

15.2.6	Statistical Analysis	15.2.6
15.2.7	Software Tools and Technologies	15.2.7
15.2.8	Discussion and Summary	15.2.8
15.3	References	15.3

Part Two Introduction to Molecular Biology

16.1	Introduction	16.1
16.2	The Living Cell	16.2
16.2.1	Introduction	16.2.1
16.2.2	Prokaryotes and Eukaryotes	16.2.2
16.2.3	Multicellularity	16.2.3
16.2.4	Unicellular and Multicellular Organisms	16.2.4
16.2.5	Stem Cells and Cell Differentiation	16.2.5
16.2.6	Cell Components	16.2.6
16.2.7	Plasma Membrane and Transport	16.2.7
16.2.8	Cytoplasm	16.2.8
16.2.9	Extracellular Matrix	16.2.9
16.2.10	Centrosome and Mitochondrion	16.2.10
16.2.11	Actin Filaments and the Cytoskeleton	16.2.11
16.2.12	Nucleus	16.2.12
16.2.13	Ribosomes	16.2.13
16.2.14	Endoplasmic Reticulum	16.2.14
16.2.15	Golgi Apparatus	16.2.15
16.2.16	Lysosome	16.2.16
16.2.17	Peroxisome	16.2.17
16.2.18	Discussion and Summary	16.2.18
16.3	References	16.3

Preface

xix	
-----	--

List of Symbols

xxi	
-----	--

About the Authors

xxiii	
-------	--

Part One Introduction

1	
---	--

1 Introduction to Bioinformatics

3	
---	--

1.1	Introduction	1.1
-----	--------------	-----

1.2	The "Omics" Era	1.2
-----	-----------------	-----

1.3	The Scope of Bioinformatics	1.3
-----	-----------------------------	-----

1.3.1	Areas of Molecular Biology Subject to Bioinformatics Analysis	1.3.1
-------	---	-------

1.3.2	Data Storage, Retrieval and Organisation	1.3.2
-------	--	-------

1.3.3	Data Analysis	1.3.3
-------	---------------	-------

1.3.4	Statistical Analysis	1.3.4
-------	----------------------	-------

1.3.5	Presentation	1.3.5
-------	--------------	-------

1.4	What Do Information Engineers and Biologists Need to Know?	1.4
-----	--	-----

1.5	Discussion and Summary	1.5
-----	------------------------	-----

	References	
--	------------	--

2 Computational Methods in Bioinformatics

9	
---	--

2.1	Introduction	2.1
-----	--------------	-----

2.2	Machine Learning and Data Mining	2.2
-----	----------------------------------	-----

2.2.1	Supervised Learning	2.2.1
-------	---------------------	-------

2.2.2	Unsupervised Learning	2.2.2
-------	-----------------------	-------

2.3	Optimisation	2.3
-----	--------------	-----

2.4	Image Processing: Bioimage Informatics	2.4
-----	--	-----

2.5	Network Analysis	2.5
-----	------------------	-----

2.6	Statistical Analysis	15
2.7	Software Tools and Technologies	15
2.8	Discussion and Summary	16
	References	17
	Part Two Introduction to Molecular Biology	19
3	The Living Cell	21
3.1	Introduction	21
3.2	Prokaryotes and Eukaryotes	21
3.3	Multicellularity	22
3.3.1	Unicellular and Multicellular Organisms	22
3.3.2	Stem Cells and Cell Differentiation	23
3.4	Cell Components	24
3.4.1	Plasma Membrane and Transport Proteins	25
3.4.2	Cytoplasm	25
3.4.3	Extracellular Matrix	26
3.4.4	Centrosome and Microtubules	26
3.4.5	Actin Filaments and the Cytoskeleton	27
3.4.6	Nucleus	27
3.4.7	Vesicles	27
3.4.8	Ribosomes	28
3.4.9	Endoplasmic Reticulum	28
3.4.10	Golgi Apparatus	29
3.4.11	Mitochondrion and the Energy of the Cell	29
3.4.12	Lysosome	30
3.4.13	Peroxisome	31
3.5	Discussion and Summary	31
	References	32
4	Central Dogma of Molecular Biology	33
4.1	Introduction	33
4.2	Central Dogma of Molecular Biology Overview	33
4.3	Proteins	34
4.4	DNA	37
4.5	RNA	39
4.6	Genes	40
4.7	Transcription and Post-transcriptional Processes	41
4.7.1	Post-transcriptional Processes	43
4.7.2	Gene-specific TFs	44
4.7.3	Post-transcriptional Regulation	44
4.8	Translation and Post-translational Processes	45
4.8.1	The Genetic Code	45
4.8.2	tRNA and Ribosomes	46
4.8.3	The Steps of Translation	47
4.8.4	Polyribosomes (Polysomes)	49
4.8.5	Post-translational Processes	50

4.9	Discussion and Summary	50
	References	51
Part Three	Data Acquisition and Pre-processing	53
5	High-throughput Technologies	55
5.1	Introduction	55
5.2	Microarrays	56
5.2.1	DNA Microarrays	56
5.2.2	Protein Microarrays	58
5.2.3	Carbohydrate Microarrays (Glycoarrays)	59
5.2.4	Other Types of Microarrays	59
5.3	Next-generation Sequencing (NGS)	60
5.3.1	DNA Sequencing	60
5.3.2	RNA Sequencing (Transcriptome Analysis)	62
5.3.3	Metagenomics	63
5.3.4	Other Applications of Sequencing	63
5.4	ChIP on Microarrays and Sequencing	63
5.5	Discussion and Summary	64
	References	65
6	Databases, Standards and Annotation	67
6.1	Introduction	67
6.2	NCBI Databases	67
6.2.1	Literature Databases (PubMed, PMC, the Bookshelf and MeSH)	68
6.2.2	GenBank (Nucleotide Database)	69
6.2.3	Reference Sequences (RefSeq) Database	69
6.2.4	Gene Database	70
6.2.5	Protein Database	70
6.2.6	Gene Expression Omnibus	71
6.2.7	Taxonomy and HomoloGene Databases	71
6.2.8	Sequence Read Archive	72
6.2.9	Genomic and Epigenomic Variations	72
6.2.10	Other NCBI Databases	72
6.3	The EBI Databases	73
6.4	Species-specific Databases	75
6.4.1	Animals	75
6.4.2	Plants	76
6.4.3	Fungi	76
6.4.4	Archaea and Bacteria	77
6.4.5	Viruses	78
6.5	Discussion and Summary	78
	References	82
7	Normalisation	87
7.1	Introduction	87
7.2	Issues Tackled by Normalisation	88

7.2.1	Within-slide and Between-slides Normalisation	88
7.2.2	Normalisation Based on Non-differentially Expressed Genes	88
7.2.3	Background Correction	89
7.2.4	Logarithmic Transformation	90
7.2.5	Intensity-dependent Bias – (MA) Plots	91
7.2.6	Replicates and Summarisation	91
7.3	Normalisation Methods	92
7.3.1	Microarray Suite 5 (MAS 5.0)	92
7.3.2	Robust Multi-array Average (RMA)	92
7.3.3	Quantile Normalisation	95
7.3.4	Locally Weighted Scatter-plot Smoothing (Lowess) Normalisation	96
7.3.5	Scaling Methods	98
7.3.6	Model-based Expression Index (MBEI)	99
7.3.7	Other Normalisation Methods	100
7.4	Discussion and Summary	103
	References	104
8	Feature Selection	109
8.1	Introduction	109
8.2	FS and FG – Problem Definition	110
8.3	Consecutive Ranking	111
8.3.1	Forward Search (Most Informative First Admitted)	111
8.3.2	Backward Elimination (Least Useful First Eliminated)	112
8.4	Individual Ranking	112
8.4.1	Information Content	112
8.4.2	SNR Criteria	113
8.5	Principal Component Analysis	115
8.6	Genetic Algorithms and Genetic Programming	115
8.7	Discussion and Summary	115
	References	117
9	Differential Expression	119
9.1	Introduction	119
9.2	Fold Change	120
9.3	Statistical Hypothesis Testing – Overview	120
9.3.1	<i>p</i> -Values and Volcano Plots	121
9.3.2	The Multiple-hypothesis Testing Problem	121
9.4	Statistical Hypothesis Testing – Methods	123
9.4.1	<i>t</i> -Statistic, Modified <i>t</i> -Statistics and the Analysis of Variance (ANOVA)	123
9.4.2	<i>B</i> -Statistic	125
9.4.3	Fisher's Exact Test	126
9.4.4	Likelihood Ratio Test	128
9.4.5	Methods for Over-dispersed Poisson Distribution	129
9.5	Discussion and Summary	129
	References	131

Part Four Clustering Methods	133
10 Clustering Forms	135
10.1 Introduction	135
10.2 Proximity Measures	136
10.2.1 Distance Metrics for Discrete Feature Objects	136
10.2.2 Distance Metrics for Continuous Feature Objects	137
10.3 Clustering Families	137
10.3.1 Partitional Clustering	137
10.3.2 Hierarchical Clustering	138
10.3.3 Fuzzy Clustering	139
10.3.4 Neural Network-based Clustering	139
10.3.5 Mixture Model Clustering	139
10.3.6 Graph-based Clustering	139
10.3.7 Consensus Clustering	140
10.3.8 Biclustering	140
10.4 Clusters and Partitions	140
10.5 Discussion and Summary	140
References	141
11 Partitional Clustering	143
11.1 Introduction	143
11.2 k -Means and its Applications	144
11.2.1 Principles	144
11.2.2 Variations	146
11.2.3 Applications	150
11.3 k -Medoids and its Applications	151
11.3.1 Principles	151
11.3.2 Variations	151
11.3.3 Applications	152
11.4 Discussion and Summary	153
References	154
12 Hierarchical Clustering	157
12.1 Introduction	157
12.2 Principles	158
12.2.1 Agglomerative Methods	158
12.2.2 Divisive Methods	162
12.3 Discussion and Summary	164
References	165
13 Fuzzy Clustering	167
13.1 Introduction	167
13.2 Principles	168
13.2.1 Fuzzy c -Means	168
13.2.2 Probabilistic c -Means	170

13.2.3	Hybrid <i>c</i> -Means	171
13.2.4	Gustafson-Kessel Algorithm	174
13.2.5	Gath-Geva Algorithm	174
13.2.6	Fuzzy <i>c</i> -Shell	175
13.2.7	FANNY	176
13.2.8	Other Fuzzy Clustering Algorithms	176
13.3	Discussion	177
	References	177
14	Neural Network-based Clustering	181
14.1	Introduction	181
14.2	Algorithms	182
14.2.1	SOM	182
14.2.2	GLVQ	184
14.2.3	Neural-gas	185
14.2.4	ART	186
14.2.5	OPOTC	189
14.2.6	SOON	190
14.3	Discussion	193
	References	194
15	Mixture Model Clustering	197
15.1	Introduction	197
15.2	Finite Mixture Models	199
15.2.1	Various Mixture Models	199
15.2.2	Non-Bayesian Methods	204
15.2.3	Bayesian Methods	212
15.3	Infinite Mixture Models	220
15.3.1	DPM Model	220
15.3.2	CRP Mixture Model	222
15.3.3	SBP Mixture Model	222
15.4	Discussion	223
	References	224
16	Graph Clustering	227
16.1	Introduction	227
16.2	Basic Definitions	228
16.2.1	Graph and Adjacency Matrix	228
16.2.2	Measures and Metrics	229
16.2.3	Similarity Matrices	232
16.3	Graph Clustering	233
16.3.1	Graph Cut Clustering	233
16.3.2	Spectral Clustering	234
16.3.3	AP Clustering	236
16.3.4	Modularity-based Clustering	238

16.3.5	Multilevel Graph Partitioning and Hypergraph Partitioning	242
16.3.6	Markov Cluster Algorithm	243
16.4	Resources	243
16.5	Discussion	244
	References	245
17	Consensus Clustering	247
17.1	Introduction	247
17.2	Overview	248
17.3	Consensus Functions	249
17.3.1	P-P Comparison	249
17.3.2	C-C Comparison	253
17.3.3	MIC Voting	254
17.3.4	M-M Co-occurrence	257
17.4	Discussion	261
	References	262
18	Biclustering	265
18.1	Introduction	265
18.2	Overview	266
18.2.1	Statement of the Biclustering Problem	266
18.2.2	Types of Biclusters	266
18.2.3	Classification of Biclustering	267
18.3	Biclustering Methods	268
18.3.1	Variance-minimisation Biclustering Methods	268
18.3.2	Correlation-maximisation Biclustering Methods	270
18.3.3	Two-way Clustering Methods	273
18.3.4	Probabilistic and Generative Methods	274
18.4	Discussion	278
	References	280
19	Clustering Methods Discussion	283
19.1	Introduction	283
19.2	Hierarchical Clustering	283
19.2.1	Yeast Cell Cycle Data	284
19.2.2	Breast Cancer	284
19.2.3	Diffuse Large B-Cell Lymphoma	286
19.3	Fuzzy Clustering	287
19.3.1	DNA Motifs Clustering	287
19.3.2	Microarray Gene Expression	287
19.4	Neural Network-based Clustering	289
19.5	Mixture Model-based Clustering	290
19.5.1	Examples of Finite Mixture Models	291
19.5.2	Examples of Infinite Mixture Models	292
19.6	Graph-based Clustering	293

19.7	Consensus Clustering	295
19.8	Biclustering	296
19.9	Summary	297
	References	298
		303
Part Five Validation and Visualisation		
20	Numerical Validation	305
20.1	Introduction	305
20.2	External Criteria	306
20.2.1	Rand Index	306
20.2.2	Adjusted Rand Index	307
20.2.3	Jaccard Index	308
20.2.4	Normalised Mutual Information	308
20.3	Internal Criteria	308
20.3.1	Adjusted Figure of Merit	309
20.3.2	CLEST	310
20.4	Relative Criteria	310
20.4.1	Minimum Description Length	311
20.4.2	Minimum Message Length	311
20.4.3	Bayesian Information Criterion	312
20.4.4	Akaike's Information Criterion	312
20.4.5	Partition Coefficient	312
20.4.6	Partition Entropy	312
20.4.7	Fukuyama-Sugeno Index	313
20.4.8	Xie-Beni Index	313
20.4.9	Calinski-Harabasz Index	313
20.4.10	Dunn's Index	313
20.4.11	Davies-Bouldin Index	314
20.4.12	I Index	314
20.4.13	Silhouette	314
20.4.14	Object-based Validation	315
20.4.15	Geometrical Index	316
20.4.16	Validity Index	316
20.4.17	Generalised Parametric Validity	317
20.5	Discussion and Summary	318
	References	320
21	Biological Validation	323
21.1	Introduction	323
21.2	GO Analysis	323
21.2.1	GO Term Enrichment	326
21.3	Upstream Sequence Analysis	331
21.4	Gene-network Analysis	333
21.5	Discussion and Summary	337
	References	338

22 Visualisations and Presentations	339
22.1 Introduction	339
22.2 Methods and Examples	339
22.2.1 Profile Patterns	339
22.2.2 Bar Chart	341
22.2.3 Error Bar	342
22.2.4 Pie Chart	345
22.2.5 Box Plot	345
22.2.6 Histogram	346
22.2.7 Scatter Plot	347
22.2.8 Venn Diagram	348
22.2.9 Tree View	350
22.2.10 Heat Map	353
22.2.11 Network Graph	354
22.2.12 Low-dimension Display	355
22.2.13 Receiver Operating Characteristics Curves	355
22.2.14 Kaplan-Meier Plot	356
22.2.15 Block Diagram	358
22.3 Summary	359
References	361
Part Six New Clustering Frameworks Designed for Bioinformatics	363
23 Splitting-Merging Awareness Tactics (SMART)	365
23.1 Introduction	365
23.2 Related Work	366
23.2.1 SSCL	366
23.2.2 SSMCL	367
23.2.3 ULFMM	367
23.2.4 VBGM	367
23.2.5 PFClust	368
23.2.6 DBSCAN	368
23.3 SMART Framework	369
23.4 Implementations	370
23.4.1 SMART-CL	370
23.4.2 SMART-FMM	372
23.4.3 SMART-MFA	374
23.5 Enhanced SMART	377
23.6 Examples	378
23.7 Discussion	383
References	383
24 Tightness-tunable Clustering (UNCLES)	385
24.1 Introduction	385
24.2 Bi-CoPaM Method	385
24.2.1 Partition Generation	386

24.2.2	Relabelling	386
24.2.3	Fuzzy Consensus Partition Matrix Generation	388
24.2.4	Binariisation	388
24.3	UNCLES Method - Other Types of External Specifications	390
24.4	M-N Scatter Plots Technique	391
24.5	Parameter-free UNCLES with M-N Plots	393
24.6	Discussion and Summary	393
	References	394
	Appendix	395
	Index	409
24.2.1	Introduction	386
24.2.2	Methods and Examples	388
24.2.3	Box Plot	391
24.2.4	Box Plot	393
24.2.5	Box Plot	393
24.2.6	Box Plot	394
24.2.7	Box Plot	395
24.2.8	Box Plot	395
24.2.9	Box Plot	395
24.2.10	Box Plot	395
24.2.11	Box Plot	395
24.2.12	Box Plot	395
24.2.13	Box Plot	395
24.2.14	Box Plot	395
24.2.15	Box Plot	395
24.2.16	Box Plot	395
24.2.17	Box Plot	395
24.2.18	Box Plot	395
24.2.19	Box Plot	395
24.2.20	Box Plot	395
24.2.21	Box Plot	395
24.2.22	Box Plot	395
24.2.23	Box Plot	395
24.2.24	Box Plot	395
24.2.25	Box Plot	395
24.2.26	Box Plot	395
24.2.27	Box Plot	395
24.2.28	Box Plot	395
24.2.29	Box Plot	395
24.2.30	Box Plot	395
24.2.31	Box Plot	395
24.2.32	Box Plot	395
24.2.33	Box Plot	395
24.2.34	Box Plot	395
24.2.35	Box Plot	395
24.2.36	Box Plot	395
24.2.37	Box Plot	395
24.2.38	Box Plot	395
24.2.39	Box Plot	395
24.2.40	Box Plot	395
24.2.41	Box Plot	395
24.2.42	Box Plot	395
24.2.43	Box Plot	395
24.2.44	Box Plot	395
24.2.45	Box Plot	395
24.2.46	Box Plot	395
24.2.47	Box Plot	395
24.2.48	Box Plot	395
24.2.49	Box Plot	395
24.2.50	Box Plot	395
24.2.51	Box Plot	395
24.2.52	Box Plot	395
24.2.53	Box Plot	395
24.2.54	Box Plot	395
24.2.55	Box Plot	395
24.2.56	Box Plot	395
24.2.57	Box Plot	395
24.2.58	Box Plot	395
24.2.59	Box Plot	395
24.2.60	Box Plot	395
24.2.61	Box Plot	395
24.2.62	Box Plot	395
24.2.63	Box Plot	395
24.2.64	Box Plot	395
24.2.65	Box Plot	395
24.2.66	Box Plot	395
24.2.67	Box Plot	395
24.2.68	Box Plot	395
24.2.69	Box Plot	395
24.2.70	Box Plot	395
24.2.71	Box Plot	395
24.2.72	Box Plot	395
24.2.73	Box Plot	395
24.2.74	Box Plot	395
24.2.75	Box Plot	395
24.2.76	Box Plot	395
24.2.77	Box Plot	395
24.2.78	Box Plot	395
24.2.79	Box Plot	395
24.2.80	Box Plot	395
24.2.81	Box Plot	395
24.2.82	Box Plot	395
24.2.83	Box Plot	395
24.2.84	Box Plot	395
24.2.85	Box Plot	395
24.2.86	Box Plot	395
24.2.87	Box Plot	395
24.2.88	Box Plot	395
24.2.89	Box Plot	395
24.2.90	Box Plot	395
24.2.91	Box Plot	395
24.2.92	Box Plot	395
24.2.93	Box Plot	395
24.2.94	Box Plot	395
24.2.95	Box Plot	395
24.2.96	Box Plot	395
24.2.97	Box Plot	395
24.2.98	Box Plot	395
24.2.99	Box Plot	395
24.3	UNCLES Method - Other Types of External Specifications	390
24.4	M-N Scatter Plots Technique	391
24.5	Parameter-free UNCLES with M-N Plots	393
24.6	Discussion and Summary	393
	References	394
	Appendix	395
	Index	409
24.2.1	Introduction	386
24.2.2	Methods and Examples	388
24.2.3	Box Plot	391
24.2.4	Box Plot	393
24.2.5	Box Plot	393
24.2.6	Box Plot	394
24.2.7	Box Plot	395
24.2.8	Box Plot	395
24.2.9	Box Plot	395
24.2.10	Box Plot	395
24.2.11	Box Plot	395
24.2.12	Box Plot	395
24.2.13	Box Plot	395
24.2.14	Box Plot	395
24.2.15	Box Plot	395
24.2.16	Box Plot	395
24.2.17	Box Plot	395
24.2.18	Box Plot	395
24.2.19	Box Plot	395
24.2.20	Box Plot	395
24.2.21	Box Plot	395
24.2.22	Box Plot	395
24.2.23	Box Plot	395
24.2.24	Box Plot	395
24.2.25	Box Plot	395
24.2.26	Box Plot	395
24.2.27	Box Plot	395
24.2.28	Box Plot	395
24.2.29	Box Plot	395
24.2.30	Box Plot	395
24.2.31	Box Plot	395
24.2.32	Box Plot	395
24.2.33	Box Plot	395
24.2.34	Box Plot	395
24.2.35	Box Plot	395
24.2.36	Box Plot	395
24.2.37	Box Plot	395
24.2.38	Box Plot	395
24.2.39	Box Plot	395
24.2.40	Box Plot	395
24.2.41	Box Plot	395
24.2.42	Box Plot	395
24.2.43	Box Plot	395
24.2.44	Box Plot	395
24.2.45	Box Plot	395
24.2.46	Box Plot	395
24.2.47	Box Plot	395
24.2.48	Box Plot	395
24.2.49	Box Plot	395
24.2.50	Box Plot	395
24.2.51	Box Plot	395
24.2.52	Box Plot	395
24.2.53	Box Plot	395
24.2.54	Box Plot	395
24.2.55	Box Plot	395
24.2.56	Box Plot	395
24.2.57	Box Plot	395
24.2.58	Box Plot	395
24.2.59	Box Plot	395
24.2.60	Box Plot	395
24.2.61	Box Plot	395
24.2.62	Box Plot	395
24.2.63	Box Plot	395
24.2.64	Box Plot	395
24.2.65	Box Plot	395
24.2.66	Box Plot	395
24.2.67	Box Plot	395
24.2.68	Box Plot	395
24.2.69	Box Plot	395
24.2.70	Box Plot	395
24.2.71	Box Plot	395
24.2.72	Box Plot	395
24.2.73	Box Plot	395
24.2.74	Box Plot	395
24.2.75	Box Plot	395
24.2.76	Box Plot	395
24.2.77	Box Plot	395
24.2.78	Box Plot	395
24.2.79	Box Plot	395
24.2.80	Box Plot	395
24.2.81	Box Plot	395
24.2.82	Box Plot	395
24.2.83	Box Plot	395
24.2.84	Box Plot	395
24.2.85	Box Plot	395
24.2.86	Box Plot	395
24.2.87	Box Plot	395
24.2.88	Box Plot	395
24.2.89	Box Plot	395
24.2.90	Box Plot	395
24.2.91	Box Plot	395
24.2.92	Box Plot	395
24.2.93	Box Plot	395
24.2.94	Box Plot	395
24.2.95	Box Plot	395
24.2.96	Box Plot	395
24.2.97	Box Plot	395
24.2.98	Box Plot	395
24.2.99	Box Plot	395