



A BIOLOGIST'S
GUIDE TO

ANALYSIS OF
DNA
MICROARRAY
DATA

STEEN KNUDSEN

Contents

<i>Preface</i>	<i>xi</i>
<i>Acknowledgments</i>	<i>xiii</i>
1 <i>Introduction</i>	1
1.1 <i>Hybridization</i>	1
1.2 <i>Affymetrix GeneChip Technology</i>	3
1.3 <i>Spotted Arrays</i>	6
1.4 <i>Serial Analysis of Gene Expression (SAGE)</i>	8
1.5 <i>Example: Affymetrix vs. Spotted Arrays</i>	9
1.6 <i>Summary</i>	11
1.7 <i>Further Reading</i>	13
2 <i>Overview of Data Analysis</i>	15
3 <i>Basic Data Analysis</i>	17
3.1 <i>Absolute Measurements</i>	17
3.2 <i>Scaling</i>	18
3.2.1 <i>Example: Linear and Nonlinear Scaling</i>	20
3.3 <i>Detection of Outliers</i>	20
3.4 <i>Fold Change</i>	21
	vii

3.5	<i>Significance</i>	22
3.5.1	<i>Nonparametric Tests</i>	24
3.5.2	<i>Correction for Multiple Testing</i>	24
3.5.3	<i>Example I: t-Test and ANOVA</i>	25
3.5.4	<i>Example II: Number of Replicates</i>	26
3.6	<i>Summary</i>	28
3.7	<i>Further Reading</i>	29
4	<i>Visualization by Reduction of Dimensionality</i>	33
4.1	<i>Principal Component Analysis</i>	33
4.2	<i>Example 1: PCA on Small Data Matrix</i>	35
4.3	<i>Example 2: PCA on Real Data</i>	37
4.4	<i>Summary</i>	37
4.5	<i>Further Reading</i>	39
5	<i>Cluster Analysis</i>	41
5.1	<i>Hierarchical Clustering</i>	41
5.2	<i>K-means Clustering</i>	43
5.3	<i>Self-Organizing Maps</i>	44
5.4	<i>Distance Measures</i>	45
5.4.1	<i>Example: Comparison of Distance Measures</i>	47
5.5	<i>Normalization</i>	49
5.6	<i>Visualization of Clusters</i>	50
5.6.1	<i>Example: Visualization of Gene Clusters in Bladder Cancer</i>	50
5.7	<i>Summary</i>	50
5.8	<i>Further Reading</i>	52
6	<i>Beyond Cluster Analysis</i>	55
6.1	<i>Function Prediction</i>	55
6.2	<i>Discovery of Regulatory Elements in Promoter Regions</i>	56
6.2.1	<i>Example 1: Discovery of Proteasomal Element</i>	57
6.2.2	<i>Example 2: Rediscovery of Mlu Cell Cycle Box (MCB)</i>	57
6.3	<i>Integration of data</i>	58
6.4	<i>Summary</i>	59
6.5	<i>Further Reading</i>	59

7	<i>Reverse Engineering of Regulatory Networks</i>	63
7.1	<i>The Time-Series Approach</i>	63
7.2	<i>The Steady-State Approach</i>	64
7.3	<i>Limitations of Network Modeling</i>	65
7.4	<i>Example 1: Steady-State Model</i>	65
7.5	<i>Example 2: Steady-State Model on Real Data</i>	66
7.6	<i>Example 3: Steady-State Model on Real Data</i>	68
7.7	<i>Example 4: Linear Time-Series Model</i>	68
7.8	<i>Further Reading</i>	71
8	<i>Molecular Classifiers</i>	75
8.1	<i>Classification Schemes</i>	76
8.1.1	<i>Nearest Neighbor</i>	76
8.1.2	<i>Neural Networks</i>	76
8.1.3	<i>Support Vector Machine</i>	76
8.2	<i>Example I: Classification of Cancer Subtypes</i>	77
8.3	<i>Example II: Classification of Cancer Subtypes</i>	78
8.4	<i>Summary</i>	79
8.5	<i>Further Reading</i>	79
9	<i>Selection of Genes for Spotting on Arrays</i>	81
9.1	<i>Gene Finding</i>	82
9.2	<i>Selection of Regions Within Genes</i>	82
9.3	<i>Selection of Primers for PCR</i>	83
9.4	<i>Selection of Unique Oligomer Probes</i>	83
9.4.1	<i>Example: Finding PCR Primers for Gene AF105374</i>	83
9.5	<i>Experimental Design</i>	84
9.6	<i>Further Reading</i>	84
10	<i>Limitations of Expression Analysis</i>	87
10.1	<i>Relative Versus Absolute RNA Quantification</i>	88
10.2	<i>Further Reading</i>	88
11	<i>Genotyping Chips</i>	91
11.1	<i>Example: Neural Networks for GeneChip prediction</i>	91
11.2	<i>Further Reading</i>	93
12	<i>Software Issues and Data Formats</i>	95

x **CONTENTS**

12.1	<i>Standardization Efforts</i>	96
12.2	<i>Standard File Format</i>	97
12.2.1	<i>Example: Small Scripts in Awk</i>	97
12.3	<i>Software for Clustering</i>	98
12.3.1	<i>Example: Clustering with ClustArray</i>	99
12.4	<i>Software for Statistical Analysis</i>	99
12.4.1	<i>Example: Statistical Analysis with R</i>	99
12.4.2	<i>The affyR Software Package</i>	103
12.4.3	<i>Commercial Statistics Packages</i>	103
12.5	<i>Summary</i>	103
12.6	<i>Further Reading</i>	104
13	<i>Commercial Software Packages</i>	105
14	<i>Bibliography</i>	109
	<i>Index</i>	123