

# Contents

Preface	v
1 Introduction	1
1.1 Brief history of bioinformatics	3
1.2 Database application in bioinformatics	6
1.3 Web tools and services for sequence homology	8
Alignment	
1.3.1 Web tools and services for protein functional site identification	9
1.3.2 Web tools and services for other biological data	10
1.4 Pattern analysis	10
1.5 The contribution of information technology	11
1.6 Chapters	12
2 Introduction to Unsupervised Learning	15
3 Probability Density Estimation Approaches	24
3.1 Histogram approach	24
3.2 Parametric approach	25
3.3 Non-parametric approach	28
3.3.1 K-nearest neighbour approach	28
3.3.2 Kernel approach	29
Summary	36
4 Dimension Reduction	38
4.1 General	38

4.2	Principal component analysis	39
4.3	An application of PCA	42
4.4	Multi-dimensional scaling	46
4.5	Application of the Sammon algorithm to gene data	48
	Summary	50
5	Cluster Analysis	52
5.1	Hierarchical clustering	52
5.2	K-means	55
5.3	Fuzzy C-means	58
5.4	Gaussian mixture models	60
5.5	Application of clustering algorithms to the <i>Burkholderia pseudomallei</i> gene expression data	64
	Summary	67
6	Self-organising Map	69
6.1	Vector quantization	69
6.2	SOM structure	73
6.3	SOM learning algorithm	75
6.4	Using SOM for classification	79
6.5	Bioinformatics applications of VQ and SOM	81
6.5.1	Sequence analysis	81
6.5.2	Gene expression data analysis	83
6.5.3	Metabolite data analysis	86
6.6	A case study of gene expression data analysis	86
6.7	A case study of sequence data analysis	88
	Summary	90
7	Introduction to Supervised Learning	92
7.1	General concepts	92
7.2	General definition	94
7.3	Model evaluation	96
7.4	Data organisation	101
7.5	Bayes rule for classification	103
	Summary	103

8	Linear/Quadratic Discriminant Analysis and K-nearest Neighbour	104
8.1	Linear discriminant analysis	104
8.2	Generalised discriminant analysis	109
8.3	K-nearest neighbour	111
8.4	KNN for gene data analysis	118
	Summary	118
9	Classification and Regression Trees, Random Forest Algorithm	120
9.1	Introduction	120
9.2	Basic principle for constructing a classification tree	121
9.3	Classification and regression tree	125
9.4	CART for compound pathway involvement prediction	126
9.5	The random forest algorithm	128
9.6	RF for analyzing <i>Burkholderia pseudomallei</i> gene expression profiles	129
	Summary	132
10	Multi-layer Perceptron	133
10.1	Introduction	133
10.2	Learning theory	137
10.2.1	Parameterization of a neural network	137
10.2.2	Learning rules	137
10.3	Learning algorithms	145
10.3.1	Regression	145
10.3.2	Classification	146
10.3.3	Procedure	147
10.4	Applications to bioinformatics	148
10.4.1	Bio-chemical data analysis	148
10.4.2	Gene expression data analysis	149
10.4.3	Protein structure data analysis	149
10.4.4	Bio-marker identification	150
10.5	A case study on <i>Burkholderia pseudomallei</i> gene expression data	150
	Summary	153

11	Basis Function Approach and Vector Machines	154
11.1	Introduction	154
11.2	Radial-basis function neural network (RBFNN)	156
11.3	Bio-basis function neural network	162
11.4	Support vector machine	168
11.5	Relevance vector machine	173
	Summary	176
12	Hidden Markov Model	177
12.1	Markov model	177
12.2	Hidden Markov model	179
12.2.1	General definition	179
12.2.2	Handling HMM	183
12.2.3	Evaluation	184
12.2.4	Decoding	188
12.2.5	Learning	189
12.3	HMM for sequence classification	191
	Summary	194
13	Feature Selection	195
13.1	Built-in strategy	195
13.1.1	Lasso regression	196
13.1.2	Ridge regression	199
13.1.3	Partial least square regression (PLS) algorithm	200
13.2	Exhaustive strategy	204
13.3	Heuristic strategy – orthogonal least square approach	204
13.4	Criteria for feature selection	208
13.4.1	Correlation measure	209
13.4.2	Fisher ratio measure	210
13.4.3	Mutual information approach	210
	Summary	212
14	Feature Extraction (Biological Data Coding)	213
14.1	Molecular sequences	214
14.2	Chemical compounds	215

14.3	General definition	216
14.4	Sequence analysis	216
14.4.1	Peptide feature extraction	216
14.4.2	Whole sequence feature extraction	222
	Summary	224
15	Sequence/Structural Bioinformatics Foundation – Peptide Classification	225
15.1	Nitration site prediction	225
15.2	Plant promoter region prediction	230
	Summary	237
16	Gene Network – Causal Network and Bayesian Networks	238
16.1	Gene regulatory network	238
16.2	Causal networks, networks, graphs	241
16.3	A brief review of the probability	242
16.4	Discrete Bayesian network	245
16.5	Inference with discrete Bayesian network	246
16.6	Learning discrete Bayesian network	247
16.7	Bayesian networks for gene regulatory networks	247
16.8	Bayesian networks for discovering peptide patterns	248
16.9	Bayesian networks for analysing <i>Burkholderia</i> <i>pseudomallei</i> gene data	249
	Summary	252
17	S-Systems	253
17.1	Michealis-Menten change law	253
17.2	S-system	256
17.3	Simplification of an S-system	259
17.4	Approaches for structure identification and parameter estimation	260
17.4.1	Neural network approach	260
17.4.2	Simulated annealing approach	261
17.4.3	Evolutionary computation approach	262

17.5 Steady-state analysis of an S-system	262
17.6 Sensitivity of an S-system	267
Summary	268
18 Future Directions	269
18.1 Multi-source data	270
18.2 Gene regulatory network construction	272
18.3 Building models using incomplete data	274
18.4 Biomarker detection from gene expression data	275
Summary	278
References	279
Index	319