

Chapman & Hall/CRC  
Data Mining and Knowledge Discovery Series

# Biological Data Mining

Edited by  
Jake Y. Chen  
Stefano Lonardi



**CRC Press**

Taylor & Francis Group

Boca Raton London New York

---

CRC Press is an imprint of the  
Taylor & Francis Group, an **informa** business  
A CHAPMAN & HALL BOOK

---

# Contents

|                                                                                                 |            |
|-------------------------------------------------------------------------------------------------|------------|
| Preface                                                                                         | ix         |
| Editors                                                                                         | xiii       |
| Contributors                                                                                    | xv         |
| <b>Part I Sequence, Structure, and Function</b>                                                 | <b>1</b>   |
| <b>1 Consensus Structure Prediction for RNA Alignments</b>                                      | <b>3</b>   |
| <i>Junilda Spirollari and Jason T. L. Wang</i>                                                  |            |
| <b>2 Invariant Geometric Properties of Secondary Structure Elements in Proteins</b>             | <b>27</b>  |
| <i>Matteo Comin, Concettina Guerra, and Giuseppe Zanotti</i>                                    |            |
| <b>3 Discovering 3D Motifs in RNA</b>                                                           | <b>49</b>  |
| <i>Alberto Apostolico, Giovanni Ciriello, Concettina Guerra, and Christine E. Heitsch</i>       |            |
| <b>4 Protein Structure Classification Using Machine Learning Methods</b>                        | <b>69</b>  |
| <i>Yazhene Krishnaraj and Chandan Reddy</i>                                                     |            |
| <b>5 Protein Surface Representation and Comparison: New Approaches in Structural Proteomics</b> | <b>89</b>  |
| <i>Lee Sael and Daisuke Kihara</i>                                                              |            |
| <b>6 Advanced Graph Mining Methods for Protein Analysis</b>                                     | <b>111</b> |
| <i>Yi-Ping Phoebe Chen, Jia Rong, and Gang Li</i>                                               |            |
| <b>7 Predicting Local Structure and Function of Proteins</b>                                    | <b>137</b> |
| <i>Huzefa Rangwala and George Karypis</i>                                                       |            |

|                 |                                                                                                     |            |
|-----------------|-----------------------------------------------------------------------------------------------------|------------|
| <b>Part II</b>  | <b>Genomics, Transcriptomics,<br/>and Proteomics</b>                                                | <b>161</b> |
| <b>8</b>        | <b>Computational Approaches for Genome Assembly<br/>Validation</b>                                  | <b>163</b> |
|                 | <i>Jeong-Hyeon Choi, Haixu Tang, Sun Kim, and Mihai Pop</i>                                         |            |
| <b>9</b>        | <b>Mining Patterns of Epistasis in Human Genetics</b>                                               | <b>187</b> |
|                 | <i>Jason H. Moore</i>                                                                               |            |
| <b>10</b>       | <b>Discovery of Regulatory Mechanisms from Gene<br/>Expression Variation by eQTL Analysis</b>       | <b>205</b> |
|                 | <i>Yang Huang, Jie Zheng, and Teresa M. Przytycka</i>                                               |            |
| <b>11</b>       | <b>Statistical Approaches to Gene Expression Microarray<br/>Data Preprocessing</b>                  | <b>229</b> |
|                 | <i>Megan Kong, Elizabeth McClellan, Richard H. Scheuermann,<br/>and Monnie McGee</i>                |            |
| <b>12</b>       | <b>Application of Feature Selection and Classification<br/>to Computational Molecular Biology</b>   | <b>257</b> |
|                 | <i>Paola Bertolazzi, Giovanni Felici, and Giuseppe Lancia</i>                                       |            |
| <b>13</b>       | <b>Statistical Indices for Computational and Data Driven<br/>Class Discovery in Microarray Data</b> | <b>295</b> |
|                 | <i>Raffaele Giancarlo, Davide Scaturro, and Filippo Utró</i>                                        |            |
| <b>14</b>       | <b>Computational Approaches to Peptide Retention Time<br/>Prediction for Proteomics</b>             | <b>337</b> |
|                 | <i>Xiang Zhang, Cheolhwan Oh, Catherine P. Riley,<br/>Hyeyoung Cho, and Charles Buck</i>            |            |
| <b>Part III</b> | <b>Functional and Molecular Interaction<br/>Networks</b>                                            | <b>351</b> |
| <b>15</b>       | <b>Inferring Protein Functional Linkage Based on Sequence<br/>Information and Beyond</b>            | <b>353</b> |
|                 | <i>Li Liao</i>                                                                                      |            |
| <b>16</b>       | <b>Computational Methods for Unraveling Transcriptional<br/>Regulatory Networks in Prokaryotes</b>  | <b>377</b> |
|                 | <i>Dongsheng Che and Guojun Li</i>                                                                  |            |
| <b>17</b>       | <b>Computational Methods for Analyzing and Modeling<br/>Biological Networks</b>                     | <b>397</b> |
|                 | <i>Nataša Pržulj and Tijana Milenković</i>                                                          |            |

|                                                                                                                                                                             |            |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|
| <b>18 Statistical Analysis of Biomolecular Networks</b>                                                                                                                     | <b>429</b> |
| <i>Jing-Dong J. Han and Chris J. Needham</i>                                                                                                                                |            |
| <b>Part IV Literature, Ontology, and Knowledge Integration</b>                                                                                                              | <b>447</b> |
| <b>19 Beyond Information Retrieval: Literature Mining for Biomedical Knowledge Discovery</b>                                                                                | <b>449</b> |
| <i>Javed Mostafa, Kazuhiro Seki, and Weimao Ke</i>                                                                                                                          |            |
| <b>20 Mining Biological Interactions from Biomedical Texts for Efficient Query Answering</b>                                                                                | <b>485</b> |
| <i>Muhammad Abulaish, Lipika Dey, and Jahiruddin</i>                                                                                                                        |            |
| <b>21 Ontology-Based Knowledge Representation of Experiment Metadata in Biological Data Mining</b>                                                                          | <b>529</b> |
| <i>Richard H. Scheuermann, Megan Kong, Carl Dahlke, Jennifer Cai, Jamie Lee, Yu Qian, Burke Squires, Patrick Dunn, Jeff Wiser, Herb Hagler, Barry Smith, and David Karp</i> |            |
| <b>22 Redescription Mining and Applications in Bioinformatics</b>                                                                                                           | <b>561</b> |
| <i>Naren Ramakrishnan and Mohammed J. Zaki</i>                                                                                                                              |            |
| <b>Part V Genome Medicine Applications</b>                                                                                                                                  | <b>587</b> |
| <b>23 Data Mining Tools and Techniques for Identification of Biomarkers for Cancer</b>                                                                                      | <b>589</b> |
| <i>Mick Correll, Simon Beaulah, Robin Munro, Jonathan Sheldon, Yike Guo, and Hai Hu</i>                                                                                     |            |
| <b>24 Cancer Biomarker Prioritization: Assessing the in vivo Impact of in vitro Models by in silico Mining of Microarray Database, Literature, and Gene Annotation</b>      | <b>615</b> |
| <i>Chia-Ju Lee, Zan Huang, Hongmei Jiang, John Crispino, and Simon Lin</i>                                                                                                  |            |
| <b>25 Biomarker Discovery by Mining Glycomic and Lipidomic Data</b>                                                                                                         | <b>627</b> |
| <i>Hairu Tang, Mehmet Dalkilic, and Yehia Mechref</i>                                                                                                                       |            |
| <b>26 Data Mining Chemical Structures and Biological Data</b>                                                                                                               | <b>649</b> |
| <i>Glenn J. Myatt and Paul E. Blower</i>                                                                                                                                    |            |
| <b>Index</b>                                                                                                                                                                | <b>689</b> |