

Contents

Preface	xi
CHAPTER 1 ■ Introduction	1
1.1 PIPELINES TO ANALYZE “OMICS” DATA	1
1.2 RNA-Seq GENE EXPRESSION IN S2-DRSC CELLS	3
1.3 MICROARRAY GENE EXPRESSION IN YEAST CELLS AND IN PROSTATE SAMPLES	3
1.4 DNA METHYLATION IN NORMAL AND COLON/RECTAL ADENOCARCINOMA SAMPLES	4
CHAPTER 2 ■ Genome-scale gene expression data	5
2.1 MICROARRAY GENE EXPRESSION DATA	5
2.1.1 Data generation	5
2.1.2 Preprocessing and quality control of microarray data	7
2.2 DATA FROM NEXT GENERATION SEQUENCING	13
2.2.1 Data generation	14
2.2.2 Preprocessing and quality control of bulk RNA- Seq data	14
CHAPTER 3 ■ Genome-scale epigenetic data	17
3.1 DATA GENERATION	17
3.2 QUALITY CONTROL AND PREPROCESSING OF DNA METHYLATION DATA	18

3.2.1	The control probe adjustment and reduction of global correlation pipeline (CPACOR)	18
3.2.2	Quantile normalization with ComBat	21
3.3	CELL TYPE COMPOSITION INFERENCES	22
3.3.1	Reference-based methods	23
3.3.2	Reference-free methods	27
3.4	APPENDIX – MODIFIED PROGRAMS IN THE CPACOR WITH AN APPLICATION	31
CHAPTER 4	■ Screening genome-scale genetic and epigenetic data	39
4.1	SCREENING VIA TRAINING AND TESTING SAMPLES	40
4.2	SCREENING INCORPORATING SURROGATE VARIABLES	41
4.3	SURE INDEPENDENCE SCREENING	46
4.3.1	Correlation learning	47
4.4	NON- AND SEMI-PARAMETRIC SCREENING TECHNIQUES	50
4.4.1	Random forest	50
4.4.2	Support vector machine	53
CHAPTER 5	■ Cluster Analysis in Data mining	57
5.1	NON-PARAMETRIC CLUSTER ANALYSIS METHODS	57
5.1.1	Distances	58
5.1.2	Partitioning-based methods	59
5.1.3	Hierarchical clustering	64
5.1.4	Hybrids of partitioning-based and hierarchical clustering	69
5.1.5	Examples – clustering to detect gene expression patterns	75
5.2	CLUSTER ANALYSES IN LINEAR REGRESSIONS	85
5.3	BICLUSTER ANALYSES	91
5.4	JOINT CLUSTER ANALYSIS	96

CHAPTER	6 ■ Methods to select genetic and epigenetic factors based on linear associations	101
6.1	FREQUENTIST APPROACHES	102
6.1.1	Elastic net	102
6.1.2	Adaptive LASSO	105
6.1.3	Smoothly clipped absolute deviation (SCAD)	106
6.2	BAYESIAN APPROACHES	108
6.2.1	Zellner's g -prior	109
6.2.2	Extension of Zellner's g -prior to multi-components G -prior	111
6.2.3	The spike-and-slab prior	113
6.3	EXAMPLES – SELECTING IMPORTANT EPIGENETIC FACTORS	118
CHAPTER	7 ■ Non- and semi-parametric methods to select genetic and epigenetic factors	125
7.1	VARIABLE SELECTION BASED ON SPLINES	126
7.2	OVERVIEW OF THE ANOVA-BASED APPROACH	128
7.3	VARIABLE SELECTION BUILT UPON REPRODUCING KERNELS	129
7.4	EXAMPLES	132
7.4.1	Selecting important epigenetic factors	132
7.4.2	Selecting variables with known underlying truth	139
CHAPTER	8 ■ Network construction and analyses	145
8.1	UNDIRECTED NETWORKS	145
8.1.1	The two-stage graphs selection method	146
8.1.2	The GGMselect package and gene expression examples	147
8.2	CORRELATION NETWORKS	152

8.3	BAYESIAN NETWORKS	158
8.4	NETWORK COMPARISONS	161
8.4.1	Comparing undirected networks	162
8.4.2	Comparing Bayesian networks	163
	Bibliography	165
	Index	183