

# CONTENTS

|                                                                                                      |             |
|------------------------------------------------------------------------------------------------------|-------------|
| <b>Contributors</b>                                                                                  | <b>xiii</b> |
| <b>Preface</b>                                                                                       | <b>xvii</b> |
| <b>1 Current Methods for Protein Secondary-Structure Prediction Based on Support Vector Machines</b> | <b>1</b>    |
| <i>Hae-Jin Hu, Robert W. Harrison, Phang C. Tai, and Yi Pan</i>                                      |             |
| 1.1 Traditional Methods                                                                              | 2           |
| 1.1.1 Statistical Approaches                                                                         | 2           |
| 1.1.2 Machine Learning Approaches                                                                    | 2           |
| 1.2 Support Vector Machine Method                                                                    | 8           |
| 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          |
| <b>2 Comparison of Seven Methods for Mining Hidden Links</b>                                         | <b>27</b>   |
| <i>Xiaohua Hu, Xiaodan Zhang, and Xiaohua Zhou</i>                                                   |             |
| 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        |
| <b>3</b> | <b>Voting Scheme-Based Evolutionary Kernel Machines for Drug Activity Comparisons</b>                                                     | <b>45</b> |
|          | <i>Bo Jin and Yan-Qing Zhang</i>                                                                                                          |           |
| 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        |
| <b>4</b> | <b>Bioinformatics Analyses of <i>Arabidopsis thaliana</i> Tiling Array Expression Data</b>                                                | <b>57</b> |
|          | <i>Trupti Joshi, Jinrong Wan, Curtis J. Palm, Kara Juneau, Ron Davis, Audrey Southwick, Katrina M. Ramonell, Gary Stacey, and Dong Xu</i> |           |
| 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        |

|          |                                                                                                       |           |
|----------|-------------------------------------------------------------------------------------------------------|-----------|
| 4.5      | Identification of Nonprotein Coding mRNA                                                              | 68        |
| 4.6      | Summary                                                                                               | 69        |
|          | Acknowledgments                                                                                       | 69        |
|          | References                                                                                            | 70        |
| <b>5</b> | <b>Identification of Marker Genes from High-Dimensional Microarray Data for Cancer Classification</b> | <b>71</b> |
|          | <i>Jiexun Li, Hua Su, and Hsinchun Chen</i>                                                           |           |
| 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        |
| <b>6</b> | <b>Patient Survival Prediction from Gene Expression Data</b>                                          | <b>89</b> |
|          | <i>Huiqing Liu, Limsoon Wong, and Ying Xu</i>                                                         |           |
| 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       |

|                                                            |            |
|------------------------------------------------------------|------------|
| <b>7 RNA Interference and microRNA</b>                     | <b>113</b> |
| <i>Shibin Qiu and Terran Lane</i>                          |            |
| 7.1 Mechanisms and Applications of RNA Interference        | 114        |
| 7.1.1 Mechanism of RNA Interference                        | 114        |
| 7.1.2 Applications of RNAi                                 | 117        |
| 7.1.3 RNAi Computational and Modeling Issues               | 120        |
| 7.2 Specificity of RNA Interference                        | 121        |
| 7.2.1 Computational Representation of RNAi                 | 121        |
| 7.2.2 Definition of Off-Target Error Rates                 | 122        |
| 7.2.3 Feature Maps of Mismatch, Bulge, and Wobble          | 124        |
| 7.2.4 Positional Effect                                    | 125        |
| 7.2.5 Results for RNAi Specificity                         | 125        |
| 7.2.6 Silencing Multiple Genes                             | 128        |
| 7.3 Computational Methods for microRNAs                    | 129        |
| 7.3.1 Prediction of microRNA Genes                         | 130        |
| 7.3.2 Prediction of miRNA Targets                          | 131        |
| 7.4 siRNA Silencing Efficacy                               | 132        |
| 7.4.1 siRNA Design Rules                                   | 132        |
| 7.4.2 Efficacy Prediction with Support Vector Regression   | 134        |
| 7.5 Summary and Open Questions                             | 136        |
| 7.5.1 siRNA Efficacy and Target mRNA Secondary Structures  | 137        |
| 7.5.2 Dynamics of Target mRNA and siRNA                    | 137        |
| 7.5.3 Integration of RNAi into Network Models              | 137        |
| Appendix: Glossary                                         | 138        |
| References                                                 | 140        |
| <b>8 Protein Structure Prediction Using String Kernels</b> | <b>145</b> |
| <i>Huzefa Rangwala, Kevin DeRonne, and George Karypis</i>  |            |
| 8.1 Protein Structure: Granularities                       | 146        |
| 8.1.1 Secondary-Structure Prediction                       | 146        |
| 8.1.2 Protein Tertiary Structure                           | 148        |
| 8.2 Learning from Data                                     | 149        |
| 8.2.1 Kernel Methods                                       | 150        |
| 8.3 Structure Prediction: Capturing the Right Signals      | 150        |
| 8.4 Secondary-Structure Prediction                         | 151        |
| 8.4.1 YASSPP Overview                                      | 152        |
| 8.4.2 Input Sequence Coding                                | 153        |
| 8.4.3 Profile-Based Kernel Functions                       | 154        |
| 8.4.4 Performance Evaluation                               | 154        |
| 8.5 Remote Homology and Fold Prediction                    | 157        |
| 8.5.1 Profile-Based Kernel Functions                       | 158        |
| 8.5.2 Performance Evaluation                               | 161        |
| 8.6 Concluding Remarks                                     | 165        |
| References                                                 | 165        |

|                                                                                                                        |            |
|------------------------------------------------------------------------------------------------------------------------|------------|
| <b>9 Public Genomic Databases: Data Representation, Storage, and Access</b>                                            | <b>169</b> |
| <i>Andrew Robinson, Wenny Rahayu, and David Taniar</i>                                                                 |            |
| 9.1 Data Representation                                                                                                | 170        |
| 9.1.1 FASTA Format                                                                                                     | 170        |
| 9.1.2 Genbank Format                                                                                                   | 171        |
| 9.1.3 Swiss-Prot Format                                                                                                | 172        |
| 9.1.4 XML Format                                                                                                       | 176        |
| 9.2 Data Storage                                                                                                       | 180        |
| 9.2.1 Multidatabase Repositories                                                                                       | 180        |
| 9.3 Data Access                                                                                                        | 183        |
| 9.3.1 Single-Database Access Point                                                                                     | 183        |
| 9.3.2 Cross-Reference Databases                                                                                        | 186        |
| 9.3.3 Multiple-Database Access Points                                                                                  | 186        |
| 9.3.4 Tool-Based Interfaces                                                                                            | 192        |
| 9.4 Discussion                                                                                                         | 194        |
| 9.5 Conclusions                                                                                                        | 194        |
| References                                                                                                             | 194        |
| <b>10 Automatic Query Expansion with Keyphrases and POS Phrase Categorization for Effective Biomedical Text Mining</b> | <b>197</b> |
| <i>Min Song and Il-Yeol Song</i>                                                                                       |            |
| 10.1 Keyphrase Extraction-Based Pseudo-Relevance Feedback                                                              | 198        |
| 10.1.1 Keyphrase Extraction Procedures                                                                                 | 199        |
| 10.1.2 Keyphrase Ranking                                                                                               | 200        |
| 10.1.3 Query Translation into DNF                                                                                      | 202        |
| 10.2 Query Expansion with WordNet                                                                                      | 203        |
| 10.3 Experiments on Medline Data Sets                                                                                  | 203        |
| 10.4 Conclusions                                                                                                       | 205        |
| References                                                                                                             | 206        |
| <b>11 Evolutionary Dynamics of Protein-Protein Interactions</b>                                                        | <b>209</b> |
| <i>L. S. Swapna, B. Offmann, and N. Srinivasan</i>                                                                     |            |
| 11.1 Class I Glutamine Amidotransferase-Like Superfamily                                                               | 211        |
| 11.1.1 DJ-1/PfpI Family                                                                                                | 213        |
| 11.1.2 Comparison of Quaternary Structures of DJ-1 Family Members                                                      | 214        |
| 11.2 Drifts in Interfaces of Close Homologs                                                                            | 214        |
| 11.2.1 Comparison of Quaternary Structures of Intracellular Protease and Hypothetical Protein YhbO                     | 216        |
| 11.2.2 Comparison of Quaternary Structures of Intracellular Protease and DJ-1                                          | 218        |

|           |                                                                                                                          |            |
|-----------|--------------------------------------------------------------------------------------------------------------------------|------------|
| 11.2.3    | Tracing the Shift Between the Topologically Conserved Intracellular Protease and DJ-1 Interfaces                         | 219        |
| 11.2.4    | Tracing the Shift Between the Nonconserved Intracellular Protease and DJ-1 Interfaces                                    | 223        |
| 11.3      | Drifts in Interfaces of Divergent Members                                                                                | 223        |
| 11.3.1    | Comparison of the Quaternary Structures of Putative Sigma Cross-Reacting Protein and DJ-1                                | 225        |
| 11.3.2    | Tracing the Shift Between the Interfaces of Putative Sigma Cross-Reacting Protein and DJ-1                               | 226        |
| 11.4      | Drifts in Interfaces at Extreme Divergence                                                                               | 226        |
| 11.4.1    | Comparison of Quaternary Structures of Intracellular Protease and Catalase                                               | 227        |
| 11.4.2    | Tracing the Drift in Protein-Protein Interaction Surface from One Family to Another: Intracellular Protease and Catalase | 228        |
| 11.5      | Conclusions                                                                                                              | 229        |
|           | Acknowledgments                                                                                                          | 230        |
|           | References                                                                                                               | 230        |
| <b>12</b> | <b>On Comparing and Visualizing RNA Secondary Structures</b>                                                             | <b>233</b> |
|           | <i>Jason T. L. Wang, Dongrong Wen, and Jianghui Liu</i>                                                                  |            |
| 12.1      | Background                                                                                                               | 234        |
| 12.2      | RSmatch                                                                                                                  | 236        |
| 12.2.1    | Secondary-Structure Decomposition                                                                                        | 237        |
| 12.2.2    | Structure Alignment Formalization                                                                                        | 238        |
| 12.2.3    | Algorithmic Framework                                                                                                    | 239        |
| 12.2.4    | Basic Concepts and Definitions                                                                                           | 240        |
| 12.2.5    | Initialization                                                                                                           | 240        |
| 12.2.6    | Filling in the Scoring Table                                                                                             | 241        |
| 12.2.7    | Running-Time Analysis                                                                                                    | 244        |
| 12.3      | RSview                                                                                                                   | 244        |
| 12.4      | Conclusions                                                                                                              | 247        |
|           | References                                                                                                               | 248        |
| <b>13</b> | <b>Integrative Analysis of Yeast Protein Translation Networks</b>                                                        | <b>251</b> |
|           | <i>Daniel D. Wu and Xiaohua Hu</i>                                                                                       |            |
| 13.1      | Protein Biosynthesis and Translation                                                                                     | 252        |
| 13.2      | Methods                                                                                                                  | 253        |
| 13.2.1    | Graph Notation                                                                                                           | 253        |
| 13.2.2    | Data Sets                                                                                                                | 254        |
| 13.2.3    | Analysis of Network Topology                                                                                             | 254        |
| 13.3      | Results                                                                                                                  | 255        |
| 13.3.1    | Global Properties of Protein Translation Networks Within the Full Yeast Interactome                                      | 255        |

|           |                                                                                                             |            |
|-----------|-------------------------------------------------------------------------------------------------------------|------------|
| 13.3.2    | Synthesis of Translation Networks                                                                           | 261        |
| 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        |
| <b>14</b> | <b>Identification of Transmembrane Proteins Using Variants of the Self-Organizing Feature Map Algorithm</b> | <b>267</b> |
|           | <i>Mary Qu Yang, Jack Y. Yang, and Craig W. Codrington</i>                                                  |            |
| 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        |
| <b>15</b> | <b>TRICLUSTER: Mining Coherent Clusters in Three-Dimensional Microarray Data</b>                            | <b>295</b> |
|           | <i>Lizhuang Zhao and Mohammed J. Zaki</i>                                                                   |            |
| 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 Biclusters 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                                           | 310        |
| 15.5      | Conclusions                                                                     | 315        |
|           | Acknowledgments                                                                 | 315        |
|           | References                                                                      | 315        |
| <b>16</b> | <b>Clustering Methods in a Protein-Protein Interaction Network</b>              | <b>319</b> |
|           | <i>Chuan Lin, Young-Rae Cho, Woo-Chang Hwang, Pengjun Pei, and Aidong Zhang</i> |            |
| 16.1      | Protein-Protein Interaction                                                     | 320        |
| 16.1.1    | Proteome in Bioinformatics                                                      | 320        |
| 16.1.2    | Significance of Protein-Protein Interaction                                     | 321        |
| 16.1.3    | Experimental Approaches for PPI Detection                                       | 321        |
| 16.1.4    | Computational Methods to Predict Protein-Protein Interaction                    | 324        |
| 16.2      | Properties of PPI Networks                                                      | 325        |
| 16.2.1    | PPI Network Representation                                                      | 325        |
| 16.2.2    | Characteristics of Protein-Protein Networks                                     | 326        |
| 16.3      | Clustering Approaches                                                           | 327        |
| 16.3.1    | Significance of Clustering in a PPI Network                                     | 327        |
| 16.3.2    | Challenges of Clustering in PPI Networks                                        | 327        |
| 16.3.3    | Distance-Based Clustering                                                       | 328        |
| 16.3.4    | Graph-Based Clustering                                                          | 335        |
| 16.4      | Validation                                                                      | 348        |
| 16.4.1    | Validation Based on Agreement with Annotated Protein Function Databases         | 349        |
| 16.4.2    | Validation Based on the Definition of Clustering                                | 349        |
| 16.4.3    | Validation Based on the Reliability of Clusters                                 | 350        |
| 16.4.4    | Validation for Protein Function Prediction: Leave-One-Out Method                | 350        |
| 16.5      | Conclusions                                                                     | 351        |
|           | References                                                                      | 351        |
|           | <b>Index</b>                                                                    | <b>357</b> |