

# Contents

<b>1</b>	<b>Introduction to Pattern Recognition and Bioinformatics. . . . .</b>	<b>1</b>
1.1	Introduction . . . . .	1
1.2	Basics of Molecular Biology . . . . .	3
1.2.1	Nucleic Acids . . . . .	5
1.2.2	Proteins . . . . .	5
1.3	Bioinformatics Tasks for Biological Data . . . . .	6
1.3.1	Alignment and Comparison of DNA, RNA, and Protein Sequences. . . . .	6
1.3.2	Identification of Genes and Functional Sites from DNA Sequences . . . . .	7
1.3.3	Prediction of Protein Functional Sites . . . . .	8
1.3.4	DNA and RNA Structure Prediction . . . . .	9
1.3.5	Protein Structure Prediction and Classification. . . . .	9
1.3.6	Molecular Design and Molecular Docking. . . . .	10
1.3.7	Phylogenetic Trees for Studying Evolutionary Relationship. . . . .	11
1.3.8	Analysis of Microarray Expression Data . . . . .	11
1.4	Pattern Recognition Perspective . . . . .	15
1.4.1	Pattern Recognition . . . . .	16
1.4.2	Relevance of Soft Computing . . . . .	20
1.5	Scope and Organization of the Book. . . . .	22
	References . . . . .	26

## Part I Classification

<b>2</b>	<b>Neural Network Tree for Identification of Splice Junction and Protein Coding Region in DNA . . . . .</b>	<b>45</b>
2.1	Introduction . . . . .	45
2.2	Neural Network Based Tree-Structured Pattern Classifier . . . . .	47
2.2.1	Selection of Multilayer Perceptron . . . . .	49
2.2.2	Splitting and Stopping Criteria. . . . .	50

2.3	Identification of Splice-Junction in DNA Sequence . . . . .	51
2.3.1	Description of Data Set . . . . .	52
2.3.2	Experimental Results . . . . .	52
2.4	Identification of Protein Coding Region in DNA Sequence . . .	53
2.4.1	Data and Method . . . . .	56
2.4.2	Feature Set . . . . .	57
2.4.3	Experimental Results . . . . .	59
2.5	Conclusion and Discussion . . . . .	64
	References . . . . .	64
<b>3</b>	<b>Design of String Kernel to Predict Protein Functional</b>	
	<b>Sites Using Kernel-Based Classifiers . . . . .</b>	<b>67</b>
3.1	Introduction . . . . .	67
3.2	String Kernel for Protein Functional Site Identification . . . . .	69
3.2.1	Bio-Basis Function . . . . .	69
3.2.2	Selection of Bio-Basis Strings Using Mutual Information . . . . .	72
3.2.3	Selection of Bio-Basis Strings Using Fisher Ratio . . .	74
3.3	Novel String Kernel Function . . . . .	75
3.3.1	Asymmetry of Biological Dissimilarity . . . . .	75
3.3.2	Novel Bio-Basis Function . . . . .	76
3.4	Biological Dissimilarity Based String Selection Method . . . . .	77
3.4.1	Fisher Ratio Using Biological Dissimilarity . . . . .	78
3.4.2	Nearest Mean Classifier . . . . .	80
3.4.3	Degree of Resemblance . . . . .	81
3.4.4	Details of the Algorithm . . . . .	82
3.4.5	Computational Complexity . . . . .	83
3.5	Quantitative Measure . . . . .	83
3.5.1	Compactness: $\alpha$ Index . . . . .	83
3.5.2	Cluster Separability: $\beta$ Index . . . . .	84
3.5.3	Class Separability: $\gamma$ Index . . . . .	84
3.6	Experimental Results . . . . .	85
3.6.1	Support Vector Machine . . . . .	86
3.6.2	Description of Data Set . . . . .	87
3.6.3	Illustrative Example . . . . .	89
3.6.4	Performance of Different String Selection Methods . . . . .	90
3.6.5	Performance of Novel Bio-Basis Function . . . . .	98
3.7	Conclusion and Discussion . . . . .	99
	References . . . . .	100

## Part II Feature Selection

<b>4</b>	<b>Rough Sets for Selection of Molecular Descriptors to Predict Biological Activity of Molecules.</b>	105
4.1	Introduction	105
4.2	Basics of Rough Sets	108
4.3	Rough Set-Based Molecular Descriptor Selection Algorithm	111
4.3.1	Maximum Relevance-Maximum Significance Criterion	112
4.3.2	Computational Complexity	114
4.3.3	Generation of Equivalence Classes	114
4.4	Experimental Results	115
4.4.1	Description of QSAR Data Sets	115
4.4.2	Support Vector Regression Method	116
4.4.3	Optimum Number of Equivalence Classes	117
4.4.4	Performance Analysis	117
4.4.5	Comparative Performance Analysis	122
4.5	Conclusion and Discussion	125
	References	126
<b>5</b>	<b><i>f</i>-Information Measures for Selection of Discriminative Genes from Microarray Data</b>	131
5.1	Introduction	131
5.2	Gene Selection Using <i>f</i> -Information Measures	133
5.2.1	Minimum Redundancy-Maximum Relevance Criterion	134
5.2.2	<i>f</i> -Information Measures for Gene Selection	135
5.2.3	Discretization	138
5.3	Experimental Results	138
5.3.1	Gene Expression Data Sets	139
5.3.2	Class Prediction Methods	139
5.3.3	Performance Analysis	140
5.3.4	Analysis Using Class Separability Index	144
5.4	Conclusion and Discussion	149
	References	150
<b>6</b>	<b>Identification of Disease Genes Using Gene Expression and Protein-Protein Interaction Data</b>	155
6.1	Introduction	155
6.2	Integrated Method for Identifying Disease Genes	157
6.3	Experimental Results	159
6.3.1	Gene Expression Data Set Used	160
6.3.2	Identification of Differentially Expressed Genes	160

6.3.3	Overlap with Known Disease-Related Genes . . . . .	160
6.3.4	PPI Data and Shortest Path Analysis. . . . .	163
6.3.5	Comparative Performance Analysis of Different Methods . . . . .	165
6.4	Conclusion and Discussion . . . . .	167
	References . . . . .	167
<b>7</b>	<b>Rough Sets for Insilico Identification of Differentially Expressed miRNAs. . . . .</b>	<b>171</b>
7.1	Introduction . . . . .	171
7.2	Selection of Differentially Expressed miRNAs. . . . .	174
7.2.1	RSMRMS Algorithm . . . . .	175
7.2.2	Fuzzy Discretization . . . . .	176
7.2.3	B.632+ Error Rate . . . . .	179
7.3	Experimental Results . . . . .	180
7.3.1	Data Sets Used. . . . .	180
7.3.2	Optimum Values of Different Parameters . . . . .	181
7.3.3	Importance of B.632+ Error Rate . . . . .	182
7.3.4	Role of Fuzzy Discretization Method . . . . .	185
7.3.5	Comparative Performance Analysis. . . . .	186
7.4	Conclusion and Discussion . . . . .	189
	References . . . . .	191

### Part III Clustering

<b>8</b>	<b>Grouping Functionally Similar Genes from Microarray Data Using Rough-Fuzzy Clustering. . . . .</b>	<b>197</b>
8.1	Introduction . . . . .	197
8.2	Clustering Algorithms and Validity Indices . . . . .	200
8.2.1	Different Gene Clustering Algorithms. . . . .	200
8.2.2	Quantitative Measures. . . . .	205
8.3	Grouping Functionally Similar Genes Using Rough-Fuzzy C-Means Algorithm . . . . .	207
8.3.1	Rough-Fuzzy C-Means . . . . .	207
8.3.2	Initialization Method. . . . .	210
8.3.3	Identification of Optimum Parameters. . . . .	211
8.4	Experimental Results . . . . .	212
8.4.1	Gene Expression Data Sets Used . . . . .	212
8.4.2	Optimum Values of Different Parameters . . . . .	213
8.4.3	Importance of Correlation-Based Initialization Method. . . . .	214
8.4.4	Performance Analysis of Different C-Means Algorithms. . . . .	216

8.4.5	Comparative Performance of CLICK, SOM, and RFCM . . . . .	216
8.4.6	Eisen Plots. . . . .	216
8.4.7	Biological Significance Analysis . . . . .	217
8.4.8	Functional Consistency of Clustering Result . . . . .	220
8.5	Conclusion and Discussion . . . . .	221
	References . . . . .	221
<b>9</b>	<b>Mutual Information Based Supervised Attribute Clustering for Microarray Sample Classification . . . . .</b>	<b>225</b>
9.1	Introduction . . . . .	225
9.2	Clustering Genes for Sample Classification . . . . .	227
9.2.1	Gene Clustering: Supervised Versus Unsupervised . . . . .	227
9.2.2	Criteria for Gene Selection and Clustering. . . . .	228
9.3	Supervised Gene Clustering Algorithm . . . . .	229
9.3.1	Supervised Similarity Measure . . . . .	229
9.3.2	Gene Clustering Algorithm . . . . .	232
9.3.3	Fundamental Property . . . . .	235
9.3.4	Computational Complexity . . . . .	235
9.4	Experimental Results . . . . .	236
9.4.1	Gene Expression Data Sets Used . . . . .	236
9.4.2	Optimum Value of Threshold. . . . .	237
9.4.3	Qualitative Analysis of Supervised Clusters. . . . .	238
9.4.4	Importance of Supervised Similarity Measure . . . . .	239
9.4.5	Importance of Augmented Genes . . . . .	240
9.4.6	Performance of Coarse and Finer Clusters. . . . .	243
9.4.7	Comparative Performance Analysis. . . . .	246
9.4.8	Biological Significance Analysis . . . . .	249
9.5	Conclusion and Discussion . . . . .	249
	References . . . . .	250
<b>10</b>	<b>Possibilistic Biclustering for Discovering Value-Coherent Overlapping <math>\delta</math>-Biclusters. . . . .</b>	<b>253</b>
10.1	Introduction . . . . .	253
10.2	Biclustering and Possibilistic Clustering . . . . .	256
10.2.1	Basics of Biclustering . . . . .	256
10.2.2	Possibilistic Clustering . . . . .	258
10.3	Possibilistic Biclustering Algorithm . . . . .	259
10.3.1	Objective Function . . . . .	259
10.3.2	Bicluster Means . . . . .	261
10.3.3	Convergence Condition . . . . .	262
10.3.4	Details of the Algorithm . . . . .	263
10.3.5	Termination Condition . . . . .	265
10.3.6	Selection of Initial Biclusters . . . . .	265

10.4	Quantitative Indices . . . . .	266
10.4.1	Average Number of Genes . . . . .	266
10.4.2	Average Number of Conditions . . . . .	267
10.4.3	Average Volume . . . . .	267
10.4.4	Average Mean Squared Residue . . . . .	267
10.4.5	Degree of Overlapping . . . . .	268
10.5	Experimental Results . . . . .	268
10.5.1	Optimum Values of Different Parameters . . . . .	269
10.5.2	Analysis of Generated Biclusters . . . . .	270
10.5.3	Comparative Analysis of Different Methods . . . . .	272
10.6	Conclusion and Discussion . . . . .	273
	References . . . . .	274
<b>11</b>	<b>Fuzzy Measures and Weighted Co-Occurrence Matrix for Segmentation of Brain MR Images . . . . .</b>	<b>277</b>
11.1	Introduction . . . . .	277
11.2	Fuzzy Measures and Co-Occurrence Matrix. . . . .	279
11.2.1	Fuzzy Set . . . . .	279
11.2.2	Co-Occurrence Matrix. . . . .	280
11.2.3	Second Order Fuzzy Correlation. . . . .	281
11.2.4	Second Order Fuzzy Entropy . . . . .	281
11.2.5	Second Order Index of Fuzziness . . . . .	282
11.2.6	2D S-Type Membership Function. . . . .	282
11.3	Thresholding Algorithm . . . . .	283
11.3.1	Modification of Co-Occurrence Matrix . . . . .	283
11.3.2	Measure of Ambiguity . . . . .	285
11.3.3	Strength of Ambiguity. . . . .	286
11.4	Experimental Results . . . . .	291
11.5	Conclusion and Discussion . . . . .	295
	References . . . . .	295
	<b>About the Authors. . . . .</b>	<b>299</b>
	<b>Index . . . . .</b>	<b>301</b>

Scalable Pattern Recognition Algorithms  
Applications in Computational Biology and  
Bioinformatics

Maji, P.; Paul, S.

2014, XXII, 304 p. 55 illus., 10 illus. in color., Hardcover

ISBN: 978-3-319-05629-6