Protein Interaction Networks: Computational Analysis

The analysis of protein–protein interactions is fundamental to the understanding of cellular organization, processes, and functions. Proteins seldom act as single isolated species; rather, proteins involved in the same cellular processes often interact with each other. Functions of uncharacterized proteins may be predicted through comparison with the interactions of similar known proteins. Recent large-scale investigations of protein–protein interactions using such techniques as two-hybrid systems, mass spectrometry, and protein microarrays have enriched the available protein interaction data and facilitated the construction of integrated protein–protein interaction networks. The resulting large volume of protein–protein interaction data has posed a challenge to experimental investigation.

This book provides a comprehensive understanding of the computational methods available for the analysis of protein–protein interaction networks. It offers an in-depth survey of a range of approaches, including statistical, topological, data-mining, and ontology-based methods. The author discusses the fundamental principles underlying each of these approaches and their respective benefits and drawbacks, and she offers suggestions for future research.

Aidong Zhang is a professor in the Department of Computer Science and Engineering at the State University of New York at Buffalo and the director of the Buffalo Center for Biomedical Computing (BCBC). She is an author of more than 200 research publications and has served on the editorial boards of the *International Journal of Bioinformatics Research and Applications* (IJBRA), *ACM Multimedia Systems*, the *International Journal of Multimedia Tools and Applications*, the *International Journal of Distributed and Parallel Databases*, and *ACM SIGMOD DiSC* (Digital Symposium Collection). Dr. Zhang is a recipient of the National Science Foundation CAREER Award and SUNY (State University of New York) Chancellor’s Research Recognition Award. Dr. Zhang is an IEEE Fellow.
PROTEIN INTERACTION NETWORKS

Computational Analysis

Aidong Zhang
State University of New York, Buffalo
To my daughter, Cathy
# Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preface</td>
<td></td>
<td>xiii</td>
</tr>
<tr>
<td>1</td>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1</td>
<td>Rapid Growth of Protein–Protein Interaction Data</td>
<td>1</td>
</tr>
<tr>
<td>1.2</td>
<td>Computational Analysis of PPI Networks</td>
<td>3</td>
</tr>
<tr>
<td>1.2.1</td>
<td>Topological Features of PPI Networks</td>
<td>4</td>
</tr>
<tr>
<td>1.2.2</td>
<td>Modularity Analysis</td>
<td>5</td>
</tr>
<tr>
<td>1.2.3</td>
<td>Prediction of Protein Functions in PPI Networks</td>
<td>6</td>
</tr>
<tr>
<td>1.2.4</td>
<td>Integration of Domain Knowledge</td>
<td>7</td>
</tr>
<tr>
<td>1.3</td>
<td>Significant Applications</td>
<td>7</td>
</tr>
<tr>
<td>1.4</td>
<td>Organization of this Book</td>
<td>9</td>
</tr>
<tr>
<td>1.5</td>
<td>Summary</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>Experimental Approaches to Generation of PPI Data</td>
<td>11</td>
</tr>
<tr>
<td>2.1</td>
<td>Introduction</td>
<td>11</td>
</tr>
<tr>
<td>2.2</td>
<td>The Y2H System</td>
<td>11</td>
</tr>
<tr>
<td>2.3</td>
<td>Mass Spectrometry (MS) Approaches</td>
<td>13</td>
</tr>
<tr>
<td>2.4</td>
<td>Protein Microarrays</td>
<td>15</td>
</tr>
<tr>
<td>2.5</td>
<td>Public PPI Data and Their Reliability</td>
<td>15</td>
</tr>
<tr>
<td>2.5.1</td>
<td>Experimental PPI Data Sets</td>
<td>15</td>
</tr>
<tr>
<td>2.5.2</td>
<td>Public PPI Databases</td>
<td>16</td>
</tr>
<tr>
<td>2.5.3</td>
<td>Functional Analysis of PPI Data</td>
<td>17</td>
</tr>
<tr>
<td>2.6</td>
<td>Summary</td>
<td>20</td>
</tr>
<tr>
<td>3</td>
<td>Computational Methods for the Prediction of PPIs</td>
<td>21</td>
</tr>
<tr>
<td>3.1</td>
<td>Introduction</td>
<td>21</td>
</tr>
<tr>
<td>3.2</td>
<td>Genome-Scale Approaches</td>
<td>21</td>
</tr>
<tr>
<td>3.3</td>
<td>Sequence-Based Approaches</td>
<td>25</td>
</tr>
<tr>
<td>3.4</td>
<td>Structure-Based Approaches</td>
<td>26</td>
</tr>
<tr>
<td>3.5</td>
<td>Learning-Based Approaches</td>
<td>27</td>
</tr>
<tr>
<td>3.6</td>
<td>Network Topology-Based Approaches</td>
<td>29</td>
</tr>
<tr>
<td>3.7</td>
<td>Summary</td>
<td>32</td>
</tr>
</tbody>
</table>
4 Basic Properties and Measurements of Protein Interaction Networks

4.1 Introduction
4.2 Representation of PPI Networks
4.3 Basic Concepts
4.4 Basic Centralities
4.4.1 Degree Centrality
4.4.2 Distance-Based Centralities
4.4.3 Current-Flow-Based Centrality
4.4.4 Random-Walk-Based Centrality
4.4.5 Feedback-Based Centrality
4.5 Characteristics of PPI Networks
4.6 Summary

5 Modularity Analysis of Protein Interaction Networks

5.1 Introduction
5.2 Useful Metrics for Modular Networks
5.2.1 Cliques
5.2.2 Cores
5.2.3 Degree-Based Index
5.2.4 Distance (Shortest Paths)-Based Index
5.3 Methods for Clustering Analysis of Protein Interaction Networks
5.3.1 Traditional Clustering Methods
5.3.2 Nontraditional Clustering Methods
5.4 Validation of Modularity
5.4.1 Clustering Coefficient
5.4.2 Validation Based on Agreement with Annotated Protein Function Databases
5.4.3 Validation Based on the Definition of Clustering
5.4.4 Topological Validation
5.4.5 Supervised Validation
5.4.6 Statistical Validation
5.4.7 Validation of Protein Function Prediction
5.5 Summary

6 Topological Analysis of Protein Interaction Networks

With Woo-chang Hwang

6.1 Introduction
6.2 Overview and Analysis of Essential Network Components
6.2.1 Error and Attack Tolerance of Complex Networks
6.2.2 Role of High-Degree Nodes in Biological Networks
6.2.3 Betweenness, Connectivity, and Centrality
6.3 Bridging Centrality Measurements
6.3.1 Performance of Bridging Centrality with Synthetic and Real-World Networks
6.3.2 Assessing Network Disruption, Structural Integrity, and Modularity
6.4 Network Modularization Using the Bridge Cut Algorithm 84
6.5 Use of Bridging Nodes in Drug Discovery 87
  6.5.1 Biological Correlates of Bridging Centrality 88
  6.5.2 Results from Drug Discovery-Relevant Human Networks 92
  6.5.3 Comparison to Alternative Approaches: Yeast Cell Cycle State Space Network 94
  6.5.4 Potential of Bridging Centrality as a Drug Discovery Tool 95
6.6 PathRatio: A Novel Topological Method for Predicting Protein Functions 97
  6.6.1 Weighted PPI Network 97
  6.6.2 Protein Connectivity and Interaction Reliability 98
  6.6.3 PathStrength and PathRatio Measurements 99
  6.6.4 Analysis of the PathRatio Topological Measurement 100
  6.6.5 Experimental Results 101
6.7 Summary 108

7 Distance-Based Modularity Analysis 109
  7.1 Introduction 109
  7.2 Topological Distance Measurement Based on Coefficients 109
  7.3 Distance Measurement by Network Distance 112
    7.3.1 PathRatio Method 112
    7.3.2 Averaging the Distances 113
  7.4 Ensemble Method 114
    7.4.1 Similarity Metrics 115
    7.4.2 Base Algorithms 116
    7.4.3 Consensus Methods 116
    7.4.4 Results of the Ensemble Methods 118
  7.5 UVCLUSTER 118
  7.6 Similarity Learning Method 120
  7.7 Measurement of Biological Distance 124
    7.7.1 Sequence Similarity-Based Measurements 124
    7.7.2 Structural Similarity-Based Measurements 125
    7.7.3 Gene Expression Similarity-Based Measurements 127
  7.8 Summary 128

8 Graph-Theoretic Approaches to Modularity Analysis 130
  8.1 Introduction 130
  8.2 Finding Dense Subgraphs 130
    8.2.1 Enumeration of Complete Subgraphs 130
    8.2.2 Monte Carlo Optimization 131
    8.2.3 Molecular Complex Detection 132
    8.2.4 Clique Percolation 133
    8.2.5 Merging by Statistical Significance 134
    8.2.6 Super-Paramagnetic Clustering 136
  8.3 Finding the Best Partition 137
    8.3.1 Recursive Minimum Cut 137
    8.3.2 Restricted Neighborhood Search Clustering (RNSC) 138
8.3.3 Betweenness Cut 140
8.3.4 Markov Clustering 140
8.3.5 Line Graph Generation 143
8.4 Graph Reduction-Based Approach 144
8.4.1 Graph Reduction 144
8.4.2 Hierarchical Modularization 146
8.4.3 Time Complexity 147
8.4.4 $k$ Effects on Graph Reduction 147
8.4.5 Hierarchical Structure of Modules 149
8.5 Summary 150

9 Flow-Based Analysis of Protein Interaction Networks 152
  9.1 Introduction 152
  9.2 Protein Function Prediction Using the FunctionalFlow Algorithm 153
  9.3 CASCADE: A Dynamic Flow Simulation for Modularity Analysis 155
    9.3.1 Occurrence Probability and Related Models 156
    9.3.2 The CASCADE Algorithm 158
    9.3.3 Analysis of Prototypical Data 160
    9.3.4 Significance of Individual Clusters 162
    9.3.5 Analysis of Functional Annotation 164
    9.3.6 Comparative Assessment of CASCADE with Other Approaches 169
    9.3.7 Analysis of Robustness 175
    9.3.8 Analysis of Computational Complexity 175
    9.3.9 Advantages of the CASCADE Method 176
  9.4 Functional Flow Analysis in Weighted PPI Networks 177
    9.4.1 Functional Influence Model 178
    9.4.2 Functional Flow Simulation Algorithm 179
    9.4.3 Time Complexity of Flow Simulation 180
    9.4.4 Detection of Overlapping Modules 181
    9.4.5 Detection of Disjoint Modules 189
    9.4.6 Functional Flow Pattern Mining 191
  9.5 Summary 198

10 Statistics and Machine Learning Based Analysis of Protein Interaction Networks 199
  With Pritam Chanda and Lei Shi
  10.1 Introduction 199
  10.2 Applications of Markov Random Field and Belief Propagation for Protein Function Prediction 200
  10.3 Protein Function Prediction Using Kernel-Based Statistical Learning Methods 207
  10.4 Protein Function Prediction Using Bayesian Networks 211
Contents

10.5 Improving Protein Function Prediction Using Bayesian Integrative Methods 213
10.6 Summary 214

11 Integration of GO into the Analysis of Protein Interaction Networks 216
With Young-rae Cho

11.1 Introduction 216
11.2 GO structure 217
   11.2.1 GO Annotations 217
11.3 Semantic Similarity-Based Integration 218
   11.3.1 Structure-Based Methods 219
   11.3.2 Information Content-Based Methods 220
   11.3.3 Combination of Structure and Information Content 221
11.4 Semantic Interactivity-Based Integration 223
11.5 Estimate of Interaction Reliability 223
   11.5.1 Functional Co-Occurrence 224
   11.5.2 Topological Significance 225
   11.5.3 Protein Lethality 226
11.6 Functional Module Detection 227
   11.6.1 Statistical Assessment 227
   11.6.2 Supervised Validation 229
11.7 Probabilistic Approaches for Function Prediction 231
   11.7.1 GO Index-Based Probabilistic Method 231
   11.7.2 Semantic Similarity-Based Probabilistic Method 235
11.8 Summary 241

12 Data Fusion in the Analysis of Protein Interaction Networks 243

12.1 Introduction 243
12.2 Integration of Gene Expression with PPI Networks 243
12.3 Integration of Protein Domain Information with PPI Networks 244
12.4 Integration of Protein Localization Information with PPI Networks 245
12.5 Integration of Several Data Sources with PPI Networks 247
   12.5.1 Kernel-Based Methods 247
   12.5.2 Bayesian Model-Based Method 249
12.6 Summary 249

13 Conclusion 251

Bibliography 255
Index 273

Color plates follow page 82
I am pleased to offer the research community my second book-length contribution to the field of bioinformatics. My first book, *Advanced Analysis of Gene Expression Microarray Data*, was published in 2006 by World Scientific as part of its Science, Engineering, and Biology Informatics (SEBI) series. I first became involved in the study of bioinformatics in 1998 and, over the ensuing decade, have been struck by the enormous quantity of data being generated and the need for effective approaches to its analysis.

The analysis of protein–protein interactions (PPIs) is fundamental to the understanding of cellular organizations, processes, and functions. It has been observed that proteins seldom act as single isolated species in the performance of their functions; rather, proteins involved in the same cellular processes often interact with each other. Therefore, the functions of uncharacterized proteins can be predicted through comparison with the interactions of similar known proteins. A detailed examination of a PPI network can thus yield significant new insights into protein functions. These interactions have traditionally been examined via intensive small-scale investigations of a small set of proteins of interest, each yielding information about a limited number of PPIs. The existing databases of PPIs have been compiled from such small-scale screens, presented in individual research papers. Because these data were subject to stringent controls and evaluation in the peer-review process, they can be considered to be fairly reliable. However, each experiment observes only a few interactions and yields a data set of very limited size. Recent large-scale investigations of PPIs using such techniques as two-hybrid systems, mass spectrometry, and protein microarrays have enriched the available protein interaction data and facilitated the construction of integrated PPI networks. The resulting large volume of PPI data has posed a challenge to experimental investigation. Consequently, computational analysis of the networks has become a necessary tool for the determination of functionally associated proteins.

This book is intended to provide a comprehensive understanding of the computational methods available for the analysis of PPI networks. It offers an in-depth survey of a range of approaches to this analysis, including statistical, topological, data minining, and ontology-based methods. The fundamental principles underlying each of
these approaches are discussed, along with their respective benefits and drawbacks. Suggestions for future research are also offered. In total, this book is intended to offer bioinformatics researchers a comprehensive and practical guide to the analysis of PPI networks, which will assist and stimulate their further investigation.

Some knowledge on the part of the reader in the fields of molecular biology, data mining, and statistics is assumed. Apart from this, the book is designed to be self-contained, as it includes introductions to the fundamental concepts underlying data generation and analysis. Thus, this book is expected to be of interest to a variety of researchers. It can be used as a textbook for advanced graduate courses in bioinformatics, and most of its content has been tested in the author’s graduate-level course in this field. In addition, it can serve as a resource for graduate students seeking topics for investigation. The book will also be useful to researchers involved in computational biology in universities, organizations, and industry. For this audience, it will provide guidance on the techniques available for analysis of PPI networks. Research professionals interested in expanding their knowledge base can draw upon the material presented here to gain an understanding of principles and methods involved in this growing and highly significant field.

ACKNOWLEDGMENTS

I would like to express my deepest thanks to my doctoral students, Pritam Chanda, Young-rae Cho, Woo-chang Hwang, Taehyong Kim, and Lei Shi, for their excellent technical contributions. I am also highly appreciative of the editorial work of Rachel Ramadhyani.

The inspiration for this book was an invitation from Ms. Lauren Cowles, a senior editor from Cambridge University Press. I would like to express my special thanks to her.

Aidong Zhang
Buffalo, New York