

# CONTENTS

<b>Preface</b>	<b>vii</b>
<b>Part I OVERVIEW</b>	<b>1</b>
<b>Chapter 1 Bioinformatics: Mining the Massive Data from High Throughput Genomics Experiments</b>	
<i>Haixu Tang and Sun Kim</i>	
1 Introduction	3
2 Recent Development of Classical Topics	5
2.1 Sequence alignment	5
2.2 Genome sequencing and fragment assembly	8
2.3 Gene annotation	9
2.4 RNA folding	11
2.5 Motif finding	12
2.6 Protein structure prediction	13
3 Emerging Topics from New Genome Technologies	14
3.1 Comparative genomics: beyond genome comparison	15
3.2 Pathway reconstruction	16
3.3 Microarray analysis	17
3.4 Proteomics	18
3.5 Protein-protein interaction	20
4 Conclusion	20
<b>Chapter 2 An Introduction to Soft Computing</b>	
<i>Amit Konar and Swagatam Das</i>	
1 Classical AI and its Pitfalls	25
2 What is Soft Computing?	27
3 Fundamental Components of Soft Computing	28
3.1 Fuzzy sets and fuzzy logic	28

3.2	Neural networks	31
3.3	Genetic algorithms	36
3.4	Belief networks	39
4	Synergism in Soft Computing	44
4.1	Neuro-fuzzy synergism	44
4.2	Neuro-GA synergism	44
4.3	Fuzzy-GA synergism	45
4.4	Neuro-belief network synergism	45
4.5	GA-belief network synergism	45
4.6	Neuro-fuzzy-GA synergism	46
5	Some Emerging Areas of Soft Computing	46
5.1	Artificial life	46
5.2	Particle swarm optimization (PSO)	47
5.3	Artificial immune system	48
5.4	Rough sets and granular computing	49
5.5	Chaos theory	50
5.6	Ant colony systems (ACS)	51
6	Summary	52

## **Part II BIOLOGICAL SEQUENCE AND STRUCTURE ANALYSIS 57**

### **Chapter 3 Reconstructing Phylogenies with Memetic Algorithms and Branch-and-Bound**

*José E. Gallardo, Carlos Cotta and Antonio J. Fernández*

1	Introduction	59
2	A Crash Introduction to Phylogenetic Inference	60
3	Evolutionary Algorithms for the Phylogeny Problem	65
4	A BnB Algorithm for Phylogenetic Inference	66
5	A Memetic Algorithm for Phylogenetic Inference	69
6	A Hybrid Algorithm	73
7	Experimental Results	75
7.1	Experimental setting	76
7.2	Sensitivity analysis on the hybrid algorithm	76
7.3	Analysis of results	77
8	Conclusions	80

## **Chapter 4 Classification of RNA Sequences with Support Vector Machines**

*Jason T. L. Wang and Xiaoming Wu*

1	Introduction	85
2	Count Kernels and Marginalized Count Kernels	88
2.1	RNA sequences with known secondary structures	88
2.2	RNA sequences with unknown secondary structures	92
3	Kernel Based on Labeled Dual Graphs	94
3.1	Labeled dual graphs	94
3.2	Marginalized kernel for labeled dual graphs	95
4	A New Kernel	97
4.1	Extracting features for global structural information	98
4.2	Extracting features for local structural information	100
5	Experiments and Results	102
5.1	Data and parameters	102
5.2	Results	104
6	Conclusion	106

## **Chapter 5 Beyond String Algorithms: Protein Sequence Analysis using Wavelet Transforms**

*Arun Krishnan and Kuo-Bin Li*

1	Introduction	109
1.1	String algorithms	110
1.2	Sequence analysis	110
1.3	Wavelet transform	111
2	Motif Searching	114
2.1	Introduction	114
2.2	Methods	115
2.3	Results	116
2.4	Allergenicity prediction	118
3	Transmembrane Helix Region (HTM) Prediction	121
4	Hydrophobic Cores	122
5	Protein Repeat Motifs	122
6	Sequence Comparison	123
7	Prediction of Protein Secondary Structures	125

8	Disease Related Studies	126
9	Other Functional Prediction	126
10	Conclusion	126

## **Chapter 6 Filtering Protein Surface Motifs Using Negative Instances of Active Sites Candidates**

*Nripendra L. Shrestha and Takenao Ohkawa*

1	Introduction	133
2	Protein Structural Data and Surface Motifs	135
2.1	Protein structural data	135
2.2	Protein molecular surface data	136
2.3	Functions of a protein and structural motifs	137
3	Overview of SUMOMO	138
3.1	Surface motif extraction	139
3.2	Filtering using similarity between local surfaces	140
3.3	Problems with SUMOMO	142
4	Filtering Surface Motifs using Negative Instances of Protein Active Sites Candidates	142
4.1	Survey on the features to distinguish real active sites from the active sites candidates	143
4.2	Ranking active sites candidates	147
5	Evaluations	148
6	Conclusions and Future Works	151

## **Chapter 7 Distill: A Machine Learning Approach to *Ab Initio* Protein Structure Prediction**

*Gianluca Pollastri, Davide Baú and Alessandro Vullo*

1	Introduction	153
2	Structural Features	155
2.1	One-dimensional structural features	155
2.2	Two-dimensional structural features	157
3	Review of Statistical Learning Methods Applied	159
3.1	RNNs for undirected graphs	159
3.2	1D DAG-RNN	161
3.3	2D DAG-RNN	163

4	Predictive Architecture	164
4.1	Data set generation	165
4.2	Training protocols	165
4.3	One-dimensional feature predictors	166
4.4	Two-dimensional feature predictors	168
5	Modeling Protein Backbones	169
5.1	Protein representation	170
5.2	Constraints-based pseudo energy	170
5.3	Optimization algorithm	171
6	Reconstruction Results	173
7	Conclusions	178

## **Chapter 8 In Silico Design of Ligands using Properties of Target Active Sites**

*Sanghamitra Bandyopadhyay, Santanu Santra,  
Ujjwal Maulik and Heinz Muehlenbein*

1	Introduction	184
2	Relevance of Genetic Algorithm for Drug Design	186
3	Basic Issues	187
3.1	Core formation	187
3.2	Chromosome representation	190
3.3	Fitness computation	191
4	Main Algorithm	192
5	Experimental Results	193
6	Discussion	199

## **Part III GENE EXPRESSION AND MICROARRAY DATA ANALYSIS 203**

### **Chapter 9 Inferring Regulations in a Genomic Network from Gene Expression Profiles**

*Nasimul Noman and Hitoshi Iba*

1	Introduction	205
2	Modeling Gene Regulatory Networks by S-system	208
2.1	Canonical model description	208

2.2	Genetic network inference problem by S-system	209
2.3	Decoupled S-system model	210
2.4	Fitness function for skeletal network structure	211
3	Inference Method	212
3.1	Trigonometric Differential Evolution (TDE)	213
3.2	Proposed algorithm	214
3.3	Local search procedure	217
4	Simulated Experiment	217
4.1	Experiment 1: inferring small scale network in noise free environment	217
4.2	Experiment 2: inferring small scale network in noisy environment	219
4.3	Experiment 3: inferring medium scale network in noisy environment	220
5	Analysis of Real Gene Expression Data	222
5.1	Experimental data set	224
6	Discussion	226
7	Conclusion	227

## **Chapter 10 A Reliable Classification of Gene Clusters for Cancer Samples Using a Hybrid Multi-Objective Evolutionary Procedure**

*Kalyanmoy Deb, A. Raji Reddy and Shamik Chaudhuri*

1	Introduction	232
2	Class Prediction Procedure	233
2.1	Two-class classification	234
2.2	Multi-class classification	235
3	Evolutionary Gene Selection Procedure	236
3.1	The optimization problem	237
3.2	A multi-objective evolutionary algorithm	237
3.3	A multi-modal NSGA-II	238
3.4	Genetic operators and modified domination operator	240
3.5	NSGA-II search using a fixed classifier size	241
3.6	Overall procedure	241
4	Simulation Results	242
4.1	Complete leukemia study	244
4.2	Diffuse large B-cell lymphoma dataset	248

4.3	Colon cancer dataset	250
4.4	NCI60 multi-class tumor dataset	252
5	Conclusions	255

## **Chapter 11 Feature Selection for Cancer Classification using Ant Colony Optimization and Support Vector Machines**

*A. Gupta, V. K. Jayaraman and B. D. Kulkarni*

1	Introduction	259
2	Ant Colony Optimization	262
3	Support Vector Machines	263
4	Proposed Ant Algorithm	266
4.1	State transition rules	266
4.2	Evaluation procedure	268
4.3	Global updating rule	268
4.4	Local updating rule	269
5	Algorithm Outline	269
6	Experiments	270
6.1	Datasets	270
6.2	Preprocessing	272
6.3	Experimental setup	273
7	Results and Discussion	274
8	Conclusions	277

## **Chapter 12 Sophisticated Methods for Cancer Classification using Microarray Data**

*Sung-Bae Cho and Han-Saem Park*

1	Introduction	281
2	Backgrounds	282
2.1	DNA microarray	282
2.2	Feature selection methods	283
2.3	Base classifiers	284
2.4	Classifier ensemble methods	285
3	Sophisticated Methods for Cancer Classification	286
3.1	Ensemble with negatively correlated features	286
3.2	Combinatorial ensemble	289
3.3	Searching optimal ensemble with GA	291

4	Experiments	293
4.1	Datasets	293
4.2	Ensemble with negative correlated features	294
4.3	Combinatorial ensemble	296
4.4	Optimal ensemble with GA	298
5	Conclusions	300
<b>Chapter 13 Multiobjective Evolutionary Approach to Fuzzy Clustering of Microarray Data</b>		
<i>Anirban Mukhopadhyay, Ujjwal Maulik and Sanghamitra Bandyopadhyay</i>		
1	Introduction	304
2	Structure of Gene Expression Data Sets	306
3	Cluster Analysis	306
3.1	K-means	307
3.2	K-medoids	308
3.3	Fuzzy C-means	308
3.4	Hierarchical agglomerative clustering	309
4	Multiobjective Genetic Algorithms	310
5	The Multiobjective Fuzzy Clustering Technique	312
5.1	Chromosome representation and population initialization	312
5.2	Computation of objective functions	312
5.3	Selection, crossover and mutation	313
5.4	Choice of objectives	314
5.5	Distance measures	314
6	Experimental Results	315
6.1	Yeast sporulation data	315
6.2	Human fibroblasts serum data	316
6.3	Performance validation	316
6.4	Input parameter values	318
6.5	Quantitative assessments	318
6.6	Visualization of results	320
6.7	Biological interpretation	322
7	Conclusions and Discussions	326
	<b>Index</b>	<b>329</b>