Complexes of physically interacting proteins constitute fundamental functional units that drive almost all biological processes within cells. A faithful reconstruction of the entire set of protein complexes (the “complexosome”) is therefore important not only to understand the composition of complexes but also the higher level functional organization within cells. Advances over the last several years, particularly through the use of high-throughput proteomics techniques, have made it possible to map substantial fractions of protein interactions (the “interactomes”) from model organisms including Arabidopsis thaliana (a flowering plant), Caenorhabditis elegans (a nematode), Drosophila melanogaster (fruit fly), and Saccharomyces cerevisiae (budding yeast). These interaction datasets have enabled systematic inquiry into the identification and study of protein complexes from organisms. Computational methods have played a significant role in this context, by contributing accurate, efficient, and exhaustive ways to analyze the enormous amounts of data. These methods have helped to compensate for some of the limitations in experimental datasets including the presence of biological and technical noise and the relative paucity of credible interactions.

In this book, we systematically walk through computational methods devised to date (approximately between 2000 and 2016) for identifying protein complexes from the network of protein interactions (the protein-protein interaction (PPI) network). We present a detailed taxonomy of these methods, and comprehensively evaluate them for protein complex identification across a variety of scenarios including the absence of many true interactions and the presence of false-positive interactions (noise) in PPI networks. Based on this evaluation, we highlight challenges faced by the methods, for instance in identifying sparse, sub-, or small complexes and in discerning overlapping complexes, and reveal how a combination of strategies is necessary to accurately reconstruct the entire complexosome.
Computational Prediction of Protein Complexes from Protein Interaction Networks
ACM Books

Editor in Chief
M. Tamer Özsu, University of Waterloo

ACM Books is a new series of high-quality books for the computer science community, published by ACM in collaboration with Morgan & Claypool Publishers. ACM Books publications are widely distributed in both print and digital formats through booksellers and to libraries (and library consortia) and individual ACM members via the ACM Digital Library platform.

Computational Prediction of Protein Complexes from Protein Interaction Networks
Sriganesh Sridhara, The University of Queensland Institute for Molecular Bioscience
Chern Han Yong, Duke-National University of Singapore Medical School
Limsoon Wong, National University of Singapore
2017

Shared-Memory Parallelism Can Be Simple, Fast, and Scalable
Julian Shun, University of California, Berkeley
2017

The Handbook of Multimodal-Multisensor Interfaces, Volume 1: Foundations, User Modeling, and Common Modality Combinations
Editors: Sharon Oviatt, Inca Designs
Björn Schuller, University of Passau and Imperial College London
Philip R. Cohen, Voicebox Technologies
Daniel Sonntag, German Research Center for Artificial Intelligence (DFKI)
Gerasimos Potamianos, University of Thessaly
Antonio Krüger, German Research Center for Artificial Intelligence (DFKI)
2017

Communities of Computing: Computer Science and Society in the ACM
Thomas J. Misa, Editor, University of Minnesota
2017

Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining
ChengXiang Zhai, University of Illinois at Urbana-Champaign
Sean Massung, University of Illinois at Urbana-Champaign
2016

An Architecture for Fast and General Data Processing on Large Clusters
Matei Zaharia, Massachusetts Institute of Technology
2016
Reactive Internet Programming: State Chart XML in Action
Franck Barbier, University of Pau, France
2016

Verified Functional Programming in Agda
Aaron Stump, The University of Iowa
2016

The VR Book: Human-Centered Design for Virtual Reality
Jason Jerald, NextGen Interactions
2016

Ada’s Legacy: Cultures of Computing from the Victorian to the Digital Age
Robin Hammerman, Stevens Institute of Technology
Andrew L. Russell, Stevens Institute of Technology
2016

Edmund Berkeley and the Social Responsibility of Computer Professionals
Bernadette Longo, New Jersey Institute of Technology
2015

Candidate Multilinear Maps
Sanjam Garg, University of California, Berkeley
2015

Smarter Than Their Machines: Oral Histories of Pioneers in Interactive Computing
John Cullinane, Northeastern University; Mossavar-Rahmani Center for Business
and Government, John F. Kennedy School of Government, Harvard University
2015

A Framework for Scientific Discovery through Video Games
Seth Cooper, University of Washington
2014

Trust Extension as a Mechanism for Secure Code Execution on Commodity
Computers
Bryan Jeffrey Parno, Microsoft Research
2014

Embracing Interference in Wireless Systems
Shyamnath Gollakota, University of Washington
2014
Computational Prediction of Protein Complexes from Protein Interaction Networks

Sriganesh Srihari
The University of Queensland Institute for Molecular Bioscience

Chern Han Yong
Duke-National University of Singapore Medical School

Limsoon Wong
National University of Singapore
Dedicated to the Honors, Masters, and Ph.D. students who worked over the years on the different aspects of PPI networks by being part of the computational biology group at the Department of Computer Science, National University of Singapore.
Contents

Preface xi

Chapter 1 Introduction to Protein Complex Prediction 1
1.1 From Protein Interactions to Protein Complexes 6
1.2 Databases for Protein Complexes 11
1.3 Organization of the Rest of the Book 13

Chapter 2 Constructing Reliable Protein-Protein Interaction (PPI) Networks 15
2.1 High-Throughput Experimental Systems to Infer PPIs 15
2.2 Data Sources for PPIs 23
2.3 Topological Properties of PPI Networks 25
2.4 Theoretical Models for PPI Networks 27
2.5 Visualizing PPI Networks 31
2.6 Building High-Confidence PPI Networks 34
2.7 Enhancing PPI Networks by Integrating Functional Interactions 50

Chapter 3 Computational Methods for Protein Complex Prediction from PPI Networks 59
3.1 Basic Definitions and Terminologies 60
3.2 Taxonomy of Methods for Protein Complex Prediction 60
3.3 Methods Based Solely on PPI Network Clustering 61
3.4 Methods Incorporating Core-Attachment Structure 78
3.5 Methods Incorporating Functional Information 85

Chapter 4 Evaluating Protein Complex Prediction Methods 91
4.1 Evaluation Criteria and Methodology 91
4.2 Evaluation on Unweighted Yeast PPI Networks 93
4.3 Evaluation on Weighted Yeast PPI Networks 95
4.4 Evaluation on Human PPI Networks 99
4.5 Case Study: Prediction of the Human Mechanistic Target of Rapamycin Complex 103
4.6 Take-Home Lessons from Evaluating Prediction Methods 105

Chapter 5 Open Challenges in Protein Complex Prediction 107
5.1 Three Main Challenges in Protein Complex Prediction 107
5.2 Identifying Sparse Protein Complexes 112
5.3 Identifying Overlapping Protein Complexes 118
5.4 Identifying Small Protein Complexes 124
5.5 Identifying Protein Sub-complexes 134
5.6 An Integrated System for Identifying Challenging Protein Complexes 136
5.7 Recent Methods for Protein Complex Prediction 138
5.8 Identifying Membrane-Protein Complexes 141

Chapter 6 Identifying Dynamic Protein Complexes 145
6.1 Dynamism of Protein Interactions and Protein Complexes 145
6.2 Identifying Temporal Protein Complexes 149
6.3 Intrinsic Disorder in Proteins 156
6.4 Intrinsic Disorder in Protein Interactions and Protein Complexes 157
6.5 Identifying Fuzzy Protein Complexes 162

Chapter 7 Identifying Evolutionarily Conserved Protein Complexes 165
7.1 Inferring Evolutionarily Conserved PPIs (Interologs) 166
7.2 Identifying Conserved Complexes from Interolog Networks, I 170
7.3 Identifying Conserved Complexes from Interolog Networks, II 178

Chapter 8 Protein Complex Prediction in the Era of Systems Biology 185
8.1 Constructing the Network of Protein Complexes 185
8.2 Identifying Protein Complexes Across Phenotypes 189
8.3 Identifying Protein Complexes in Diseases 192
8.4 Enhancing Quantitative Proteomics Using PPI Networks and Protein Complexes 208
8.5 Systems Biology Tools and Databases to Analyze Biomolecular Networks 221

Chapter 9 Conclusion 225

References 233
Authors’ Biographies 279
Preface

The suggestion and motivation to write this book came from Limsoon, who thought that it would be a great idea to compile our (Sriganesh’s and Chern Han’s) Ph.D. research conducted at National University of Singapore on protein complex prediction from protein-protein interaction (PPI) networks into a comprehensive book for the research community. Since we (Sriganesh and Chern Han) completed our Ph.D.s not long ago, the timing could not have been better for writing this book while the topic is still fresh in our minds and the empirical set up (datasets and software pipelines) for evaluating the methods is still in a “quick-to-run” form. However, although we had our Ph.D. theses to our convenient disposal and reference, it is only after we started writing this book that we realized the real scale of the task that we had embarked upon.

The problem of protein complex prediction may be just one of the plethora of computational problems that have opened up since the deluge of proteomics (protein-protein interaction; PPI) data over the last several years. However, in reality this problem encompasses or directly relates to several important and open problems in the area—in particular, the fundamental problems of modeling, visualizing, and denoising of PPI networks, prediction of PPIs (novel as well as evolutionarily conserved), and protein function prediction from PPI data. Therefore, to write a comprehensive self-contained book, we had to cover even these closely related problems to some extent or at least allude to or reference them in the book. We had to do so without missing the connection between these problems and our central problem of protein complex prediction in the book.

The early tone to write the book in this manner was set by our review article in a 2015 special issue of *FEBS Letters*, where we covered a number of protein complex prediction methods which are based on a diverse range of topological, functional, temporal, structural, and evolutionary information. However, being only a single-volume article, the description of the methods was brief, and to compile
this description in the form of a book we had to delve a lot deeper into the algorithmic underpinnings of each of the methods, highlight how each method utilized the information (topological, functional, temporal, structural, and evolutionary) on which it was based in its own unique way, and evaluate and study the applications of the methods across a diverse range of datasets and scenarios. To do this well, we had to: (i) cover in substantial detail the preliminaries such as the experimental techniques available to infer PPIs, the limitations of each of these techniques, PPI network topology, modeling, and denoising, PPI databases that are available, and how functional, temporal, structural, and evolutionary information of proteins can be integrated with PPI networks; and (ii) we had to categorize protein complex prediction methods into logical groups based on some criteria, and dedicate a separate chapter for each group to make our description comprehensive. In the book, we cover (i) in Chapter 2 and in the form of independent sections within each of the other Chapters 3, 5, 6, 7, and 8. We cover (ii) by allocating Chapters 3 and 4 for “classical” methods and their comprehensive evaluation, Chapter 5 for methods that predict certain kinds of “challenging complexes” which the classical methods do not predict well, Chapter 6 for methods that utilize temporal and structural information, Chapter 7 for methods that utilize information on evolutionary conservation, and Chapter 8 for methods that integrate other kinds of omics datasets to predict “specialized” complexes—e.g., protein complexes in diseases.

The requirement for a book exclusively dedicated to the problem of protein complex prediction from PPI networks at this point in time cannot be understated. Over the last two decades, a major focus of high-throughput experimental technologies and of computational methods to analyze the generated data has been in genomics—e.g., in the analysis of genome sequencing data. It is relatively recently that this focus has started to shift toward proteomics and computational methods to analyze proteomics data. For example, while the complete sequence of the human genome was assembled more than a decade ago, it is only over the last three years that there have been similar large-scale efforts to map the human proteome. The ProteomicsDB (http://www.proteomicsdb.org/), Human Proteome Map (http://humanproteomemap.org/), and the Human Protein Atlas projects (http://www.proteinatlas.org/) have all appeared only over the last three years. Similarly is The Cancer Proteome Atlas (TCPA) project (http://app1.bioinformatics.mdanderson.org/tcpa/_design/basic/index.html) which complements The Cancer Genome Atlas project (TCGA). This means that developing more effective solutions for fundamental problems such as protein complex prediction has become all the more important today, as we try to apply these solutions to larger and more complex datasets arising from these newer technologies and projects. In this respect,
Preface

we had to write the book not just by considering protein complex prediction methods (i.e., their algorithmic details) as important in their own right but also by giving significant importance to the applications of these methods in the light of today's complex datasets and research questions. There are several sections within each of the Chapters 6, 7, and 8 that play this dual role, e.g., a section in Chapter 7 discusses the evolutionary conservation of core cellular processes based on conservation patterns of protein complexes, and a section in Chapter 8 discusses the dysregulation of these processes in diseases based on rewiring of protein complexes between normal and disease conditions.

In the end, we hope that we have done justice to what we intend this book to be. We hope that this book provides valuable insights into protein complex prediction and inspires further research in the area especially for tackling the open challenges, as well as inspires new applications in diverse areas of biomedicine.

Acknowledgments

Although this book is primarily concerned with the problem of protein complex prediction, the book also covers several other aspects of PPI networks. We would like to therefore dedicate this book to the students—Honors, Masters, and Ph.D. students—who worked on these different aspects of PPI networks by being part of the computational biology group at the Department of Computer Science, National University of Singapore, over the years. Several of the methods covered in this book are a result of the extensive research conducted by these students. Sriganesh would like to thank Hon Wai Leong (Professor of Computer Science, National University of Singapore) under whom he conducted his Ph.D. research on protein complex prediction; Mark Ragan (Head of Division of Genomics of Development and Disease at Institute for Molecular Bioscience, The University of Queensland) under whom he conducted his postdoctoral research, a substantial portion of which was on identifying protein complexes in diseases; and Kum Kum Khanna (Senior Principal Research Fellow and Group Leader at QIMR-Berghofer Medical Research Institute) whose guidance played a significant part in his understanding of biological aspects of protein complexes. Sriganesh is grateful to Mark for passing him an original copy of a 1977 volume of *Progress in Biophysics and Molecular Biology* in which G. Rickey Welch makes a consistent principled argument that “multienzyme clusters” are advantageous to the cell and organism because they enable metabolites to be channeled within the clusters and protein expression to be co-regulated [Welch 1977]—a possession which Sriganesh will deeply cherish. Chern Han would like to thank his coauthors: Sriganesh for doing the heavy lifting in writing, editing, and
driving this project and Limsoon Wong for guiding him through his Ph.D. journey on protein complexes. He would also like to acknowledge the support of Bin Tean Teh (Professor with Program in Cancer and Stem Cell Biology, Duke-NUS Medical School), who currently oversees his postdoctoral research. Limsoon would like to acknowledge Chern Han and Sriganesh for doing the bulk of the writing for this book, and especially thank Sriganesh for taking the overall lead on the project. When he suggested the book to Chern Han and Sriganesh, he had not imagined that he would eventually be a co-author.

We are indebted also to the Editor-in-Chief of ACM Books, Tamer Özsu, Executive Editor Diane Cerra, Production Manager Paul C. Anagnostopoulos, and the entire team at ACM Books and Morgan & Claypool Publishers for their encouragement and for producing this book so beautifully.

Sriganesh Srihari
Chern Han Yong
Limsoon Wong
May 2017
Introduction to Protein Complex Prediction

*Unfortunately, the proteome is much more complicated than the genome.*
—Carol Ezzell [Ezzel et al. 2002]

In an early survey, American biochemist Bruce Alberts termed large assemblies of proteins as protein machines of cells [Alberts et al. 1998]. Protein assemblies are composed of highly specialized parts that coordinate to execute almost all of the biochemical, signaling, and functional processes in cells [Alberts et al. 1998]. It is not hard to see why protein assemblies are more advantageous to cells than individual proteins working in an uncoordinated manner. Compare, for example, the speed and elegance of the DNA replication machinery that simultaneously replicates both strands of the DNA double helix with what could ensue if each of the individual components—DNA helicases for separating the double-stranded DNA into single stands, DNA polymerases for assembling nucleotides, DNA primase for generating the primers, and the sliding clamp to hold these enzymes onto the DNA—acted in an uncoordinated manner [Alberts et al. 1998]. Although what might seem like individual parts brought together to perform arbitrary functions, protein assemblies can be very specific and enormously complicated. For example, the spliceosome is composed of 5 small nuclear RNAs (snRNAs or “snurps”) and more than 50 proteins, and is thought to catalyze an ordered sequence of more than 10 RNA rearrangements at a time as it removes an intron from an RNA transcript [Alberts et al. 1998, Baker et al. 1998]. The discovery of this intron-splicing process won Phillip A. Sharp and Richard J. Roberts the 1993 Nobel Prize in Physiology or Medicine.¹

Protein assemblies are known to be in the order of hundreds even in the simplest of eukaryotic cells. For example, more than 400 protein assemblies have been identified in the single-celled eukaryote *Saccharomyces cerevisiae* (budding yeast) [Pu et al. 2009]. However, our knowledge of these protein assemblies is still fragmentary, as is our conception of how each of these assemblies work together to constitute the “higher level” functional architecture of cells. A faithful attempt toward identification and characterization of all protein assemblies is therefore crucial to elucidate the functioning of the cellular machinery.

To identify the entire complement of protein assemblies, it is important to first crack the proteome—a concept so novel that the word “proteome” first appeared only around 20 years ago [Wilkins et al. 1996, Bryson 2003, Cox and Mann 2007]. The proteome, as defined in the UniProt Knowledgebase, is the entire complement of proteins expressed or derived from protein-coding genes in an organism [Bairoch and Apweiler 1996, UniProt 2015]. With the introduction of high-throughput experimental (proteomics) techniques including mass spectrometric [Cox and Mann 2007, Aebersold and Mann 2003] and protein quantitative trait locus (QTL) technologies [Foss et al. 2007], mapping of proteins on a large scale has become feasible. Just like how genomics techniques (including genome sequencing) were first demonstrated in model organisms, proteome-mapping has progressed initially and most rapidly for model prokaryotes including *Escherichia coli* (bacteria) and model eukaryotes including *Saccharomyces cerevisiae* (budding or baker’s yeast), *Drosophila melanogaster* (fruit fly), *Caenorhabditis elegans* (a nematode), and *Arabidopsis thaliana* (a flowering plant). Table 1.1 summarizes the numbers of proteins or protein-coding genes identified from these organisms. Of these, the proportions of protein-coding genes that are essential (genes that are thought to be critical for the survival of the cell or organism; “fitness genes”) range from ~2% in *Drosophila* to ~6.5% in *Caenorhabditis* and ~18% in *Saccharomyces* [Cherry et al. 2012, Chen et al. 2012]. Recent landmark studies using large-scale proteomics [Wilhelm et al. 2014, Kim et al. 2014, Uhlén et al. 2010, Uhlén et al. 2015] on *Homo sapiens* (human) cells have characterized >17,000 (or >90%) putative protein-coding genes from ≥40 tissues and organs in the human body. An encyclopedic resource on these proteins covering their levels of expression and abundance in different human tissues is available from the ProteomicsDB (http://www.proteomicsdb.org/) [Wilhelm et al. 2014], The Human Proteome Map (http://humanproteomemap.org/) [Kim et al. 2014], and The Human Protein Atlas (http://www.proteinatlas.org/) [Uhlén et al. 2010, Uhlén et al. 2015] projects. GeneCards (http://www.genecards.org/) [Safran et al. 2002, Safran et al. 2010] aggregates information on human protein-coding genes from >125 Web sources.
Table 1.1 Examples of proteome resources for some model and higher-order organisms (as of December 2015), covering also *Danio rerio* (Zebrafish), *Mus musculus* (house mouse), *Rattus norvegicus* (Norwegian rat), *Schizosaccharomyces pombe* (fission yeast), and *Xenopus laevis* (African clawed frog)

<table>
<thead>
<tr>
<th>Organism</th>
<th>No. of Proteins/Protein-Coding Genes</th>
<th>Source</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>R. norvegicus</em></td>
<td>~29,600</td>
<td><a href="http://rgd.mcw.edu/">http://rgd.mcw.edu/</a></td>
<td>[Shimoyama et al. 2015]</td>
</tr>
</tbody>
</table>
Chapter 1  Introduction to Protein Complex Prediction

and presents the information in an integrative user-friendly manner. The expression levels of nearly 200 proteins that are essential for driving different human cancers are available from The Cancer Proteome Atlas (TCPA) project (http://app1.bioinformatics.mdanderson.org/tcpa/_design/basic/index.html) [Li et al. 2013], measured from more than 3,000 tissue samples across 11 cancer types studied as part of The Cancer Genome Atlas (TCGA) project (http://cancergenome.nih.gov/). Short-hairpin RNA (shRNA)-mediated knockdown [Paddison et al. 2002, Lambeth and Smith 2013], clustered regularly interspaced short palindromic repeats (CRISPR)/Cas9-based gene editing [Sanjana et al. 2014, Baltimore et al. 2015, Shalem et al. 2015], and disruptive mutagenesis [Bökel 2008] screening using MCF-10A (near-normal mammary), MDA-MB-435 (breast cancer), KBM7 (chronic myeloid leukemia), HAP1 (haploid), A375 (melanoma), HCT116 (colorectal cancer), and HUES62 (human embryonic stem) cells have characterized 1,500–1,880 (or 8–10%) “core” protein-coding genes as essential in human cells [Marcotte et al. 2016, Silva et al. 2008, Wang et al. 2014, Hart et al. 2015, Hart et al. 2014, Wang et al. 2015, Blomen et al. 2015].

Comparative analyses of proteomes from different species have revealed interesting insights into the evolution and conservation of proteins. For example, it is estimated that the genomes (proteomes) of human and budding yeast diverged about 1 billion years ago from a common ancestor [Douzery et al. 2014], and these share several thousand genes accounting for more than one-third of the yeast genome [O’Brien et al. 2005, Ostlund et al. 2010]. Yeast and human orthologs are highly diverged; the amino-acid sequence similarity between human and yeast proteins ranges from 9–92%, with a genome-wide average of 32%. But, sequence similarity predicts only a part of the picture [Sun et al. 2016]. Recent studies [Kachroo et al. 2015, Laurent et al. 2015] have reported that 414 (or nearly half of the) essential protein-coding genes in yeast could be “replaced” by human genes, with replaceability depending on gene (protein) assemblies: genes in the same process tend to be similarly replaceable (e.g., sterol biosynthesis) or not replaceable (e.g., DNA replication initiation).

Irrespective of whether in a lower-order model or a higher-order complex organism, a protein has to physically interact with other proteins and biomolecules to remain functional. Estimates in human suggest that over 80% of proteins do not function alone, but instead interact to function as macromolecular assemblies [Berggard et al. 2007]. This organization of individual proteins into assemblies is tightly regulated in cellular space and time, and is supported by protein conformational changes, posttranslational modifications, and competitive binding [Gibson and Goldberg 2009]. On the basis of the stability (area of interaction surface and du-
ration of interaction) and partner specificity, the interactions between proteins are classified as homo- or hetero-oligomeric, obligate or non-obligate, and permanent or transient [Zhang 2009, Nooren and Thornton 2003]. Proteins in obligate interactions cannot exist as stable structures on their own and are frequently bound to their partners upon translation and folding, whereas proteins in non-obligate interactions can exist as stable structures in bound and unbound states. Obligate interactions are generally permanent or constitutive, which once formed exist for the entire lifetime of the proteins, whereas non-obligate interactions may be permanent, or alternatively transient, wherein the protein interacts with its partners for a brief time period and dissociates after that. Depending on the functional, spatial, and temporal context of the interactions, protein assemblies are classified as protein complexes, functional modules, and biochemical (metabolic) and signaling pathways.

Protein complexes are the most basic forms of protein assemblies and constitute fundamental functional units within cells. Complexes are stoichiometrically stable structures and are formed from physical interactions between proteins coming together at a specific time and space. Complexes are responsible for a wide range of functions within cells including formation of cytoskeleton, transportation of cargo, metabolism of substrates for the production of energy, replication of DNA, protection and maintenance of the genome, transcription and translation of genes to gene products, maintenance of protein turn over, and protection of cells from internal and external damaging agents. Complexes can be permanent—i.e., once assembled can function for the entire lifetime of cells (e.g., ribosomes)—or transient—i.e., assembled temporarily to perform a specific function and are disassembled after that (e.g., cell-cycle kinase-substrate complexes formed in a cell-cycle dependent manner).

Functional modules are formed when two or more protein complexes interact with each other and often other biomolecules (viz. nucleic acids, sugars, lipids, small molecules, and individual proteins) at a specific time and space to perform a particular function and disassociate after that. This molecular organization has been termed “protein sociology” [Robinson et al. 2007]. For example, the DNA replication machinery, highlighted earlier, is formed by a tightly coordinated assembly of DNA polymerases, DNA helicase, DNA primase, the sliding clamp and other complexes within the nucleus to ensure error-free replication of the DNA during cell division.

Pathways are formed when sets of complexes and individual proteins interact via an ordered sequence of interactions to transduce signals (signaling pathways) or metabolize substrates from one form to another (metabolic pathways). For
example, the MAPK pathway is composed of a sequence of microtubule-associated protein kinases (MAPKs) that transduce signals from the cell membrane to the nucleus, to induce the transcription of specific genes within the nucleus. Unlike complexes and functional modules, pathways do not require all components to co-localize in time and space.

1.1 From Protein Interactions to Protein Complexes

Physical interactions between proteins are fundamental to the formation of protein complexes. Therefore, mapping the entire complement of protein interactions (the “interactome”) occurring within cells (in vivo) is crucial for identifying and characterizing complexes. However, inferring all interactions occurring during the entire lifetime of cells in an organism is challenging, and this challenge increases multifold as the complexity of the organism increases—e.g., for multicellular organisms made up of multiple cell types.

The development of high-throughput proteomics technologies including yeast two-hybrid- (Y2H) [Fields and Song 1989], co-immunoprecipitation (Co-IP) [Golemis and Adams 2002] and affinity-purification (AP)-based [Rigaut et al. 1999] screens have revolutionized our ability to interrogate protein interactions on a massive scale, and have enabled global surveys of interactomes from a number of organisms. In particular, up to 70% of the interactions from model organisms including yeast [Ito et al. 2000, Uetz et al. 2000, Ho et al. 2002, Gavin et al. 2002, Gavin et al. 2006, Krogan et al. 2006], fly [Guruharsha et al. 2011], and nematode [Butland et al. 2005, Li et al. 2004] have been mapped, and the identification of interactions from higher-order multicellular organisms including species of flowering plant Arabidopsis, fish Danio (zebrafish), and several mammals—Mus musculus (house mouse), Rattus norvegicus (Norwegian rat), and humans—is rapidly underway; the interactions are cataloged in large public databases [Stark et al. 2011, Rolland et al. 2014].

The earliest and most widely used experimental techniques to capture binary interacting proteins on a high-throughput scale were mostly yeast two-hybrid (Y2H) [Fields and Song 1989]. However, datasets of protein interactions inferred from Y2H screens were found to have significant numbers of spurious interactions [Von Mering et al. 2002, Bader and Hogue 2002, Bader et al. 2004]. This is attributed in part to the nature of the Y2H protocol in which all potential interactors are tested within the same compartment (nucleus) even though some of these do not meet during their lifetimes due to compartmentalization (different subcellular localizations) within living cells.
1.1 From Protein Interactions to Protein Complexes

Co-immunoprecipitation or affinity-purification (Co-IP/AP) techniques were introduced later and these are more specific in detecting interactions between co-complexed proteins [Golemis and Adams 2002, Rigaut et al. 1999, Köcher and Superti-Furga 2007]. In these protocols, cohesive groups or complexes of proteins are “pulled down,” from which the binary interactions between the proteins are individually inferred. However, this indirect inference could lead to over- or under-estimation of protein interactions. In the tandem affinity purification (TAP) procedure [Rigaut et al. 1999, Puig et al. 2001], proteins of interest (“baits”) are TAP-tagged and purified in an affinity column with potential interaction partners (“preys”). The pulled-down complexes are subjected to mass spectrometric (MS) analysis to identify individual components within the complexes. However, although more reliable than Y2H, the TAP/MS procedure can be elaborate and with the inclusion of MS, it can be expensive too. The exhaustiveness of TAP/MS depends on the baits used—there is no way to identify all possible complexes unless all possible baits are tested. Proteins which do not interact directly with the chosen bait but interact with one or more of the preys, might also get pulled down as part of the purified complex. In some cases, these proteins are indeed part of the real complex whereas in other cases these proteins are not (i.e., they are contaminants); therefore multiple purifications are required, possibly with each protein as a bait and as a prey, to identify the correct set of proteins within the complex. The TAP procedure therefore offers two successive affinity purifications so that the chance of retained contaminants reduces significantly. Conversely, a chosen bait might form a real complex with a set of proteins without actually interacting directly with every protein from the set, and therefore some proteins might not get pulled down as part of the purified complex. In these cases, multiple baits would need to be tested to assemble the complete complex. Moreover, since some proteins participate in more than one complex, multiple independent purifications are required to identify all hosting complexes for these proteins.

Binary interactions between the proteins in a pulled-down protein complex are inferred using two models: matrix and spoke. In the matrix model, a binary interaction is inferred between every pair of proteins within the complex, whereas in the spoke model interactions are inferred only between the bait and all its preys. Since all pairs of proteins within a complex do not necessarily interact, the matrix model is usually an overestimation of the total number of binary interactions, whereas the spoke model is an underestimation. Therefore, usually a balance is struck between the two models that is close enough to the estimated total number of interactions for the species or organism.
Despite differences in procedures and technologies, the use of different experimental protocols can effectively complement one another in detecting interactions. While TAP can be more specific and detect mainly stable (co-complexed) protein interactions, Y2H can be more exhaustive and detect even transient and between-complex interactions. Based on BioGrid version 3.4.130 (November 2015) [Stark et al. 2011, Chatr-Aryamontri et al. 2015], the numbers of mapped physical interactions range from 99 in \textit{E. coli} to \(~\sim\) 82,300 in \textit{S. cerevisiae} and \(~\sim\) 230,900 in \textit{H. sapiens} (summarized in Table 1.2). It remains to be seen how many of these interactions actually occur in the physiological contexts of living cells or cell types, how many are subject to genetic and physiological variations, and how many still remain to be mapped.

The binary interactions inferred from the different experiments are assembled into a \textit{protein-protein interaction network}, or simply, \textit{PPI network}. The PPI network presents a global or “systems” view of the interactome, and provides a mathematical (topological) framework to analyze these interactions. Protein complexes are expected to be embedded as \textit{modular} structures within the PPI network [Hartwell et al. 1999, Spirin and Mirny 2003]. Topologically, this modularity refers to densely connected subsets of proteins separated by less-dense regions in the network [Newman

Table 1.2  Numbers of mapped physical interactions between proteins across different model and higher-order organisms

<table>
<thead>
<tr>
<th>Organism</th>
<th>No. of Interactions</th>
<th>No. of Proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{A. thaliana}</td>
<td>34,320</td>
<td>9,240</td>
</tr>
<tr>
<td>\textit{C. elegans}</td>
<td>5,783</td>
<td>3,269</td>
</tr>
<tr>
<td>\textit{D. rerio}</td>
<td>188</td>
<td>181</td>
</tr>
<tr>
<td>\textit{D. melanogaster}</td>
<td>36,741</td>
<td>8,071</td>
</tr>
<tr>
<td>\textit{E. coli}</td>
<td>99</td>
<td>104</td>
</tr>
<tr>
<td>\textit{H. sapiens}</td>
<td>230,843</td>
<td>20,006</td>
</tr>
<tr>
<td>\textit{M. musculus}</td>
<td>18,465</td>
<td>8,611</td>
</tr>
<tr>
<td>\textit{R. norvegicus}</td>
<td>4,537</td>
<td>3,328</td>
</tr>
<tr>
<td>\textit{S. cerevisiae}</td>
<td>82,327</td>
<td>6,278</td>
</tr>
<tr>
<td>\textit{S. pombe}</td>
<td>9,492</td>
<td>2,944</td>
</tr>
<tr>
<td>\textit{X. laevis}</td>
<td>532</td>
<td>471</td>
</tr>
</tbody>
</table>

Based on BioGrid version 3.4.130 (November 2015) [Stark et al. 2011, Chatr-Aryamontri et al. 2015].
Biologically, this modularity represents division of labor among the complexes, and provides robustness against disruptions to the network from internal (e.g., mutations) and external (e.g., chemical attacks) agents. Computational methods developed to identify protein complexes therefore mine for modular subnetworks in the PPI network. While this strategy appears reasonable in general, limitations in PPI datasets, arising due to the shortcomings highlighted above in experimental protocols, severely restrict the feasibility of accurately predicting complexes from the network. Specifically, the limitations in existing PPI datasets that directly impact protein complex prediction include:

1. presence of a large number of spurious (noisy) interactions;
2. relative paucity of interactions between “complexed” proteins; and
3. missing contextual—e.g., temporal and spatial—information about the interactions.

These limitations translate to the following three main challenges currently faced by computational methods for protein complex prediction:

1. difficulty in detecting sparse complexes;
2. difficulty in detecting small (containing fewer than four proteins) and subcomplexes; and
3. difficulty in deconvoluting overlapping complexes (i.e., complexes that share many proteins), especially when these complexes occur under different cellular contexts.

While the interactome coverage can be improved by integrating multiple PPI datasets, the lack of agreement between the datasets from different experimental protocols [Von Mering et al. 2002, Bader et al. 2004], and the multifold increase in accompanying noise (spurious interactions), tend to cancel out the advantage gained from the increased coverage. Consequently, the confidence of each interaction has to be assessed (confidence scoring) and low-confidence interactions have to be first removed from the datasets (filtering) before performing any downstream analysis. To summarize, computational identification of protein complexes from interaction datasets follows these steps (Figure 1.1):

1. integrating interactions from multiple experiments and stringently assessing the confidence (reliability) of these interactions;
2. constructing a reliable PPI network using only the high-confidence interactions;
Figure 1.1 Identification of protein complexes from protein interaction data. (a) A high-confidence PPI network is assembled from physical interactions between proteins after discarding low-confidence (potentially spurious) interactions. (b) Candidate protein complexes are predicted from this PPI network using network-clustering approaches. The quality of the predicted complexes is validated against bona fide complexes, whereas novel complexes are functionally assessed and assigned new roles where possible.
identifying modular subnetworks from the PPI network to generate a candidate list of protein complexes; and

4. evaluating these candidate complexes against bona fide complexes, and validating and assigning roles for novel complexes.

As we shall see in the following chapters, several sophisticated approaches have been developed over the years to overcome some of the above-mentioned challenges.

Computational methods have co-evolved with proteomics technologies, and over the last ten years a plethora of computational methods have been developed to predict complexes from PPI networks, which is the subject of this book. In general, computational methods complement experimental approaches in several ways. These methods have helped counter some of the limitations arising in proteomic studies, e.g., by eliminating spurious interactions via interaction scoring, and by enriching true interactions via prediction of missing interactions. The novel interactions and protein complexes predicted from these methods have been added back to proteomics databases, and these have helped to further enhance our resources and knowledge in the field.

### 1.2 Databases for Protein Complexes

Several high-quality resources for protein complexes have been developed over the years covering both lower-order model and higher-order organisms (summarized in Table 1.3). In total, Aloy [Aloy et al. 2004], CYC2008 [Pu et al. 2009], and MIPS [Mewes et al. 2008] contain over 450 manually curated complexes from *S. cerevisiae* (budding yeast). CORUM [Reu Jpp et al. 2008, 2010] contains ∼3,000 mammalian complexes of which ∼1,970 are protein complexes identified from human cells. The European Molecular Biology Laboratory (EMBL) and European Bioinformatics Institute (EBI) maintain a database of manually curated protein complexes from 18 different species including *C. elegans, H. sapiens, M. musculus, S. cerevisiae*, and *S. pombe* [Meldal et al. 2015].

### Table 1.3: Publicly available databases for protein complexes

<table>
<thead>
<tr>
<th>Database</th>
<th>Organisms</th>
<th>No. of Complexes</th>
<th>Source</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMBL-EBI Complex Portal</td>
<td></td>
<td>1,564</td>
<td><a href="http://www.ebi.ac.uk/intact/index">http://www.ebi.ac.uk/intact/index</a></td>
<td>[Meldal et al. 2015]</td>
</tr>
<tr>
<td>hu.MAP</td>
<td><em>H. sapiens</em></td>
<td>&gt;4,600</td>
<td><a href="http://proteincomplexes.org">http://proteincomplexes.org</a></td>
<td>[Drew et al. 2017]</td>
</tr>
<tr>
<td>Metazoan conserved&lt;sup&gt;c&lt;/sup&gt;</td>
<td>Metazoa</td>
<td>344–490</td>
<td><a href="http://metazoa.med.utoronto.ca/">http://metazoa.med.utoronto.ca/</a></td>
<td>[Wan et al. 2015]</td>
</tr>
<tr>
<td>MIPS—CORUM</td>
<td>Mammals</td>
<td>2,837</td>
<td><a href="http://mips.helmholtz-muenchen.de/genre/proj/corum/">http://mips.helmholtz-muenchen.de/genre/proj/corum/</a></td>
<td>[Reuett et al. 2008, Ruepen et al. 2010]</td>
</tr>
<tr>
<td>MIPS—Yeast</td>
<td><em>S. cerevisiae</em></td>
<td>313</td>
<td><a href="http://www.helmholtz-muenchen.de/en/ibis/">http://www.helmholtz-muenchen.de/en/ibis/</a></td>
<td>[Mewes et al. 2008]</td>
</tr>
</tbody>
</table>

<sup>a</sup> No. of complexes as of 2016.<br>
<sup>b</sup> COMPLEAT includes protein complexes from *D. melanogaster, H. sapiens*, and *S. cerevisiae*. The EMBL-EBI portal includes protein complexes from 18 different species of which are *C. elegans* (16 complexes), *H. sapiens* (441), *M. musculus* (404), *S. cerevisiae* (399), and *S. pombe* (16). CORUM includes mammalian protein complexes, mainly from *H. sapiens* (64%), *M. musculus* (house mouse) (15%) and *R. norvegicus* (12%) (Norwegian rat).

<sup>c</sup> Includes mainly conserved complexes among the metazoans, *C. elegans, D. melanogaster, H. sapiens, M. musculus*, and *Stronglylocentrotus purpuratus* (purple sea urchin), consisting of 344 complexes with entirely ancient proteins and 490 complexes with largely ancient proteins conserved ubiquitously among eukaryotes.
dataset includes \( \sim 300 \) complexes composed of entirely ancient proteins (evolutionarily conserved from lower-order organisms), and \( \sim 500 \) complexes composed of largely ancient proteins conserved ubiquitously among eukaryotes. Drew et al. [2017] present a comprehensive catalog of \( >4,600 \) computationally predicted human protein complexes covering \( >7,700 \) proteins and \( >56,000 \) interactions by analyzing data from \( >9,000 \) published mass spectrometry experiments. Vinayagam et al. [2013] present COMPLEAT (http://www.flyrnai.org/compleat/), a database of \( 3,077, 3,636, \) and \( 2,173 \) literature-curated protein complexes from \( D. \ melanogaster, H. \ sapiens, \) and \( S. \ cerevisiae \), respectively. Ori et al. [2016] combined mammalian complexes from CORUM and COMPLEAT to generate a dataset of \( 279 \) protein complexes from mammals.

**Organization of the Rest of the Book**

The rest of this book reads as follows. Chapter 2 discusses important concepts underlying PPI networks and presents prerequisites for understanding subsequent chapters. We discuss different high-throughput experimental techniques employed to infer PPIs (including the Y2H and AP/MS techniques mentioned earlier), explaining briefly the biological and biochemical concepts underlying these techniques and highlighting their strengths and weaknesses. We explain computational approaches that denoise (PPI weighting) and integrate data from multiple experiments to construct reliable PPI networks. We also discuss topological properties of PPI networks, theoretical models for PPI networks, and the various databases and software tools that catalog and visualize PPI networks. Chapter 3 forms the main crux of this book as it introduces and discusses in depth the algorithmic underpinnings of some of the classical (seminal) computational methods to identify protein complexes from PPI networks. While some of these methods work solely on the topology of the PPI network, others incorporate additional biological information—e.g., in the form of functional annotations—with PPI network topology to improve their predictions. Chapter 4 presents a comprehensive empirical evaluation of six widely used protein complex prediction methods available in the literature using unweighted and weighted PPI networks from yeast and human. Taking a known human protein complex as an example, we discuss how the methods have fared in recovering this complex from the PPI network. Based on this evaluation, we explain in Chapter 5 the shortcomings of current methods in detecting certain kinds of protein complexes, e.g., protein complexes that are sparse or that overlap with other complexes. Through this, we highlight the open challenges that need to be tackled to improve coverage and accuracy of protein
complex prediction. We discuss some recently proposed methods that attempt to tackle these open challenges and to what extent these methods have been successful. Chapter 6 is dedicated to an important class of protein complexes that are *dynamic* in their protein composition and assembly. While some of these protein complexes are temporal in nature—i.e., assemble at a specific timepoint and dissociate after that—others are structurally variable—e.g., change their 3D structure and/or composition—based on the cellular context. Quite obviously, it is not possible to detect dynamic complexes solely by analyzing the PPI network; methods that integrate gene or protein expression and 3D structural information are required. These more-sophisticated methods are covered here. Chapter 7 discusses methods to identify protein complexes that are conserved between organisms or species; these *evolutionarily conserved* complexes provide important insights into the conservation of cellular processes through the evolution. Finally, in today’s era of *systems biology* where biological systems are studied as a complex interplay of multiple (biomolecular) entities, we explain how protein complex prediction methods are playing a crucial role in shaping up the field; these applications are covered in Chapter 8. We discuss the application of these methods for predicting dysregulated or dysfunctional protein complexes, identifying rewiring of interactions within complexes, and in discovery of new disease genes and drug targets. We conclude the book in Chapter 9 by reiterating the diverse applications of protein complex prediction methods and thereby the importance of computational methods in driving this exciting field of research.
The identification of PPIs yields insights into functional relationships between proteins. Over the years, a number of different experimental techniques have been developed to infer PPIs. This inference of PPIs is orthogonal, but also complementary, to experiments inferring genetic interactions; both provide lists of candidate interactions and implicate functional relationships between proteins [Morris et al. 2014].

2.1 High-Throughput Experimental Systems to Infer PPIs

Physical interactions between proteins are inferred using different biochemical, biophysical, and genetic techniques (summarized in Table 2.1). Yeast two-hybrid (Y2H; less commonly, YTH) [Ito et al. 2000, Uetz et al. 2000, Ho et al. 2002] and protein-fragment complement assays [Michnick 2003, Remy and Michnick 2004, Remy et al. 2007] enable identification of direct binary physical interactions between the proteins, whereas co-immunoprecipitation or affinity purification assays [Golemis and Adams 2002, Rigaut et al. 1999, Köcher and Superti-Furga 2007, Dunham et al. 2012] enable pull down of whole protein complexes from which the binary interactions are inferred. Protein-fragment complementation assay (PCA)
## Table 2.1 Experimental techniques for screening protein interactions; these techniques can be employed in a high-throughput manner to screen whole protein libraries for potential interactors

<table>
<thead>
<tr>
<th>Experimental Technique</th>
<th>Cell Assay</th>
<th>Interaction Type</th>
<th>Key References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yeast two-hybrid (Y2H)</td>
<td>In vivo (yeast, mammalian)</td>
<td>Binary interactions</td>
<td>[Ito et al. 2000, Uetz et al. 2000, Ho et al. 2002]</td>
</tr>
<tr>
<td>Co-immunoprecipitation precipitation followed by mass spectrometry (Co-IP/MS)</td>
<td>In vitro</td>
<td>Co-complex relationships</td>
<td>[Golemis and Adams 2002, Rigaut et al. 1999]</td>
</tr>
</tbody>
</table>
coupled with biomolecular fluorescence complementation (BIFC) [Grinberg et al. 2004] enables mapping of interaction surfaces of proteins, and is thus a good tool to confirm protein binding. Membrane YTH and mammalian membrane YTH (MaMTH) [Lalonde et al. 2008, Kittanakom et al. 2009, Lalonde et al. 2010, Petschnigg et al. 2014, Yao et al. 2017] enable identification of interactions involving membrane or membrane-bound proteins which are typically difficult to identify using traditional Y2H and AP techniques. Techniques inferring genetic interactions [Brown et al. 2006] enable detection of functional associations or genetic relationships between the proteins (genes), but these associations do not always correspond to physical interactions. Here, we present only an overview of each of the experimental techniques; for a more descriptive survey, the readers are referred to Brückner et al. [2009], Shoemaker and Panchenko [2007], and Snider et al. [2015].

**Yeast Two-Hybrid (Y2H) Screening System**

Y2H was first described by Fields and Song [1989] and is based on the modularity of binding domains in eukaryotic transcription factors. Eukaryotic transcription factors have at least two distinct domains: (1) the DNA binding domain (BD) which directs binding to a promoter DNA sequence (upstream activating sequence (UAS)); and (2) the transcription activating domain (AD) which activates the transcription of target reporter genes. Splitting the BD and AD domains inactivates transcription, but even indirectly connecting AD and BD can restore transcription resulting in activation of specific reporter genes. Plasmids are engineered to produce a protein product (chimeric or “hybrid”) in which the BD fragment is in-frame fused onto a protein of interest (the bait), while the AD fragment is in-frame fused onto another protein (the prey) (Figure 2.1). The plasmids are then transfected into cells chosen for the screening method, usually from yeast. If the bait and prey proteins interact, the AD and BD domains are indirectly connected, resulting in activation of reporters within nuclei of cells. Typically, multiple independent yeast colonies are assayed for each combination of plasmids to account for the heterogeneity in protein expression levels and their ability to activate reporter transcription.

This basic Y2H technique has been improved over the years to enable large library screening [Chien et al. 1991, Dufree et al. 1993, Gyuris et al. 1993, Finley and Brent 1994]. Interaction mating is one such protocol that can screen more than one bait against a library of preys, and can save considerable time and materials. In this protocol, the AD- and BD-fused proteins begin in two different haploid yeast strains with opposite mating types. These proteins are brought together by mating, a process in which two haploid cells fuse to form a single diploid cell. The diploids are then tested using conventional reporter activation for possible interactors.
Chapter 2  Constructing Reliable Protein-Protein Interaction (PPI) Networks

Figure 2.1  Schematic representation of the yeast two-hybrid protocol to detect interaction between bait and prey proteins. If the bait and prey proteins interact, the DNA binding domain (BD) fused to the bait and the transcription activating domain (AD) fused to the prey are indirectly connected resulting in the activation of the reporter gene. UAS: upstream activating sequence (promoter) of the reporter gene.

Therefore, different bait-expressing strains can be mated with a library of prey-expressing strains, and the resulting diploids can be screened for interactors. It is important to know how many viable diploids have arisen and to determine the false-positive frequency of the detected interactions. True interactors tend to come up in a timeframe specific for each given bait, with false positives clustering at a different timepoint. Multiple yeast colonies are assayed to confirm the interactors.

Y2H screens have been extensively used to detect protein interactions among yeast proteins, with two of the earliest studies reporting 692 [Uetz et al. 2000] and 841 [Ito et al. 2000] interactions for \( S.\) cerevisiae. In the bacteria Helicobacter pylori, one of the first applications of Y2H identified over 1,200 interactions, covering about 47% of the bacterial proteome [Rain et al. 2001]. Applications on fly proceeded on an even greater scale when Giot et al. [2003] identified 10,021 protein interactions involving 4,500 proteins in \( D.\) melanogaster. More recently, Vo et al. [2016] used Y2H to map binary interactions in the yeast \( S.\) pombe (fission yeast). This network, called FissionNet, consisted of 2,278 interactions covering 4,989 protein-coding genes in \( S.\) pombe. The Y2H system has also been applied for humans, with two initial studies [Rual et al. 2005, Stelzl et al. 2005] yielding over 5,000 interactions among human proteins. More recently, Rolland et al. [2014] employed Y2H to characterize nearly 14,000 human interactions.
2.1 High-Throughput Experimental Systems to Infer PPIs

However, inherent to this type of library screening, the number of detected false-positive interactions is usually high. Among the possible reasons for the generation of false positives is that the experimental compartmentalization (within the nucleus) for bait and prey proteins does not correspond to the natural cellular compartmentalization. Moreover, proteins that are not correctly folded under experimental conditions or are “sticky” may show non-specific interactions. The third source of false positives is the interaction of the preys themselves with reporter proteins, which can turn on the reporter genes. Von Mering et al. [2002] estimated the accuracy of classic Y2H to be less than 10%, with subsequent evaluations suggesting the number of false positives to be between 50% and 70% in large-scale Y2H interaction datasets for yeast [Bader and Hogue 2002, Bader et al. 2004].

Co-Immunoprecipitation/Affinity Purification (AP) Followed by Mass Spectrometry (Co-IP/AP followed by MS)

Complementing the in vivo Y2H screens are the in vitro Co-IP/AP followed by MS screens that identify whole complexes of interacting proteins, from which the binary interactions between proteins can be inferred [Golemis and Adams 2002, Rigaut et al. 1999, Köcher and Superti-Furga 2007, Dunham et al. 2012]. The Co-IP/AP followed by MS screens consist of two steps: co-immunoprecipitation/affinity purification and mass spectrometry (Figure 2.2). In the first step, cells are lysed

![Figure 2.2](image-url)  
**Figure 2.2** Schematic representation of the co-immunoprecipitation/affinity purification followed by mass spectrometry (Co-IP/AP followed by MS) protocol. The protein of interest (bait) is targeted with a specific antibody and pulled down with its interactors in a cell lysate buffer. The individual components of the pulled-down complex are identified using mass spectrometry. These days, liquid chromatography with mass spectrometry (LCMS) instead of running the gel is increasingly being used more frequently for as a combined physical-separation and MS-analysis technique [Pitt 2009].
in a radioimmunoprecipitation assay (RIPA) buffer. The RIPA buffer enables efficient cell lysis and protein solubilization while avoiding protein degradation and interference with biological activity of the proteins. A known member of the set of proteins (the protein of interest or bait) is epitope-tagged and is either immunoprecipitated using a specific antibody against the tag or purified using affinity columns recognizing the tag, giving the interacting partners (preys) of the bait. Normally, this purification step is more effective when two consecutive purification steps are used with proteins that are doubly tagged (hence called tandem affinity purification or TAP). This results in an enrichment of native multi-protein complexes containing the bait. The individual components within each such purified complex are then screened by gel electrophoresis and identified using mass spectrometry.

In one of the first applications of TAP/MS, Ho et al. [2002] expressed 10% of the coding open reading frames from yeast, and the identified interactions connected 25% of the yeast proteome as multi-protein complexes. Subsequently, Gavin et al. [2002], Gavin et al. [2006], and Krogan et al. [2006] purified 1,993 and 2,357 TAP-tagged proteins covering 60% and 72% of the yeast proteome, and identified 7,592 and 7,123 protein interactions from yeast, respectively. One of the first proof-of-concept studies for humans applied AP/MS to characterize interactors using 338 bait proteins that were selected based on their putative involvement in diseases, and the study identified 6,463 interactions between 2,235 proteins [Ewing et al. 2007].

**Comparison of Y2H and AP/MS Experimental Techniques**

A majority of the interaction data collected so far has come from Y2H screening. For example, approximately half of the data available in databases including IntAct [Hermjakob et al. 2004, Kerrien et al. 2012] and MINT [Zanzoni et al. 2002, Chatr-Aryamontri et al. 2007] are from Y2H screens [Brückner et al. 2009] (more sources of PPI data are listed in Table 2.2). This could in part be attributed to the inaccessibility of mass spectrometry due to the expensive large equipment that is required. But, in general, Y2H and AP/MS techniques are complementary in the kind of interactors they detect. If a set of proteins form a stable complex, then an AP/MS screen can determine all the proteins within the complex, but may not necessarily confirm every interacting pair (the binary interactions) within the complex. On the other hand, a Y2H screen can detect whether any given two proteins directly interact. While stable interactions between co-complexed proteins can be accurately determined using AP/MS techniques, Y2H techniques are useful for identifying transient interactions between the proteins. However, due to considerable func-
tional cross-talk within cells, Y2H can also report an interaction even when the proteins are not directly connected. In addition, some types of interactions can be missed in Y2H due to inherent limitations in the technique—e.g., interactions involving membrane proteins, or proteins requiring posttranslational modifications to interact—but these limitations may also occur with AP/MS-based approaches [Brückner et al. 2009]. Therefore, only a combination of different approaches that necessarily also includes computational methods (to filter out the incorrectly detected interactions) will eventually lead to a fairly complete characterization of all physiologically relevant interactions in a given cell or organism.

**Protein-Fragment Complementation Assay (PCA)**

PCA is a relatively new technique which can detect *in vivo* protein interactions as well as their modulation or spatial and temporal changes [Michnick 2003, Morell et al. 2009, Tarassov et al. 2008]. Similar to Y2H, PCA is based on the principle of splitting a reporter protein into two fragments, each of which cannot function alone [Michnick 2003]. However, unlike Y2H, PCA is based on the formation of a biomolecular complex between the bait and prey, where both are fused to the split domains of the reporter. Importantly, the formation of this complex occurs in competition with alternative endogenous interaction partners present within the cell. The interaction brings the two split fragments in proximity enabling their non-covalent reassembly, folding, and recovery of protein reporter function [Morell et al. 2009]. Typically, the reporter proteins are fluorescent proteins, and the formation of biomolecular complexes is visualized using biomolecular fluorescence complementation (BIFC). BIFC can also be used to map the interaction surfaces of these complexes. This enables investigation of competitive binding between mutually exclusive interaction partners as well as comparison of their intracellular distributions [Grinberg et al. 2004].

PCA can be used as a screening tool to identify potential interaction partners of a specific protein [Remy and Michnick 2004, Remy et al. 2007], or to validate the interactions detected from other techniques such as Y2H [Vo et al. 2016]. In one of the first applications of PCA on a genome-wide *in vivo* scale, Tarassov et al. [2008] identified 2,770 interactions among 1,124 proteins from *S. cerevisiae*. Vo et al. [2016] used PCA as an orthogonal assay to reconfirm the interactions detected in *S. pombe* (from the FissionNet network consisting of 2,278 interactions; discussed earlier). PCA has also been employed to validate interactions between membrane proteins or membrane-associated proteins [Babu et al. 2012, Shoemaker and Panchenko 2007] (discussed next).
Techniques for Inferring Membrane-Protein Interactions

Membrane proteins are attached to or associated with membranes of cells or their organelles, and constitute approximately 30% of the proteomes of organisms [Carpenter et al. 2008, Von Heijne 2007, Byrne and Iwata 2002]. Being non-polar (hydrophobic), membrane proteins are difficult to crystallize using traditional X-ray crystallography compared to soluble proteins, and are the least studied among all proteins using high-throughput proteomics techniques [Carpenter et al. 2008].

Membrane proteins are involved in the transportation of ions, metabolites, and larger molecules such as proteins, RNA, and lipids across membranes, in sending and receiving chemical signals and propagating electrical impulses across membranes, in anchoring enzymes and other proteins to membranes, in controlling membrane lipid composition, and in organizing and maintaining the shape of organelles and the cell itself [Lodish et al. 2000]. In humans, the G-protein-coupled-receptors (GPCRs), which are membrane proteins involved in signal transduction across membranes, alone account for 15% of all membrane proteins; and 30% of all drug targets are GPCRs [Von Heijne 2007]. Due to the key roles of membrane proteins, identifying interactions involving these proteins has important applications especially in drug development.

Membrane protein complexes are notoriously difficult to study using traditional high-throughput techniques [Lalonde et al. 2008]. Intact membrane-protein complexes are difficult to pull down using conventional AP/MS systems. This is due in part to the hydrophobic nature of membrane proteins as well as the ready dissociation of subunit interactions, either between trans-membrane subunits or between trans-membrane and cytoplasmic subunits [Barrera et al. 2008]. Further, membrane protein structure is difficult to study by commonly used high-resolution methods including X-ray crystallography and NMR spectroscopy.

A major avenue by which one can understand membrane proteins and their complexes is by mapping the membrane-protein “subinteractome”—the subset of interactions involving membrane proteins. Conventional Y2H system is confined to the nucleus of the cell thereby excluding the study of membrane proteins. New biochemical techniques have been developed to facilitate the characterization of interactions among membrane proteins. Among these is the split-ubiquitin membrane yeast two-hybrid (MYTH) system [Miller et al. 2005, Kittanakom et al. 2009, Stagljar et al. 1998, Petschnigg et al. 2012]. This system is based on ubiquitin, an evolutionarily conserved 76-amino acid protein that serves as a tag for proteins targeted for degradation by the 26S proteasome. The presence of ubiquitin is recognized by ubiquitin-specific proteases (UBPs) located in the nucleus and cytoplasm of all eukaryotic cells. Ubiquitin can be split and expressed as two halves: the amino-
terminal (N) and the carboxyl terminal (C). These two halves have a high affinity for each other in the cell and can reconstitute to form pseudo-ubiquitin that is recognizable by UBPs.

In MYTH, the bait proteins are fused to the C-terminal of a split-ubiquitin, and the prey proteins are fused to the N-terminal. The two halves reconstitute into a pseudo-ubiquitin protein if there is affinity between the bait and prey proteins. This pseudo-ubiquitin is recognized by UBPs, which cleaves after the C-terminus of ubiquitin to release the transcription factor, which then enters the nucleus to activate reporter genes.

Two of the earliest studies using the MYTH screens reported a fair number of interactions among membrane proteins from yeast: 343 interactions among 179 proteins by Lalonde et al. [2010], and 808 interactions among 536 proteins by Miller et al. [2005]. PCA has also been adopted to identify and/or verify membrane-protein interactions. For example, Babu et al. [2012] used PCA to validate and integrate 1,726 yeast membrane-protein interactions obtained from multiple studies, and these encompassed 501 putative membrane protein complexes.

The mammalian version of membrane yeast two-hybrid, MaMTH, is also based on the split-ubiquitin assay and is derived from the MYTH assay. Stagljar and colleagues [Petschnigg et al. 2014, Yao et al. 2017] used MaMTH to probe interactions involving the epidermal growth factor receptor/receptor tyrosine-protein kinase (RTK) ErbB-1 (EGFR/ERBB1), Erb-B2 receptor tyrosine kinase 2 (ERBB2), and other RTKs that localize to the plasma membrane in human cells. When applied to human lung cancer cells, the assay identified 124 interactors for wild-type and mutant EGFR [Petschnigg et al. 2014].

### 2.2 Data Sources for PPIs

Several public and proprietary databases now catalog protein interactions from both lower-order model and higher-order organisms (summarized in Table 2.2). These databases contain PPI data in an acceptable format required for data deposition, such as IMEx (http://www.imexconsortium.org/submit-your-data) [Orchard et al. 2012]. The Biomolecular Interaction Network Database (BIND) [Bader et al. 2003], now called Biomolecular Object Network Database (BOND), includes experimentally determined protein-protein, protein-small molecule, and protein-nucleic acid interactions. BioGrid [Stark et al. 2011] catalogs physical and genetic interactions inferred from multiple high-throughput experiments. The Database of Interacting Proteins (DIP) [Xenarios et al. 2002] contains experimentally determined protein interactions with a “core” subset of interactions that have passed quality
<table>
<thead>
<tr>
<th>PPI Database</th>
<th>Source</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIND</td>
<td><a href="http://bind.ca">http://bind.ca</a></td>
<td>[Bader et al. 2003]</td>
</tr>
<tr>
<td>CYGD</td>
<td><a href="http://mips.helmholtz-muenchen.de/genre/proj/yeast/">http://mips.helmholtz-muenchen.de/genre/proj/yeast/</a></td>
<td>[Güldener et al. 2005]</td>
</tr>
<tr>
<td>OPHID/IID</td>
<td><a href="http://ophid.utoronto.ca/">http://ophid.utoronto.ca/</a></td>
<td>[Brown and Jurisica 2005]</td>
</tr>
<tr>
<td>InnateDB</td>
<td><a href="http://www.innatedb.com/">http://www.innatedb.com/</a></td>
<td>[Lynn et al. 2008]</td>
</tr>
<tr>
<td>MIPS</td>
<td><a href="http://mips.helmholtz-muenchen.de/proj/ppi/">http://mips.helmholtz-muenchen.de/proj/ppi/</a></td>
<td>[Mewes et al. 2008]</td>
</tr>
<tr>
<td>MPPI</td>
<td><a href="http://mips.helmholtz-muenchen.de/proj/ppi/">http://mips.helmholtz-muenchen.de/proj/ppi/</a></td>
<td>[Pagel et al. 2005]</td>
</tr>
<tr>
<td>STRING</td>
<td><a href="http://string-db.org/">http://string-db.org/</a></td>
<td>[Von Mering et al. 2003, Szklarczyk et al. 2011]</td>
</tr>
</tbody>
</table>

2.3 Topological Properties of PPI Networks

A simple yet effective way to represent interaction data is in the form of an undirected network called a protein-protein interaction network or simply PPI network, given as $G = (V, E)$, where $V$ is the set of proteins and $E$ is the set of physical interactions between the proteins. Such a network presents a global or “systems” view of the entire set of proteins and their interactions, and provides a topological (mathematical) framework to interrogate the interactions. In the definitions throughout this book, we also use $V(G)$ and $E(G)$ to refer to the set of proteins and interactions of a (sub)network of $G$. For a protein $v \in V$, the set $N(v)$ or $N_v$ includes all immediate neighbors of $v$, and $\text{deg}(v) = |N(v)|$ is the degree of $v$. These neighbors together with their interactions, $E_v = E(v) = \{(v, u) : u \in N(v)\} \cup \{(u, w) : u \in N(v), w \in N(v) \cap N(u)\}$, constitute the local (immediate) neighborhood subnetwork of $v$.

PPI networks, like most real-world networks, have characteristic topological properties which are distinct from that of random networks. But, to understand this distinction we need to first understand what are random networks. Traditionally, random networks have been described using the Erdős-Rényi (ER) model, in which $G(n, p)$ is a random network with $|V| = n$ nodes and each possible edge connecting pairs of these nodes has probability $p$ of existing [Erdős 1960, Bollobás 1985]. The expected number of edges in the network is $\binom{n}{2}p$, and the expected mean degree is $np$. Alternatively, a random network is defined as a network chosen uniformly at random from the collection $\binom{\binom{n}{2}}{m}$ of all possible networks with $n$ nodes and $m$ edges. If $p$ is the probability for the existence of an edge, the probability for each network...