expressed genes	145
6.2.1.2	Gene pairs	146
6.2.1.3	Virtual genes	148
6.2.1.4	Genetic algorithms	150
6.2.2	Unsupervised Approaches	152
6.2.2.1	PCA: Principal component analysis	152
6.2.2.2	Gene shaving	154
6.3	Class Prediction	155

6.3.1	Linear Discriminant Analysis	155
6.3.2	Instance-Based Classification	158
6.3.2.1	KNN: k -Nearest Neighbor	158
6.3.2.2	Weighted voting	159
6.3.3	Decision Trees	160
6.3.4	Support Vector Machines	162
6.4	Class Discovery	163
6.4.1	Problem statement	165
6.4.2	CLIFF: CLustering via Iterative Feature Filtering	165
6.4.2.1	The sample-partition process	166
6.4.2.2	The gene-filtering process	167
6.4.3	ESPD: Empirical Sample Pattern Detection	168
6.4.3.1	Measurements for phenotype structure detection	168
6.4.3.2	Algorithms	173
6.4.3.3	Experimental results	184
6.5	Classification Validation	190
6.5.1	Prediction Accuracy	190
6.5.2	Prediction Reliability	191
6.6	Summary	192
7.	Pattern-Based Analysis	195
7.1	Introduction	195
7.2	Mining Association Rules	197
7.2.1	Concepts of Association-Rule Mining	198
7.2.2	The Apriori Algorithm	200
7.2.3	The FP-Growth Algorithm	201
7.2.4	The CARPENTER Algorithm	202
7.2.5	Generating Association Rules in Microarray Data	204
7.2.5.1	Rule filtering	205
7.2.5.2	Rule grouping	206
7.3	Mining Pattern-Based Clusters in Microarray Data	207
7.3.1	Heuristic Approaches	208
7.3.1.1	Coupled two-way clustering (CTWC)	208
7.3.1.2	Plaid model	209
7.3.1.3	Biclustering and δ -Clusters	210
7.3.2	Deterministic Approaches	211
7.3.2.1	δ -pCluster	211
7.3.2.2	OP-Cluster	213

7.4	Mining Gene-Sample-Time Microarray Data	214
7.4.1	Three-dimensional Microarray Data	214
7.4.2	Coherent Gene Clusters	215
7.4.2.1	Problem description	217
7.4.2.2	Maximal coherent sample sets	219
7.4.2.3	The mining algorithms	222
7.4.2.4	Experimental results	227
7.4.3	Tri-Clusters	232
7.4.3.1	The tri-cluster model	232
7.4.3.2	Properties of tri-clusters	234
7.4.3.3	Mining tri-clusters	235
7.5	Summary	238
8.	Visualization of Microarray Data	239
8.1	Introduction	239
8.2	Single-Array Visualization	241
8.2.1	Box Plot	242
8.2.2	Histogram	243
8.2.3	Scatter Plot	244
8.2.4	Gene Pies	246
8.3	Multi-Array Visualization	247
8.3.1	Global Visualizations	247
8.3.2	Optimal Visualizations	249
8.3.3	Projection Visualization	250
8.4	VizStruct	251
8.4.1	Fourier Harmonic Projections	253
8.4.1.1	Discrete-time signal paradigm	253
8.4.1.2	The Fourier harmonic projection algorithm	254
8.4.2	Properties of FHPs	257
8.4.2.1	Basic properties	257
8.4.2.2	Advanced properties	258
8.4.2.3	Harmonic equivalency	260
8.4.2.4	Effects of harmonic twiddle power index	261
8.4.3	Enhancements of Fourier Harmonic Projections	263
8.4.4	Exploratory Visualization of Gene Profiling	265
8.4.4.1	Microarray data sets for visualization	265
8.4.4.2	Identification of informative genes	265
8.4.4.3	Classifier construction and evaluation	265
8.4.4.4	Dimension arrangement	267

8.4.4.5	Visualization of various data sets	270
8.4.4.6	Comparison of FFHP to Sammon's mapping	275
8.4.5	Confirmative Visualization of Gene Time-series	277
8.4.5.1	Data sets for visualization	277
8.4.5.2	The harmonic projection approach	278
8.4.5.3	Rat kidney data set	278
8.4.5.4	Yeast-A data set	279
8.4.5.5	Yeast-B data set	282
8.5	Summary	282
9.	New Trends in Mining Gene Expression Microarray Data	285
9.1	Introduction	285
9.2	Meta-Analysis of Microarray Data	285
9.2.1	Meta-Analysis of Differential Genes	286
9.2.2	Meta-Analysis of Co-Expressed Genes	287
9.3	Semi-Supervised Clustering	288
9.3.1	General Semi-Supervised Clustering Algorithms	289
9.3.2	A Seed-Generation Approach	291
9.3.2.1	Seed-generation methods	291
9.3.2.2	Pattern-selection rules	292
9.3.2.3	The framework for the seed-generation approach	295
9.4	Integration of Gene Expression Data with Other Data	296
9.4.1	A Probabilistic Model for Joint Mining	299
9.4.2	A Graph-Based Model for Joint Mining	300
9.5	Summary	304
10.	Conclusion	305
	<i>Bibliography</i>	307
	<i>Index</i>	331