# CONTENTS

<table>
<thead>
<tr>
<th>Contributors</th>
<th>xiii</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preface</td>
<td>xvii</td>
</tr>
</tbody>
</table>

## 1 Current Methods for Protein Secondary-Structure Prediction Based on Support Vector Machines

_Hae-Jin Hu, Robert W. Harrison, Phang C. Tai, and Yi Pan_

1.1 Traditional Methods
   - 1.1.1 Statistical Approaches 2
   - 1.1.2 Machine Learning Approaches 2

1.2 Support Vector Machine Method
   - 1.2.1 Introduction to SVM 8
   - 1.2.2 Encoding Profile 10
   - 1.2.3 Kernel Functions 11
   - 1.2.4 Tertiary Classifier Design 15
   - 1.2.5 Accuracy Measure of SVM 20

1.3 Performance Comparison of SVM Methods 22
1.4 Discussion and Conclusions 23
References 23

## 2 Comparison of Seven Methods for Mining Hidden Links

_Xiaohua Hu, Xiaodan Zhang, and Xiaohua Zhou_

2.1 Analysis of the Literature on Raynaud’s Disease 27
2.2 Related Work 29
2.3 Methods 30
  2.3.1 Information Measures 31
  2.3.2 Ranking Methods 31
  2.3.3 Seven Methods 32

2.4 Experiment Results and Analysis 37
  2.4.1 Data Set 37
  2.4.2 Chi-Square, Chi-Square Association Rule, and Mutual Information Link ABC Methods Compared 38
  2.4.3 Chi-Square ABC Method: Semantic Check for Mining Implicit Connections 38
  2.4.4 Chi-Square and Mutual Information Link ABC Methods 40

2.5 Discussion and Conclusions 43
Acknowledgments 43
References 44

3 Voting Scheme–Based Evolutionary Kernel Machines for Drug Activity Comparisons 45
Bo Jin and Yan-Qing Zhang

3.1 Granular Kernel and Kernel Tree Design 46
  3.1.1 Definitions 46
  3.1.2 Granular Kernel Properties 47
3.2 GKTSEs 48
3.3 Evolutionary Voting Kernel Machines 51
3.4 Simulations 53
  3.4.1 Data Set and Experimental Setup 53
  3.4.2 Experimental Results and Comparisons 53
3.5 Conclusions and Future Work 54
Acknowledgments 55
References 55

4 Bioinformatics Analyses of Arabidopsis thaliana Tiling Array Expression Data 57
Trupti Joshi, Jinrong Wan, Curtis J. Palm, Kara Juneau, Ron Davis, Audrey Southwick, Katrina M. Ramonell, Gary Stacey, and Dong Xu

4.1 Tiling Array Design and Data Description 58
  4.1.1 Data 58
  4.1.2 Tiling Array Expression Patterns 59
  4.1.3 Tiling Array Data Analysis 59
4.2 Ontology Analyses 61
4.3 Antisense Regulation Identification 63
  4.3.1 Antisense Silencing 63
  4.3.2 Antisense Regulation Identification 63
4.4 Correlated Expression Between Two DNA Strands 67
5 Identification of Marker Genes from High-Dimensional Microarray Data for Cancer Classification 71

Jiexun Li, Hua Su, and Hsinchun Chen

5.1 Feature Selection 73
  5.1.1 Taxonomy of Feature Selection 73
  5.1.2 Evaluation Criterion 73
  5.1.3 Generation Procedure 76
5.2 Gene Selection 78
  5.2.1 Individual Gene Ranking 78
  5.2.2 Gene Subset Selection 79
  5.2.3 Summary of Gene Selection 82
5.3 Comparative Study of Gene Selection Methods 83
  5.3.1 Microarray Data Descriptions 83
  5.3.2 Gene Selection Approaches 83
  5.3.3 Experimental Results 84
5.4 Conclusions and Discussion 85
Acknowledgments 85
References 85

6 Patient Survival Prediction from Gene Expression Data 89

Huiqing Liu, Limsoon Wong, and Ying Xu

6.1 General Methods 91
  6.1.1 Kaplan–Meier Survival Analysis 91
  6.1.2 Cox Proportional-Hazards Regression 93
6.2 Applications 95
  6.2.1 Diffuse Large-B-Cell Lymphoma 95
  6.2.2 Lung Adenocarcinoma 97
  6.2.3 Remarks 98
6.3 Incorporating Data Mining Techniques to Survival Prediction 98
  6.3.1 Gene Selection by Statistical Properties 99
  6.3.2 Cancer Subtype Identification via Survival Information 100
6.4 Selection of Extreme Patient Samples 103
  6.4.1 Short- and Long-Term Survivors 103
  6.4.2 SVM-Based Risk Scoring Function 103
  6.4.3 Results 104
6.5 Summary and Concluding Remarks 108
Acknowledgments 109
References 109
9 Public Genomic Databases: Data Representation, Storage, and Access

Andrew Robinson, Wenny Rahayu, and David Taniar

9.1 Data Representation
9.1.1 FASTA Format
9.1.2 Genbank Format
9.1.3 Swiss-Prot Format
9.1.4 XML Format

9.2 Data Storage
9.2.1 Multidatabase Repositories

9.3 Data Access
9.3.1 Single-Database Access Point
9.3.2 Cross-Reference Databases
9.3.3 Multiple-Database Access Points
9.3.4 Tool-Based Interfaces

9.4 Discussion
9.5 Conclusions

References

10 Automatic Query Expansion with Keyphrases and POS Phrase Categorization for Effective Biomedical Text Mining

Min Song and Il-Yeol Song

10.1 Keyphrase Extraction-Based Pseudo-Relevance Feedback
10.1.1 Keyphrase Extraction Procedures
10.1.2 Keyphrase Ranking
10.1.3 Query Translation into DNF

10.2 Query Expansion with WordNet

10.3 Experiments on Medline Data Sets

10.4 Conclusions

References

11 Evolutionary Dynamics of Protein–Protein Interactions

L. S. Swapna, B. Offmann, and N. Srinivasan

11.1 Class I Glutamine Amidotransferase–Like Superfamily
11.1.1 DJ-1/PfpI Family
11.1.2 Comparison of Quaternary Structures of DJ-1 Family Members

11.2 Drifts in Interfaces of Close Homologs
11.2.1 Comparison of Quaternary Structures of Intracellular Protease and Hypothetical Protein YhbO
11.2.2 Comparison of Quaternary Structures of Intracellular Protease and DJ-1
13.3.2 Synthesis of Translation Networks 258
13.3.3 Essentiality of Proteins in Translation Networks 261
13.3.4 Cellular Localization of Proteins in Translation Networks 261
13.3.5 Translation Networks and Protein Phosphorylation 261
13.4 Conclusions 264
Acknowledgments 265
References 265

14 Identification of Transmembrane Proteins Using Variants of the Self-Organizing Feature Map Algorithm 267
Mary Qu Yang, Jack Y. Yang, and Craig W. Codrington

14.1 Physiochemical Analysis of Proteins 268
14.1.1 Hydropathy 270
14.1.2 Polarity 271
14.1.3 Flexibility 272
14.1.4 Electronic Polarizability and van der Waals Volume 273
14.1.5 Bulkiness 273
14.1.6 Electronic Effects 273
14.1.7 Summary 275
14.2 Variants of the SOM Algorithm 275
14.2.1 The SOGR Algorithm 276
14.2.2 Variants of the SOGR Algorithm 277
14.2.3 Improving Classifier Performance by Filtering 283
14.3 Results 286
14.4 Discussion and Conclusions 288
References 290

15 triCLUSTER: Mining Coherent Clusters in Three-Dimensional Microarray Data 295
Lizhuang Zhao and Mohammed J. Zaki

15.1 Preliminary Concepts 296
15.2 Related Work 300
15.2.1 Time-Based Microarray Clustering 300
15.2.2 Feature- and Graph-Based Clustering 301
15.2.3 Pattern-Based Clustering 302
15.3 The triCLUSTER Algorithm 303
15.3.1 Constructing a Range Multigraph 303
15.3.2 Mining Biclusters from a Range Multigraph 305
15.3.3 Getting Triclusters from a Bicluster Graph 307
15.3.4 Merging and Pruning Clusters 308
15.3.5 Complexity Analysis 309
15.4 Experiments 309
15.4.1 Results from Synthetic Data Sets 310
15.4.2 Results from Real Microarray Datasets

15.5 Conclusions

Acknowledgments

References

16 Clustering Methods in a Protein–Protein Interaction Network

Chuan Lin, Young-Rae Cho, Woo-Chang Hwang, Pengjun Pei, and Aidong Zhang

16.1 Protein–Protein Interaction

16.1.1 Proteome in Bioinformatics

16.1.2 Significance of Protein–Protein Interaction

16.1.3 Experimental Approaches for PPI Detection

16.1.4 Computational Methods to Predict Protein–Protein Interaction

16.2 Properties of PPI Networks

16.2.1 PPI Network Representation

16.2.2 Characteristics of Protein–Protein Networks

16.3 Clustering Approaches

16.3.1 Significance of Clustering in a PPI Network

16.3.2 Challenges of Clustering in PPI Networks

16.3.3 Distance-Based Clustering

16.3.4 Graph-Based Clustering

16.4 Validation

16.4.1 Validation Based on Agreement with Annotated Protein Function Databases

16.4.2 Validation Based on the Definition of Clustering

16.4.3 Validation Based on the Reliability of Clusters

16.4.4 Validation for Protein Function Prediction: Leave-One-Out Method

16.5 Conclusions

References

Index