

Contents

Acknowledgements	xiii
Authors	xv
Preface	xvii
1 Introduction	1
1.1 Introduction	1
1.2 Central Dogma	2
1.3 Measuring Gene Expression	2
1.4 Representation of Gene Expression Data	4
1.5 Gene Expression Data Analysis: Applications	6
1.6 Machine Learning	8
1.7 Statistical and Biological Evaluation	10
1.8 Gene Expression Analysis Approaches	11
1.8.1 Preprocessing in Microarray and RNAseq Data	12
1.8.2 Co-Expressed Pattern-Finding Using Machine Learning	16
1.8.3 Co-Expressed Pattern-Finding Using Network-Based Approaches	20
1.9 Differential Co-Expression Analysis	21
1.10 Differential Expression Analysis	21
1.11 Tools and Systems for Gene Expression Data Analysis	22
1.11.1 (Diff) Co-Expression Analysis Tools and Systems	22
1.11.2 Differential Expression Analysis Tools and Systems	23
1.12 Contribution of This Book	23
1.13 Organization of This Book	24

2	Information Flow in Biological Systems	27
2.1	Concept of Systems Theory	27
2.1.1	A Brief History of Systems Thinking	27
2.1.2	Areas of Application of Systems Theory in Biology	28
2.2	Complexity in Biological Systems	28
2.2.1	Hierarchical Organization of Biological Systems from Macroscopic Levels to Microscopic Levels	28
2.2.2	Information Flow in Biological Systems	29
2.2.3	Top-Down and Bottom-Up Flow	30
2.3	Central Dogma of Molecular Biology	30
2.3.1	DNA Replication	31
2.3.2	Transcription	32
2.3.3	Translation	33
2.4	Ambiguity in Central Dogma	34
2.4.1	Reverse Transcription	35
2.4.2	RNA Replication	36
2.5	Discussion	37
2.5.1	Biological Information Flow from a Computer Science Perspective	37
2.5.2	Future Perspective	37
3	Gene Expression Data Generation	39
3.1	History of Gene Expression Data Generation	39
3.2	Low-Throughput Methods	41
3.2.1	Northern Blotting	41
3.2.2	Ribonuclease Protection Assay	41
3.2.3	qRT-PCR	42
3.2.4	SAGE	42
3.3	High-Throughput Methods	43
3.3.1	Microarray	43
3.3.2	RNA-Seq	44
3.3.3	Types of RNA-Seq	46
3.3.4	Gene Expression Data Repositories	48
3.3.5	Standards in Gene Expression Data	50
3.4	Chapter Summary	52
4	Statistical Foundations and Machine Learning	53
4.1	Introduction	53
4.2	Statistical Background	53
4.2.1	Statistical Modeling	53

4.2.2	Probability Distributions	54
4.2.3	Hypothesis Testing	54
4.2.4	Exact Tests	55
4.2.5	Common Data Distributions	56
4.2.6	Multiple Testing	64
4.2.7	False Discovery Rate	64
4.2.8	Maximum Likelihood Estimation	65
4.3	Machine Learning Background	67
4.3.1	Significance of Machine Learning	68
4.3.2	Machine Learning and Its Types	70
4.3.3	Supervised Learning Methods	73
4.3.4	Unsupervised Learning Methods	84
4.3.5	Outlier Mining	124
4.3.6	Association Rule Mining	128
4.4	Chapter Summary	140
4.4.1	Statistical Modeling	140
4.4.2	Supervised Learning: Classification and Regression Analysis	140
4.4.3	Proximity Measures	141
4.4.4	Unsupervised Learning: Clustering	141
4.4.5	Unsupervised Learning: Biclustering	142
4.4.6	Unsupervised Learning: Triclustering	142
4.4.7	Outlier Mining	143
4.4.8	Unsupervised Learning: Association Mining	143
5	Co-Expression Analysis	145
5.1	Introduction	145
5.2	Gene Co-Expression Analysis	147
5.2.1	Types of Gene Co-Expression	148
5.2.2	An Example	148
5.3	Measures to Identify Co-Expressed Patterns	151
5.4	Co-Expression Analysis Using Clustering	152
5.4.1	CEA Using Clustering: A Generic Architecture	153
5.4.2	Co-Expressed Pattern Finding Using 1-Way Clustering	163
5.4.3	Subspace or 2-way Clustering in Co-Expression Mining	178
5.4.4	Co-Expressed Pattern-Finding Using 3-Way Clustering	186
5.5	Network Analysis for Co-Expressed Pattern-Finding	192
5.5.1	Definition of CEN	193
5.5.2	Analyzing CENs: A Generic Architecture	193
5.6	Chapter Summary and Recommendations	215

6	Differential Expression Analysis	219
6.1	Introduction	219
6.1.1	Importance of DE Analysis	220
6.2	Differential Expression (DE) of a Gene	221
6.2.1	Differential Expression of a Gene: An Example	221
6.3	Differential Expression Analysis (DEA)	222
6.3.1	A Generic Framework	223
6.3.2	Preprocessing	223
6.3.3	DE Genes Identification	230
6.3.4	DE Gene Analysis	243
6.3.5	Statistical Validation	247
6.3.6	Discussion	249
6.4	Biomarker Identification Using DEA: A Case Study	250
6.4.1	Problem Definition	251
6.4.2	Dataset Used	251
6.4.3	Preprocessing	251
6.4.4	Framework of Analysis Used	252
6.4.5	Results	254
6.4.6	Discussion	256
6.5	Summary and Recommendations	257
7	Tools and Systems	261
7.1	Introduction	261
7.1.1	Generic Characteristics of a Systems Biology Tool	261
7.1.2	Target Systems Biology Activities	262
7.2	Systems Biology Tools	265
7.2.1	A Taxonomy	265
7.2.2	Pre-Processing Tools	266
7.3	Gene Expression Data Analysis Tools	278
7.3.1	Co-Expression Analysis	279
7.3.2	Differential Co-Expression Analysis	283
7.3.3	Differential Expression Analysis	283
7.4	Visualization	284
7.5	Validation	285
7.5.1	Statistical Validation	286
7.6	Biological Validation	288
7.7	Chapter Summary and Concluding Remarks	289

8 Concluding Remarks and Research Challenges	295
8.1 Concluding Remarks	295
8.2 Some Issues and Research Challenges	296
Bibliography	301
Glossary	347
Index	355

This humble work would not have been possible without the constant support, encouragement and constructive criticism of a large number of people. Special thanks and sincere appreciation are due to the following dedicated regular members: Dr. Rajan V. Ahir, Dr. Ravi Sarna, Dr. Anand Kishor, Dr. Prakash Mahanta, Dr. Swarny Ray, Dr. K. C. Baslyn, and Dr. Debesh Nath. We are extremely grateful for the constant support extended by our sincere students and research scholars: Mr. Husein A. Chowdhury, Mr. Chinnoye Bariah, Mr. Akash Das, Mr. Anur Sahoo, Ms. Pallabi Patwary, Ms. Koyel Mandal, Ms. Rachayica Bhattacharya, Mr. Prajal Kalita, and Mihir Gokwad.

We are grateful to the panel of reviewers for their constructive suggestions and critical evaluation. The constant support and cooperation received from our colleagues and students during the period of writing this book is sincerely acknowledged.

Pankaj Barah
 Divya Kumar Bhattacharyya
 Jagat Kumar Kalita