Learning of Protein Interaction Networks

Yanjun Qi

May 2008

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
qyj@cs.cmu.edu

A dissertation submitted to Carnegie Mellon University
in partial fulfillment of the requirements
for the degree of Doctor Of Philosophy

Thesis Committee:

Ziv Bar-Joseph (Carnegie Mellon University, Chair)
Judith Klein-Seetharaman (University of Pittsburgh & Carnegie Mellon University, Chair)
Christos Faloutsos (Carnegie Mellon University)
Jaime Carbonell (Carnegie Mellon University)
Baldo Oliva (Universitat Pompeu Fabra, Spain)

Keywords: Biological Network Mining, Relational Network Mining, Information Integration, Supervised Learning, Supervised Classification, Graph Mining, Group Detection, Protein-Protein Interactions, Protein Complexes, Protein Interaction Networks

To Mom, Dad, Ruifeng and Sophia for love and support.

Abstract

Protein-protein interactions (PPI) play a key role in determining the outcome of most cellular processes. Correctly identifying and characterizing protein interactions and the networks they comprise is critical for understanding the molecular mechanisms within the cell.

Large-scale biological experimental methods can directly and systematically detect the set of interacting proteins within an organism. Unfortunately, the resulting datasets are often incomplete and exhibit high false positive and false negative rates. In addition to the direct experimental data, a number of large biological datasets also provide indirect evidence about protein-interaction relationships. Thus computational approaches could be utilized to combine multiple information sources in order to predict the sets of interacting protein pairs and identify important biological substructures in this network.

In this dissertation, we first carry out a systematic study of the efficacy of using supervised learning methods to integrate direct and indirect biological evidence for predicting pairwise protein interactions. The results indicate that the utility of information, the way the data is encoded as features, the target types of protein interactions and the computational approaches used are all significant for predicting such interactions. We then propose four learning algorithms for deriving PPI networks from different perspectives.

(I) A combined computational and experimental approach is proposed for predicting interaction partners of human membrane receptors. The random forest binary classifier is employed to determine if a potential receptor-human pair interacts or not. Biological feedback is used to optimize feature encoding and improve the accuracy of predictions. The resulting receptor PPI network is then analyzed through graph property analysis, graph module identification and protein-family related network pattern search. Several novel predictions are further experimentally validated. Our proposed framework shows that focusing on specific subnetworks generates better predictions. The predicted network provides the most reliable dataset on the network of interactions involving human membrane receptors to date.

(II) Considering that PPI networks are highly sparse graphs and there is no large negative reference set (non-interacting pairs) available, we design a ranking approach to identify

candidate interaction pairs that are "similar" to known interacting pairs. Robust similarity estimation is especially important here because of high noise rates and the problem of many missing values in biological data. Our ranking method determines the degree of similarity between protein pairs using a trained random forest model. The similarity is, then, used by a weighted k-Nearest-Neighbor algorithm to rank candidate protein pairs. Applying the algorithm on yeast data produces robust performance results that compare favorably with previously suggested methods.

(III) A multiple-view learning strategy (referred to as "Mixture of Feature Experts") is further proposed for predicting PPIs that takes into account the heterogeneous nature of feature properties. First, features are split into roughly homogeneous groups. Then, each individual group (called "expert") gives classification opinions and their scores are combined using weighted voting. Different experts have different degrees of influence on the prediction depending on the available features. When applied to yeast and human species, this method improves upon the generally used methods, and the weighting of the experts provides a means to evaluate the prediction based on high scoring feature experts.

(IV) "Protein complex" (a special group formation) is one typical pattern contained in protein-protein interaction networks. We present an algorithm for inferring protein complexes based on graph topological patterns and biological properties. Each complex subgraph is modeled by using a probabilistic Bayesian Network. The derived log-likelihood ratio is then used to score subgraphs in the protein interaction graph and to identify new complexes. We apply this method to protein interaction data in yeast. Our algorithm recovers known complexes much better than previous clique-based algorithms.

In summary, our proposed algorithms provide strong computational tools for predicting and analyzing protein-protein interaction networks. They have been applied successfully in yeast and human, and have generated promising results. For instance, without the novel interaction between rhodopsin and chemokines found by our computational approach, the important functional implication of rhodopsin in the immune system would not have been possibly discovered.

Acknowledgements

First and foremost, I would like to thank my advisors, Ziv Bar-Joseph and Judith Klein-Seetharaman who brought me into the research field of computational biology. Their advice, support and encouragement helped me finish this dissertation. Ziv inspired me to actively investigate a lot of interesting and significant learning problems. Judith guided me to look at and understand a broader field of biological challenges. I would also like to thank my other committee members, Christos Faloutsos, Jaime Carbonell and Baldo Oliva. Christos gave me tremendous helps and suggestions in investigating graph mining techniques. Jaime gave me insightful advice and helped me define a better scope of this dissertation. Baldo's suggestions and comments greatly improved my understanding of the protein interaction prediction task.

This work would not have been possible without the supports and interests of many individuals at CMU and U.Pitt. I would like to thank John Lafferty, Eric Xing, Roni Rosenfeld, Alexander Hauptmann and Andrew Moore for their valuable discussions, suggestions and/or helps. I would also like to thank my colleagues in system biology group and the biological language modeling group for their valuable comments and contributions. Those who deserve of specific mentions include Harpreet Dhiman, Ivan Budyak, Eric Gardner, Neil Bhola, Naveena Yanamala, David Man, Arpana Dutta, Kalyan Tirupula, Fernanda Balem, Oznur Tastan, Yong Lu, Rebecca Hutchinson and Jason Ernst.

Many thanks go to my friends and fellow graduate students in Pittsburgh: Ulas Bardak, Vasco Pedro, Betty Cheng who were the best officemates; Brooke Hyatt, Monica Hopes and Stacey Young who gave the best staff support; Qin Jin, Jie Lu, Chun Jin, Yi Zhang, Xiaojing Zhu, Rong Jin, Shimin Chen, Jeongwoo Ko, Jean Oh, Ting Liu, Jian Zhang, Fan Li, Yan Liu, Jiazhi Ou, Minglong Shao, Yanghai Tsin, Xiaofang Wang, Yan Li, Ke Yang, Rong Yan, Fan Li, Zhenzhen Kou, Kaiping Chen, Xinghua An, Yanhua Hu, Jimeng Sun, Yang Liu, Joy Zhang and many others for making my graduate life more fun and enjoyable.

Finally, my deepest gratitude goes to my family. My parents, Zhongqiang Qi and Huiyu Li, give me endless love and persistent supports. My husband, Ruifeng Sun, has encouraged me all the times with his love, patience, and understanding during this long CMU journey. My brother Jianjun Qi always believes in and supports me. Last but not the least, my lovely daughter Sophia has cheered me up to finish this thesis with her sweet smiles.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In recent years, the human and other genome sequencing projects have generated vast amounts of data that identified the existence of thousands of new gene products whose functions and interrelationships are not yet known. The overall molecular architecture of all organisms is largely mediated both structurally and functionally through the elaborate coordination of protein-protein interactions. In particular, the distortion of protein interactions may lead to the development of diseases. Thus correctly identifying the interrelationship between proteins at the system level is urgent and necessary, since it would lead to a better understanding of the functional properties that define real world behaviors of most complex biological systems [1].

1.1 Motivation

Currently most protein interactions remain to be discovered [2, 3]. A number of large-scale (or high-throughput) experimental approaches have been applied to define sets of interacting proteins on a proteome-wide scale (details in Chapter 2). However, the generated interaction datasets are often incomplete and highly noisy [2]. Also there is surprisingly little convergence in the data generated by different detection methods, suggesting that they are non-saturating, erroneous, or both [2].

Considering these limitations in experimental data and the urgent need to identify protein interrelationships at the system level, additional approaches are needed to accelerate

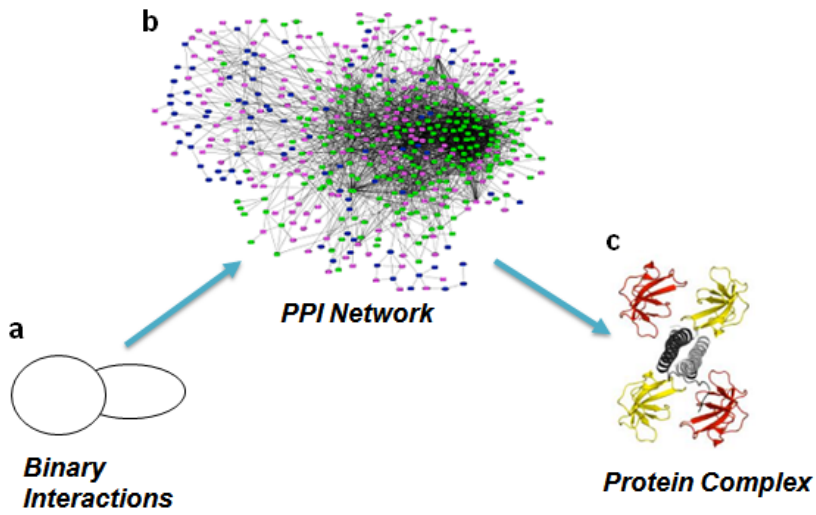

Figure 1.1: Three computational challenges in learning of protein interaction networks. (a) With the majority of interactions missing, how to recover pairwise interaction edges in the network is the first problem to solve. (b) Binary protein interactions predicted on the proteome-wide scale could be assembled into large network topologies. It is important to investigate the global properties of PPI networks and the biological implication behind network properties. (c) Proteins often collaborate together as a unit (called "complex") to exert specific functions. Finding complexes on PPI graphs is an important challenge.

the recovery of complex protein-interaction systems. Given the vast amount of available biological evidence and the current representative ability of mathematical models, computational methods are gaining importance in almost all related research areas.

Figure 1.1 outlines an overview of three computational challenges covered in this dissertation. The challenges, as explained below, correspond to different portions of the entire PPI graph. Careful data analysis and correct problem formulation are necessary to build successful computational algorithms for these tasks.

- (a) The majority of protein interactions are still unknown. The first crucial problem to be solved is how to recover missing edges (pairwise protein-protein interactions).

- (b) Binary protein interactions predicted on a proteome-wide scale could be assembled into large network topologies. The global properties of PPI networks across species are needed to decipher the biological implications behind network properties. Related topics of interests also include the network implications for the stability

of biological systems.

- (c) Proteins often collaborate together and form stable associations as a group (called "complex"). Subgroups or complexes within the interaction network often appear as subgraph patterns. Identification of such complexes on PPI graphs is an important challenge in understanding the cell.

1.2 Thesis Overview

Computationally, protein-protein interaction (PPI) networks can be conveniently modeled as undirected graphs, where the nodes are proteins and any two nodes are connected by an undirected edge, if the corresponding proteins bind physically. Currently, this graph contains many noisy edges, a large portion of which are missing. However since a number of large scale biological datasets provide indirect evidence for protein interaction relationships, we could integrate the available information sources to recover and analyze PPI networks.

This dissertation focuses primarily on three challenges (Figure 1.1): (1) To infer how likely each possible pair of proteins interacts; (2) To identify significant substructures (called "complexes") on the PPI graphs; (3) To analyze the global properties of the interaction graphs.

Thesis Statement: *This dissertation provides a systematic computational framework for discovering protein-protein interactions (PPI) and for identifying important patterns within PPI networks. The computational predictions yielded by this framework suggest a number of novel biological hypotheses that have been verified with subsequent laboratory experimentations.*

A systematic study and four novel methods are proposed and have been applied in multiple species. These techniques can be very useful for choosing potential targets for experimental screening or for validating experimental data.

Systematic Study of PPI Prediction through Information Fusion By transforming multiple direct and indirect biological data sources into a feature vector representing every pair of proteins, the task of predicting pairwise protein interactions can be formalized as a binary classification problem. Many different research groups have independently suggested using supervised learning methods for predicting protein interactions. However, the data sources, approaches and the means of implementations have varied widely. The protein interaction prediction task itself can be sub-divided into predictions of (1) physical interaction, (2) co-complex relationship and (3) pathway co-membership. To systematically investigate the utility of different data sources and how the data is encoded as features for predicting each of these types of protein interactions, a large set of biological features was assembled and their encoding was varied for use in each of the three prediction tasks. Different classifiers were used to assess the accuracy in predicting interactions. For all classifiers, the three prediction tasks had different success rates. Independently of prediction task, the random forest classifier consistently ranked as one of the top two methods. In addition, the importance of different biological datasets also varies across specific interaction tasks and styles for encoding features. (Related paper: [4])

Subnetwork PPI Prediction by A Combined Computational and Experimental Approach Membrane receptor-activated signal transduction pathways play an essential role in both cellular functions and disease mechanisms of humans. Thus far, identification of the full set of proteins interacting with membrane receptors by high-throughput experimental means has been impossible because methods used to directly identify protein interactions can't generally be applied to membrane proteins. We instead design a combined framework to investigate protein-protein interactions related to human membrane receptors both computationally and experimentally. First we extract features from diverse biological data sources, including sequence, structure, function and genomic information. We predict specific interactions involving receptors using the random forest applied to the binary classification task of whether two proteins interact or not. Biological feedbacks have been used both to optimize feature encoding and to improve the predictions. The interactions determined for all human membrane receptors make up the human membrane receptor network. By analyzing global-level properties of this network we then identify receptor hubs,

reveal strongly interacting clusters, highlight the abundance of receptor-receptor interactions, and identify ligands shared between receptors. Finally, we have validated some of the predictions made by the classifier experimentally. Our prosed framework clearly shows that focusing on specific subnetworks generates better predictions than treating this PPI prediction in human generally. The predicted network provides a reliable dataset on the network of interactions involving human membrane receptors. (Related papers: [5, 6, 7]).

PPI Prediction through Ranking The distribution between PPI pairs and the non-interacting protein pairs in the above classification setting is highly skewed. Also when considering that there is no large negative reference set (non-interacting pairs) available, we present a ranking approach to identify candidate interaction pairs that are ”similar” to known interacting protein pairs. Estimation of robust similarity is especially important because of high noise rates and heavy missing value problems associated with the biological features used. A novel method is proposed to compute such similarities in order to classify pairs of proteins as either interacting or not. Our method uses direct and indirect information about interaction pairs to construct a random decision forest from a training set. The resulting forest is then used to determine the similarity between protein pairs and this similarity is incorporated into a classification algorithm (a modified k-Nearest-Neighbor) to classify protein pairs. Testing the algorithm on yeast data indicates that the performance of this approach compares favorably with previously suggested methods for this task, verifying the importance of robust similarity as well. (Related paper: [8])

PPI Prediction from Multiple-View Learning When integrating direct and indirect data to predict interactions, most proposed methods utilize a common classifier for all pairs. However, due to missing data and high redundancy among the features used, different protein pairs may benefit from the use of different features based on the set of attributes available. In addition, in many cases it is hard to directly determine which of the data sources contribute to a prediction. This information is important for biologists using these predications in the design of new experiments. To address these challenges we propose a multiple-view classification strategy for protein-protein interaction prediction, called ”Mixture of Feature Experts”. We split the features into roughly homogeneous sets of feature

groups (called ”experts”). Logistic regression is used on each individual expert and the resulting scores (opinions) are combined through weighted voting to generate the final decision (using another logistic regression). When combining the scores, the weighting of each expert depends on the set of input attributes available for that pair. Thus, different experts will have more or less influence on the prediction depending on the available features. We have applied our method to predict the set of interacting proteins in yeast and human. Our results improve upon the results obtained using previous methods for this task. In addition, the weighting of the experts provides ways to evaluate a specific prediction based on high scoring feature experts. (Related papers: [9, 10])

Complex Identification by Supervised Graph Clustering Protein complexes integrate multiple gene products to coordinate various biological functions. Given a graph representing all pairwise protein interactions, one can search for subgraphs to identify protein complexes. Previous methods for performing such a search were based on the assumption that complexes would form a clique in the PPI graph. While this assumption is true for some complexes, it does not hold for many others. New algorithms are required to recover complexes having other types of topological structure. We present an algorithm for inferring protein complexes from weighted interaction graphs. By using graph topological patterns and biological properties as features, we model each complex subgraph by a probabilistic Bayesian Network (BN). We then use a training set of known complexes to learn the parameters of this BN model. The log-likelihood ratio derived from the BN is then used to score subgraphs in the protein interaction graph and identify new complexes. A heuristic local search strategy is proposed to identify potential complexes. We have applied our method to protein interaction data in yeast. As we will show our algorithm achieves a considerable improvement over clique based algorithms in terms of its ability to recover known complexes. We then investigate some of the new complexes predicted by our algorithm and find that they likely represent true complexes. (Related paper: [11])

1.3 Thesis Organization

Chapter 2 provides a brief introduction to biological background on protein interactions and PPI networks. Chapter 3 summarizes the related literature and describes our systematic comparison of predicting pairwise PPIs through information fusion. Chapter 4 presents a combined approach for detecting PPIs for human membrane receptors and explains how we chose computational predictions for biological validations. In Chapter 5, we describe a ranking strategy for PPI predictions. Chapter 6 illustrates the "Mixture of Feature Experts" method which identifies PPIs through a weighted voting of multiple views. It provides guiding information to help the design of laboratory experiments relating to PPIs. Chapter 7 presents our work on identifying protein complexes in the protein interaction network through supervised graph analysis. Chapter 8 summarizes the contributions of the research described in this dissertation and discusses potential improvements and interesting future projects.

Chapter 2

Biological Background

A brief introduction of the motivation of this dissertation has been made in Chapter 1. Then in this chapter, we present basic information about proteins, protein-protein interactions, protein interaction networks, and other related biological interactions.

The term ”protein-protein interactions” (PPI) refers to the association of protein molecules with each other. The associations are interesting from multiple perspectives such as general research areas like biochemistry and biophysics, specific biological processes and pathways such as signal transduction pathways, and system-level studies of networks on the organism-wide scale.

2.1 Proteins and Protein Function

Proteins are biosynthetic polymers composed of covalently connected amino acid units. They are involved in practically every function performed by a cell. Several important functional classes include [12]: (1) enzymes, which catalyze, for example, the many of the reactions of metabolism; (2) structural proteins, such as collagen which is the main protein of connective tissue in animals; (3) regulatory proteins, such as transcription factors that regulate the transcription of genes; (4) signalling molecules, such as certain hormones, like insulin, and their receptors; and (5) defensive proteins such as antibodies of the immune system.

Owing to the advent of high-throughput sequencing techniques, the complete sequences

of several genomes are now known. However, the biological function of a large proportion of sequenced proteins remains to be identified. Moreover, a given protein may have more than one function, so many proteins that are known to be in some class may have as yet undiscovered functionalities. Predicting protein functions is one of the most important challenges of current computational biology research. To facilitate such research, various biological data could be used, including sequence, gene expression patterns, phylogenetic profiles, domain fusions and so on.

Protein-protein interactions operate at almost every level of cellular functions. Thus, implications about function can often be made via protein-protein interaction studies. These inferences are based on the premise that the function of unknown proteins may be discovered through studying their interaction with a known protein target having a known function [13, 14, 15]. The study of protein interactions will help us understand how proteins function within the cell.

2.2 Protein-Protein Interaction (PPI)

Though some percent of proteins can be expected to work in relative isolation, the majority operate in coordination with other proteins in PPI networks to arrange the processes revolving around cellular structure and function. These processes include cell cycle control, differentiation, protein folding, signaling, transcription, translation, post-translational modification, and transport [15]. Protein interactions play key roles in these processes. For instance, signals from the exterior of a cell to the inside of that cell are conveyed by protein-protein interactions of the signaling molecules. This process, called signal transduction, plays a fundamental role in many biological processes and in many diseases (e.g., cancer). A protein may modify another protein via interaction. For example, a protein kinase will add a phosphate to a target protein. Such modification of proteins can itself change protein-protein interactions. Given protein-protein interactions are of central importance for virtually every process in a living cell, information about these interactions improves our understanding of diseases and can provide the basis for new therapeutic approaches [16].

To properly understand the significance of protein-protein interactions in the cell, one needs to identify different types of interactions, understand the extent to which they take place in the cell, and determine their consequences.

Types of Protein Interactions Protein interactions can be classified based on a number of different features [16]:

- Their strength: stable or transient. Stable and transient interactions can be either strong or weak [14]. (1) *Stable* interactions are usually associated with proteins that are purified as multi-subunit complexes. Stable interactions are best studied by co-immunoprecipitation, pull-down or far-Western methods [13] (Table 2.1). (2) *Transient* interactions are believed to control the majority of cellular processes. As the name implies, transient interactions are on/off or temporary in nature and typically require a set of conditions that stimulate the interaction. Transient interactions can be captured by cross-linking or label-transfer methods [13] (Table 2.1).

- Their specificity: specific or nonspecific. A specific interaction means that one protein could only interact with another specific protein partner.

- The similarity between interacting subunits: homo-oligomers or hetero-oligomers. A protein complex made of several different protein subunits is called a hetero-oligomer. When only one type of protein subunit is used in the complex, it is called homo-oligomer.

Protein Interaction Network To understand complex phenomena of protein-protein interactions in the cell, researchers often assemble the available protein-protein interactions of a certain species and analyze the system (also known as ”protein interactome”) from several different perspectives, which are illustrated by Figure 2.1 (adapted from [1]).

- (a) To gain a comprehensive understanding of the mechanisms underlying a biological function, researchers often investigate one or several protein interactions at a time. These are referred as *small-scale* experiments.

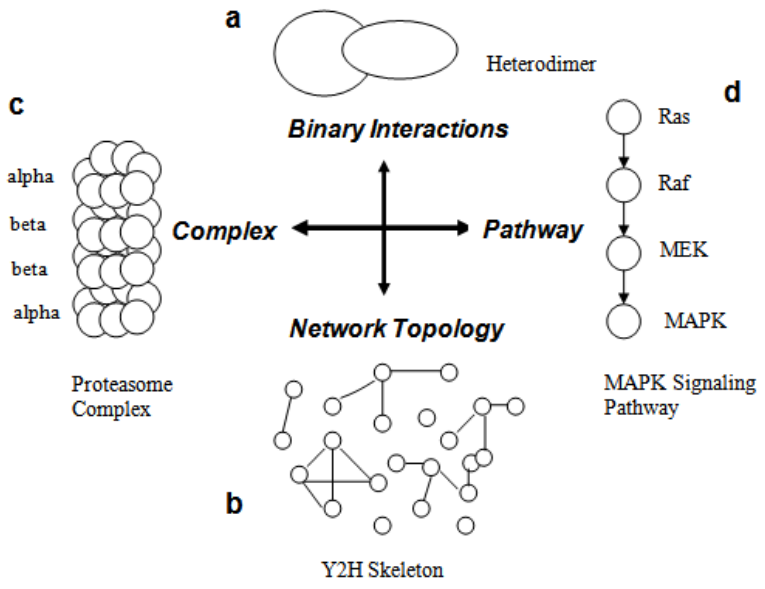

Figure 2.1: Overview of the protein interaction network (Modified from Figure 1 of [1], included as background information only.). (a) Binary protein interactions can be assessed at small scale for a comprehensive understanding of the mechanisms and functional purpose of an interaction. (b) Experiments on a genome-wide scale could test hundreds/thousands of protein interactions at one time and be assembled into large network topologies. (c) Some biological functions require the formation of stable complexes of multiple protein units, as is the case (for instance the proteasome). (d) Some interactions occur only if other interactions have taken place prior. This is often the case during signal transduction processes and such logically connected interactions are referred to as pathways (such as the MAPK signaling pathway).

- (b) Recently *large-scale* yeast two-hybrid (Y2H) screening analysis has been applied to several model species to detect interactions across the entire proteome (all possible proteins) of an organism. The resulting thousands of protein interactions allow to construct a topological skeleton of the entire protein interaction network.

- (c) Within the interaction network, some proteins form associations with multiple protein binding partners to build what is referred to as complexes. In a complex, proteins are in close proximity to allow them to work together (for instance, the proteasome complex). Often these complexes can be stable units, that do not change

significantly in composition over time, while others are highly dynamic and the composition changes as a function of the state in which a cell is. Large-scale proteomic studies that used comprehensive methods of affinity tagging and purification have been remarkably successful at identifying the components of complexes.

- (d) Some interactions such as can be observed in many signaling pathways (like the MAPK signaling pathway) follow logically on another. As the example shown in Figure 2.1d, MEK only phosphorylates MAPK, if MEK was activated by Raf and Raf prior by Ras. In such pathways, interactions are often transient and occur only under the right circumstances. Transient interactions are generally under-studied by most large-scale experimental procedures.

2.3 Biological Experiments for PPI Detection

Because of their importance in development and disease, protein-protein interactions have been the object of intense research in recent years. The interactions among proteins can take on many forms (e.g., be subject to the same regulation, have an impact on functions of one another, or occur in a common pathway), and many proteins only operate in complexes and through physical contact with other proteins. These factors have prompted the development of various experimental methods for detecting protein-protein interactions [13, 16].

2.3.1 Small-scale PPI Experiments

Traditionally, protein interactions have been studied individually through the use of genetic, biochemical and biophysical techniques (also termed *small-scale* methods) [13]. Small-scale experiments to select and detect proteins that bind to other proteins could be performed by measuring the natural affinity of binding partners through either *in vitro* approaches or *in vivo* by the yeast two-hybrid system [16].

In vitro methods: *In vitro* means performing a given experiment in a controlled environment outside of a living organism. All methods have their advantages and disadvantages

Table 2.1: Widely employed *in vitro* biological experimental methods for identifying protein-protein interactions [13, 16].

Method Name	**Short Description**
Protein Arrays	Antibody-based or bait-based arrays detect interactions of proteins from complex mixtures
Surface Plasmon Resonance	Relate binding to small changes in laser light reflected from gold surfaces where a bait protein is attached
Co-Immunoprecipitations	A purification procedure to determine if two different proteins interact
FRET	Fluorescence Resonance Energy Transfer (FRET) studies the transfer of two interacting proteins carrying fluorescence labels
Label Transfer	Tag a known protein with a detectable label, then detect interaction partners by the presence of the label
Far Western	Employ non-antibody proteins to detect the protein(s) of interest on the blot
NMR	Nuclear Magnetic Resonance (NMR) provides insights into the dynamic interaction of proteins in solution
X-ray Crystallography	Crystallization of the interacting complex allows definition of the interaction structure

and generally provide complementary information. Together they can provide valuable insights into protein interactions [13]. Widely employed methods are briefly listed below and in Table 2.1.

- Experimental techniques used in identifying novel/unknown interactions involve surface plasmon resonance and/or pull down assays (e.g., co-immunoprecipitations).
- Techniques used for confirming known/predicted interactions include FRET (fluorescence resonance energy transfer), label-transfer and far-western analysis.
- Nuclear Magnetic Resonance (NMR) and X-ray Crystallography, are employed to investigate interactions at the atomic level (the vast majority binary interactions).

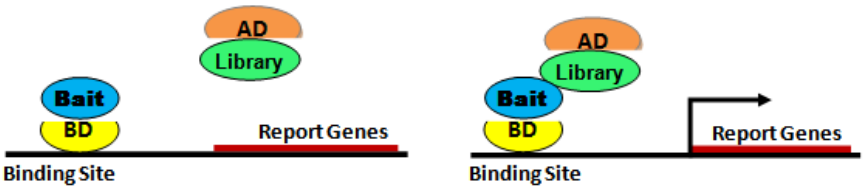

Figure 2.2: The yeast-two-hybrid system. (a). The DNA-binding domain hybrid does not activate transcription if protein "Bait" does not contain an activation domain. The activation domain hybrid does not activate transcription either because it does not localize to the DNA-binding site. (b). Interaction between "Bait" and other proteins form a libray brings the activation domain into close proximity to the DNA-binding site and results in transcription of a reporter gene.

In vivo systems: *In vivo* refers to a reaction that is taking place inside an organism. The most widely used *in vivo* system to study protein interactions is the "yeast two-hybrid" (Y2H) system. The Y2H uses the transcription process to identify protein interactions (see further details below). Interactions indicated by this approach often require the confirmation from *in vitro* methods to increase the confidence in the interactions.

The principle of the Y2H method is presented in Figure 2.2. Pairs of proteins to be tested for interaction are expressed as fusion proteins (hybrids) in yeast. The bait protein is fused to a transcription factor DNA binding domain, the other protein, the prey protein, is fused to a transcription factor activation domain. When expressed in a yeast cell containing the appropriate reporter gene, interaction of the bait with the prey brings the DNA binding domain and the activation domain into close proximity, creating a functional transcription factor. This triggers transcription of the reporter gene. The interaction can then be detected by expression of the linked reporter genes [14, 13, 16]. Numerous variations of Y2H have been developed, including systems with several reporter genes or systems having one-hybrid and three-hybrids.

The Y2H technique has been used extensively both on the large-scale (see Section 2.3.2 below) and for individual interaction experiments. It has been successfully applied to several organisms (details in Section 2.3.2).

Usually, a combination of different techniques is necessary to validate, characterize

and confirm protein interactions. Previously unknown proteins may be discovered by their association with one or more known proteins. Protein interaction analysis may also uncover unique, or unforeseen functional roles for well-known proteins [14, 16].

2.3.2 Large-scale PPI Experiments

The speed at which new proteins are being discovered or predicted has created a need for methods that can detect *high-throughput* or '*large-scale*' interaction data. In the last several years, methods that can globally tackle the problem have been introduced, resulting in a vast amount of new interaction data [2]. The Y2H assay and complex purification detection techniques using mass spectrometry are the two most popular approaches successfully applied on a large scale.

Yeast-Two-Hybrid (Y2H) assay Y2H screens [17, 18, 19, 20], the principle of which is described above and in Figure 2.2, have been used to detect pairwise binary interactions systematically at large scale. For screening entire genomes, two main approaches were used to extend the small-scale Y2H method: matrix-based and library-based [16].

- In the *matrix* approach, a matrix of prey clones is created in which each clone expresses a particular prey protein at one position of a plate. Then, each bait is mated with an array of prey strains and those diploids where two proteins interact are selected based on the expression of a reporter gene and the position on a plate.
- In the *library* approach, each bait is screened against an undefined prey library containing random cDNA fragments. Diploid positives are selected based on their ability to grow on specific substrates; and interacting proteins are determined by DNA sequencing techniques.

The first two large-scale Y2H analyses were carried out in yeast and revealed 692 and 841 putative interactions, respectively [17, 18]. The overlap between these two experimental studies was very small, limited to only 141 interactions (20%). Y2H array approaches have recently been extended to fly, worm and human proteins [21, 20]. This method is

an *in vivo* technique, where transient and unstable interactions can be detected. However, this technique could easily miss certain interactions due to insufficient depth of screening and misfolding of the fusion proteins. In addition, it takes place in the nucleus, so many proteins are not in their native compartment.

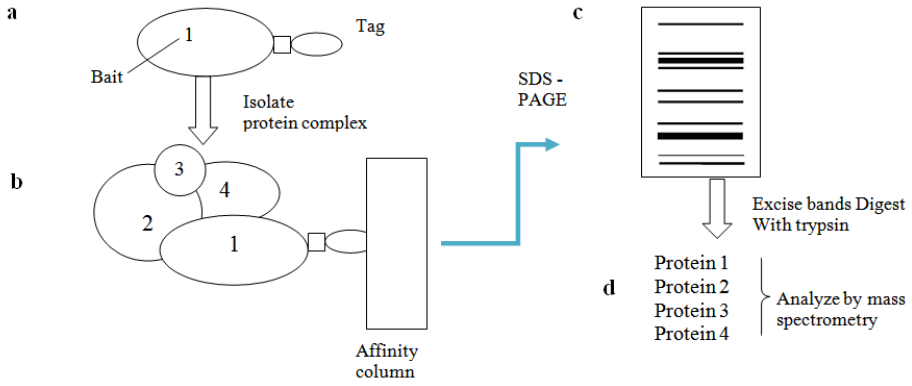

Figure 2.3: Protein complex purification and identification techniques using mass spectrometry by [22, 23, 24]. The approach involves four steps: (a) An 'affinity tag' is first attached to a target protein. (b) Bait proteins are systematically precipitated. (c) Purified protein complexes are resolved, so that proteins become separated according to mass. (d) Proteins are detected and analyzed by mass spectrometry techniques.

Mass spectrometry of purified complexes Protein complex purification and identification techniques using mass spectrometry [22, 23, 24] are employed to reveal the components of protein complexes, i.e. multiple proteins that interact with each other mostly directly but also indirectly. Figure 2.3 describes the process of this method: (a) Individual proteins are tagged and used as baits to biochemically purify whole protein complexes. (b) Bait proteins are systematically precipitated, along with any associated proteins, on an "affinity column". (c) Purified protein complexes are resolved by one-dimensional SDS-PAGE (a technique that involves running an electric charge through the complexes on a gel, so that proteins become separated according to mass). (d) Proteins are excised from the gel, digested with an enzyme, typically trypsin, and the digest is analyzed by mass spectrometry. Database-search algorithms are then used to identify specific proteins from their mass spectra.

For large-scale mass spectrometry based protein complex purification techniques, their advantages include: several members of a complex can be tagged at once by this technique, and it detects real complexes in physiological settings. However, these methods may miss complexes that are not present under the given conditions. Also tagging may disturb complex formation, and weakly associated components may dissociate and escape detection.

In general, large-scale experiments have generated promising results and are chiefly responsible for the relatively large amount of direct protein protein interaction evidences. However, these datasets are often incomplete and noisy [2]. It is fair to say that high-throughput interaction studies are generally difficult to reproduce. For these reasons, large-scale interaction studies are frequently criticized.

2.4 Availability of PPI Data

Even though much effort has been spent on the study of the interaction between proteins, current experimental PPI data is still preliminary, both in terms of the quality as well as quantity.

Quality of high-throughput data von Mering et al. [2] undertook a comprehensive analysis to compare different yeast PPI sets with each other and with a reference set of previously reported protein interactions. Their goal was to measure the accuracy and potential as well as to identify biases, strengths and weaknesses. They found that among approximately 80,000 interactions between yeast proteins available from different high-throughput methods, only a surprisingly small number (about 2,400 pairs) were supported by more than one method. This suggests that either the methods may not have reached saturation, or that many of the methods produced a high proportion of false positives. They also estimated that more than half of all current high-throughput data were spurious (see Figure 2.4). For example as shown in Figure 2.4, a filtered yeast two-hybrid dataset showed medium accuracy as compared to the benchmark and its fraction of false positives was predicted to be about 50%. Different methods may have difficulties for certain types of interactions. Thus, to increase the coverage and to improve the confidence in detected or predicted protein

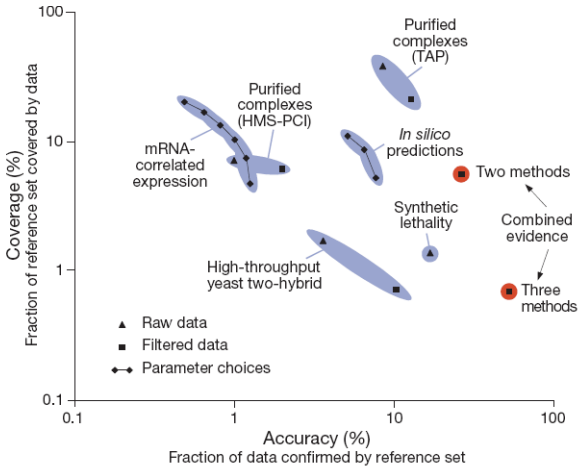

Figure 2.4: Quantitative comparison of interaction data sets in yeast. Figure from [2]'s Figure 2, included as background information only. The various Yeast data sets were benchmarked against a reference set of 10,907 trusted interactions. Each dot in the graph represented an entire interaction data set, and its position specifies the coverage and accuracy (on a logClog scale).

interactions, as many complementary methods as possible should be used.

Yeast PPI databases The Yeast Proteome Database (YPD) [25, 2] represented the first systematic effort to compile protein-interaction and other data from the literature; YPD is now available only on a subscription basis. A number of other important databases (Table 2.2) that curate protein and genetic interactions of yeast from the literature have been developed, including the Munich Information Center for Protein Sequences (MIPS) database [26], the Molecular Interactions (MINT) database [25], the IntAct database [25], the Database of Interacting Proteins (DIP), the Biomolecular Interaction Network Database (BIND) [27], and the BioGRID database [28].

Human PPI databases A number of public repositories for human PPIs are currently available, including the databases: BIND, DIP, IntAct, MINT and MIPS databases [3], described above for yeast. There exists also a specific database for human protein interactions, the Human Protein Reference Database (HPRD) [29]. These databases are listed in Table 2.2. Each of these databases has its own unique features with a large variation in the

Table 2.2: Recent Public PPI Databases. The third column describes the type of PPI data contained in each database: H (high-throughput experimental data), M (manual curation), F (functional predictions). More details and databases related to other interaction information are listed in [25].

Database Name	**Num. of PPIs**	**Type**	**URL**
BioGRID	116,000	H,M	http://www.thebiogrid.org
DIP	55,733	H	http://dip.doe-mbl.ucla.edu
BIND	83,517	H,M	http://bind.ca
MIPS	15,488	H,M,F	http://mips.gsf.de/services/ppi
HPRD	33,710	H,M	http://www.hprd.org
MINT	71,854	H,M	http://mint.bio.uniroma2.it/mint
IntAct	68,165	H,M	http://www.ebi.ac.uk/intact
STRING	730,000	H,F	http://string.embl.de

type and depth of their annotations. Currently available human PPI sets can be divided into three classes [3]: (1) Obtained from literature search [30, 31]; (2) Derived from interactions between orthologous proteins in other organisms [32, 33]; (3) Based on large scans using yeast-two-hybrid (Y2H) assays [19, 34]. Each of these different strategies has its advantages as well as disadvantages. For example, Y2H-based mapping approaches offer rapid screens containing thousands of proteins, but the data may have high false-positive rates. All interaction maps implied considerable selection and detection biases [3]. A comparison between these human PPI datasets reveals that there is only a small, but nonetheless statistically significant overlap.

Shoemaker and Panchenko wrote a valuable review of PPI databases in 2007 [25]. They found that interactions recorded in all these databases represented only part of the primary literature.

2.5 Related Biological Interactions

Most of a cell's biological characteristics arise from interactions between its numerous constituents, including the proteins, small molecules, membranes DNA and RNA. Therefore, a key challenge for biology is to understand the structure and the dynamics of the complex intercellular graph of interactions that contribute to the structure and function of a living cell [35]. Besides protein-protein interactions, there are other types of biological interactions also important for the cell.

Genetic interactions Genetic interactions combine functional relationships among genes revealed by the phenotype of cells carrying combined mutations of those genes. Reguly et al. [28] have divided the phenotypes into eight categories (dosage growth defect, dosage lethality, dosage rescue, phenotypic enhancement, phenotypic suppression, synthetic growth defect, synthetic lethality, synthetic rescue). The synthetic genetic array (SGA) and synthetic lethal analysis by microarray (dSLAM) methods were used to systematically uncover synthetic lethal genetic interactions, in which a group of non-lethal gene mutations combine to cause inviability. The BioGRID database [28] provides a comprehensive curation of reliable genetic interactions from the current primary biomedical literature.

Protein-DNA or protein-RNA interactions Protein-RNA and protein-DNA interactions are involved in several processes essential to normal cell function [14]. These interactions are integrated into key cellular processes including transcription, translation, regulation of gene expression, recognition, replication, recombination, repair, and etc. DNA, as the genetic repository of information, requires interaction with proteins for the genetic information to be extracted and utilized timely within the cell. DNA or RNA-binding proteins are commonly used to recognize and manipulate DNA or RNA structures. Transcription complex formation, initiation of transcription, and translation of messenger RNA to the proteins, all involve formation of protein to nucleic acid complexes containing either DNA or RNA. These complexes naturally play an essential role in the regulation of protein expression.

Note: Since we are focusing on protein-protein interactions in this dissertation, terms like ”protein interaction” or ”protein interaction networks” in the following chapters refer to protein-protein interactions only.

Chapter 3

Related Work

The biological background of PPIs has been introduced in the previous chapter (Chapter 2). Then, this chapter summarizes the literature related to prediction of protein interactions in general and presents a systematic study of key issues related to pairwise PPI predictions using information-integration strategy.

3.1 PPI Prediction by Information Integration

To understand the working mechanism of the cell, it is vital to accurately characterize the set of protein interactions in a given proteome. The basic units of protein interaction networks are binary edges which represent physical binding between the members of each pair. Thus, to study the whole PPI network, we start by first identifying each of the edges in possible protein pairs.

It has been pointed out that high-throughput experimental interaction data can exhibit high false positive and false negative rates. Traditional small-scale experiments are costly and laborious [13]. As a result, most of the possible protein interactions have not be discovered experimentally.

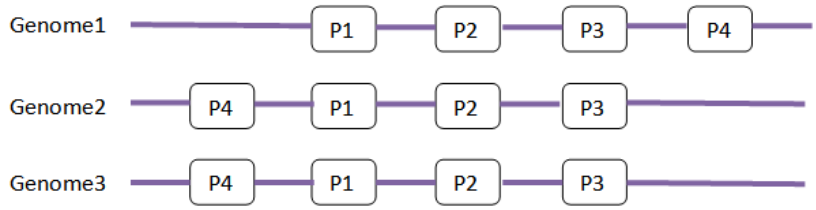

Figure 3.1: An example of protein interaction prediction from indirect evidence: Gene neighborhood approach.

3.1.1 Prediction from Indirect Evidence

In addition to experiments that directly test for PPIs, there are many indirect sources that may contain information about PPIs as well. For example, it has been shown that many interacting pairs are co-expressed [2] and that expression of proteins found in the same complex in some cases can be controlled by the same transcription factor(s) [36]. Sequence data has also been used to infer such interactions (for example by relying on domain-domain interactions or structure information [37]). Many other characteristics of a gene or protein pair also have predictive values. Each of these datasets provides partial information about the interacting pairs. A series of recent approaches investigated the utility of various indirect datasets for the inference of PPI pairs. These include:

- Over-represented domain pairs or motif pairs observed in interacting protein pairs have been studied and used to infer PPI interactions [38, 39, 40, 41, 42, 43].
- Protein structural information have been incorporated for predicting potential PPIs [44, 45]. Conservations of pairs of sequence patches involved in PPI interface were used in [44].
- Several methods have helped to infer protein interactions based on the conservation of gene neighborhood (Figure 3.1), conservation of gene order, gene fusion events, or the co-evolution of interacting protein pair sequences [25, 2]. For instance as shown in Figure 3.1, gene-neighborhood provides very strong signals for functional association between gene products within and across species [2].

The findings derived from these approaches suggest that direct measurements on protein interactions can be combined with indirect information to improve the quality of protein interaction prediction.

3.1.2 Prediction by Information Integration

Based on the above observations, a number of researchers have suggested that direct data on protein interactions can be combined with indirect data in a supervised learning framework [46, 47, 48, 49, 50, 51, 52, 53, 54, 55]. Studies using this approach all use a classification algorithm to integrate diverse biological datasets. A classifier is trained to distinguish between positive examples of truly interacting protein pairs from the negative examples of non-interacting pairs. Each protein pair is encoded as a feature vector where features may represent a particular information source regarding either protein interactions, related mRNA expressions, domain compositions, or evidence coming from various experimental methods.

Prediction for yeast Von Mering et al. [2] were among the first to discuss the problem of accurately inferring protein interactions from high-throughput data sources. The proposed solution, which used the intersection of direct high-throughput experimental results, achieved a very low false positive rate. However, the coverage was also very low. Less than 3 percent of known interacting pairs were recovered using this method. The ”STRING” database built by these authors [2] created functional associated pairs derived from computational integration of known protein-protein associations, co-expression pairs and pairs transferred across organisms, a database that is widely used today.. Jansen et al. [48] proposed the use of a naive Bayes classifier on a summary feature set relying on the MIPS complexes catalog as gold standard. Lin et al. [50] repeated the experiments in [48] with two other classifiers: RF and LR. They also discussed the importance of different features and concluded that the MIPS and Gene Ontology functional categories were the most informative. Bader et al. [46] used LR to estimate the posterior probability that a pair of proteins will interact. The features used in their work were derived directly from the high-throughput experiments in summary style. Zhang et al. [52] used a decision tree

with a detailed feature set and the MIPS complex data as the gold standard. Yamanishi et al. [51] presented a method to infer protein interaction networks using a variant of kernel canonical correlation analysis. They had relied on a detailed dataset and used pathway data from KEGG as their gold standard. Lee et al. [49] integrated diverse functional genomic data by reinterpreting experiments to provide numerical likelihoods that genes are functionally linked. They relied on a summary type of feature set and used pathway data from KEGG as their gold standard. All approaches above considered protein pairs independently when inferring the presence of PPIs. Differently from these methods, Jaimovich et al. [56] considered the neighborhood interaction pairs together and employed a 'Relational Markov Random Field' approach for collective inference of PPIs in yeast.

Prediction for human Compared to yeast, the human proteome is significantly more complex due to the larger number of proteins, their splice isoforms, post-translational modifications and mechanisms of dynamic regulations. Because a number of data sets available for yeast are not yet available for human, there are fewer previous studies related to learning human PPIs through multiple data integration. Rhodes et al. [54] employed a strategy using the sum of likelihood ratio scores strategy to predict human protein interaction confidence. These likelihood ratio scores are derived based on homologous PPI, gene expression, the GO Process and domain based sequence evidence. Brown and Jurisica [33] tried to get a better human PPI set from the evolutionary point of view. Based on the homology relationship, they used high-throughput interactions in other model organisms to infer millions of potential human PPIs. Guo et al. [57] recently assessed the capability of GO information to predict protein interactions involved in human regulatory pathways. They showed that the functional similarity of proteins within known pathways decays rapidly as their path length increases. Arun et al. [31] developed and applied natural-language processing and literature-mining algorithms to recover interactions among human proteins from Medline abstracts. Scott and Barton [58] extended the probabilistic framework for the prediction of human protein-protein interactions with more features, which include co-expression, orthology to known interacting proteins and the full-Bayesian combination of subcellular localization, co-occurrence of domains and post-translational modifications.

All of the above methods were shown to improve the success of PPI prediction when

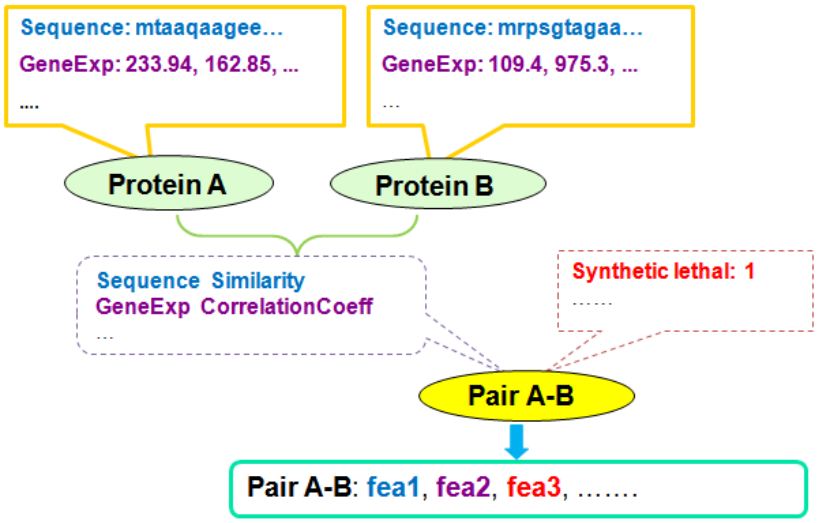

Figure 3.2: The process of combining data from multiple biological sources and then converting them to feature vectors describing protein-protein pairs. For each gene/protein specific feature, we find a natural way to transform it to describe a protein-protein pair. For example, for gene expression data, we use the correlation coefficient as the feature for a protein-protein pair.

compared to direct data alone. The improvements are not just from the perspective of predicting novel interactions but also for the purpose of stratifying the many candidate interactions with confidence.

Feature extraction for pairwise protein-protein pairs As described above, each protein pair can be encoded as a feature vector where features represent a particular information source regarding protein pairs in the information integration framework. However, each type of biological information has its own representative form. For example, protein sequence takes the form of a character string, corresponding to the order of amino acids as they occur in a polypeptide chain. Gene expression data is usually a vector of expression values across multiple time points for a specific gene. Synthetic lethal data describes that a pair of genes having mutations together would render the cells either inviable or viable.

We present the method we used for feature extraction in Figure 3.2. For each data set that represents a certain gene / protein's property, we designed a biologically meaningful way to calculate the similarity between two genes / proteins with respect to the specific

Table 3.1: Symbols used for PPI prediction

Symbol	Description
X	Feature vector
x	Feature vector of an example Each example in PPI prediction task is a protein-protein pair.
d	The total number of attributes
X_i	The i-th feature item of an example.
Y	Class label of an example. In PPI prediction task, $Y \in$ {Interact (1), not-Interact (-1)}
$(x^{(j)}, y^{(j)})$	The j-th training example
N	Number of examples in the data set
θ,ν, ω	Model parameters

evidence. For instance, for two proteins' sequence information, we use the BlastP [59] sequence alignment E-value as one feature for this protein-protein pair from the protein sequence evidence. For other data sources, similar procedures are pursued to determine the features for a protein pair. Concatenating these features together then give us the feature vector describing a protein-protein pair.

Many biological data sets may be directly or indirectly related to PPIs. We try to collect as many as possible for yeast and human. The extracted features are described in detail in the following two chapters. Furthermore, we want to emphasize that this framework could not be applied on predicting homo-dimers because of the feature-extraction strategy. Since most features used here are gene-specific, the corresponding feature items of self protein pairs would thus have no distinctive ability to predict homo-dimer interactions.

3.1.3 Systematic Comparison of Controlling Factors

To correctly identify the interrelationship between proteins, many different research groups (Section 3.1.2) have independently suggested the use of supervised learning methods to integrate direct and indirect biological data sources for the protein interaction prediction task. While these approaches are related in using the same information integration framework , they differ in three key aspects:

- The gold standard data sets used for training and testing;

- The set of features used for prediction and the way these features were encoded;
- The learning method employed.

These differences make it hard to directly compare approaches and to identify features that perform well on the different types of protein interaction prediction tasks. These are important questions, especially when designing experiments to infer protein interactions in organisms other than yeast. For example, identifying the set of important features can help determine if enough data exists for such a prediction task in a particular organism, and to indicate which type of data is most useful. With this goal in mind, we undertake a systematical study [4] to investigate how differences in the three aspects noted above affect the prediction performance. Specifically we consider the following settings:

Gold standard datasets Three gold standard datasets were previously used to train and test algorithms for protein-protein interaction prediction. Each of them matches a subtask of PPI predictions. The three tasks are the prediction of (1) physical interaction, (2) co-complex relationship and (3) pathway co-membership. In predicting direct physical interaction between protein pairs, the Database of Interacting Proteins (DIP) ("small-scale" subset [60]) is used [61, 8]. A broader definition of protein interaction is the co-complex relationship in which proteins are considered pairs even if they do not directly interact but are connected through other proteins. The Munich Information Center for Protein Sequences (MIPS) complex catalogue [26] has been used as the gold standard dataset for this prediction task [48, 52, 50]. Finally, the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [62] has provided the gold standard for inferring pathway networks [49].

Feature encoding Two fundamentally different and general types of feature encoding were used in the past: a "detailed" encoding style, in which every experiment is considered separately [52], and a "summary" style, in which similar types of experiments, such as all expression experiments, are grouped together and yield a single value [48, 50, 8].

Classification methods Many different classifiers were suggested for the protein interaction prediction task, including Logistic regression (LR) [46], Naive Bayes (NB) classifier

[48], Random Forest (RF) [50, 8], Decision Tree (DT) [52], Kernel methods [51, 53] and Bayesian Scoring approaches [49].

We carried out our comparison by assembling a large set of biological features and we varied their encoding for use in each of the three prediction tasks. In [4], we systematically analyzed the effect of varying each of the different design issues. We compared prediction performance by testing on all possible combinations of feature encoding styles, reference datasets and classifiers. A constant set of features was used, with two different encodings, and standard implementation of classification algorithms. For all classifiers, the three prediction tasks had different success rates and co-complex prediction appears to be an easier task than the other two. Independently of prediction tasks, however, the RF classifier consistently ranked as one of the top two classifiers for all combinations of feature sets. Therefore, we used this classifier to study the importance of different biological datasets. First, we used the splitting function of the RF tree structure, the Gini index, to estimate feature importance. Second, we determined classification accuracy when only the top-ranking features were used as an input in the classifier. We find that the importance of different features depends on the specific prediction task and the way they are encoded. Strikingly, gene expression is consistently the most important feature for all three prediction tasks, while the protein interactions identified using the yeast-two-hybrid system were not among the top-ranking features under any condition [4].

There may be several factors that contribute to the success of RF in the comparison study when compared with other classifiers: (a) The currently available direct and indirect protein interaction data is inherently noisy and contains many missing values. The randomization and ensemble strategies within RF make it more robust to noise when compared to LR. (b) Biological datasets are often correlated with each other and thus should not be treated as independent sources. Linear and non linear regression models assume independence and may therefore perform worse than other classifiers in tasks where correlations among features are strong. In contrast, the RF classifier does not make any assumptions about the relationship between the data, which makes it more appropriate for the type of data available for the protein interaction prediction task.

3.2 Protein Complex Identification

When analyzing PPI networks at the level of binary interactions, much information is lost, because proteins often perform their functions together in groups or as part of particular patterns. Understanding these interaction groups (*complexes*) and patterns (*pathways*) are essential for systematically modeling the behavior of cellular networks. Graph analysis algorithms can help us understand how proteins are logically connected. The connections between proteins can be best represented on a graph in which the nodes correspond to proteins and the edges correspond to the interactions. Thus the identification of complexes or pathways is simply the computational problems of locating important subgraphs. This kind of analysis can help produce valuable insights into both the topological properties and functional organizations of protein networks in cells. In this dissertation we solve only the complex detection problem, leaving the pathway detection for future studies.

Many cellular functions are performed by complexes containing multiple protein interaction partners. Predicting molecular complexes, one of the fundamental units in PPI networks, is one of the most important tasks in the analysis of protein interaction networks. High-throughput experimental approaches [24, 22, 23] that are used to determine the set of protein complexes on a proteome-wide scale often suffer from high false positive and false negative rates [2]. Thus, there have been various previous computational attempts trying to identify complexes or related functional modules. Most previous methods [63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74] for automatic complex identification (or related functional module detection) have employed the unsupervised graph clustering style and try to discover similarly or densely connected subgraphs of nodes (clusters) [75]. Below we split the related literature into five types.

Graph segmentation Several studies have attempted to segment the PPI graph into disjoint highly connected clusters (complexes). King et al. [63] partitioned the nodes of a given graph into distinct clusters, based on their neighboring interactions, using a cost based local search algorithm. Dunn et al. [64], divided the network into clusters by removing the edges with the highest centralities. The edge-removal process iteratively recalculated betweenness until a fixed number of edges have been removed.

Overlapping clustering Since some proteins are part of multiple complexes or functional modules, a number of approaches allow overlapping clusters. With the approach named "MCODE", Bader et al. [67] tried to detect densely connected regions in large PPI networks using vertex weights to represent local neighborhood density. Pereira-Leal et al. [66] used the line graph strategy of the network (in which a node represents an interaction between two proteins and edges share interactors between interactions) to produce an overlapping graph partitioning of the original PPI network. Adamcsek et al. [68] identified overlapping densely interconnected groups in a given undirected graph using the k-clique percolation clusters in the network. Spirin et al. [69] analyzed the multi-body structure of the PPI network to discover molecular modules that are densely connected with themselves but sparsely connected with the rest of the network. The authors claimed that two types of modules were found: protein complexes and dynamic functional units. Zotenko et al. [76] used a graph-theoretical approach to identify functional groups and represent overlaps between functional groups in the form of the "tree of complexes". In [77], Brohee and van Helden made a comparative assessment for protein-protein interaction networks of four clustering algorithms: Markov Clustering (MCL), Restricted Neighborhood Search Clustering (RNSC), Super Paramagnetic Clustering (SPC), and Molecular Complex Detection (MCODE). They found that MCL and RNSC were more robust to graph alterations than the other two algorithms.

Conservation across species Sharan et al. [72] used conservation alignment to find protein complexes that are common between yeast and bacteria. They formulated a log likelihood ratio model to represent individual edges shared by proteins and used a clique structure to represent a protein complex. A heuristic local-search strategy was used for searching the conserved complexes as the heavy subgraphs, in which nodes corresponded to orthologous protein pairs.

Considering spatial constraints Utilizing the spatial aspects of complex formation, Scholtens et al. [73] applied a local modeling method to better estimate the protein complex membership from direct mass spectrometry complex data and Y2H binary interaction data. They

claimed to achieve a finer level of detail than that obtained by using only the mass spectrometry data. Chu et al. [74] proposed an infinite latent feature model to identify protein complexes and their constituents from datasets derived from large-scale affinity based mass spectrometry techniques.

New similarity measures Rives et al. [70] applied standard clustering algorithms to group similar nodes on the interaction graph. The cluster similarity is calculated based on vectors of nodes' attributes, such as their shortest path distances to other nodes. Aranu et al. [71] used hierarchical clustering of proteins to define a new similarity measure based on the stability of node pair composition.

All previous methods have presumed that complexes correspond to the dense regions of networks. While this is true for some complexes, there are many other topological structures that may also represent a complex. One example is a "hub" or "star" model, in which all vertices connect to a "hub" protein [46]. Another possible topology is a structure that links several small connected components. This topology is especially attractive for large complexes: given the space constraints, it is unlikely that all proteins in a large complex would be able to interact with all others. In Chapter 7 we present a computational framework to identify complexes without making strong assumptions about their topology [11].

3.3 Network Topology Analysis of PPI Graphs

The topology of a network refers to the relative connectivity of its nodes. Different topologies affect specific network properties. The topological structures have been analyzed for the following reasons [78]:

- It has been realized that the architectural features of molecular interaction networks within a cell are often reflected to a large degree in other complex systems as well, such as the Internet, world wide web (WWW) or organizational networks. The unexpected similarity indicates that similar laws may govern most complex networks in nature. This enables the expertise from large and well-mapped non-biological systems to be utilized for characterizing the complicated inter-relationships that govern

Table 3.2: Symbols used for graph analysis

Symbol	Description
G	Graph. $G = (V, E)$
E	Set of edges in the graph
V	Set of nodes in the graph
N	Number of nodes in the graph
W	Weights of edges in the graph
A	Adjacent matrix of the graph
k	Degree of a node
c_k	Count of nodes with degree k
r	Power law exponent
CC	Clustering coefficient of the graph

cellular functions [35].

- Cellular function is a contextual attribute of complex interaction patterns between cellular constituents [35]. The quantifiable tools of network theory offer possibilities for providing insights into properties of the cell's organization, evolution and stability.

- The relative positions of proteins within the interaction networks might indicate their functional importance. For instance a positive correlation between biological essentiality and graphical connectivity has been demonstrated [78], suggesting a relationship between topological centrality and functional essentiality.

Thus it is important to understand and model the topological and dynamic properties of various biological networks in a quantifiable manner. There are various types of interaction networks in the cell, (including protein-protein interaction, metabolic, signalling and transcription-regulatory networks). None of them function independently. Rather together they form a "network of networks" which is responsible for the behavior of the cell [35].

As the literature on topological analysis of real networks is vast, this section will briefly discuss just a few related studies as a guide to this topic. Comprehensive reviews can be found in [35, 79, 80, 81]. Some graph patterns are seen repeatedly in real networks. The main ones include [80]: (1) Power law distribution; (2) Small world effect; (3) Community effects. They are described in detail below.

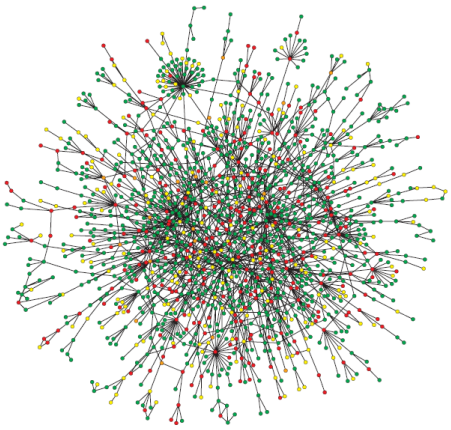

Figure 3.3: A map of protein-protein interactions in yeast, which was based on early yeast two-hybrid measurements. Figure obtained from Figure 2 of [35], included as background information only. A few highly connected nodes (which are also known as hubs) hold the network together [35].

Power law distribution The most elementary characteristic of a graph node is its *degree* (or connectivity). Degree k measures how many links a node has to other nodes. For undirected networks like the PPI graph, k refers to the number of edges a node relates to. [35]. The degree distribution of an undirected graph is a plot of the count c_k of nodes with degree k, versus the degree k, typically on a log-log scale [80].

Very often, the degree distribution of real networks follow a power-law distribution [80], which means that the number of nodes c_k with degree k is related to k by:

$$c_k = c * k^{-r} \tag{3.1}$$

where c and r are positive constants. The degree distribution appears linear when plotted on the log-log scale (see Figure 3.4d). The constant r is often called the power law exponent. The significance of power law distributions has to do with their being heavy-tailed, which means that they decay more slowly than exponential or Gaussian distributions (referred to as "random networks", see Figure 3.4c). Thus, a power law degree distribution would be much more likely to have nodes with a very high degree (much higher than the mean) than the other two distributions [80] (Figure 3.4).

Many cellular interaction networks have been shown to be scale-free [35], meaning

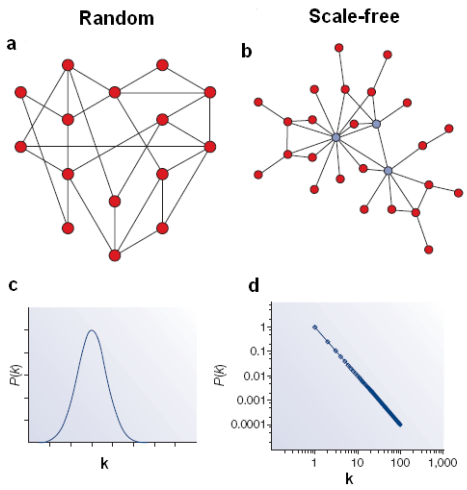

Figure 3.4: Degree distribution of random network versus scale-free network. Figure modified from Box 2 of [35], included for background information only. (a) A schematic representation of a random network; (b) A schematic representation of a scale-free network. The degree distribution of random network (c) obey a Gaussian distribution, whereas the degree distribution of scale-free networks (d) obey a power-law distribution.

the probability of a protein having k links follow a power-law distribution with a degree exponent r in the range between 2 and 3. Such a distribution indicates that most proteins in the network participate in only a few interactions, while a few proteins participate in many (*hubs*).

Scale-free networks are resistant to random failure but are vulnerable to targeted attacks, specifically against hubs [78]. This property has been found to account for the robustness of biological networks to perturbations like mutations and environmental stress. Thus identification of high-degree proteins (hubs) may produce a strategy for therapeutic mediation of signaling pathways that associate with cancers [78]. Such a strategy would have less impact if the true topology were exponential, and would be not operable if the true topology were random [78].

Figure 3.3 shows a protein interaction map of the yeast as predicted by previous systematic two-hybrid screens [35]. Most proteins participate in only a few interactions, and only a few participate in dozens: this is typical of scale-free networks [35, 34].

Small world effect A common feature of all complex networks is that any two nodes can be connected through a path of a few links only [35]. This so-called "small-world effect", was originally observed in the research on social networks and is often characterized as the famous "six degrees of separation" [80]. Scale-free networks are ultra small, which means their path length is much shorter than predicted by the small-world effect for random networks. Within the cell, the ultra-small-world effect was first found for metabolism, where paths of only three to four reactions can link most pairs of metabolites [35]. The short path length indicates that local perturbations in metabolite concentrations could reach the whole network very quickly.

Figure 3.3 shows that cellular networks [35] are different from social networks in terms of connections between hub nodes. In protein interaction networks, highly connected nodes (hubs) avoid linking directly to each other and instead connect to proteins with only a few interactions. Whereas in social networks, well-connected people tend to know each other [35].

Community effects A community is informally defined as a set of nodes where each node is "closer" to the other nodes within the community than to nodes outside [80]. This property is believed to exist in many real-world structures, including biological networks. Community effects are investigated in two contexts [80]:

- First, they could be studied through local neighborhoods, as characterized by the clustering coefficient. The clustering coefficient (CC) provides a measure for the inter-connectivity in the neighborhood of a protein and is based on the number of edges connecting the neighbors of the node divided by the maximum number of such edges. The clustering coefficient of the entire graph can be found by averaging over all of the nodes in the graph.

  For a protein node p, with n as the number of links connecting its k_p neighbor nodes to each other [34], its clustering coefficient equals to:

$$CC_p = \frac{2n}{k_p * (k_p - 1)}, \tag{3.2}$$

- Second, community effects could be studied through node groups with potentially longer paths between members. Various methods could be used including graph partition or bipartite core (for instance, through bi-clustering). Graph partitioning techniques typically break the graph into disjoint partitions (or communities) by optimizing some measure [80].

Most real-world graphs (including PPI networks) exhibit strong community effects, which are also reflected in the clustering coefficients of these graphs: they are almost always much larger than in random graphs of the same approximate size [80]. The cellular networks studied to date, including protein interaction and protein domain networks, have a high graph clustering coefficient, indicating that high clustering is a generic feature of biological networks [35].

Biological network motifs Not all subgraphs are equally significant or important in real networks. *Motifs* are subgraphs that occur significantly more often in the given network than expected by chance alone. A number of previous studies [82, 83] pointed out the existence of simple building network motifs in PPI graphs and transcription regulation networks. Recently, a number of efficient tools have been designed to facilitate the detection of motifs [82, 84, 83]. Many complex networks have been shown to have certain structural design principles. "Network motifs" could help researchers better understand the basic structural elements of a specific network [75]. There is a high degree of evolutionary conservation of network motifs in the yeast protein interaction network. The convergent evolution towards the same motif types has also been seen in the transcription-regulatory network of diverse species. All these observations further indicate that motifs are indeed of direct biological relevance [35].

Besides the graph based studies above, there are also a series of studies analyzing the protein interaction networks from the perspective of incompleteness, which is a common problem in current PPI datasets.

Topologies of subnetworks Currently available protein-protein interactions (PPIs) cover only a fraction of the complete PPI networks. These partial networks display scale-free

topologies. Han et al. [78] analyzed whether the scale-free topologies of the partial networks could be used to accurately infer the topology of the complete PPI networks. The authors generated four theoretical network models of different topologies (random, exponential, power law, truncated normal) and found that partial sampling of these networks resulted in sub-networks with topological characteristics that were virtually indistinguishable from those of currently available Y2H-derived partial maps. Based on these results, they concluded that given the limited coverage levels, the observed scale-free topology of existing PPI maps cannot be confidently extrapolated to complete PPI networks [78].

Another study [75] indicated the possibility to extrapolate from subnets to the properties of a whole network only if the degree distributions of the whole network and randomly sampled subnets reflect the same family of probability distributions. However this condition is not satisfied for scale-free degree distributions [85]. Moreover, limited sampling alone may also create apparent scale-free topologies, irrespective of the original network topology [85, 75]. These results suggest that interpretation of global properties of the complete network should be made with caution if based on the current (still limited in accuracy and coverage) partially observed networks.

Recent variants of PPI networks Recently a number of important studies tried to incorporate more biological context for the PPI networks by extracting or expanding PPI graphs into specific subsets or supersets. Goh et al. [86] presented the 'human disease network' which included disorders and disease genes that are linked by known disorder-gene associations. This network tried to explore all known phenotype and disease gene associations in a single graph theoretical framework, indicating the common genetic origin of many diseases. Kim et al. [87] characterized protein interactions by using atomic-resolution information from three-dimensional protein structures. Using a proposed structural measure, the study subdivided PPI hubs and provided insight into their evolutionary rate. Lage et al. [88] created a phenome-interactome network by integrating quality-controlled interactions of human proteins with a validated, computationally derived phenotype similarity score. This network permitted the identification of previously unknown complexes likely to be associated with disease. Linding et al. [89] developed an approach called NetworKIN

to augment motif-based predictions in the context of kinases and phosphoproteins. This method claimed to pinpoint kinases responsible for specific phosphorylations and yielded a significant improvement in the accuracy with which phosphorylation networks can be constructed. These studies of PPI networks data focused on specific objectives and provide great insights both biologically and computationally.

3.4 Summary

Protein-protein interaction maps provide a valuable framework for a better understanding of the functional organization of the cell. Computational predictions could suggest new biological hypotheses regarding unexplored new interactions or groups of interacting pairs. In this chapter we briefly reviewed the related literature on three topics covered in this dissertation.

- Pairwise PPI prediction through integration. Previous studies differed in terms of classifiers, feature sets and their encodings and gold-standard datasets used. We performed a systematic comparison how these issues affect the ability to make accurate predictions.
- Searching for protein complexes on the protein interaction graph which could be treated as a subgraph identification task. A series of computational methods using the graph analysis concepts and techniques were proposed to handle this task.
- Global analysis of biological network topologies. These kinds of studies could provide insights into the biological properties related to evolution, function, stability, and dynamic responses.

Learning of protein interaction networks is an important topic, for both its biological significance and the generality of related computational methodologies. From a broader perspective, the first two problems appear to have close connections to an active machine learning topic called ”Statistical Relational Learning” [90]. Similarly, for the third topic, a large number of related studies exist in current social science and graph mining [91]. We

omit these discussions. However it should be realized that high similarities exist between the various methodologies and problems covered in this chapter and those fields.

Chapter 4

A Combined Approach for Subnetwork PPI Predictions

We have discussed the biological basics of PPIs in Chapter 1. The literature related to computational learning of PPI graphs was also introduced in Chapter 2. It is believed that protein-protein interaction maps could provide better understandings of the functional organization of the cell and computational tools have been proved to be useful in improving the quality of current PPI data sets.

Naturally computational methods for predicting PPI networks could assist experimental efforts by either prioritizing protein interactions to be tested or by validating (or refuting) high-throughput screens. In this chapter we propose a combined framework to integrate computational PPI learning, network analysis, in vitro experimentation, and biological expertise.

4.1 Introduction

Human membrane receptor proteins are attractive drug targets because they mediate the communication between the cell and its environment. There are two types of membrane receptors. (1) Type I receptors refer to a broad group of diverse families of membrane receptors that directly or indirectly activate enzymatic activity, such as tyrosine kinase. (2) Type II receptors refer to the large G protein coupled receptor (GPCR) family.

Signaling mechanisms initiated by membrane receptors are complex and involve numerous proteins [92]. In addition, different receptor pathways cross-talk with each other. Therefore to fully understand signaling pathways and the crosstalk between them would require identification of the repertoire of all proteins that interact with membrane receptors. It is expected that such understanding would provide a useful resource in the study of complex diseases.

A recent survey of the human genome has identified approximately 1000 membrane receptors [94]. Data from small scale experiments, identified approximately 2500 pairs of interacting proteins, where at least one of the proteins in the interacting pair is a receptor [29]. The high-throughput experimental yeast-two-hybrid (Y2H) method has been used to identify several thousands of new interactions in human [19, 34]. However, this method suffers from high false positive and false negative rates [2] and is less appropriate for membrane protein interactions. The receptor related interactions are also under represented in mass spectrometry based protein interaction screens [95] due to the experimental difficulties arising from the need to maintain a hydrophobic environment for structural integrity of membrane proteins [3]. Consequently, the only two available Y2H datasets for human protein interactions contain no membrane receptors at all and the mass spectrometry screen identified only 136 pairwise interactions involving 27 membrane receptors [95]. In addition, methods specifically designed to detect membrane protein interactions have not yet yielded results on a large scale for human as well [96].

In addition to direct experimental methods, computational approaches have been proven useful in cataloguing the human protein interactions in a variety of ways (Section 3.1.3). It has been clearly established that using direct and indirect data together as features in a supervised learning framework improves the success in predicting yeast protein interactions when compared to direct data alone [97]. Previous computational studies focused on general protein interaction predictions covering the entire interaction network. Here we concentrate on identification of the interactions of all known human membrane receptors with all human proteins (referred to as ”the membrane receptor interactome” or ”the membrane receptor interaction network” throughout this chapter), as opposed to interaction between all human proteins.

Figure 4.1 provides an overview of the approaches we take to identify the membrane

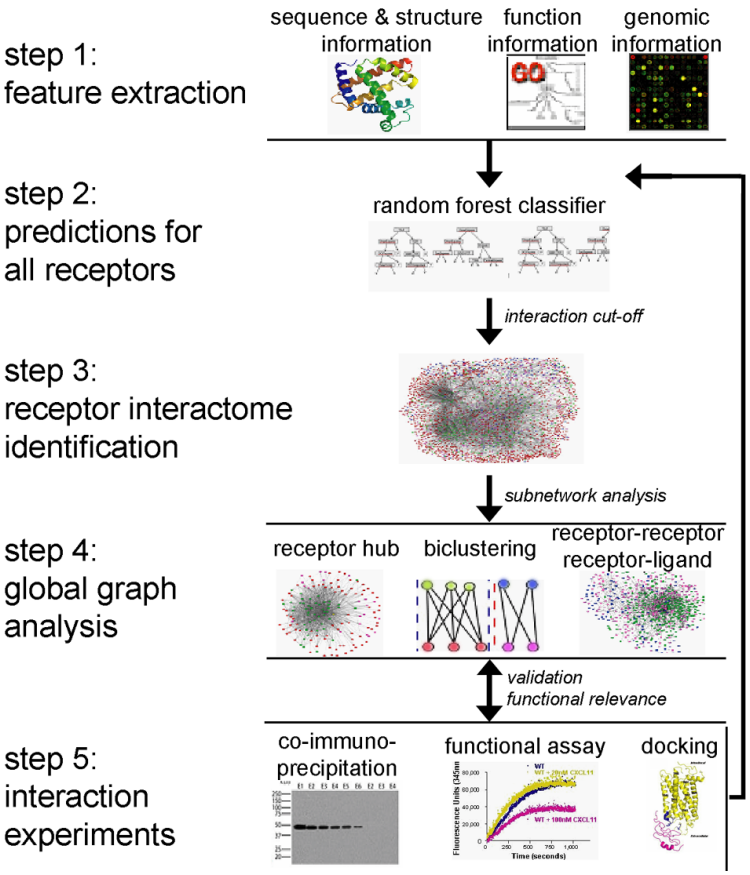

Figure 4.1: Illustration of our combined computational-experimental approach to investigate the human membrane receptor interactome. Step 1. Feature Extraction. Step 2. Prediction for all receptors. Evidence was integrated using a random forest classifier for protein-protein interaction prediction. Step 3. Receptor interactome identification. Visualizations were done with Cytoscope [93]. Nodes are drawn in different colors, with green representing type I receptors, blue for GPCR receptors, pink for ligands and red for other human gene products. Step 4. Global graph analysis. Four types of analyses were carried out: receptor hub identification, biclustering, receptor-receptor and receptor-ligand identification. Step 5. Interaction validation. Specific pairs with high likelihood of interaction based on random forest score were validated by co-immunoprecipitation, functional assays and protein docking.

receptor interactome. Protein-protein interactions related to human membrane receptors are investigated both computationally and experimentally.

- First we extract biologically meaningful features from diverse biological data sources,

including sequence, structure, function and genomic information.

- We predict specific interactions involving receptors using the random forest classifier applied to the binary classification task of whether two proteins interact or not. The determined interactions for all human membrane receptors make up the human membrane receptor interaction network.
- The analysis of global level properties of this network identifies receptor hubs, reveals strongly interacting clusters, highlights the abundance of receptor-receptor interactions, and identifies ligands shared between receptors.
- We have validated some of the predictions made by the classifier experimentally by co-immunoprecipitation, functional assays and protein docking.

4.2 Integration of Evidence for Membrane Receptor PPI Prediction

Combining evidence from many different sources as features in a supervised learning framework has been proven a successful strategy in predicting protein interactions in yeast [97, 8] and in human [54, 58]. Here, we employ the random forest binary classification approach [98] for integrating multiple data sets to predict interactions for human membrane receptors de novo. The evidence sources are converted into features describing reference examples of known positives and negatives (also named ”gold-standards”).

4.2.1 Extraction of Features

Feature attributes for each protein pair are extracted from data sets that may be related to interactions. These include sequence information, gene expression, functional annotation, tissue location, homologous interactions, and domain based association evidence.

We collect a total of 27 feature attributes from 8 different data sources (Table 4.1). Specifically,

Table 4.1: Feature Set for Human Membrane Receptor Pairwise Protein-Protein Interaction Prediction. We collected a total of 27 features from 8 different data sources. The first column lists the index number. The second column lists the name of the feature source. The third column lists the numbers of features in each source. The fourth column presents the percentage of pairs for which information is available using this feature.

No.	**Feature Source**	**Size**	**Coverage(%)**	**Reference**
1	GO Function	1	39.1	[99]
2	GO Component	1	36.3	[99]
3	GO Process	1	37.6	[99]
4	Tissue	1	57.1	[100]
5	Gene Expression	16	34.0	[101]
6	Sequence similarity	1	100	[59]
7	Yeast Homology PPI	5	100	[102, 60, 4]
8	Domain interaction	1	37.7	[103]

- Features 1-3: GO ontology. Three 'similarity' measures were derived from Gene Ontology (GO) [99], according to the proteins' positions in the three ontology hierarchies: biological process, molecular function and cellular component. For each candidate protein pair the feature describes how many times both proteins are in the same functional class of the GO slim level [99].

- Feature 4: Tissue distribution. The tissue in which a protein is present is an important property of human proteins. To represent whether two proteins appear in the same human tissues or not, we count the number of tissues in which both are expressed [100] and use this number as the feature.

- Features 5-21: Gene co-expression. Features were derived from sixteen expression sets downloaded from the NCBI Gene Expression Omnibus (GEO) [101] database. The gene expression sets have been normalized in the GEO database already. Pearson's correlation between two genes' expression values are calculated and used as features.

- Feature 22: Sequence. The protein sequence alignment score was used as another similarity feature source. We used NCBI's BLAST method [59] to align the two

sequences of each pair. All BLASTP hits with E-values less than or equal to 0.001 were used. The actual E-value was used as the feature, based on the notion that the lower the E-value, the more significant the hit.

- Feature 23-26: Homologous interactions in yeast. Homologous PPIs were derived based on if a candidate pair's homologous proteins bind to each other in another species or not. We derived homology pairs from yeast PPI pairs here. The homology relationship between human proteins and yeast proteins is based on the sequence alignment scores from PSI-BLAST [104] with five iterations of runs. The yeast PPI data sets used include interactions from the DIP database and four other predicted yeast PPI data sets from [4].

- Feature 27: Domain-domain interactions. These features were derived based on the hypergeometric distribution of domain-domain co-occurrence in protein interaction pairs. For a new candidate protein pair we used the smallest p-value from their related domain-domain pairs as features, the smaller the value the more significant the domain pair.

There are many possible ways to encode evidence sources into feature attributes and it is an important factor for the reliability of the computational predictions. For instance, initially we applied detailed encoding of available features which leads to 130 attributes for each pair. While the overall performance of a prediction system based on these features was reasonable (data not shown), biological insight was used to improve the predictions (Figure 4.1, back arrow). For example, in manual inspection of specific predictions, it appeared that functional similarity dominated the selected binding partners and we therefore reduced the number of feature items derived from functional similarity. Biological feedback was also used in optimizing the feature similarity measures. We finally settled on 27 feature attributes for each protein pair as in Table 4.1. Biological feedback significantly improved computational performance and predicted better putative binding partners (data not shown).

4.2.2 Random Forest Classifier

The PPI prediction task is tackled by a supervised learning style. A d-dimensional feature vector $x^{(i)}$ is derived from multiple data sources for every pair of proteins. d describes the total number of attributes and $d = 27$ here. Given these vectors, the task of protein interaction prediction can be presented as a binary classification problem. That is, given feature vector $x^{(i)}$, does this i-th pair interact ($y^{(i)}$=1) or not ($y^{(i)}$= 0)?

There are a number of unique characteristics in this classification task. The features are noisy, have many missing values, exhibit different value types and some features are redundant. In addition, there exists the skewed distribution between the positive class (interacting pairs) and the negative class (non-interacting pairs). In order to overcome these difficulties, we employ the random forest (RF) classifier ([98]) proposed by Breiman for this task.

Random forest [98] uses a collection of independent decision trees instead of one tree. Denote by Θ the set of possible feature attributes and by $h(x, \Theta)$ a tree grown to classify a vector x. Using these notations a random forest f is defined as:

$$f = \{h(x, \Theta_k)\}, k = 1, 2, ..., K \tag{4.1}$$

Where $\Theta_k \subseteq \Theta$. That is, a random forest is a collection of trees, where each tree is grown using a subset of the possible examples. For the k-th tree, Θ_k is randomly selected, and is independent of the past random vectors $\Theta_1, \Theta_2, ..., \Theta_{k-1}$. In order to classify, we count each of the trees' "votes" for one of the classes and the most popular class is assigned to an input x.

Specifically, the random forest is created in the following way (Figure 4.2): Each tree is grown on a bootstrap sample of the training set (this helps in avoiding overfitting). A number $m < d$ is specified, and for each node in the tree, the split is chosen from m variables that are selected at random out of the total d attributes. Once the trees are grown, they can be used to estimate missing values by an iterative algorithm. For the extreme unbalanced positive to negative situation, RF sets different weights for the classes to balance the overall error rate. In general, we grow 200-300 trees in our experiments. For m, we use the default value that was equal to the square root of the feature dimension.

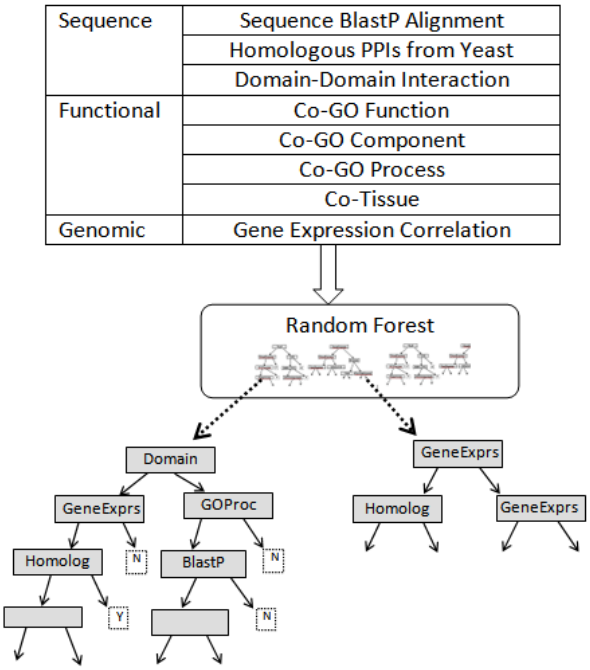

Figure 4.2: Diverse biological data sets are collected and used as evidence to predict PPIs for receptors via evidence integration with the random forest (RF) classifier. To generate the random forest, we select a subset of training data for each tree. Next, for every node in these trees a random subset of the attributes is chosen and the attribute achieving the best division is selected. Once model trees are grown, testing protein pairs are propagated down and the 'votes' from all trees are used as the resulting interaction score.

From the bias and variance analysis of the RF, Breiman claimed that the RF method could be seen as an adaptive nearest neighbor algorithm [105] with the following characteristics: (1) The randomization process works to reduce the variance. (2) RF adapts to a linear loss function by having the narrowest widths in the terminal nodes corresponding to the largest components of the loss function. (3) RF automatically adapts to its sample size. (4) The optimal value of the m parameter does not depend on the sample size. As commonly known, the NN method has low bias and high variance depending on the model parameters [106]. Additionally compared to NN, RF uses the bootstrapping process to reduce model variance at the same time.

4.2.3 Experimental Setup and Evaluation

Reference Set (Gold Standard Set)

In the supervised framework we need a reference set (also called gold-standard set) to evaluate performance of different algorithms. As Jansen et al., (2003) [107] pointed out, the gold-standard data set to train the classifier on should ideally be (a) generated independently from the evidence sources, (b) sufficiently large for reliable statistics, and (c) free of systematic bias. The gold-standard positives are extracted from the Human Protein Reference Database (HPRD) [103]. This data set contained 2522 high-confidence pair-wise protein interactions, where at least one of the interacting proteins is a receptor. The interactions reported in the HPRD were detected by low-throughput approaches revealing physical bindings. A list of 904 human receptor proteins from the Human Plasma Membrane Receptome (HPMR) database [94] was used to filter the HPRD for these positive interactions.

Identification of gold-standard negatives is less straight-forward. Because of the nature of laboratory experiments, it is very difficult to prove that two proteins do not interact and a negative dataset is, therefore, not available. (1) One strategy to handle this issue is to sample random pairs from those possible pairs for which an interaction is not reported in the HPRD. Considering the small fraction of interacting pairs in the total set of potential protein pairs (estimated to be less than 0.1%), the error for contamination is expected to be low [4]. We thus use a random set of receptor-protein pairs excluding all known HPRD pairs as another negative training set. (2) We also explore a filtering of the random negative list as a second strategy, using a random list of receptor-protein pairs not in HPRD and with similar molecular functions. The drawback of the pure random negative set is that random proteins may be very easily distinguishable from interacting proteins simply because of their different functions. We want to measure if classifiers can learn the fine distinctions between functionally related interacting and non-interacting proteins or not.

There is also the concern of homologous proteins in constructing the negative reference set. Since our feature extraction strategy is per protein pair based, two protein pairs with each partner having its homologous protein in another pair would make two very similar

feature vectors. This might cause some problems in the cross validation evaluations. Considering that these pairs are in a small number, we do not filter the negative reference set with homology concern. However this is a nice strategy we would try to use in our future work.

Evaluation Measures

Prediction accuracy versus Sensitivity (also called Precision vs. Recall) curves are used to evaluate computational performance. Prediction accuracy (Precision) refers to the fraction of interacting pairs predicted by the classifier to be truly interacting. Sensitivity (or Recall) measures the fraction of known pairs of interacting proteins have been identified by the learning model. The Prediction accuracy vs. Sensitivity (Precision vs. Recall curve) is then plotted for different cut-offs on the predicted score.

Based on the histogram distribution of the number of interacting partners each receptor has in HPRD, we estimate that on average, only 1 in 1000 human proteins would interact with a receptor (which means that receptors have an average of 25 binding partners). The gold standard reference set is thus constructed using this ratio between positives and negatives. Performance comparisons are based on the train-test procedures: We randomly sample a training set containing 80,000 protein pairs (maintaining the ratio between positive and negative gold standard pairs as listed above) to learn the prediction model. Then we sample a test set (another 80,000 pairs) from the remaining protein pairs, and use the trained model to evaluate the performance of the classifier. The above steps are repeated 10 times for each classifier and average values are reported. Parameter optimization is carried out in all cases using separate train-test runs.

Performance Comparison

Relying on the benchmark data, we statistically evaluate the performance of our approach using precision vs. recall curves and the results are shown in Figure 4.3A-D.

We first describe the results obtained with the random negative gold standard. Comparing the random forest classifier to several other popular classification algorithms (Naive Bayes, Logistic Regression and Support Vector Machine classifiers), we found that the RF

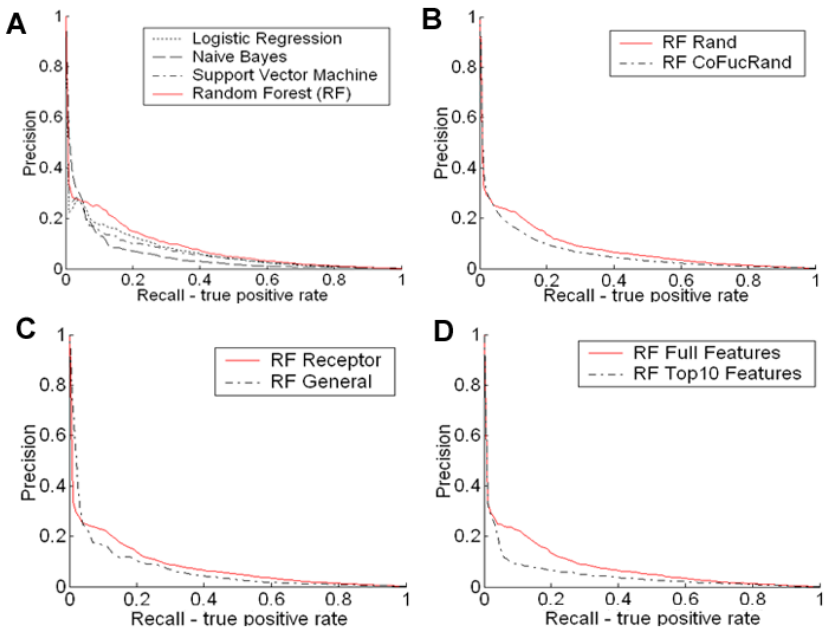

Figure 4.3: Statistical comparison of performance in the human membrane receptor protein interaction prediction task. A. The RF classifier was compared to three other classifiers: Support Vector Machine, Naive Bayes, and Logistic Regression. Prediction accuracy refers to the fraction of predictions that are known to be correct. Sensitivity refers to the fraction of known interactions that are correctly predicted. The prediction accuracy versus sensitivity curve is then plotted for different cutoffs on the predicted scores. B. Evaluation of two different negative gold standards. The negative pairs were either sampled entirely at random or using random receptor-protein pairs with similar functions. C. Performance comparison between receptor interactome identification task and general human PPI prediction task. Classifiers trained only using membrane receptors outperform those trained on the global interaction data. D. Performance comparison between using the full feature attributes and using just the top 10 ranked features based on the Gini criterion. RF classifier is used for subfigures B-D.

method performs the best (Figure 4.3A, red line). Several factors possibly contribute to the success of RF when compared with other classifiers.

- The currently available direct and indirect protein interaction data is inherently noisy and contains many missing values. The randomization and bagging strategies within RF make it more robust to noise when compared to other classifiers.

- Biological datasets are often correlated with each other and thus should not be treated as independent sources. Linear and non-linear regression models assume independence and may therefore perform worse than other classifiers in tasks where correlations among features are strong. In contrast, the RF classifier does not make any assumptions about the relationship between the data, which makes it more appropriate for the type of data available for the protein interaction prediction task.

- It is also important for the method to consider the feature correlation and missing value problems together. If a pair has values for one redundant feature but not the other, the RF method can still use this feature in the prediction process.

- In the PPI detection task there are potentially many more non-interacting pairs compared to interacting pairs (1000:1 ratio estimated). At the same time there are no available large negative sets because it is impossible to prove that two proteins do not interact, since one may simply not have looked under the ”right” conditions. The resulting highly skewed class distribution and the problem of no negative reference set makes the PPI classification task very hard. It is not quite feasible to assume a linear boundary or some other shape of boundaries under these situations. As Breiman claimed in [105] that RF could be seen as an adaptive nearest neighbor (NN) algorithms. This means that the RF model has low bias and high variance potentially. The low bias feature could be one of the reasons why the RF is more successful here.

To address the problem of not having a negative gold standard dataset, we compare the performance with the negative gold-standard set derived by two strategies: random non-interacting pairs and functionally related random non-interacting proteins. Figure 4.3 shows that using the functional gold standard negative achieves somewhat less prediction power as compared to the random gold standard negative.

To predict receptor-protein pairs, an alternative strategy is to predict the general human interactome first, and then extract membrane receptor interactions from this general set. This is a viable option since the features used in the training of our approach are not membrane specific. However, as Figure 4.3C clearly shows, the precision and recall of the receptor interactome is significantly higher when training on the receptor-only gold standard as opposed to training on the entire human gold standard. Clearly focusing on the

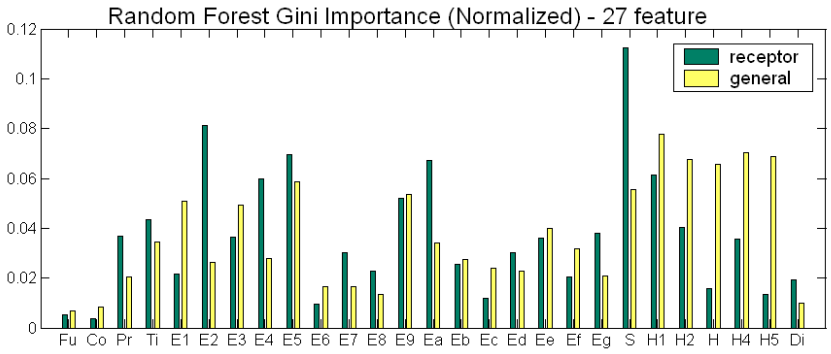

Figure 4.4: Distribution of normalized RF-Gini importance [98] under two cases: receptor interactome identification task and general human PPI prediction task. Clearly features contribute differently for these two tasks. X-axis maps to the 27 features in Table 4.1.

membrane receptor subset for PPI identifications allows us to generate better predictions for a group of receptor proteins that are experimentally very difficult to study. This is probably due to the ability of the classifier to highlight features that are uniquely important for classifying membrane receptors. Figure 4.4 presents the RF-Gini importance (proposed in [98] with details in the next paragraph) of all 27 features for these two cases of classification. We could see that features contribute differently. To predict partners for membrane receptors, the sequence similarity, the biological process evidence, three gene expression sets seem to be more important than for the general human PPI task.

Finally, we investigate which features are most informative for the membrane receptor interaction prediction task. Figure 4.3D shows the performance change when only the top ten most discriminative features were used to train/test by the RF. Top-ranking features were selected using the Gini criterion proposed in [98]. This criterion uses the decrease in the sum of the impurity values (from parent to children) for each feature over all trees in the forest as a simple and reliable estimate of the feature importance for this prediction task. The sequence alignment was ranked as the highest feature. Among the top ten Gini ranked features, five of them are similarities of gene expressions. The other four features ranked in the top ten include domain-domain interaction features, the homology PPI features from yeast, tissue positions and the biological process from GO. While the top ten features achieve reasonable predictions, the performance is significantly less than when

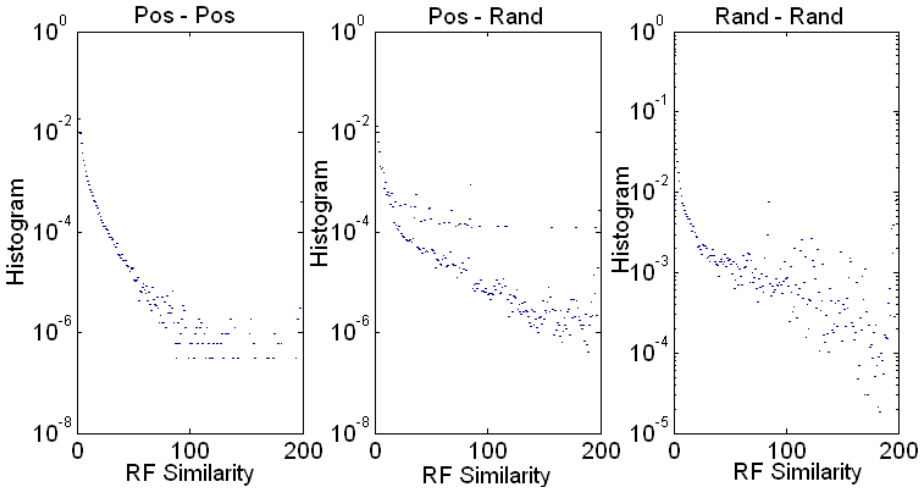

Figure 4.5: The pairwise RF similarity for human receptor PPIs. Three histograms of pairwise similarities between all positive pairs (left) all random pairs (right) and between all positive and all random pairs (center). Note that while the random set is fairly tight, the positive set exhibits a greater diversity and is also far (on average) from most random samples (see Chapter 5 for details of the applied similarity measure).

using all 27 features, suggesting that despite feature overlap they contain highly complementary information.

In conclusion, our performance was best when using receptor pairs only for training, the randomly generated pairs as the negative gold standard and the full set of optimally encoded features. Under these conditions, the RF achieves a prediction accuracy (the fraction of predictions that are known to be correct) of 20% at a sensitivity (the fraction of known interactions that are correctly predicted) of 16%. This performance is comparable to large-scale experimental PPI data sets in general [2] and superior to the receptor related pairs extracted from previous general human interactomes predicted [54, 58] and experimentally determined [95].

Although the 20% accuracy seems not high, our generated results could still help the experimentalists greatly, when searching for interaction partners for receptor proteins. As estimated in human, only one protein-protein pair among 1000 random pairs would be an interacting pair on average. This means, if choosing randomly, biologists need to check 1000 potential pairs by biological experiments (on average) for one PPI. But if using our predicted PPI list (with 20% accuracy), only 5 pairs are needed for finding one PPI on

average. So even though this accuracy value is low, the predictions would be quite useful.

To further investigate the data property, we also draw three histograms of pairwise similarity in Figure 4.5. The pairwise similarity would be defined in Section 5 and represents how similar two protein-protein pairs are. Figure 4.5 describes the similarity distribution between all positive receptor pairs (left), between negative receptor pairs (center), and between positive to negative receptor pairs (right). Clearly the pairwise similarities between negative reference pairs are stronger than the other two cases.

4.3 Global Analysis of Receptor Interaction Network

To estimate what score cut-off we should use to generate a reliable membrane receptor interaction graph, we investigated the distribution of predicted scores based on known HPRD pairs and the remaining random receptor-protein pairs (Figure 4.7). From this graph, it can be seen that a cut-off of 2.0 is stringent in the sense that it is well able to separate the two classes. We therefore generated the membrane receptor interactome using this cut-off. The derived network contains 9100 edges, and includes 559 membrane receptors and 1750 non-receptors (Figure 4.6A). Of the 9100 edges, 1462 edges are already in the HPRD (which achieves 16% accuracy).

One issue is worth to be mentioned that the cut-offs for RF scores are different when used for different purposes. As described above, in order to create a receptor network graph easy for visualization, we use the stringent cut-off 2.0. In the following section 4.4, we choose a lower cut-off 0.0 to select predicted top interaction pairs for biological validations. At this cutoff, our derived list of PPIs achieves 10% accuracy according to the HPRD receptor pairs.

We investigate a series of network properties to describe the human receptor interaction graph at a global scale.

- Hub. We identify those receptors having the largest number of protein partners as 'hubs'. Since GPCR and type I receptors have very different properties, we list the hubs of these two families of receptors separately.

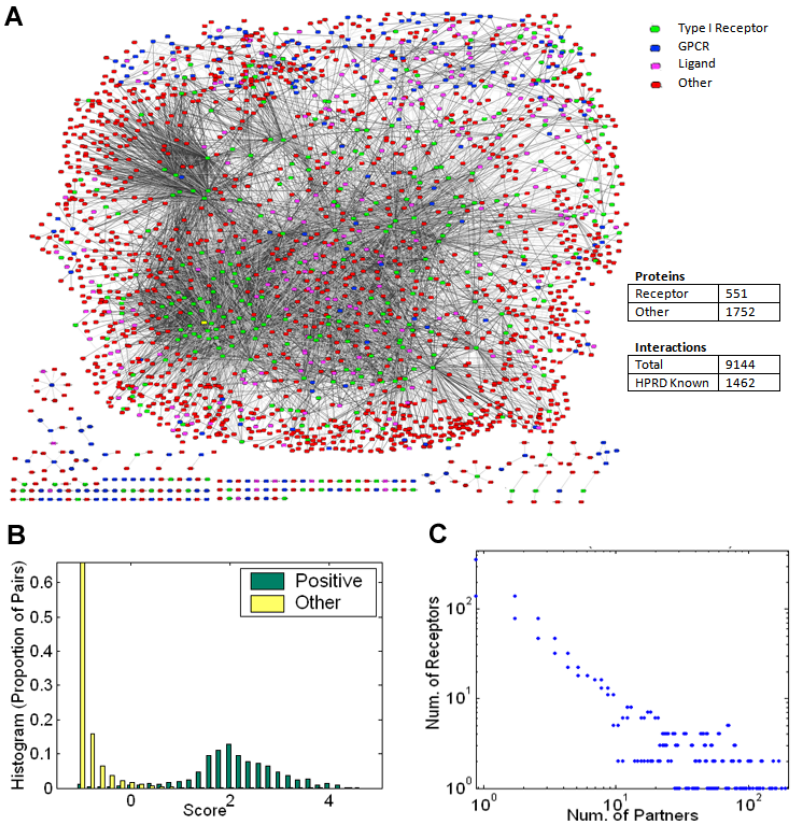

Figure 4.6: Global analysis for the human membrane receptor interaction network. A. Graphical overview of the entire network of interactions. Receptors in the GPCR family are colored blue, type I receptors are green (except EGFR, which is highlighted in yellow), ligands are pink, and other soluble proteins are red. Ligand assignments were extracted from GO. Visualization were performed using Cytoscape [93]. B. The histogram distribution of predicted scores for pairwise receptor-protein pairs. Yellow bars are for positive pairs (in HPRD) and green bars represent the remaining pairs. C. Degree distribution of receptors in the predicted receptor interaction graph.

- Degree distribution. Biological networks usually obey a power law degree distribution. This is also true for our receptor interactome. We calculate a series of degree distributions for receptors with respect to different types of interaction partners.

- Modules. We apply bi-clustering analysis to identify local network modules in the receptor interaction network. Each of these modules contains a subset of receptors and their highly connected partners.

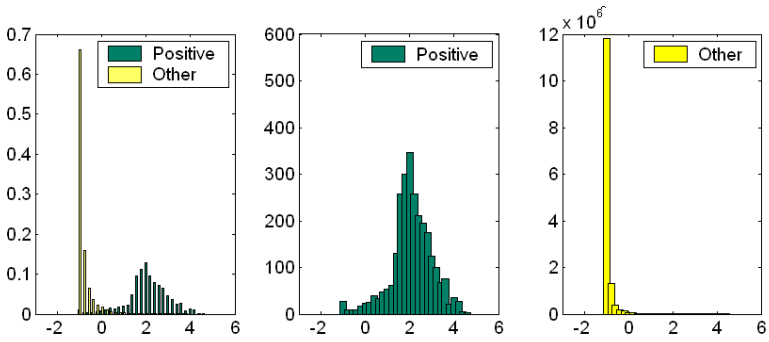

Figure 4.7: The histogram distribution of predicted RF scores for pairwise receptor-protein pairs. Yellow bars are for positive pairs (in HPRD) and green bars represent the remaining pairs. The left-most subfigure is the same as Figure 4.6 with Y-axis describing the percentage values. The middle subfigure is for positive receptor interaction pairs with Y-axis representing the numbers. The right-most subfigure is for the remaining receptor-human protein pairs with Y-axis representing the numbers.

- Subnetwork related subgraph patterns. Since there are two major families of receptors and receptors are closely connected to ligands, it is interesting to see how these families distribute on the graph.

4.3.1 Modules

The membrane receptor interactome is an undirected graph connecting two types of proteins: the receptors and their human interaction partners. It is therefore interesting to find those local modules in which a set of receptors are all highly linked to a similar set of binding partners (Figure 4.8A).

For this purpose, we carry out biclustering analysis [108]. Biclustering is a data mining technique that allows the simultaneous clustering of receptors and their interacting partners. Specifically, the Iterative Signature Algorithm (ISA) [108] is used to search for sub-matrices representing the interaction edges in our network. About 40 biclusters are detected with sizes ranging from 150 genes (47 receptors to 108 partners) to 8 genes (3 receptors with 5 partners).

We observe that the involvement of most molecules in a cluster are in a common pathway which validates the basis of the formation of clusters. The largest bicluster includes

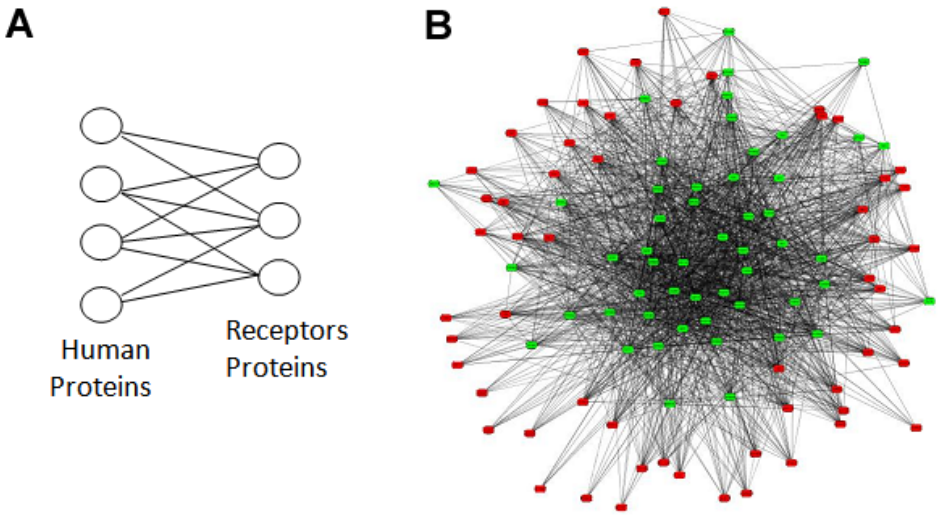

Figure 4.8: Biclustering analysis for the human membrane receptor interaction network. A. Biclustering analysis. B. The visualization of the biggest bicluster in the predicted receptor interactome. Green represents type I receptors. All other human gene nodes are red.

109 gene nodes with 1962 PPI edges between receptors and their partners (Figure 4.8B). This dense subgraph has closely connected Type I receptors and includes an interaction between EGFR and HCK and was further studied experimentally (see below).

4.3.2 Receptor Hubs

Next, we investigate if our graph contains receptor hubs, which are those membrane receptors that interact with a large number of proteins. Figure 4.10 shows the degree distribution of all receptors in the predicted receptor interactome. Degree distribution gives the probability that a selected protein has k partners with $k = \{1, 2, \ldots\}$. We found that the degree distribution of receptors roughly obeys a power law (with a heavy tail. Figure 4.6C). This finding indicates that there are a few hubs that are connected by many proteins characterized by low degree distributions (small number of binding partners) and such proteins [35] could be potential drug targets (Section 3.3). A known receptor hub is the EGFR, and its predicted interactions are shown in Figure 4.10A. The number of known binding partners at the time of downloading HPRD data for the EGFR was 91 (note that we only used interactions that are proven to be direct as opposed to within a complex). We rank-ordered

the predictions based on the random forest score. We found 81 (89%) of the 91 known interactions amongst the top 200 predicted partners, and 90 (99%) amongst the top 700 predicted partners. These results indicate that 119 proteins that are not known to interact with the EGFR have scores equal or higher to known binding partners.

We identify new receptor hubs by ranking those receptors by the number of interactions partners. Using the same stringent cut-off score of 2.0, we predict that several type I receptors may serve as hubs with potentially hundreds of interactions, some of which may even interact with more proteins than the EGFR. In contrast to type I receptors, GPCRs rarely serve as hubs in our receptor interactome. Only 19 receptors are predicted to interact with more than 10 proteins. Interestingly, the function of the majority of these receptors is related to the immune system, with many chemokine receptors being amongst the list.

4.3.3 Protein Type based Graph Patterns

We then try to analyze the interaction patterns between different classes of proteins on the receptor interaction network. Related proteins could be divided into four types: (1) Type I receptors; (2) Type II receptors (GPCR); (3) Ligands; (4) Others. Thus in the following, we investigate this network from these protein types.

Receptor-receptor interactions Many receptors are predicted to interact with other receptors. The currently known receptor to receptor interactions appear to constitute only a minuscule fraction of what may be a network of highly interconnected membrane facilitated associations. This observation is illustrated in Figure Figure 4.9A. Like the EGFR, many other type I receptors are predicted to interact with other type I receptors. While receptor-receptor interactions between GPCRs also occur in our predictions, these are significantly less likely. When predicted, they only connect two or very few receptors. In contrast, type I receptors are highly interconnected. There are also very few examples for predicted direct physical interactions between GPCR and type I receptors. This observation suggests that inhibiting receptor heterooligomerization may not be a general drug design strategy to target GPCR signaling, but may be of general value to target type I receptor signaling. By comparing the degree distributions for type I and type II receptors, we can

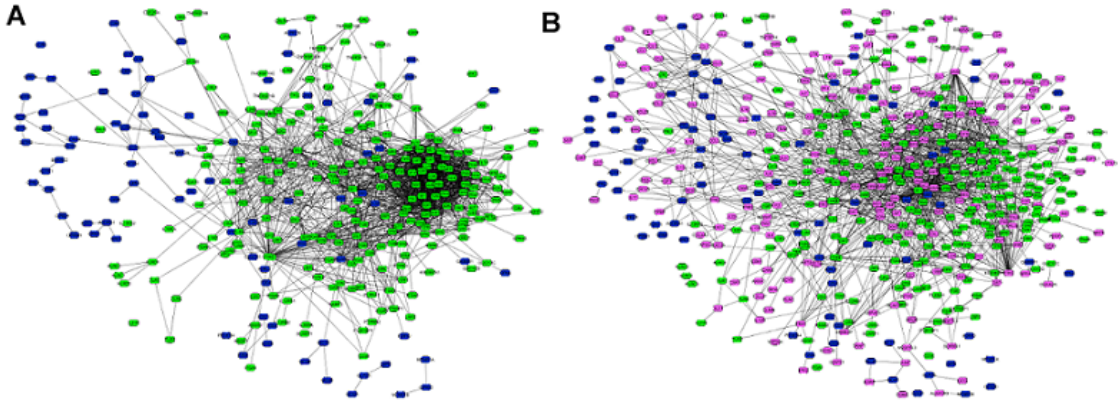

Figure 4.9: Receptor-receptor and receptor-ligand interactions in the predicted receptor interactome. Receptors in the GPCR family are colored blue, type I receptors are green, and ligands are pink. Ligand assignments were extracted from GO (Ashburner et al., 2000). Small groups of connected proteins were omitted from the graphs. A. Receptor to receptor interactions in the receptor interactome network (1721 edges). This network is dominated by the interactions between type I receptors. B. Ligand-receptor interactions in the receptor interactome network (1335 edges). The direct receptor-receptor interaction edges shown in A were omitted in B for clarity.

identify global differences in the respective sub-networks. Type I receptors display many more receptor partners as compared to GPCRs.

Receptor-ligand interactions Our global graph investigation reveals that ligands may be shared by several receptors, regardless of receptor family (Figure 4.9B). However, there appears to be fewer examples for this type of receptor crosstalk than via physical interactions between the receptors: There are many more direct receptor-receptor interactions (Figure 4.9A) than receptor-ligand-receptor interactions (Figure 4.9B). Receptor-ligand interaction predictions could be useful in identifying novel functions of receptors (see rhodopsin-chemokine interaction below), or as indicators of strong functional links (see TGF-β1-EGFR interaction below).

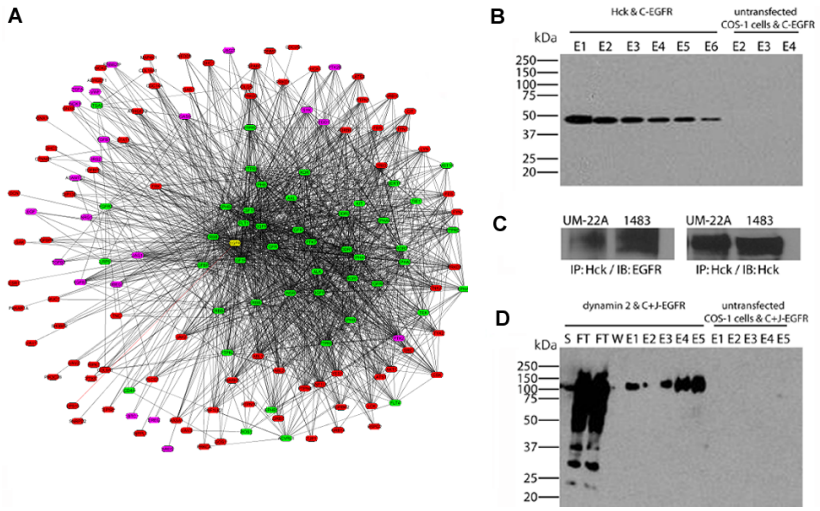

Figure 4.10: EGFR related interactions. A. EGFR related interaction network in the predicted interactome. EGFR is colored yellow, other type I receptors are green, ligands are pink, and other soluble proteins are red. Ligand assignments were extracted from GO [99]. B. Experimental validation of the predicted interaction between EGFR and Hck. EGFR was bound to an antibody column and Hck was bound and co-eluted with epitope peptide. The panel represents an anti-histidine blot to detect Hck. C. EGFR and HCK also interact in 2 HNSCC cell line models (lower panel). D. Experimental validation of the predicted interaction between EGFR and dynamin-2. EGFR were bound to an antibody column and dynamin-2 was bound and co-eluted with epitope peptide. The panel is an anti-GFP blot to detect dynamin-2.

4.4 Biological Validation

The above analysis of the predicted receptor interaction network generated a large number of hypotheses, as well as predictions of many specific interactions. Researchers with interests in any human membrane receptor will find these predictions a rich source of information ready to be exploited through experimental validations. To enable such exploitations, we developed a webservice that provides easy data access of all of our predictions (flan.blm.cs.cmu.edu/HMRI).

To provide examples for the types of experiments and novel findings the predictions stimulate, we discuss case studies related to two global properties as analyzed above: (1) receptor hub interactions and (2) receptor ligand interactions.

(I) Receptor Hub Interactions

Two cases of receptor hub interactions are validated with pull-down assays as described below.

EGFR-HCK Pair The hub receptor EGFR is identified in the most significant bicluster as shown in Figure 4.8B. This cluster also contains the Src-homology kinase Hck, a signaling protein with a role in HIV-1 pathogenesis [109] and oncogenesis [110].

To validate the prediction that Hck and the EGFR interact, co-purification experiments are first carried out with EGFR over-expressed in COS-1 cells, and Hck over-expressed in insect cells. The experiment confirms that the EGFR cytoplasmic domain interacts with Hck (Figure 4.10B). This interaction is novel and may have direct implications in cancer treatment strategies.

EGFR is a direct target for the treatment of head and neck cancer with the FDA approved drug cetuximab. To demonstrate whether Hck, as a member of the Src family kinases, interacts with EGFR in the head and neck cancer cells, a co-immunoprecipitation assay was performed. Figure 4.10C illustrates that under normal growth conditions EGFR interacts with Hck in both UM-22A and 1483 HNSCC cell lines. Interestingly, the src family kinase inhibitor, dasatinib, also has high affinity for Hck (Lombardo et al., 2004) suggesting that in cancer cell lines Hck may interact with EGFR and contribute to tumor progression pathways.

EGFR-Dynamin Pair Another highly ranked predicted interaction of the EGFR was that with dynamin-2, a protein regulating vesicle formation on lipid membranes. A functional link between EGFR and dynamin-2 was already known because catalytically inactive dynamin-2 is no longer able to internalize the EGFR [111]. However it was unknown that the two proteins physically interact. To validate this prediction, co-immunoprecipitation experiments are carried out with proteins expressed in COS-1 cells, which confirmed that the EGFR cytoplasmic domain interacts with dynamin-2 (Figure 4.10D). This suggests that the EGFR is mechanistically involved in its internalization at the molecular level and not just at the regulatory level.

(II) Receptor-ligand Interactions

To investigate the global property of ligands that are shared between different receptors (receptor-ligand interactions), we investigated two such new predictions, between the prototypical GPCR, rhodopsin, and chemokine ligands, and between the type I receptor EGFR and transforming growth factor beta 1 (TGF-β1).

EGFR-TGFβ1 Pair A functional interaction between EGFR and TGF-β1 is able to be confirmed by measuring the ligand-induced phosphorylation level increases in MAPK. PCI-37A squamous cell carcinoma of the head and neck cells were incubated with EGF and TGF-β1 and expression levels of MAPK and phospho-MAPK were detected on a western blot TGF-β1 is able to stimulate MAPK to similar levels as EGF, consistent with a strong functional link between the two pathways [112]. TGF-β1 was also found to co-immunoprecipitate with purified GST-EGFR, suggesting a physical interaction between the two proteins.

Rhodopsin-Chemokine Pair For the GPCR rhodopsin, we find that 8 known binding partners within the first 13 highest ranked predictions. There are a large number of chemokine ligands ranked relatively high: with four of the top 50 ranked interactions being chemokine ligands. A physical interaction between chemokine CXCL11 and GPCR rhodopsin are experimentally confirmed to modulates both proteins' functions.

In G protein activation assays using fluorescence spectroscopy [113] measuring rhodopsin function, a greater than 40% decrease in G protein activation was observed with 100-fold molar excess of CXCL11 over rhodopsin (Figure 4.11A). In chemotaxis assays measuring chemokine function, L1.2 murine preB-cells engineered to express CXCR3, the receptor specific for CXCL11, migrated toward CXCL11 (final concentration of 100nM) as expected. However, in the presence of 75μg of rhodopsin in asolectin lipid vesicles, corresponding to a molar ratio of 30:1 (rhodopsin:chemokine), a dramatic decrease in the number of cells that migrated is observed (Figure 4.11B), which is likely due to the depletion of soluble and available chemokine. Little inhibition is observed when asolectin lipid vesicles are added without rhodopsin. Some inhibition, but significantly less than in the presence of rhodopsin, is observed when asolectin lipid vesicles alone are used without rhodopsin.

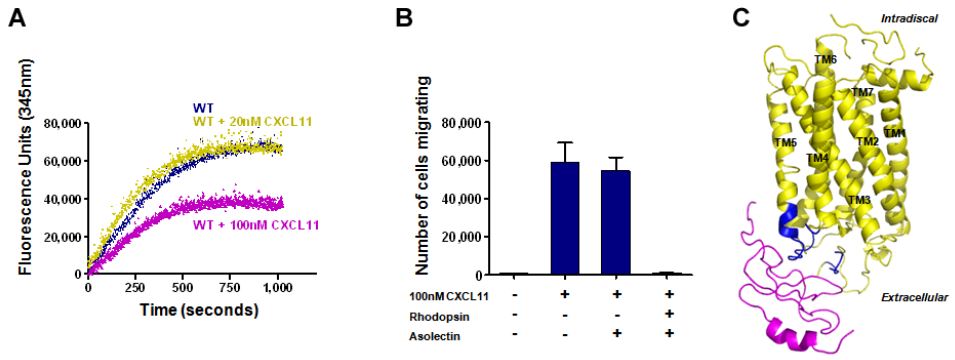

Figure 4.11: Rhodopsin interaction with chemokine CXCL11. A. Antagonist effect of CXCL11 in G protein activation by rhodopsin. G protein activation was measured using fluorescence spectroscopy as described [113]. B. Chemotaxis of L1.2 cells expressing CXCR3 towards chemokine CXCL11. C. Docking of chemokine and rhodopsin structures. CXCL11 NMR structure, pdb id code 1RJT: model 1 [114] was docked to the rhodopsin dark-state crystal structure, pdb id code 1f88 [115] using the ClusPro software [116]

Less inhibition of chemotaxis of CCR4-expressing cells is observed with CCL22.

To further confirm the interaction between CXCL11 and rhodopsin and to begin identifying the potential sites of interaction, we use the ClusPro docking software [116] for rigid docking of the CXCL11 [114] and rhodopsin [115] structures. Forty-seven related orientations of CXCL11 are observed with interfaces at the extracellular domain of rhodopsin, covering the top of its helices 4 and 5 (Figure 4.11C).

The fact that the interaction between rhodopsin and chemokines inhibits both chemokine as well as rhodopsin function suggests that in addition to its established function as the visual photoreceptor, rhodopsin may also modulate immune system function.

4.5 Enhancing Performance with Structural Evidence

An emerging new approach in protein interaction studies is to take advantage of structural information to predict physical binding. Usually interacting pairs of close homologs (proteins that are similar in their amino acid sequence) physically interact in the same way. Moreover, conservation of an interaction depends on the conservation of the interface between interacting partners. Espadaler et al. [44] explored these two hypotheses and used

them to predict new putative interactions.

First, the authors [44] made use of the conservation of pairs of sequence patches involved in protein-protein interfaces to predict putative protein interaction pairs. The interfaces were derived from the analysis of residue contacts of the ”seeding set” of protein complexes with known 3D structure. By utilizing these interface patterns, a set of new interactions were then predicted and called the *Sequence Search of Interface Patterns (SSIP) set*. Second, the authors used the hypothesis that homologous sequences share similar interaction partners, thus stating that the set of interacting partners of a given protein is enriched by its homologs [44]. Through an expansion considering both homologous relationship and known interaction relationships, new interactions were predicted and named as the *Structure Relationship (SR) set*. The intersection of the two sets of potential interactions consisting of the SSIP and those predicted by the SR were further separated into three lists of pairs, according to their relationship with known interacting proteins found in DIP [44].

We utilize the above putative interactions [44] as an extra feature, adding them into our human membrane receptor PPI prediction task, resulting in a total of 28 features. We evaluate the prediction performance by keeping the experimental setup the same as described in Section 4.2.3. When using the RF classifier, we find that the average precision-vs-recall curves of the two cases (original 27 features versus the current 28 features) are quite similar.

We also utilize another popular performance evaluation criterion, namely the AUC scores to compare the performance of our method with and without the structural feature. The area under the ROC curve (AUC) is commonly used as a summary measure of diagnostic accuracy. ROC means Receiver Operator Characteristic curves which measure the trade-off between sensitivity and specificity. In the PPI prediction task, we are predominantly concerned with the detection performance of our models when the false positive rate is low. This maps to the area under a portion of the ROC curve. For example, R50 is a partial AUC score that measures the area under the ROC curve until reaching 50 negative predictions. Similarly R100 is the partial AUC score when reaching 100 negative predictions. Table 4.2 lists the average AUC scores of two cases by the RF classifier. We could see that there is a marginal increase in the AUC scores after adding the structure feature. This observation suggests that these types of structure-based features are promising and may more significantly improve the performance when investigating small sets or specific

Table 4.2: Comparison of average AUC scores when adding a protein structural feature to enhance human membrane receptor PPI predictions.

Score	AUC	AUC100	AUC300
RF 27fea	0.9243	0.1111	0.1919
RF 28fea	0.9269	0.1108	0.1936

predicted interaction pairs.

4.6 Summary

Protein-protein interactions (PPI's) are critical for virtually any biological function. Analysis of interactions in signal transduction pathways in particular can help understand disease mechanisms and provide hypotheses on new disease targets.

In this chapter we propose a combined computational and experimental approach to predict and validate individual pairwise protein interactions in protein networks relating to human membrane receptors. Due to the experimental challenges present for this type of proteins we rely on biological datasets providing indirect evidence about interaction relationships. We develop a classification strategy to integrate evidence from different data sources for predictions of receptor-protein interactions. It has been suggested previously, that focusing on specific subnetworks may provide more reliable information [117]. We clearly show that focusing on predicting membrane receptors generates better predictions than first predicting all human protein interactions and then selecting those related to membrane receptors.

Global analysis of the resulting membrane receptor interactome suggests that it may contain several receptor hubs and numerous receptor-receptor interactions, predominantly between type I receptors. Several novel receptor-ligand interactions are also found in our predictions. We have experimentally validated both hub receptor interactions and receptor-ligand interactions, which provide novel hypotheses on protein function and mechanism of action. For example, the implication of rhodopsin in immune system function by its interaction with chemokines would not have been possible without the computational approach

presented here. Since we predict thousands of previously unknown interactions, this and the other example interactions validated serve to demonstrate the potential in generating novel and experimentally testable hypotheses.

Researchers [118, 87] recently found that protein interaction hubs could be analyzed in two types: party hub and date hub. Clear differences between party hubs (static complexes) and date hubs (transient interactions) exist. For example more date hubs contain long disordered regions than the party hubs, indicating that these regions are important for flexible binding but less important to static interactions. Because of the insufficient availability of structural data and the lack of large-scale information on time- and condition-dependence of transient interactions, we do not consider the hub distinction when analyzing receptors hubs in our global analysis. However this is an interesting and important direction to extend our work.

The integration of computational PPI prediction, network analysis, biological experimentation, and biological expert feedback presents a feasible strategy to discover novel biological hypotheses in an iterative and reliable manner.

Chapter 5

PPI Prediction Using Ranking

Chapter 4 proposed a method to combine computational PPI learning, network analysis, in vitro experimentation, and biological expertise for identifying interaction partners for human membrane receptors. In this chapter we make efforts for detecting protein-protein interactions in yeast. Here candidate interaction pairs are identified relying on the assumption that they are ”similar” to known interacting pairs according to multiple feature evidence.

5.1 Introduction

Chapter 4 dealt with the PPI prediction for human membrane receptors using classification approaches. The PPI prediction task in general has several properties that distinguish it from those tasks in a typical binary classification setting:

- First, the task has a highly skewed class distribution, which means that there are many more non-interacting pairs than interacting pairs. On average only 1 in $\sim$1000 human proteins interacts with another human protein. A similar estimation was conducted and averagely only 1 in 600 possible protein pairs actually interact in Yeast.
- Second, only a small number of positive examples (interacting pairs) are reliable. Also no available large negative set is available.
- Third, the cost for misclassifying an interacting pair is different from the cost for misclassifying a non-interacting pair.

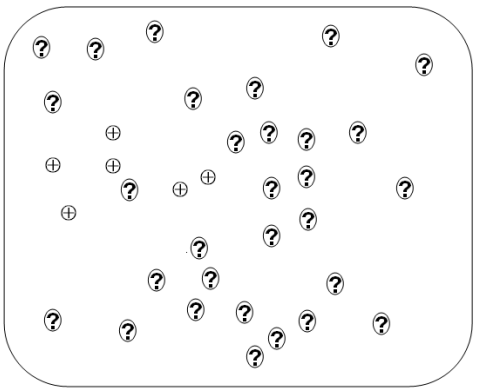

Figure 5.1: A schematic diagram of data distributions in the PPI prediction task. The task has a highly skewed class distribution, with only a small number of positive examples (circled '+') being known and no confirmed large set of negative examples (circled '?').

All these constraints make the computational problem quite hard (Section 4.2.3). Figure 5.1 gives a schematic illustration for the distribution of examples possible for this task: circled plus symbols represent positive examples and circled question marks mean the rest. Clearly assuming a linear boundary or some other shapes of boundaries between two classes (positive to the rest) does not seem appropriate.

Our goal of computational PPI predictions is to find out possible interacting pairs in a certain model organism as accurately as possible and as completely as possible. Additionally, when biologists choose candidates based on the computational predictions, due to the difficulties intrinsic in the biological experiments, it is not feasible for them to validate a large number of putative interactions. Thus, it is essential to rank unknown positive pairs as high as possible.

Considering the properties of reference sets and the above goals, we propose to handle this task in a ranking style: (1) Trying to rank the known positive items high in the prediction list; (2) Being able to rank the unknown positive items as high as possible.

As described in Chapter 3 and Chapter 4, it is important to integrate both direct evidence and indirect biological information when predicting protein-protein interactions. When combining disparate biological datasets, several facts need to be considered: (1) Biological

features are noisy and are often containing many missing values. (2) Features are heterogenous, with some being categorical while others continuous. (3) Features correlate with the interaction relationship at various levels. Some features should be weighted more heavily than others (for example, the direct experimental evidence should be weighted higher than the indirect evidence).

This chapter focuses on how to measure the similarity between protein pairs by integrating multiple biological evidence. We present a method that overcomes the above problems by using a random forest method to compute the similarity between protein interaction pairs. We construct a set of decision trees in which each tree contains a random subset of the attributes. Next, protein pairs are propagated down the trees and a similarity matrix based on leaf occupancy is calculated for all pairs. Decision trees and the randomization strategy within the random forest can accommodate heterogenous data and can automatically weight the different data sources based on their ability to distinguish between interacting and non-interacting pairs. Because the trees are generated from random subsets of the possible attributes, missing values are filled in by the use of an iterative algorithm. Finally, a weighted k-nearest-neighbor algorithm, in which distances are based on the computed similarity, is used to classify (rank) pairs as interacting or not.

5.2 Methods

Multiple high-throughput datasets were used to construct a d-dimensional vector $X^{(i)}$ for every pair of proteins. Each entry in the vector summarizes one of these datasets (asking, for example, "Are these two proteins bound by the same transcription factor?" or "What is their expression correlation?". Table 5.1 gives a complete set of attributes in each vector). Given these vectors the task of predicting protein interaction can be represented as a binary classification problem. Given $X^{(i)}$ does the i-th pair interact ($Y^{(i)} = 1$) or not ($Y^{(i)} = -1$).

As we point out above, this task has a number of properties (high noise rate, missing value problem and heterogenous nature), reference set problem (highly skewed and no negative set) and prediction objectives (ranks also matter and cost factor). In order to overcome these difficulties we divide the classification task into two steps (see Figure 5.2):

- First we compute a similarity measure between pairs of genes.

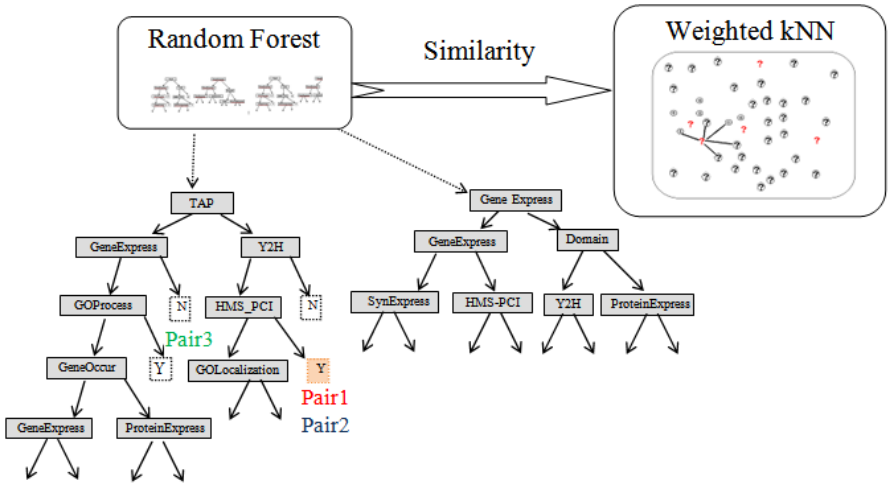

Figure 5.2: Classification process. To generate the random forest a subset of training data is selected for each tree. Next, for each node in the trees a random subset of the attributes is chosen and the attribute achieving the best division is selected. Once the trees are grown, all protein pairs (remaining training and test sets) are propagated down the trees and similarity is computed based on leaf occupancy (see text). Using the computed similarity a weighted kNN algorithm is used to rank pairs by the resulting interaction scores.

- Then this measure is used to rank protein pairs by a k-Nearest-Neighbor (kNN) approach.

Below we discuss each of these two parts in detail.

5.2.1 Random Forest Similarity

Random forest [98] method is used to determine the similarity between protein pairs here. As stated in Chapter 4, Random forest (RF) was initially introduced as a classification algorithm, though it can also be used to compute similarities. RF constructs a large set of independent decision trees. Results from these trees are combined for the classification or similarity calculation task.

A decision tree is a binary tree with nodes corresponding to attributes in the input vectors. Tree nodes are used to determine how to best propagate a given attribute set down the tree. Nodes can either be threshold nodes or categorical nodes. Decision trees also contain terminal (or leaf) nodes that are labeled as -1 or 1. In order to classify a protein

pair as interacting or non-interacting, the pair is propagated down the tree and a decision is made based on the terminal node that is reached. The Random Forest (RF) [98] consists of a collection of independent decision trees. Trees are grown using a training set. At each node the algorithm searches for the attribute that best separates all instances in that node. If the attribute perfectly classifies all instances so that all of them in one of the two descendent nodes have the same label, then this node becomes a terminal node with the appropriate label.

For a given forest f we compute the similarity between two pairs of protein pairs $X^{(1)}$ and $X^{(2)}$ in the following way. First, we propagate the value of each pair down all trees within f. Next, the terminal node position for each pair in each of the trees is recorded. Let $Z^{(1)} = (Z_1^{(1)},, Z_K^{(1)})$ be these tree node positions for $X^{(1)}$ and similarly define $Z^{(2)}$. Then the similarity between pair $X^{(1)}$ and $X^{(2)}$ is set to:

$$S(X^{(1)}, X^{(2)}) = \sum_{i=1}^{K} I(Z_i^{(1)} == Z_i^{(2)})/K \tag{5.1}$$

Where I is the indicator function. As we discuss in Results, in order to allow for cross validation tests we partition our training set to two lists. The first is used to generate the random forest. The second is used for the kNN algorithm. In order to compute similarities using the second set the following algorithm is used. Given a random forest with K trees and up to N terminal nodes in each tree we first generate a $N * K$ vector V where each entry in V contains a linked list of training set pairs that reside in that node. Given a new test pair we first propagate it down all trees (in O(N*K) time) and for each of the terminal nodes it arrives at, we find the corresponding set of training pairs from V. For each such pair we increase their similarity count by one. Thus, for a given test pair it takes only O(—S_{train}—+NK) to compute its similarity to all the training points, where S_{train} is the training set and $|S|$ means the number of elements in S.

5.2.2 Classification of Protein Pairs

We use a weighted version of the k-Nearest Neighbor (kNN) algorithm to classify pairs as interacting or not. While we have tried a number of classifiers for this data (see Results)

the main advantage of kNN for this task is its ability to classify based on both similarity and dissimilarity (as opposed to similarity alone). As can be seen in Figure 5.32, while non interacting pairs are similar to each other, the main distinguishing feature of interacting pairs is their distance from (or dissimilarity with) non interacting pairs. Due to the highly skewed distribution of interacting and non interacting pairs, it is likely that the closest pair to an interacting pair will be a non interacting pair (though their similarity might not be high). Decision trees (or RF) may use these to incorrectly classify an interacting pair as non interacting. However, kNN can take into account the magnitude of the similarity, and if it is too weak, it will be classified as interacting.

Specifically, given a set of training examples $(X^{(i)}, Y^{(i)})$, and a query point $X^{(q)}$, we calculate the interaction possibility score for q using the weighted mean of its neighbor's $Y^{(i)}$ values, where the weight depends on the similarity of each of training pairs to q:

$$f(q) = \sum_{p=1}^{k} S(X^{(q)}, X^{(neighbor(p))}) * Y^{(neighbor(p))} \tag{5.2}$$

Here $S(X^{(i)}, X^{(q)})$ is the similarity between i and q as computed by RF. The *testing* set then turns into a ranking list of protein pairs by these pairs' derived interaction scores. A cutoff t can be derived using a validation set to threshold this ranking list such that q is classified as interacting if $f(q) > t$.

When the cost factor needs to be considered, our model could be easily extended. For example, if the cost of misclassifying one positive pair is '2' and the cost of misclassifying one negative pair is '1', we could change the positive pairs' label value as $Y = 2$ instead of the original $Y = 1$ (negative pairs' label value would remain '-1' for this case). This shows how the cost factor would affect the ranking list through Equation 5.2.

5.3 Experiments and Results

We first discuss the biological data used for building the feature vectors of Yeast pairs. Next we present results for applying our classification method for determining protein interaction pairs in yeast.

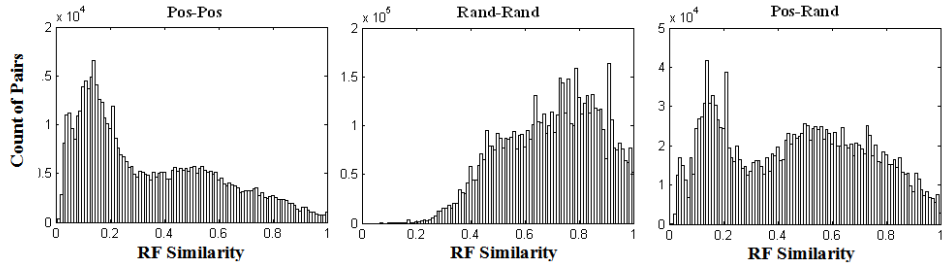

Figure 5.3: The pairwise RF similarity. Three histograms of pairwise similarities between all positive pairs (left) all random pairs (center) and between all positive and all random pairs. Note that while the random set is fairly tight, the positive set exhibits a greater diversity and is also far (on average) from most random samples. kNN can utilize this fact by relying on the actual distance to the closest neighbors (see text for details).

5.3.1 Feature Set

As mentioned in the introduction, there are many high throughput biological data sources related to protein-protein interactions. The method described here is general and can be used with any type of biological data. Thus, while we have tried to use as much data sources as we could, when a new data source (such as protein expression arrays) becomes available, the method discussed in this chapter can take advantage of that data as well. We extract a total of 15 attributes (Table 5.1) for each protein pair. Overall, these data sources can be divided into three categories: Direct experimental data sets (two-hybrid screens and mass spectrometry), indirect high throughput data sets (gene expression, protein-DNA binding etc.) and sequence based data sources (domain information, gene fusion, etc). The feature vector for yeast PPI task in this chapter are encoded with the summary feature style. In addition to combining these data sources, our method can also indicate which of the different data sources is better for predicting protein interactions as we discuss below. Biological data sets usually have missing value problems. Table 5.1 lists the degree of missing value for each feature as its forth column.

5.3.2 Performance Comparison

We compare the proposed method with several popular classifiers in this section.

Table 5.1: Feature set for Yeast PPI predictions(in summary encoding). We extract a total of 15 attributes for each protein pair. Overall, these data sources can be divided into three categories: direct experimental data set, indirect high throughput data sets and sequence based data sources. The first column lists the feature index number. The second column lists the feature name. The third column indicate the degree of feature importance. Features are listed in decreasing order using the importance ranking. The forth column lists the coverage of the feature.

No. of Features	Dataset	Importance Ranking (Counting Way, N=4)	Coverage
1	Gene Expression	0.1868	0.9586
2	Protein Expression	0.1302	0.3696
3	Domain-Domain Interaction	0.1154	0.0064
4	GO Biological Process	0.1053	0.6305
5	TAP Mass	0.0800	0.0468
6	Synthetic Lethal	0.0739	1.0000
7	GO Cellular Component	0.0570	0.7852
8	GO Molecular Function	0.0519	0.5188
9	Y2H	0.0427	0.3382
10	HMS-PCI Mass	0.0425	0.0468
11	Syn-expression	0.0385	1.0000
12	Gene Co-occur	0.0317	1.0000
13	Gene Neighborhood	0.0211	1.0000
14	Protein-DNA Binding	0.0178	1.0000
15	Gene Fusion	0.0052	1.0000

Reference set Any classification algorithm requires a training set. For the positive set (or the interacting pairs) we use a set of 4000 protein pairs derived for the database of interacting proteins (DIP) [60]. This database is composed of interacting protein pairs which have been experimentally validated, and thus can serve as a reliable positive set. Unlike positive interactions, it is rare to find a confirmed report on a non interacting pair. A random set of protein pairs (minus positive ones) is used as the negative set. This selection is justified because of the small fraction of interacting pairs in the total set of potential protein pairs. It is estimated that only 1 in 600 possible protein pairs actually interact [119, 47] and thus, over 99.8% of our random data is indeed non-interacting which is

Table 5.2: Reference set (gold standard set) for yeast PPI predictions.

Set	Num. of Pairs	Note
Positive Set	4036	Small scale PPI experiments (from DIP [60])
Random Set	2,391,420	Random minus positive

probably better than the accuracy of most training data. The basic situation of our reference set is described in Table 5.2 below.

Evaluation setting We use the precision vs. recall curves to perform the comparisons. Precision estimates the fraction (or percentage) among the pairs identified as interacting by the classifier that are truly interacting. Recall means for the known interaction pairs, what is the percentage that is identified? In other words, precision is the accuracy of our predictor whereas recall is the coverage of the classifier. Note that even 50% or lower precision can useful. For example, biologists studying a specific protein can extract a list of potential interacting partners computationally first and carry out further experiments knowing that on average 50% of their experiments will identify true interacting pairs. The ratio is much better than if the set of potential pairs was randomly selected (estimated 1 out of 600).

For our algorithm, in each cross-validation run, we divide our training set into two equal parts. The first part is used to construct the random forest and the second is used by the kNN algorithm. Thus, our algorithm uses the same amount of training data as the other algorithms we compare to (see below). In order to generate a precision-vs-recall curve we use different thresholds as discussed above. For the other classification methods, we generate the curve in a similar manner. For instance, for the naive Bayes classifier, we can use the naive Bayes prediction probability of a test point to arrive at a ranked list.

Performance Figure 5.4 shows a comparison between our method and a number of other classification algorithms. The figure on the left compares our method with a weighted kNN that uses Euclidean distance instead of the random forest similarity, with the naive Bayes method and with a single decision tree. For a wide range of high precision values our method outperforms the other methods. It is especially interesting to compare our method

with the kNN method using Euclidian distance. As can be seen, using the robust similarity estimates from the random forest results greatly improves the classification results. Figure 5.4 also includes a comparison of our algorithm to a classification method that only uses the resulting random forest (based on popular vote) to classify protein pairs and to a number of other popular classifiers including Support Vector Machine (SVM), logistic regression and Adaboost (right figure). In all cases, our algorithm performed better for precision values that are higher than 0.32. Specifically, holding precision fixed at 0.5 (or 50%), our algorithm achieved a recall rate of 20% while logistic regression achieved 14% recall, random forest and SVM achieved 11% recall and Adaboost had a 7% recall rate. Finally, we note that while the methods we have used to compare our algorithm with were inspired by previous work (such as single decision tree [52] and naive Bayes [48], we have used a slightly different feature set and a different training set compared to each of these two papers. Thus, the results reported for these methods here are not comparable to the ones earlier reported in these related papers.

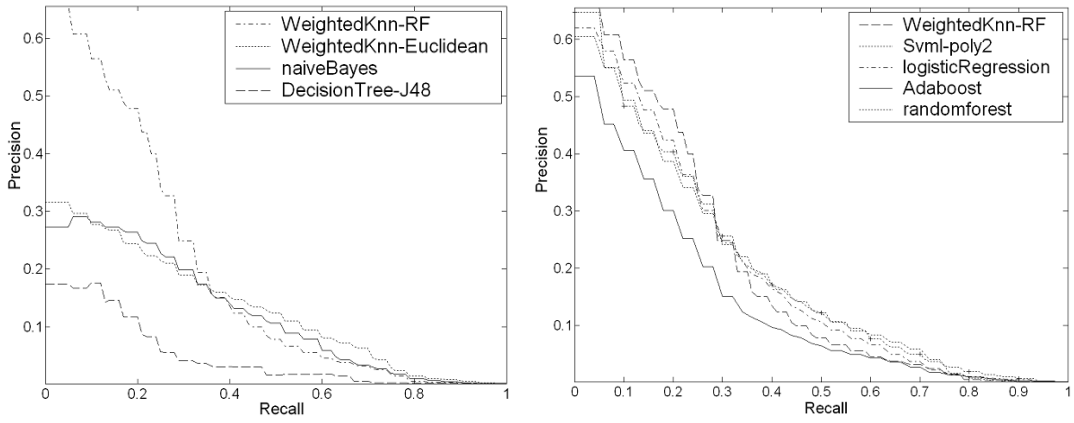

Figure 5.4: Precision vs. Recall curves. Left: Comparison of weighted kNN using random forest similarity, weighted kNN using Euclidean distance, naive Bayes and a single decision tree (J48); Right: Comparison of weighted kNN using random forest similarity, logistic regression, support vector machine, Adaboost and random forest classifier.

Important attributes Biologically, it is of particular interest to identify the attributes and data sources that contribute the most to our ability to classify protein pairs. Such

an analysis can help uncover relationships between different data sources which are not directly apparent. In addition, it can help identify what data sources should be generated for determining interaction in other species (for example, in human). One way to determine such a set using random forest is to score attributes based on the levels of nodes that use them to split the data. Since each node splits the data using the best available attribute, attributes used in higher levels in the tree contribute more than those used in lower levels.

To arrive at a rough estimate for the contribution of each attribute we have counted the percentage of nodes that use this attribute in the top four levels of all trees in our trained random forest model. Of the 15 features we used, gene co-expression had the highest score with 18% of top nodes using it to split the input data. Next came three features: protein co-expression, domain-domain interaction and GO co-process, each with $\sim$11% of the nodes. These were followed by TAP mass spectrometry data (8%) GO co-localization (6%), Y2H screens (4%) and HMS-PCI (4%) (See Table 5.1 for the complete list). Interestingly, indirect information played a very important role in the decision process though this results may result from the fact that direct experiments cover less than 30% of all protein pairs. However, mass spectrometry data are clearly more significant than Y2H data, consistent with the notion that mass spectrometric identification of protein-protein interaction is less prone to artifacts than Y2H experiments. It is particularly encouraging that co-expression and GO features contribute such strong components to the prediction, clearly supporting the notion that a large amount of indirect data measuring the biologically relevant information is helpful in predicting interaction partners.

5.3.3 Validation: Yeast Pheromone Response Pathway

To analyze the utility of our computational results in the design of new experiments, we compare the predictions of our method to their labels for one specific pathway, the yeast pheromone pathway. The yeast mating factors MAT /a bind to their cognate membrane receptors, Ste2/3, members of the G protein coupled receptor family. Subsequent binding and activation of the G protein induces a MAP kinase signaling pathway via the G protein $\beta\gamma$ subunit.

Table 5.3: Performance statistics of our validation. We compare predicted pairs with the small positive label DIP [60] set and the pair interacting relationships in KEGG [62]. The number "19" with star symbol labels those pairs having no evidence as interacting or non-interacting, which need further analysis.

	Test Set	Predicted Positive	Predicted Negative
Whole	300	44 / 300	256 / 300
POS by both	11 /300	8 / 44	3 / 256
POS by DIP only	31 /300	13 / 44	18 / 256
POS by KEGG only	8 / 300	4 / 44	4 / 256
Rest	251 / 300	19* / 44	231 / 300

We select 25 proteins that are known to participate in this pathway. We apply our algorithm (using a different training set) to classify the 300 (25*24/2) potential interacting pairs. The performance statistics are presented in Table 5.3. Our algorithm classify 44 of these pairs as interacting. 31 of these pairs (70.45%) are known to interact while only 2 (4.55%) are verified to be wrong predictions. The remaining 11 pairs (25%) are new predictions that are reasonable and would functionally make sense. They form two clusters: The first involves the possible interaction between the STE5 anchor protein and the receptors. The receptor would then lead to additional interactions due to STE5's anchoring function. The second cluster involves a possible interaction between the most downstream components of the signaling cascade including BEM1, BNI1 and FUS1, mediating cell fusion (see [8] for details). These new preliminary findings can be used to design future lab experiments.

5.4 Summary

In this chapter we presented a method for predicting protein-protein interactions by integrating diverse high-throughput biological datasets. Our method involves two steps. First, a similarity measure is computed between protein pairs. Then a classification algorithm uses the computed similarities to classify pairs as interacting or non-interacting. We have applied our algorithm to the task of classifying protein pairs in yeast. As we have shown, our algorithm outperforms previous methods suggested for this task and can also produce

meaningful biological results for known pathways in yeast.

We have used random forest to derive a similarity function between protein pairs. Recently, a number of methods have been suggested for learning distance matrices [120]. We would like to test some of these methods and see if they can improve the accuracy of our classification. It will be especially challenging to apply these methods to datasets with missing values, like those used in this research.

Interestingly, many of the features determined to be important using our method are indirect measurements. This opens the possibility of extending this method to determine interacting pairs in organisms for which little direct high throughput information is available, such as humans.

Chapter 6

PPI Prediction by Multiple View Learning

In the traditional *single-view* machine learning scenario, learners have access to the entire set of features in the domain (for example, see Chapter 4 and Chapter 5). By contrast, in the *multi-view* setting, one can partition the domain's feature in subsets (views) that are sufficient for learning the target concept. For instance, in our PPI prediction task, some researchers carried out interaction prediction based on the protein sequence and structure only; Others predict interactions on pathways through gene expression data analysis. In a multi-view learning problem, an example x is described by a different set of features in each view (we name these views as feature experts in the following). For example, in a domain with two views V_1 and V_2, a labeled example is a triple $< x_1, x_2, y >$, where y is its label, and x_1 and x_2 are its descriptions in the two views. Similarly, an unlabeled example is denoted by $< x_1, x_2, ? >$. Learning from multiple views was first proposed by Blum and Mitchell (1998) in the Co-Training [121] approach for web classification task. They proved that for a problem with two views the target concept can be learned based on a few labeled and many unlabeled examples, provided that the views are compatible and uncorrelated. Multi-view algorithms then have been successfully applied to a variety of tasks in real-world domains [122, 123, 124].

Our last two chapters handle the PPI prediction problem by *single-view* learning. In this chapter we handle this prediction task through multiple views learning using a mixture

model.

6.1 Introduction

A number of researchers recently (Chapter 3) presented methods for integrating direct and indirect data in predicting interactions. While useful, the methods do not address two important problems in this domain. First, these classification methods estimate a set of parameters that are used for all input pairs. However, the existing biological datasets contain many missing values and highly correlated features. Thus, different protein pairs may benefit from using different feature sets. The second problem is that biologists who want to use these methods to design experiments cannot easily determine which of the features contributed to a resulting prediction. Since different researchers may have different opinions regarding the reliability of the various features, it is useful if the method can indicate, for every pair, which of the features contribute the most to the classification result.

In this chapter we address the above challenges from multiple views using a mixture model that is named Mixture-of-Feature-Experts (MFE). We divide the biological datasets into several groups (each group as a view or an expert). Each group represents a specific data type and is used by a feature expert (classifier) to predict interactions. Results from all experts are combined such that the weight of each expert depends on the input sample and thus varies between input pairs. This weight can also indicate the importance of the features used by the expert for predicting a pair. We applied our method to predict PPIs in yeast and human. Using Precision vs. Recall curves and AUC scores we show that the MFE method improved upon traditional classification methods that were previously applied in predicting PPIs. For a specific Yeast pathway, the pheromone pathway, we show that it is possible to extract confidence information from the weight distribution, in addition to providing new predictions.

6.2 Methods

There are many biological data sets that may be directly or indirectly related to PPIs. We tried to collect as many as possible for yeast and human. The extracted features are described in detail in Chapter 4 and 5.

Several feature properties need to be considered when designing algorithms. First different features have varying degrees of missing values. Second, the derived features are *heterogeneous*. In addition, some of them are highly correlated features (for example expression data from two different stress response experiments). Third, there is the issue of *weighting* these different data sources. Different protein pairs may benefit from using different feature sets in the prediction process. For every pair it is useful for computational techniques to provide information about how features contribute to the classification predictions. For biologists who want to use these methods to build new hypotheses, integrating this information and the expert knowledge could assist in lab experimental design.

6.2.1 Feature Experts

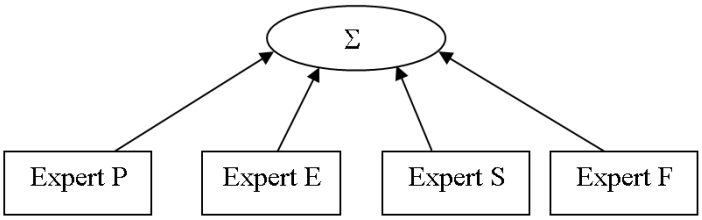

Figure 6.1: Mixture of Four Feature Experts. Graphical representation of the Mixture-of-Feature-Experts method (MFE) for yeast. For definition of P,F,S,E experts, see details in Section 6.2.1.

Overall, the biological data sources can be divided into four feature categories, which are referred to as feature experts (called as 'views' in multiple view learning) in this chapter:

1. Expert P: direct high-throughput experimental PPI data. This category contains those data sets that directly detected interaction relationships between proteins. They were

derived through high-throughput biological experiments such as Y2H screens and mass spectrometry.

2. Expert E: indirect high-throughput data. This category includes those experimental data sources that were generated through high-throughput techniques and represent certain aspects of genes/proteins other than the PPI relationship, such as gene expression and protein-DNA binding.

3. Expert S: sequence based data sources. This category includes those features that represent how similar two proteins are based on sequence or structure information. For example, this expert includes domain information and gene fusion data.

4. Expert F: functional properties of proteins. This category contains information about how similar two proteins are in terms of functional annotations such as biological process, protein localization, protein class, and essentiality.

After splitting, the features within experts are derived from similar data sources and are roughly *homogeneous* when compared with each other. Usually biologists could give opinions and make comparisons on general categories of biological evidence. Thus, it would also be useful for computational methods to provide automatic information about how several feature categories (experts) contribute to every predicted interaction pair. The derived computational importance together with biologists' expert knowledge could assist the further prediction analysis and the design of lab PPI experiments. In this work, we divided features into four experts. Apparently, the number of experts to be split into could be different. The splitting depends on the need of the application and the analysis ability of the biologists who would validate the predictions.

Features for Yeast For yeast a total of 162 feature attributes is collected from 17 different data sources (Table 6.1). Three data sources are derived from the direct high-throughput yeast PPI data sets, with two from mass spectrometry and one from high-throughput yeast-two-hybrid screens. These evidence describe pair of proteins directly and thus are used as feature items in the feature vector. Six data sources represent each gene's functional annotations. The 'similarity' features derived from them represent how similar two proteins occur

in the certain annotation space or from a specific function perspective. Four other different sources derived features that describe the similarity between two genes from sequence and structure perspectives. The remaining attributes are all based on indirect high-throughput experimental data. For example, this includes gene expression correlations. All related data sources and how they were converted into features representing pair of proteins have been described in details in [4].

Table 6.1: Feature set derived for pairwise protein-protein interaction prediction in yeast. We use a total of 162 features from 17 different data sources. The first column lists the feature expert to which the feature source was assigned. We have designed a total of four experts: P, F, S and E (for definition see Section 6.2.1). The second column lists the name of the feature source. The third column lists the number of attributes from each source. The fourth column presents the average percentage of pairs for which information is available using this feature source. All related data sources and how they were converted into features have been described in details in [4].

Expert	**Feature Source**	**Size**	**Coverage (%)**
P	HMS-PCI MS	1	8.3
P	TAP MS	1	8.8
P	Yeast-2-Hybrid	1	3.9
F	GO Function	21	80.7
F	GO Process	33	76.1
F	GO Component	23	81.5
F	Essentiality	1	100
F	MIPS protein class	25	4.6
F	MIPS mutant phenotype	11	9.4
S	Gene fusion/cooccurence	1	100
S	Sequence similarity	1	100
S	Homology derived PPI	4	100
S	Domain interaction	1	100
E	Gene Expression	20	88.9
E	Protein Expression	1	42.8
E	Trans Factor Binding	16	98.0
E	Synthetic Lethal	1	7.6

Table 6.2: Feature set derived for pairwise protein-protein interaction prediction in human. We collected a total of 27 features from 8 different data sources. The first column lists the feature expert to which the feature source was attributed to. Unlike yeast, for human we had a total of three experts: F, E and S (for definition see Section 6.2.1). The second column lists the name of the feature source. The third column lists the number of attributes from each source. The fourth column presents the average percentage of pairs for which information is available using this feature source.

Expert	Feature Source	Size	Coverage(%)
F	GO Function	1	39.1
F	GO Component	1	36.3
F	GO Process	1	37.6
F	Tissue	1	57.1
E	Gene Expression	16	34.0
S	Sequence similarity	1	100
S	Yeast Homology PPI	5	100
S	Domain interaction	1	37.7

Features for Human For human we collected a total of 27 feature attributes from 8 different data sources (Table 6.2). Collecting data for human proteins is much harder than for yeast because several data sets that are available for yeast are not yet available for human and there exist much more human proteins than yeast.

Note that in human there are only two very small Y2H data sets [34, 19] available. We therefore currently do not have a 'P' feature expert for human data. As more data sets become available, this feature expert can be generated for human as well.

6.2.2 Mixture of Feature Experts (MFE)

Using the multi-view setting, features are grouped into four (for yeast) or three (for human PPIs) categories. While the features are heterogeneous overall, within feature experts, attributes are roughly homogeneous and are derived from similar data sources. Our main intuition in using the expert-based structure is to investigate the relationship between these homogeneous feature groups in terms of predicting PPIs and to compare the importance of experts contributing to each prediction. This provides a principled way for selecting

informative feature types during the prediction process.

We design a method called Mixture-of-Feature-Experts (MFE) to achieve the above computational properties. As Figure 6.1 shows, our framework can be viewed as a single layer tree, with feature experts at the leaves. Each expert uses one of the dataset groups to predict PPIs. A root gate is used to integrate predictions from multiple feature experts. The weights assigned to each of the experts by the root gate depends on the input set for a given pair. Intuitively, this framework is analogous to the following process: each feature expert gives their opinion about how likely the investigated pair interacts and then the gate creates a final decision by the weighted sum of the experts' predictions. Moreover, these weights are local and specific to the current example pair.

In the following sections, X describes the input feature vector variable and Y represents the target output variable. Input variable X represents d-dimensional feature vectors built from features in Table 6.1 or Table 6.2. Target variable $Y \in \{-1, 1\}$ means whether a protein pair interacts (1) or not (-1).

Given our feature experts setting, the conditional probability of the target variable Y given the input variable X could be written as:

$$P(Y|X) = \sum_{M} P(M|X)P(Y|X, M) \tag{6.1}$$

where M is a set of $hidden$ data and indicates which expert was responsible for generating each example data pair. Having I experts, M is a I-dimensional indicator vector variable. That is, all entries in M are 0 except for one of the entries which is set to 1. The sum is over all configuration of variable M. In other words, target class label Y is dependent on the input data X and the choice of expert M. The choice of M is also dependent on the input X. $P(M|X)$ is modeled using the root gate, while $P(Y|M, X)$ is modeled by each feature expert in our framework. The graphical model view of MFE method is illustrated in Figure 6.2. This Bayesian network structure states that the target variable Y is dependent on the input vector variable X and the multinomial random variable M. It is essentially a modification of the probabilistic *Mixture-of-Experts* (ME) model [125].

Using a training set including N examples, the n-th example pair is described using $(x^{(n)}, y^{(n)})$. For $n = 1$ to N, each data example $(x^{(n)}, y^{(n)})$ has a corresponding vector

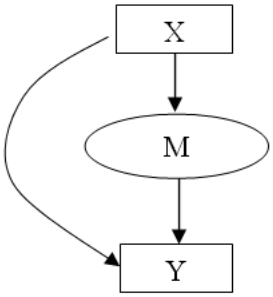

Figure 6.2: A graphical model view of the Mixture-of-Experts (ME) method. The target variable Y is dependent on the input vector X and the multinomial random variable M. $P(M|X)$ is modeled by the gate while $P(Y|X, M)$ is modeled by the experts.

$m^{(n)}$. The dimension of vector $m^{(n)}$ is equal to the number of feature experts: I ($I = 4$ for yeast and $I = 3$ for human). With $i = 1$ to I, $n = 1$ to N, each entry of this vector $m_i^{(n)}$ is as following:

$$m_i^{(n)} = \begin{cases} 1, & \text{if using feature expert } i \text{ for example } n \\ 0, & \text{otherwise.} \end{cases} \tag{6.2}$$

Thus, based on Equation (6.1) the conditional probability $P(y^{(n)}|x^{(n)})$ is formulated specifically as:

$$P(y^{(n)}|x^{(n)}) = \sum_{i=1}^{I} P(m_i^{(n)} = 1|x^{(n)}, v) P(y^{(n)}|x^{(n)}, m_i^{(n)} = 1, \omega_i) \tag{6.3}$$

where w_i are the model parameters used for feature expert i and v contains the model parameters used for the gate.

In general each expert can take any form such that the expected value of their probability density is consistent with the form of the problem. In this work, we use binary logistic

regression for each of the feature experts. For the i-th expert ($i = 1...I$) we write:

$$P(y^{(n)}|x^{(n)}, m_i^{(n)} = 1, \omega_i) = \frac{1}{1 + exp(-y^{(n)}(w_i^T x^{(n)}))} \tag{6.4}$$

Similarly, the root gate can take any functional form that is consistent with a probability distribution. For instance [125] used multinomial logit models for the gates. Here, we extend the binary logistic regression to model the multinomial probability distribution of variable M through voting. This is analogous to using the one-versus-all strategy to transform a I-class classification into I binary logistic regression problem [106]. First binary logistic regression model is run once for each output branch of the root gate. Next, modified probability weights are calculated for each branch by combining all branch models. Each branch of the root gate controls the weighting of a certain feature expert in our work. For the i-th branch ($i = 1...I$ for our gate) v_i represent the logistic regression parameters for this branch and variable C_i represents the binomial probability distribution from this branch. Thus,

$$P(c_i^{(n)} = 1) = \frac{1}{1 + exp(-(v_i^T x^{(n)}))} \tag{6.5}$$

Then by normalizing over all branches, we get the multinomial probability distribution of variable M as below:

$$P(m_i^{(n)} = 1|x^{(n)}, v) = \frac{P(c_i^{(n)} = 1)}{\sum_{j=1}^{I} P(c_j^{(n)} = 1)} \tag{6.6}$$

This means that $P(m_i^{(n)} = 1|x^{(n)}, v)$ depends on the input attributes ($x^{(n)}$) and it represents the gate weight for expert i when predicting the n-th pair. In all of the above logistic regression steps, we apply ridge estimators to infer stable regularized parameters.

In summary within our feature experts framework the interaction prediction from MFE is a weighted sum of the opinions from each feature expert. The weights assigned to each expert are controlled by the input feature values as well as by the feature experts.

6.2.3 Expectation Maximization (EM)

Based on the probabilistic model in Equation (6.1), learning in MFE architecture is treated as a maximum likelihood problem. The model parameters include the gate parameters v and the expert parameters ω_i. We compute the log likelihood by taking the logarithm of the products of $P(y^{(n)}|x^{(n)})$ as follows,

$$ll = \sum_{n=1}^{N} log(\sum_{i=1}^{I} P(m_i^{(n)} = 1|x^{(n)}, v)P(y^{(n)}|x^{(n)}, m_i^{(n)} = 1, \omega_i)) \tag{6.7}$$

In the following we use Θ as the set of all the parameters including both experts and gate parameters.

Jordan and Jacobs [126] have proposed an expectation-maximization (EM) algorithm for adjusting parameters in ME architecture. The EM algorithm is an iterative approach for maximum likelihood estimation (MLE). Each iteration of an EM algorithm consists of two steps, the E-step and the M-step. For the t-th epoch, model parameters are represented as Θ^t.

In the E-step we compute the posterior probability $h_i^{(n)}$ using Equation (6.8). $h_i^{(n)}$ represents the posterior weight for expert i in predicting pair n once both the input and the target output are known. $h_i^{(n)}$ is derived using Bayes rule:

$$h_i^{(n)} = P(m_i^{(n)} = 1|x^{(n)}, y^{(n)}, \Theta^t) \tag{6.8}$$

$$= \frac{P(m_i^{(n)} = 1|x^{(n)}, \Theta^t)P(y^{(n)}|x^{(n)}, m_i^{(n)} = 1, \Theta^t)}{P(y^{(n)}|x^{(n)}, \Theta^t)} \tag{6.9}$$

$$= \frac{P(m_i^{(n)} = 1|x^{(n)}, v^t)P(y^{(n)}|x^{(n)}, m_i^{(n)} = 1, \omega_i^t)}{\sum_{j=1}^{I} P(m_j^{(n)} = 1|x^{(n)}, v^t)P(y^{(n)}|x^{(n)}, m_j^{(n)} = 1, \omega_j^t)} \tag{6.10}$$

By decomposition of the expected complete data-likelihood, the M-step reduces to separate maximization problems [126, 125], one for each expert and gate. In our MFE framework it solves the following maximization problems: for each expert,

$$\omega_i^{t+1} = argmax_{\omega_i} \sum_{n=1}^{N} h_i^{(n)} log(P(y^{(n)}|x^{(n)}, m_i^{(n)} = 1, \omega_i)) \tag{6.11}$$

and for the root gate,

$$v^{t+1} = argmax_v \sum_{n=1}^{N} \sum_{j=1}^{I} h_j^{(n)} log(P(m_j^{(n)} = 1|x^{(n)}, v)) \tag{6.12}$$

Each of these maximization problems are themselves maximum likelihood problems [126, 125]. Equation (6.11) is simply the general form of a weighted maximum likelihood problem in the probability density $P(y^{(n)}|x^{(n)}, m_i^{(n)} = 1, \omega_i)$. Given our expert choice, the log likelihood in Equation (6.11) is a weighted log likelihood (weighted by $h_i^{(n)}$) for the logistic regression model. An efficient algorithm known as iteratively reweighted least-squares (IRLS) is available to solve this maximum likelihood task [126].

Equation (6.12) involves maximizing the cross-entropy between the posterior probability $h_j^{(n)}$ and the prior probability $P(m_j^{(n)} = 1|x^{(n)}, v)$. This cross-entropy is the log likelihood associated with a multinomial logistic gate model in which the $h_j^{(n)}$ could be treated as an output observation. Thus the maximization in Equation (6.12) is a maximum likelihood problem for a generalized linear model and can also be solved using IRLS technique.

Overall the EM algorithm could be summarized as the following iterative process:

1. For each data pair $(x^{(n)}, y^{(n)})$, compute the posterior probability $h_i^{(n)}$ using the current values of the parameters.

2. For each expert i, solve a maximization problem in Equation (6.11) with observation $\{x^{(n)}, y^{(n)}\}_{n=1}^{N}$ and observation weights $\{h_i^{(n)}\}_{n=1}^{N}$.

3. For the root gate, solve the maximization problem in Equation (6.12) with observation $\{x^{(n)}, y^{(n)}\}_{n=1}^{N}$ and observation weights $\{\{h_i^{(n)}\}_{n=1}^{N}\}_{i=1}^{I}$.

4. Iterate by using the updated parameter values until a termination criterium is satisfied.

6.2.4 Dealing with Feature Missing Value Problem

As pointed out, biological datasets contain many missing values and this problem prevails a serious obstacle in achieving significant improvements in prediction performance.

The simplest approach to handle the missing feature items is to fill the missing entries using certain values. For example, for a real-valued feature the filled value could be the mean of the feature column or for a categorical feature we could use the most common value. In the following sections we use the term 'MFE-FM' to represent the MFE method while using mean estimates for missing values (MFE-FM: mixture of feature experts with missing values filled).

We apply a more principled strategy to handle missing feature values. Specifically, for each feature that has low feature coverage, this strategy add an extra feature column to represent the feature availability.

For $d = 1...D$ ($D = 162$ for yeast and $D = 27$ for human), X_d represents the d-th feature column and $g(X_d)$ describes the ratio of missing cases for feature X_d. If $g(X_d)$ is larger than a predefined ratio, we add a new, binary, feature column $X_{(D+1)}$ to represent the availability of feature X_d. That is, if for an example pair the feature X_d is missing, this new feature $X_{(D+1)}$ would be set to 0. Otherwise it would be set to 1. The method now uses this new feature and can learn different parameters for observed and estimated features. Totally if there are p original feature columns that have new feature columns added, the final feature vector then grows to be $D + p$ dimensional. While this strategy increases the size of our feature set, it is still very small ($\sim$200 for yeast and $\sim$50 for human) compared to the total number of protein pairs ($\sim$18M for yeast and $\sim$4000M for human).

In our MFE framework, since the weighting depends on the input features, using this adding features strategy our classifiers can use the present / absent information to modify the weights of different feature experts. Similarly this strategy could also improve the classifiers used by each feature expert. In the following sections the term 'MFE' means the MFE method when using this added extra features strategy.

6.3 Experiments and Results

We first discuss the reference sets and evaluation strategies used in performance comparisons. Next we present results for comparing the MFE method to several popular classifiers for predicting protein interaction pairs in yeast and human.

6.3.1 Experimental Setting

Reference Set (Gold Standard Set) As described in Chapter 3, any classification algorithm requires a training set. For the positive set, there are a small number of interacting protein pairs that have been reliably determined by small-scale laboratory experiments. This set serves as our positive standard for this learning problem. For yeast, ∼2900 interacting protein pairs were extracted from the database of interacting proteins (DIP) [60]. For human, ∼15,000 protein-protein interaction pairs were extracted from the Human protein reference database (HPRD) [127]. Both sets were filtered to exclude self-interactions. A random set of protein pairs are used as the negative set, excluding those interacting pairs that are known. In yeast, it is estimated that roughly only 1 in about 600 possible pairs actually interacts [8]. In human, this ratio is even smaller, roughly 1 in several thousands of possible pairs is estimated to interact. Thus, over 99.8% of our random data is indeed non-interacting, which is probably better than the accuracy of most training data. Combining the positive and negative PPI sets, a reference set (also referred to as gold standard set) is then constructed for use as training/testing sets when applying learning methods.

The extreme unbalanced ratio situation between positive and negative sets should be taken into account in designing proper computational methods for this task.

Evaluation Strategy Based on the reference set, we use again the following two measures to evaluate the performance of our predictions, Precision vs. Recall curves and AUC scores (area and partial areas under the Receiver Operator Characteristic curve).

In Precision vs. Recall curves, Precision refers to the fraction of interacting pairs predicted by the classifier that are truly interacting (true positives). Recall measures how many of the known pairs of interacting proteins have been identified by the learning model. The Precision vs. Recall curve is then plotted for different cutoffs on the predicted score.

Receiver Operator Characteristic (ROC) curves plot the true positive rate against the false positive rate for different cut-off values of the predicted score. It measures the trade-off between sensitivity and specificity. The area under the ROC curve (AUC) is commonly used as a summary measure of diagnostic accuracy. It can take values from 0.0 to 1.0. In some cases, rather than looking at the area under the entire ROC curve, it is more informative to only consider the area under a portion of the curve. In our prediction task, we are

predominantly concerned with the detection performance of our models under conditions where the false positive rate is low. For example, R50 is a partial AUC score that measures the area under the ROC curve until reaching 50 negative predictions. Similarly R100 is the partial AUC score when reaching 100 negative predictions.

6.3.2 Performance Comparison

To measure the ability of the MFE method to predict PPIs, we compared it with four other popular classifiers that have been suggested in the past for this task: Logistic Regression (LR) , Naïve Bayes (NB), Support Vector Machines (SVM) and Random Forest (RF). Our MFE method is implemented using Matlab. Standard toolkits are used for the other methods. Specifically, The SVMlight toolkit was used for SVM [128]. Logistic Regression and Naive Bayes were obtained from the WEKA machine learning tool box [129]. Random Forest was from the Berkeley RF package [98]. The input feature vectors to these methods are exactly the vectors from Table 6.1 or Table 6.2 with missing values filled.

(Note: In contrast to the summary style of PPI features used in Chapter 5, the features for Yeast PPI task in this chapter are encoded with the detailed feature style. Within our systematic study work [4], we have found that under the detailed feature style, RF similarity ranking method (Chapter 5) achieves similar performance as RF but not better. Though for the summary encoding style, it does improve the PPI predictions compared to RF [4]. We do not put the comparison result of RF similarity ranking method in this chapter. Generally speaking, the classifiers achieve different performance under the summary and detailed encoding styles.)

All comparisons were based on the following training and testing procedures. In yeast, we randomly sampled a training set containing ∼30,000 protein pairs to learn the decision model. Then we sampled a test set (another ∼30,000 pairs) from the remaining protein pairs, and used the trained model to evaluate the performance of the classifiers. The above steps were repeated 10 times for each classifier and average values are reported. Similar procedures were pursued in human where the training and the testing sets included ∼80,000 examples. For each evaluated classifier, parameter optimization was carried out in all cases in identical train-test fashion.

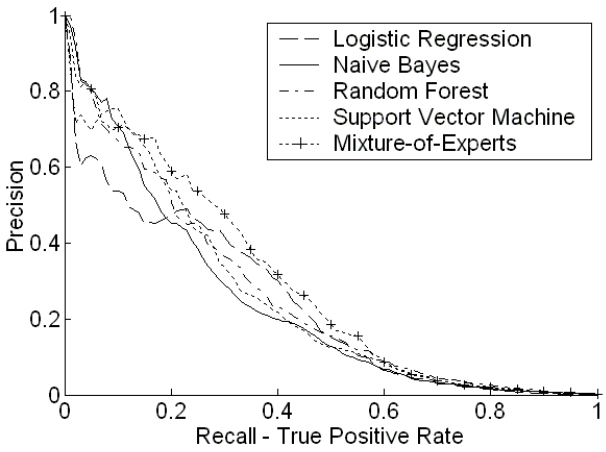

Figure 6.3: Average Precision vs. Recall curves when comparing MFE method with four other classifiers (LR/NB/RF/SVM) for PPI prediction in yeast. LR: Logistic regression; NB: Naive Bayes; RF: Random Forest; SVM: Support Vector Machine; MFE: Mixture-of-Feature-Experts. The MFE curve dominates the curves for the other four methods in most of the recall ranges.

Based on the estimated ratio of interacting versus non-interacting pairs in yeast and human, we have roughly $\sim$50 to $\sim$100 positive PPIs in each test run. For the training set, we up-sampled the positive examples in a pre-processing step, which resulted in roughly $\sim$800 positive examples for each training run in human and roughly $\sim$300 positive pairs for each yeast training. This sampling strategy reduces the problem of too few positive examples in the training set without affecting the performance significantly [130].

Figure 6.3 plots the average precision versus recall curves of these five different methods for the yeast PPIs prediction and Figure 6.4 is for human. In both figures, the curves derived from MFE approach dominate the other four methods in most of the low recall ranges.

Table 6.3 lists the average AUC score and partial AUC scores for the yeast PPI evaluation. The standard derivations for each score estimation are also listed in the table. MFE scores are highlighted and it clearly achieves better AUC/R50/R100 scores compared to the other methods. For instance, MFE improves the R50 score by $\sim$7% when compared to the other classifiers tested. Table 6.4 lists the scores for the human data set. Similarly as for yeast, the MFE method achieves better results. For example, MFE achieves $\sim$10%

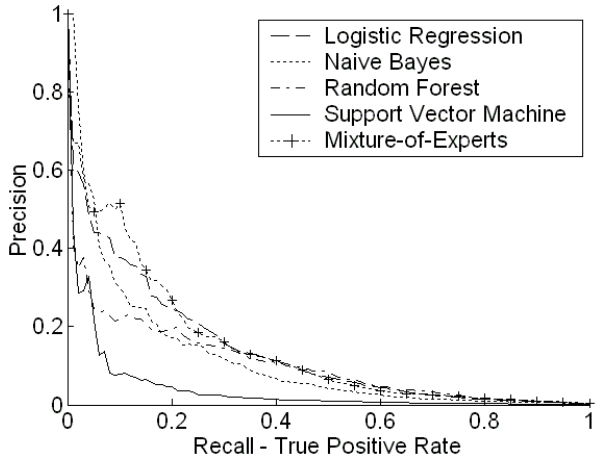

Figure 6.4: Average Precision vs. Recall curves when comparing MFE with four other classifiers (LR/NB/RF/SVM) for PPI prediction in human . LR: Logistic regression; NB: Naive Bayes; RF: Random Forest; SVM: Support Vector Machine; MFE: Mixture-of-Feature-Experts. Again, the MFE curve dominates the other four curves for most of the low recall values.

Table 6.3: Average AUC and partial AUC scores for six classification methods for PPI prediction in yeast. LR: Logistic regression; NB: Naive Bayes; RF: Random Forest; SVM: Support Vector Machine; MFE: Mixture-of-Feature-Experts; MFE-FM: Mixture-of-Feature-Experts with missing features filled. Average AUC and partial AUC scores are reported and the standard derivations for each score estimation are also listed in the table. MFE scores are highlighted and it clearly achieves better AUC/R50/R100 scores compared to the other five.

Method	**AUC mean**	AUC std	**R50 mean**	R50 std	**R100 mean**	R100 std
LR	0.8823	0.033	0.2866	0.070	0.3546	0.073
NB	0.9349	0.015	0.2486	0.047	0.3135	0.062
RF	0.9321	0.014	0.2688	0.048	0.3434	0.049
SVM	0.9159	0.024	0.2585	0.063	0.3262	0.067
MFE	**0.9463**	0.013	**0.3080**	0.078	**0.3799**	0.077
MFE-FM	0.9220	0.021	0.2918	0.061	0.3738	0.058

improvement in R50 score compared to the other classifiers used. Thus the MFE method achieves the best results for all criteria tested.

Table 6.4: Average AUC and partial AUC scores for six classification methods for PPI prediction in human. LR: Logistic regression; NB: Naive Bayes; RF: Random Forest; SVM: Support Vector Machine; MFE: Mixture-of-Feature-Experts; MFE-FM: Mixture-of-Feature-Experts with missing values filled. Average AUC and partial AUC scores are reported and the standard derivations for each score estimation are also listed in the table. MFE scores are highlighted and it again achieves better AUC/R50/R100 scores compared to the other five classifiers.

Method	**AUC mean**	AUC std	**R50 mean**	R50 std	**R100 mean**	R100 std
LR	0.9419	0.020	0.1148	0.031	0.1684	0.031
NB	0.9389	0.003	0.0964	0.031	0.1356	0.035
RF	0.9427	0.009	0.0740	0.025	0.1263	0.030
SVM	0.7645	0.091	0.0455	0.028	0.0589	0.040
MFE	**0.9608**	0.007	**0.1341**	0.023	**0.1759**	0.027
MFE-FM	0.9384	0.018	0.1297	0.023	0.1713	0.025

The last two rows of Table 6.3, list the AUC and partial AUC scores of MFE-FM and MFE methods in yeast. MFE clearly achieves better performance compared to MFE-FM ($\sim$3% increase in R50 score). This means that by explicitly indicating the availability of feature attributes our method improves the classification outcome. Similar conclusions could be drawn for human as shown in Table 6.4.

The feature-experts methodology we proposed is very general. As discussed in the 'Methods' section, the number of feature experts the heterogeneous data sets are split into could be different. The splitting essentially depends on the need of the application and the preference of the biologists who would analyze and/or validate the predictions. At the limit case, we can assign each feature to an individual expert. To test this we carried out one new experiment for the human prediction task treating every feature as its own expert. As the results (details see [10]) indicate, this does not improve the performance of the algorithm, perhaps because it leads to overfitting of the parameters.

6.4 Feature Importance Discussion

Biologically, it is of particular interest to identify the extent to which heterogeneous data sources carry information about protein interactions. An analysis of the contribution of different features can also help uncover relationships between different data sources that are not directly apparent.

Analysis of feature importance is important on the global scale as well as for the prediction and analysis of specific protein pairs. We therefore ask the following questions: (1) How do the different features affect PPI prediction performance overall? and (2) How do the different features contribute differently for each example pair? We have explored these two questions using the yeast results.

6.4.1 Global Feature Analysis

To control data collection costs, it is important to select only informative data types globally. Once informative data types are identified, one does not need to use unnecessary data sets when solving similar network inference problems for other sets of proteins or for other organisms. This can significantly speed up prediction of PPIs in new species, as well as when updating predictions on model species such as yeast and human with new data sources.

To identify overall feature importance among our feature experts, we remove feature experts one by one, and run the MFE methods on the remaining three experts. We then examine how the performance changes. Table 6.5 lists the score changes of R50 and AUC after removing the experts one by one. The less the score changes the less important the feature expert is. We found that removing the sequence expert 'S' had the least impact on both scores. The indirect high-throughput data expert 'E' ranked second from the bottom in the prediction of yeast PPI's.

It is surprising that removing expert 'E' (which contains mostly microarray expression data) does not hurt performance much. This is seemingly in contradiction to previous estimations in which tree based feature ranking methods ranked gene expression features very highly [4]. Note that, when the feature sets are not grouped, the wide availability of gene expression data and its high coverage may result in an increased use of this feature,

Table 6.5: Global feature expert importance can be measured by the decrease in AUC and R50 scores when removing the expert in the MFE method. The first column lists the four feature experts. The second and fourth column list the R50 and AUC scores when applying MFE while only using the remaining three experts. The third and fifth column list the changes between these R50 and AUC scores and the full experts version. For definition of P,F,S,E experts, see details in the 'Feature Experts' section.

	MFE R50	R50 DROP	MFE AUC	AUC DROP
P	0.2310	0.0770	0.9244	0.0219
F	0.2609	0.0471	0.8821	0.0642
S	0.3191	-0.0111	0.9459	0.0004
E	0.3022	0.0058	0.9323	0.0140
Full	0.3080		0.9463	

even though it may lead to overfitting. As our results suggest, splitting the data into more homogeneous groups (feature experts here) may help increase the prediction accuracy by decreasing its reliance on these high throughput data sources.

The possible reason of this conflict might come from how the two methods use the features in prediction. In the RF method, features correlations are investigated implicitly during the construction of tree structures and the randomization process. This means that expression related features could affect the prediction more when they are being combined with other features. When they are separated from other features (which happens in our feature expert method), they themselves are not strong evidence to make accurate protein interaction prediction. This observation is consistent with biological intuition.

6.4.2 Feature Importance for Specific Protein Pairs

For each predicted pair it would be useful for computational techniques to provide information about which features contributed to the predictions for that pair. Our MFE method naturally reveals how each feature category contributes to the interaction predictions. The posterior probability from Equation (6.8) could be treated as the level of contribution from

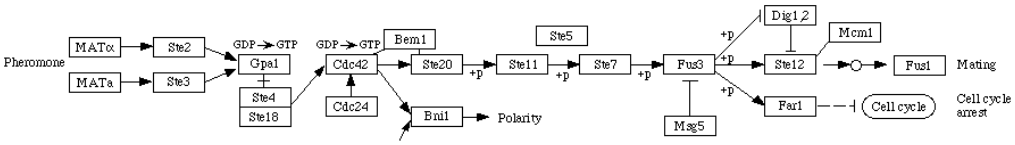

Figure 6.5: The yeast pheromone response pathway. This figure is from the KEGG [62] database.

each expert to the final prediction. Then for a specific candidate protein pair, these values could give a detailed description about how each expert contributes to the integrated prediction.

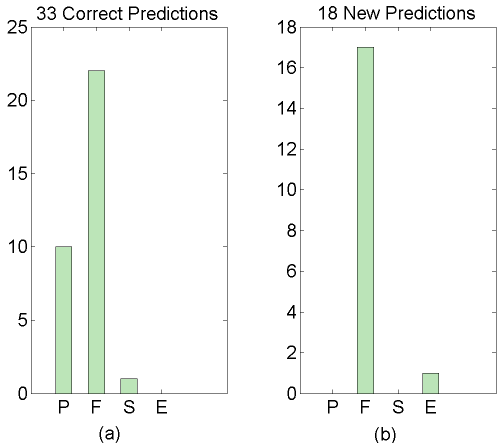

Figure 6.6: Distribution of highest scoring experts for the yeast pheromone response pathway validation. For definition of P,F,S,E experts, see the 'Feature Experts' section. (a) shows the frequency at which each of the four experts had the maximal score for the 33 known interacting pairs. (b) shows the frequency at which each of the four experts had the maximal score for the 18 new predictions.

To demonstrate the utility of this unique capability of the MFE method to reveal feature importance in specific predictions, we investigated a specific yeast pathway; the yeast pheromone response. For this pathway we compare the contribution of different experts in the known and predicted interacting pairs. Figure 6.5 presents the known interactions in this pathway as determined by the KEGG database [62]. In this pathway the yeast mating factors MAT alpha/a bind to their cognate membrane receptors Ste2/3, members of

the G protein coupled receptor family. Subsequent binding and activation of the G protein induces a MAP kinase signaling pathway via G protein activation [131].

We selected 25 proteins that are known to participate in this pathway and applied the MFE algorithm to classify the 300 (25*24/2) potentially interacting pairs. The training was built on the set including ∼500 positive pairs and ∼50000 negative random pairs. All of them have no relationship with the validated 25 proteins. The positive versus negative ratio in this set is roughly the same as the ratio we used in the above performance comparisons. The training set included 500 positive pairs and 50000 negative (random) pairs. None of these pairs contained any of the known 25 proteins in this pathway. The positive versus negative ratio in this set is roughly the same as the ratio we used for the performance comparisons. We determined a prediction threshold using the training set. 51 of the 300 pairs had scores above the threshold and were thus predicted to be interacting. Among them, 33 interactions (64.7%) had been experimentally validated. The remaining 18 pairs are new predictions.

Figure 6.6 shows the frequency at which each of the four experts showed maximal contributions among validated pairs. In line with biological intuition, the direct high-throughput evidence (expert P) and functional databases (expert F) are the predominant experts in the correct predictions. Figure 6.6 shows that the majority of the 18 new predictions are based on recommendations by expert F. Based on the reliability of expert F in making correct predictions, this result indicates that the majority of the new predictions may turn out to be correct, once experimentally tested.

6.5 Summary

One of the most important goals of computational PPI predictions is to suggest biological hypotheses regarding unexplored new interactions that are testable with subsequent experimentation. Among high scoring predictions, the most interesting ones can be chosen by an individual investigators using intuition and specialized knowledge.

This chapter addresses two important problems for the PPI prediction task. First, previous classification methods estimate a set of parameters that are used for all input pairs. However, the biological datasets used contain many missing values and highly correlated

features. Thus, different samples may benefit from using different feature sets. The second problem is that biologists who want to use these methods to select experiments cannot easily determine which of the features contributed to the resulting prediction. Since different researchers may have different opinions regarding the reliability of the various feature sources, it is useful if the method can indicate, for every pair, which feature contributes the most to the classification result.

In this work we propose a Mixture-of-Feature-Experts (MFE) approach to address the above two challenges when predicting protein-protein interactions. Diverse high-throughput biological datasets are split into homogeneous feature experts. Each expert uses a subset of the data to predict protein interactions and expert predictions are combined such that the weight of each expert depends on the input data for the predicted protein pair. This method is useful for overcoming problems in achieving high prediction performance due to missing values which are a major issue when analyzing biological datasets. In addition, the weights can be used by biologists to determine confidence in the prediction for each pair. We have shown that this algorithm improves upon previous methods suggested in yeast and human for this task. Extensions of this approach to other species are straight forward when more information becomes available.

We believe that as the prediction task becomes harder (for example, when analyzing interspecies protein interactions) the need for methods that can accommodate high levels of missing values and are directly interpretable by biologists increases. The next step will be to apply our method to interaction prediction tasks related to important types of human proteins where missing values and the small number of positive examples are major obstacles in obtaining an accurate protein interaction map.

Chapter 7

Protein Complex Identification by Supervised Graph Local Clustering

Now that we have obtained a good representation for the binary links in protein-protein interaction network from previous chapters, we would like to make use of this PPI graph for further studies. There exist many higher level patterns in these graphs as well. For example, protein complexes are important functional groups of protein interaction networks. In this chapter, we present an algorithm for inferring protein complexes from weighted interaction graphs in a supervised graph clustering style.

7.1 Introduction

Protein-protein interactions (PPI) are fundamental to the biological processes within a cell. Correctly identifying the interaction network among proteins in an organism is useful for deciphering the molecular mechanisms underlying given biological functions. Beyond individual interactions, there is a lot more systematic information contained in protein interaction graphs. Complex formation is one of the typical patterns in this graph and many cellular functions are performed by these complexes containing multiple protein interaction partners. As the number of species for which global high throughput protein interaction data is measured becomes larger [17, 18, 19, 20], methods for accurately identifying complexes from such data become a bottleneck for further analysis of the resulting interaction

graph.

High-throughput experimental approaches aiming to specifically determine the components of protein complexes on a proteome-wide scale suffer from high false positive and false negative rates [2]. In particular, mass spectrometry methods [22, 24] may miss complexes that are not present under the given conditions; tagging may disturb complex formation and weakly associated components may dissociate and escape detections. Therefore, accurately identifying protein complexes remains a challenge.

The logical connections between proteins in complexes can be best represented as a graph where the nodes correspond to proteins and the edges correspond to the interactions. Extracting the set of protein complexes from these graphs can help obtain insights into both the topological properties and functional organization of protein networks in cells. Previous attempts at automatic complex identification have mainly involved the use of binary protein-protein interaction graphs. Most methods utilized unsupervised graph clustering for this task by trying to discover densely connected subgraphs.

Automatic complex identification approaches can be divided into five categories and has been summarized in Section 3.2. These methods are based on the assumption that complexes form a clique in the interaction graph. While this is true for many complexes, there are many other topological structures that may represent a complex on a PPI graph. One example is a 'star' model, in which all vertices connect to a 'Bait' protein (termed 'spoke' model in [132]). Another possible topology is a structure that links several small densely connected components with loose linked edges. This topology is especially attractive for large complexes: due to spatial limitations, it is unlikely that all proteins in a large complex can interact with all others. See Figure 7.1 for some examples of real complexes with different topologies.

7.2 Methods

In this chapter we present a computational framework that can identify complexes without making strong assumptions about their topology. Instead of the 'cliqueness' assumption, we derive several properties from *known* complexes, and use these properties to search for new complexes. Since our method relies on real complexes, it does not assume any prior

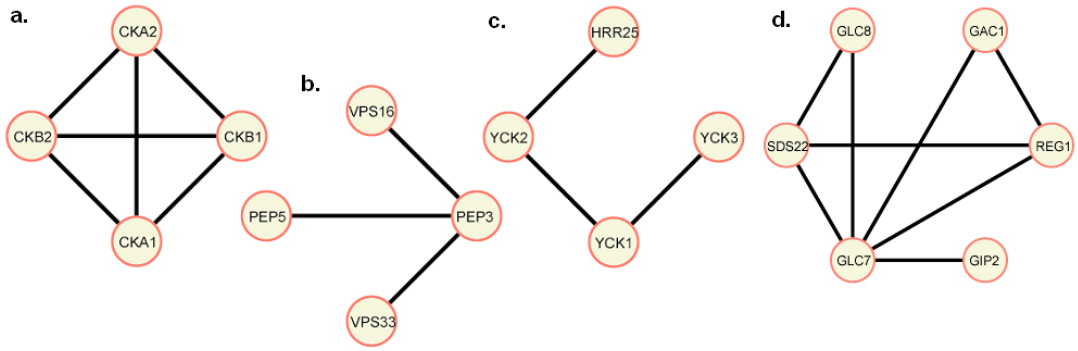

Figure 7.1: Projection of selected Yeast MIPS Complexes on our weighted PPI graph (weight thresholded). a. Example of a clique. All nodes are connected by edges. b. Example of a star-shape, also referred to as the spoke model. c. Example of a linear shape. d. Example for a hybrid shape where small cliques are connected by a common node.

model for complexes. Our algorithm is probabilistic. Following training to determine the importance of different properties, it can assign a score to any subgraph in the graph. By thresholding this likelihood ratio score we can label some of the subgraphs as complexes. Our model results in a significantly improved F1-score when compared to the density-based approaches. Using a cross validation analysis we show that the graphs discovered by our method highly coincide with complexes from the hand-curated MIPS database and a recent high confidence mass spectrometry dataset [23]. The top ranked new complexes are likely to provide novel hypotheses for the mechanism of action or definition of function of proteins within the predicted complex as we discuss in Results.

The main feature of our method is that it considers the possibility of multiple factors defining complexes in protein interaction graphs. Instead of assuming a specific topological model, we design a general framework which learns to weigh possible subgraph patterns based on the available known complexes.

Previous analysis of known PPI graphs has already revealed multiple shapes forming sub-graphs. For example, [132] proposed two topological models in the context of protein complexes. The first is the 'matrix model' which assumes that each of the members in the complex physically interact with all other members (leading to a clique-like structure). The second shape is the 'spoke model' that assumes that all proteins in a complex directly interact with one 'bait' protein leading to a star shape. Hybrids of these or other models are

also possible, resulting in more complex topologies.

Besides graph structures, there could be other features that characterize complexes. In particular, complexes have certain biological, chemical or physical properties that distinguish them from non-complexes. For example, the physical size of a complex may be an important feature. There is a physical limitation to creating large complexes because inner proteins become inacessible and therefore become more difficult to regulate. By incorporating such additional features into our supervised learning framework, the proposed model is able to integrate multiple evidence sources to identify new complexes in the PPI graph.

The input to our algorithm is a weighted graph of interacting proteins. The network is modeled as a graph, where vertexes represent proteins and edges represent interactions. Edge weight represents how likely is the interaction. Since the current data does not provide any directionality information, the PPI graph considered in this work is a weighted undirected graph. Our objective is to recover the protein complexes from this undirected protein-protein interaction (PPI) graph. Computationally speaking, complexes are one special kind of subgraphs on the PPI network. A *subgraph* represents a subset of nodes with a specific set of edges connecting them. The number of distinct subgraphs, or clusters, grows exponentially with the number of nodes.

7.2.1 Complex Features

Extracting appropriate features for subgraphs representing complexes is related to the problem of measuring the similarity between complex subgraphs. This task has been studied for other networks, specifically social networks [91, 81, 80]. In general, these previous approaches either (1) utilize properties of nodes or edges (indegree, outdegree, cliqueness, [133]) or (2) rely on comparing non-trivial substructures such as triangles or rectangles [134, 135]. We use both types to arrive at a list of properties for a feature vector that describes a subgraph in the PPI network. The properties include topological measurements about the subgraph structures and biological properties of the group of proteins in the subgraph.

Table 7.1 presents the set of features we use. We rely in part on prior work [35, 67, 80, 138, 20] to determine which features may be useful for this complex identification task.

Table 7.1: Features for representing protein complex properties. Each row represents a group of similar features. We use 33 features extracted divided into 10 groups. See supporting website for more details. The second column provides the name of the feature group and the third column provide a reference. The fourth column specifies which type of graph is used to derive the property.

No.	Group	Reference	Graph Type	Num. Features
1	Node Size	[80]	Binary	1
2	Graph Density	[80]	Binary	1
3	Degree Statistics	[136]	Binary	4
4	Edge Weight Statistics	[80]	Weight	4
5	Density wrt. Weight Cutoffs	[80]	Weight	7
6	Degree Correlation Statistics	[20]	Binary	3
7	Clustering Coefficient Statistics	[136]	Binary	3
8	Topological Coefficient Statistics	[20]	Binary	3
9	First Eigen Values	[80]	Binary	3
10	Protein Weight/Size Statistics	[137]		4

Each row in Table 7.1 represents one group of features. Totally 33 features were extracted from ten groups.

Below we briefly discuss each of the feature types used. The numbers match the numbers in Table 7.1.

1 Given a complex subgraph $G = (V, E)$, with $|V|$ vertexes and $|E|$ edges, the first property we considered is the number of nodes in the subgraph: $|V|$.

2 The density is defined as $|E|$ divided by the theoretical maximum number of possible edges $|E|_{max}$. Since we do not consider self interactions in the input weighted PPI graph, $|E|_{max} = |V| * (|V| - 1)/2$. As mentioned above, in the 'matrix' model the graph density is expected to be very high, whereas it may be lower for the 'spoke' shape.

3 This feature is calculated from the degree of nodes in the candidate subgraph. Degree is defined as the number of partners for a node. This group includes mean degree, degree variance, degree median and degree maximum.

4 The edge weight feature includes mean and variance of edge weights considering two different cases (with and without missing edges).

5 This group utilizes the densities under each case of weight cutoffs as the features. These features try to evaluate the possibility of topological changes under different weight cutoffs.

6 Degree correlation property measures the neighborhood connectivity of nodes within the subgraph. For each node it is defined as the average number of links of the nearest neighbors of the protein. We use mean, variance and maximum of this property in the feature set.

7 Clustering coefficient (CC) measures the number of triangles that go through nodes. For each node assuming it has q neighbors and there are t number of links connecting each other among the q partner nodes, thus $CC = 2t/q(q-1)$. We use mean, variance and maximum of this property in the group. 'Star' or 'linear' shapes achieve small values here while 'matrix' or 'hybrid' shapes get higher values relying on the proportions of within 'triangles'.

8 The topological coefficient (TC) is a relative measure of the extent to which a protein shares interaction partners with other proteins. It reflects the number of rectangles that pass through a node. For node p, we assume it has q partners, one of them is node s and there are $N(p,s)$ number of nodes to which both p and s are linked (plus 1 if there is a direct link between p and s). Thus $TC = average_q(N(p,s)/q)$. We use mean, variance and maximum of the property and the expected value for different shapes also varies depending on the ratio of rectangles within.

9 The first three largest singular values (SV) of the candidate subgraph's adjacency matrix. Different shapes have distinct value distributions with these three SV. For instance when comparing subgraphs with the same size, 'matrix' shape has higher value of the first SV than other shapes and 'Star' shape has lower value of the third SV (details in supporting web).

As for biological properties, we use average and maximum protein length and average and maximum protein weight of each subgraph. This feature is based on the intuition that

protein complexes are unlikely to grow indefinitely, because proteins within the center of large complexes become inaccessible to interactions with other putative partners.

Our framework of feature representation is general. It is straightforward to add other topological properties that are found to be relevant for this problem. It is also possible to add other types of features. For example information about the function of proteins can be encoded in our framework as well.

In addition, our idea of exploring subgraph patterns to model protein complexes (special groups within graph) has close connections with the *'exponential random graph model'* (ERGM [91]) which is popular in analyzing social networks for years. In ERGM, exponential models are used to represent the probability distributions of possible graphs (the whole graph with fixed node size). Extending ERGM is our framework is a possible direction to improve as well. [139] tried to use the global graph properties to model the graph probability distribution across temporal changes. This is very similar to our idea except that the subgraphs we tried to model is not require to be fixed node size.

7.2.2 Modeling Complexes with a Supervised Bayesian Network

We assume a generative probabilistic model for complexes. Figure 7.2 presents an overview of our model. Our method uses a Bayesian Network (BN) model. Features are generated, independently, based on two parameters: Whether the subgraph is a complex or not (C) and the number of nodes in the subgraph (N). The main reason we pay a special attention to N and do not model it as another complex property is because of the tendency of other properties to depend on N. For example, the larger the complex the more unlikely it is that all members will interact with each other (due to spatial constraints). Thus, the density property is directly related to the size. Similarly other properties such as 'mean of edge weight' and 'average clustering coefficient' depend on N as well. While it would have been useful to assume more dependency among other features as well, the more dependencies our model has the more data we need in order to estimate its parameters. We believe that the current model strikes a good balance between the need to encode feature dependencies and the available training data. Thus, other feature descriptors, $X_1 \ldots X_m$ are assumed to

be independent given the size and the label of the subgraph.

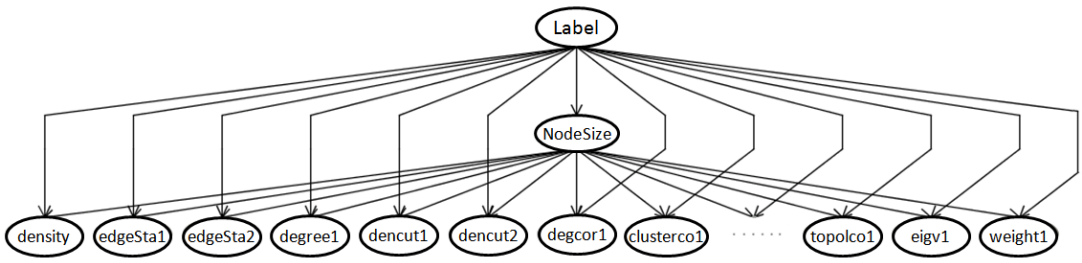

Figure 7.2: A Bayesian Probabilistic Model for representing a subgraph in our framework. The root node 'Label' is the binary indicator for complexes (1 if this graph is a complex, 0 otherwise). The second level node 'nodeSize' represents the number of nodes in the subgraph. The remaining nodes are all located in the third level and each represents a feature property described in Table 7.1.

For a subgraph in our PPI network we can compute the conditional probability of how likely it represents a complex using the following equation (4).

$$p(c_i = 1|n, x_1, x_2, ..., x_m) \tag{7.1}$$

$$= \frac{p(n, x_1, x_2, ..., x_m|c_i = 1)p(c_i = 1)}{p(n, x_1, x_2, ..., x_m)} \tag{7.2}$$

$$= \frac{p(x_1, x_2, ..., x_m|n, c_i = 1)p(n|c_i = 1)p(c_i = 1)}{p(n, x_1, x_2, ..., x_m)} \tag{7.3}$$

$$= \frac{\prod_{k=1}^{m} p(x_k|n, c_i = 1)p(n|c_i = 1)p(c_i = 1)}{p(n, x_1, x_2, ..., x_m)} \tag{7.4}$$

The second row uses Bayes rule. The third row utilizes the chain rule. The fourth equation uses the conditional independence encoded in our graphical model to decompose the probability to products of different features. Similarly, we can compute a posterior probability for a non complex by replacing 1 with 0 in the above equation.

Using these two posteriors we can compute a log likelihood ratio score for each candidate subgraph:

$$L = log\frac{p(c_1|n, x_1, x_2, ..., x_m)}{p(c_0|n, x_1, x_2, ..., x_m)} \tag{7.5}$$

$$= log\frac{p(n|c_1)p(c_1)\prod_{k=1}^{m} p(x_k|n, c_1)}{p(n|c_0)p(c_0)\prod_{k=1}^{m} p(x_k|n, c_0)} \tag{7.6}$$

Applying Bayes' rule and canceling common terms in the numerator and denominator, the only terms we need to compute for the likelihood ratio L are the prior probability $P(C_i)$ and the conditional probabilities $P(N|C)$ and $P(X_k|N, C_i)$.

Maximum likelihood estimation is used for learning these conditional dependencies from training data. We first discretized the continuous features and then used the multinomial distribution to model their probabilities (we chose to uniformly discretize each features into 10 equal width bins in the experiments presented in Results). Due to the small sample size of the training data, we apply a Bayesian Beta Prior to smooth the multinomial parameters in extreme cases [140]. As for the prior CPD $p(C = 1)$ of complexes, we assign a default value of 0.0001 which leads to good performance in cross validation experiments.

The BN structure in Figure 7.2 was manually selected. We have also tried to learn the BN structure using tree augmented structure learning techniques [129]. However, the resulting performance of the learned network is not significantly better than our proposed structure (Figure 7.2). Since our structure is simpler we omit the related results here. However potential improvements may be possible with more training examples and better BN structure learning approaches.

7.2.3 Searching for New Complexes

The above model can be used to evaluate candidate subgraphs. If the log likelihood ratio exceeds a certain threshold the subgraph is predicted to be a complex. This reduces the problem of identifying proteins complexes to the problem of searching for high scoring subgraphs in our PPI network. However, as we prove in the following lemma this problem is NP-hard.

Lemma 7.2.1. *Identifying the set of maximally scoring subgraph in our PPI graph is NP-hard*

Table 7.2: Local search for protein complex identification.

Choices for how to start:
- Interesting selected nodes;
- Top degree nodes;
- All related nodes ordered by degree;
Expand current cluster:
- Generate a subset V* from all neighbors of current cluster;
- Choose top rank M nodes as candidate to expand the current cluster.
The order of adding is based on the maximum similarity weight to the current cluster;
- Choose the node who achieves the best new cluster score among M candidates;
Search:
- Accept the new cluster candidate if with higher score;
- If lower, accept with probability $exp(l' - l)/T$;
- Temperature parameter T decreasing by a scaling factor alpha after each round;
- Accepted cluster must score higher than a threshold;
When to stop:
- Current cluster has no neighbor nodes to expand;
- Number of rounds since the last score improvements is larger than a specified number;
- The number of expanding rounds is larger a the specified number;

Proof. We prove this by reducing our search problem to max-clique, a NP hard problem [141]. To reduce our model to max-clique we will assume that we are only using one property, the graph density and that all edges in our graph have a weight of 1. Further, we set the probability of a complex given a subgraph to:

$$p(C|N, X) = N/N + 1 \quad if \quad X = 1$$

$$p(C|N, X) = 0 \quad if \quad X < 1$$

For this model, the only subgraphs with positive scores are the cliques in our graph. In addition, the bigger the clique the higher our score and so finding the highest scoring subgraph is equivalent to finding the maximal clique. □

Based on the above lemma, efficient and scalable heuristics algorithms are needed. Protein complexes are expected to be clustered locally within the PPI graph. Thus, we

turn to heuristic local search methods. There are many approaches for local graph search proposed in the literature, which include hill climbing, simulated annealing, heuristic based greedy search, or tabu-search heuristic [81]. All these strategies try to find local optima for certain fitness functions.

Table 7.3: Protein complex identification algorithm.

Input:
- Weighted protein-protein interaction matrix;
- A training set of complexes and non-complexes;

Output:
- Discovered list of protein complexes;

Complex Model Parameter Estimation:
- Extract property features from positive and negative training examples.
- Discretize the continuous features.
- Calculate the BN MLE parameters for different features properties on the multinomial distribution.

Search for Complexes:
- Starting from the seeding subgraphs, apply simulated annealing search to expand and identify candidate complexes;
- Output subgraphs with ratio scores exceeding a certain threshold;

Here we choose to employ the iterated simulated annealing (ISA) [81, 142] search, using complex ratio score as the objective function (see Equation (7.6)). The basic idea for ISA is: after each round of modifying the current cluster, we accept the new cluster candidate if it has higher score L' than the current score L, but even if the score decreases, we accept the new cluster with probability $exp((L - L')/T)$, where T is the temperature of the system. This allows the algorithm to avoid local minima in some cases. After each round, the temperature is decreased by a scaling factor α by setting $T' = \alpha T$. The initial temperature T_0, the scaling factor α, and the number of rounds are parameters of the search process. After the algorithm terminates the highest scoring subgraph is returned and the search continues. [142] pointed out that given suitable parameter setting, ISA could identify the global optimum though this setting is generally unknown and can be impractically hard

to find.

At the beginning, we connect each seeding protein to its highest weight neighbor and then use them as the starting cluster. Beginning from these clusters, we pursue the cluster modification process and the simulated annealing search. A number of heuristics could be used for modifying the current cluster. The order in which we add new proteins to the cluster is based on their impact on the cluster ratio score. We also explore the option of removing nodes from the cluster and merging of two clusters. We chose to limit the search rounds number to 20 which means the size of the complexes we search for is between 3 to 20. We use cross validation to choose best values for the temperature and scaling factor parameters. To avoid we visit the same/similar clusters studied before, we keep checking the overlapping ratio between the current cluster to the investigated clusters so far. If the ratio is higher than a threshold, we would just stop the searching for current seed.

For the searching step, we could make an estimate about the complexity. Assuming that the graph we searched on have n nodes and e edges. The maximum size of subgraphs we detected is p (chose as 20 in our evaluations). Thus, for each candidate subgraph, we need to perform the feature extractions which would cost $O(p^3)$. When searching from a seeding node, we assume the average degree of nodes in the graph as q (in sparse graphs, $q << n$) and the cluster expansion in each round is constrained to m (set as 20 in the evaluations) maximum weight nodes. Then finding the related complex for the node would be cost $O(q * p * m * p^3)$. If aiming to find all the complexes in a graph, we start from all the related n nodes, totally the algorithm cost $O(n * q * m * p^4)$. Thus, if $(n >> p^4)$, the search is polynomial to n. If not, the complexity is largely controlled by p.

The complete proposed algorithm for complex identification is presented in Table 7.2.3. Our input is the weighted PPI graph and a set of known complexes and non-complexes (random collections of genes) as training data. First, we learn model parameters for the probabilistic BN model from the training data. Next, we search for subgraphs to identify candidate complexes. The final output clusters are those clusters found to have ratio score larger than a predefined threshold.

7.2.4 Weighted Undirected PPI Graph

As discussed above, we assume that our model input is a weighted undirected graph representing PPI network. The edge weight describes how likely an interaction happens between the two related proteins. As we presented in the previous chapters indirect data can be combined with the direct interaction data to improve the accuracy of protein interaction prediction. This type of analysis usually results in an interaction probability or confidence score assigned to each protein pair. Edges in our graph are weighted using this interaction probability which is computed as follows. In previous work [4], we assembled a large set of biological features (a total of 162 features representing 17 distinct groups of biological data sources) for the task of pairwise protein interaction prediction in Yeast. Considering our current goal we remove the features derived from the two high throughput mass spectrometry data sets [22, 24]. Training is based on the small scale physical PPI data in the DIP database [60]. Based on our previous evaluation, the support vector machine (SVM) [128] classifier performs as well or better than any of the other classifiers suggested for this physical interaction task. We have thus used the results of our SVM analysis (see details in [4].) to obtain weights for edges in our graph. Weights range from minus infinity to infinity where larger values indicate a higher likelihood to be an interacting pair. To reduce the number of edges in our graph we apply a cutoff and remove all edges with weights below the cutoff. We have chosen the cutoff (1.0) such that the number of remaining edges roughly corresponds to previous estimates of the number of protein interaction pairs in yeast [2].

To further improve the quality of the PPI graph we filter the predicted weighted graph using a newly published Yeast interaction data set from [28]. For each of the remaining interactions we keep the weight learned from our integrated data analysis. This data contains a comprehensive database of genetic and protein interactions for the budding yeast Saccharomyces cerevisiae, manually curated from over 31,793 abstracts and online publications. A total of 35,244 interactions are reported, including literature curated and high throughput interactions. To allow fair comparisons we removed those interactions coming from the high-throughput mass spec experiments in this data set.

7.3 Experiments and Results

7.3.1 Reference Sets

The MIPS [26] protein complex catalog is a curated set of 260 protein complexes for yeast that was complied from the literature and is thus more accurate than large scale mass spectrometry complex data. After filtering away those complexes composed of a single or a pair of proteins we were left with 101 complexes in MIPS. The size of the complexes in MIPS is distributed as a power law, with most of the complexes having fewer than five proteins. We use the projection of the MIPS complexes on our PPI graphs as the positive training examples. See Figure 7.1 for four examples of such a projection.

Another independent positive set we used is the set of protein complexes from a newly published TAP-MS experiment [23], one of the most comprehensive genome-wide screens for complexes in budding yeast. Again, we filtered those complexes with only two proteins leading to 152 complexes that were used as a positive examples set to test our method.

Since we are using a supervised learning method we also need negative training data. In our case we generated by randomly selecting nodes in the graph. The size distribution of these non-complexes follows the same power law distribution of the known complexes in MIPS. Figure 7.3(a) presents the histogram of these distributions for each of the three reference sets: 'MIPS', 'TAP06' and 'Non-complexes'. As can be seen, all roughly follow the similar 'power law' distributions.

Figure 7.4 presents the distribution of two classes for real complexes (blue) versus negative examples (red) when projected them on the first three principal coordinates after applying SVD on features. The distribution strongly indicates that the proposed features can separate the two sets reasonably.

7.3.2 Evaluation Measures

In order to quantify the success of different methods in recovering the set of known complexes we define three sets for each *pair* of a known and predicted complex:

- A: Number of proteins only in the predicted complex

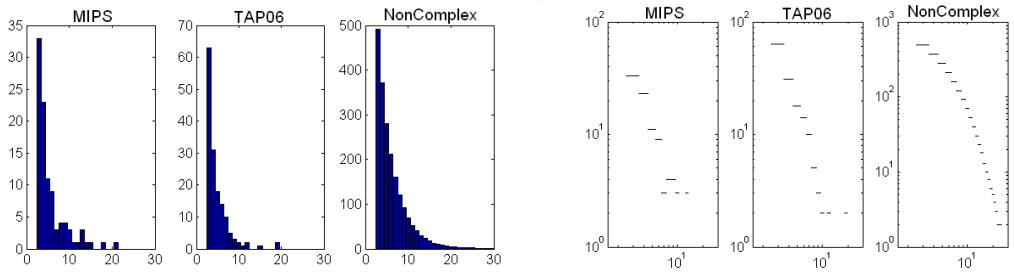

Figure 7.3: Data Distribution of Reference Sets. (a). Left subfigure: Histogram of number of proteins containing in each of the three reference sets: 'MIPS', 'TAP06' and 'Non-complexes'. All roughly obey the 'power law' distributions. Horizontal axis means the number of proteins. Vertical axis mean the number of subgraphs. (We set horizontal axis ends at 30 for all three subfigures.) (b). Right subfigure: Log-log plot of the left subfigure.

- B: Number of proteins only in the known complex
- C: Number of proteins in the overlap complex

We say that a predicted complex recovers a known complex if

$$\frac{C}{A+C} > p \quad and \quad \frac{C}{B+C} > p \tag{7.7}$$

where p is an input parameter between 0 and 1 which we set to 0.5. Thus we require that both, the majority of the proteins in the complex be recovered and that the majority of the protein in the predicted complex belong to that known complex.

Based on the above definition, three evaluation criteria are applied to quantify the quality of different protein complex identification methods:

- Recall: Measures the fraction of known complexes detected by predicted complexes, divided by the total number of positive examples in the test set.
- Precision: Measures the fraction of the predicted complexes that match the positive complexes among all predicted complexes.
- F1: The F1 score combines the precision and recall scores. It is defined as $2pr/(p+r)$.

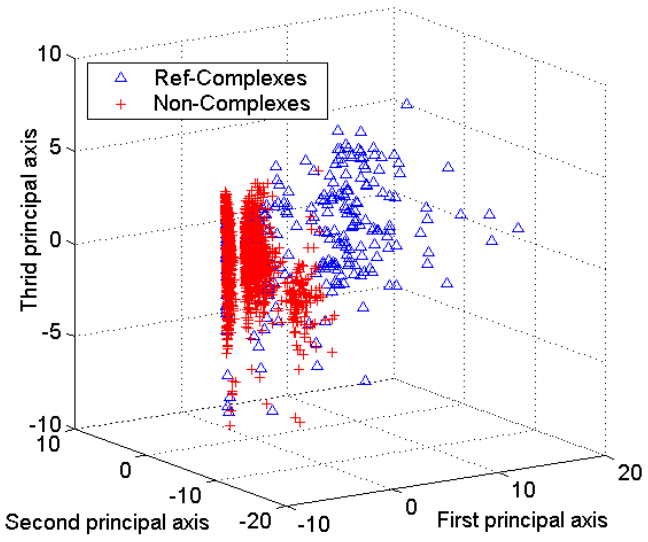

Figure 7.4: Reference examples' distribution when projected with the first three principle components after applying SVD to the features.

All three values range from 0 to 1, with 1 being the best score.

Recall quantifies the extent to which a solution set captures the labeled examples. Precision measures the accuracy of the solution set. A good protein complex detector should have both high precision and high recall. The F1 measure provides a reasonable combination for both precision and recall. These three criterions are frequently used in many computational areas [143].

7.3.3 Performance Comparison

To assess the performance of the complex identification, we conduct experiments using MIPS as positive training set and TAP06 as a test set and vice versa. There are a total of 1376 proteins in the MIPS and TAP06 complexes. Thus, we applied our train-test analysis on a PPI graph containing theses genes. The resulting graph used contains 1376 proteins and 10,918 weighted edges.

We have compared our method to three other methods suggested for complex idnetification. We term our proposed approach as

- The 'MCODE' complex detection method proposed in [67]. MCODE finds clusters

(highly interconnected regions) in any network loaded into Cytoscape. The method was developed for protein-protein interaction networks in which these clusters correspond to protein complexes [67].

- Clique (density based) methods. For this we use the same search algorithm discussed in Methods. However, unlike our method which maximizes the BN likelihood ratio, for Clique we simply try to find the maximal scoring cliques in the graph.

- Supervised learning using SVMs. This method is used to determine whether the BN structure helps in identifying complexes. It uses the same features as our method but instead of using a BN it uses a SVM [128].

The performance comparison is presented in Table 7.4. For each method, we report the precision, recall and F1, separately. As can be seen our method dominates all other methods in both precision and recall (and, of course, in F1 scores). The recall rate of our method is around 50%. This number is impressive when considering the fact that the training and testing were done on different datasets. Our precision is lower (between 20-30%). However, since many of the complexes are not included in either gold standard sets, this precision value can be the result of correct predictions that are not included in the available data. We discuss some of these complexes below. As for the other methods, surprisingly, the recall and F1 values reported by MCODE are much lower than both the 'Density' and 'SCI-SVM' methods. We investigated the clusters identified by 'MCODE' and determined that they were relatively large compared to clusters determined by other methods which may have hurt performance. Interestingly the performance of 'SCI-SVM' is not as good as 'SCI-BN'. This is largely caused by the unique way BN can handle the 'node size' feature. For the 'Density' approach, it performs reasonably well for the Recall measure but not as good in terms of precision.

7.4 Validation

Using a threshold of 1.0 for the weights of the edges, our yeast PPI network contains 5234 proteins and 19246 interaction edges linking them. To identify and validate new

Table 7.4: Performance comparison between our algorithm ("SCI-BN"), SVM with the same set of features ("SCI-SVM"), Clique based method using only the density feature ("Density") and the 'MCODE' methods [67] ("MCODE"). Evaluation is based on precision, recall and F1 measure. Experiments carried out with either MIPS as positive training set and TAP06 as test set, or vice versa.

Train	**Test**	**Method**	**Precision**	**Recall**	**F1**
MIPS	TAP06	Density	0.217	0.409	0.283
MIPS	TAP06	MCODE	0.293	0.088	0.135
MIPS	TAP06	SCI-SVM	0.247	0.377	0.298
MIPS	TAP06	SCI-BN	0.312	0.489	0.381
TAP06	MIPS	Density	0.143	0.515	0.224
TAP06	MIPS	MCODE	0.146	0.063	0.088
TAP06	MIPS	SCI-SVM	0.176	0.379	0.240
TAP06	MIPS	SCI-BN	0.219	0.537	0.312

complexes within this network graph, we trained a new BN model on all of the MIPS manual complexes as positive examples and used 2000 randomly selected non-complexes subgraphs as negative examples. Within the resulting full graph, we predict 987 complexes using the 'SCI-BN' search method.

To identify new complexes within the predicted graph, we compared the predicted clusters with those reported in five reference datasets, the manually curated MIPS dataset [[26]] and four large-scale complex datasets obtained using high-throughput experimental approaches [22, 24, 23, 144]. After filtering those clusters matching reference complexes, we are left with 570 novel predictions. These are either entirely new complexes or extensions to known complexes by adding new proteins.

We analyzed examples of new complexes determined by our algorithm and found out that they are highly likely to be true complexes (see details in [11]). Amongst the new complexes, most highly ranked were of size 3-4. The size distribution agrees with the distribution of known complexes. While many of these top scoring complexes took the shape of cliques, others displayed more diverse shapes. Examples are shown in Figure 7.5. Black edges in Figure 5 represent interactions with SVM score higher than 4.0 (indicating strong evidence for interactions between proteins). Biological analysis for each detected

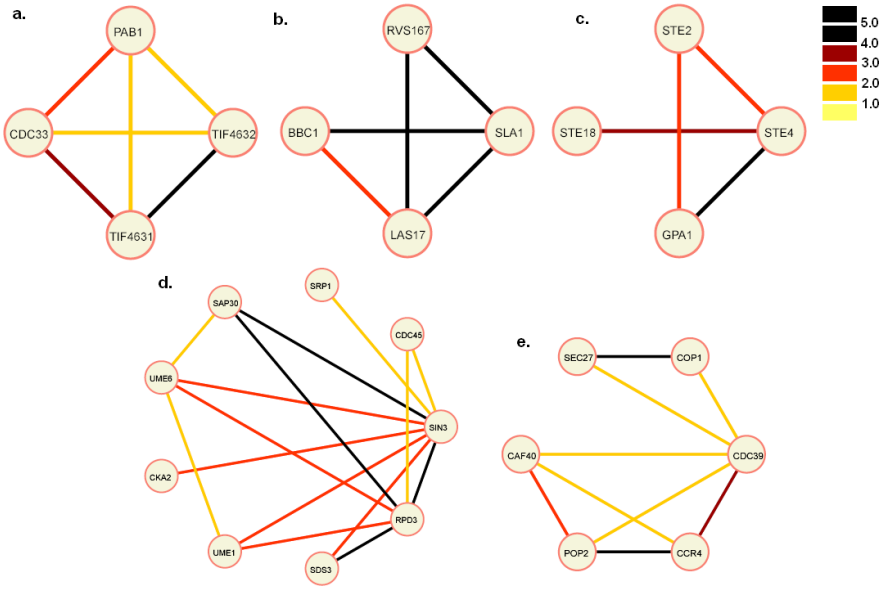

Figure 7.5: Projection of predicted complexes on our weighted PPI graph. The edge weights are thresholded and color coded. See color legend (top right corner bar) for edge weights. Descriptions for each predicted complex are provided in the 'Validation' section.

complex in Figure 7.5 are described in [11].

7.5 Summary

In this chapter we presented a probabilistic algorithm for discovering complexes in a supervised manner. Specifically we extract features that can be used to distinguish complex versus non complexes and train a classifier using these features to identify new complexes in the PPI graph. Unlike previous methods that relied on the 'dense' assumption of complex subgraphs, our algorithm integrates subgraph topologies and biological evidence, and learns the importance of each of the features from known complexes. This allows our algorithm to identify complexes with topologies that are missed by previous methods. We have shown that our algorithm can achieve better precision and recall rates for previously identified complexes. We also investigated examples of new complexes determined by our algorithm and discussed their possible biological function.

Our framework of feature representation is general. It is straightforward to add other topological properties that are found to be relevant for this problem. It is also possible to add other types of features. For example information about the function of proteins can be encoded in our framework as well. It is feasible to add structure or sequence information as well. For example information about structures can be encoded as certain statistical values describing how the structures of the proteins match to other members in the same subgraph from the perspective of binding interfaces.

We hope to extend this work and improve both feature representation and search so that we can detect other types of interactions. Besides complexes, pathways of logically connected proteins also play a major role in both cellular metabolism and signaling. How to detect interesting pathways on PPI graphs in our framework is an interesting direction to pursue. Another interesting direction is to apply this method to other species for which protein interaction data became available recently, including humans.

Chapter 8

Conclusions and Future Directions

In the previous chapters, we provide a set of computational tools that enables researchers to predict various aspects of protein-protein interaction graphs. The proposed methods and the biological results obtained provide a better understanding of current PPI networks in yeast and in human.

This chapter summarizes both the proposed computational approaches and the biological impact we obtained, and points out potential directions for future studies.

8.1 Learning of Protein Interaction Networks

Protein-protein interactions operate at every level of cellular function. Comprehensively identifying these interactions is important for systematically defining the roles played by cellular proteins for biological functions. Large-scale biological PPI experiments can directly detect hundreds or thousands of protein interactions at a time. However the resulting data sets are often incomplete and exhibit high false positive and false negative rates.

Computationally, a protein-protein interaction (PPI) network could be conveniently modeled as an undirected graph, where the nodes are proteins and two nodes are connected by an undirected edge if the corresponding proteins physically bind. However, currently this type of graph contains many noisy edges and a large portion of edges is missing for most organisms. The general goal of this dissertation is to study this graph as completely as possible, in terms of both the set of interacting pairs and important biological substructures.

Using a systematic comparison of previous approaches for PPI prediction using information integration, we found the many factors affect the final prediction performance, including the utility of different information, the way the data is encoded as features, the target types of protein interactions and computational approaches. Our study showed the importance of taking these controlling factors into account when designing new algorithms for this task. Working from various perspectives we then proposed four algorithms for the learning of yeast and human PPI networks.

(I) We have proposed a combined computational and experimental approach to predict interaction partners for human membrane receptors. The method integrates highly informative direct and indirect biological evidences to decide if a candidate receptor-human pair interacts or not. The resulting receptor PPI network is then analyzed through graph analysis at a global level. Several novel predictions have been further experimentally validated.

(II) Considering the fact that there is no negative reference set available in the above classification setting, we design a ranking approach to identify candidate interaction pairs that are ”similar” to known interacting pairs. We first use direct and indirect information relating to protein interactions to determine a robust similarity measure between protein pairs. Then this similarity is used in a weighted k-Nearest-Neighbor algorithm to rank potential protein pairs. The resulting performance on yeast indicates the feasibility and utility of the proposed similarity estimates.

(III) Taking into account the heterogeneous feature property, a multiple-view learning strategy has been designed for the PPI prediction task. Features are split into roughly homogeneous groups and each group functions as a PPI predictor (expert). The individual experts use logistic regression and their scores are combined using another logistic regression. When combining the scores the weighting of each expert depends on the set of input attributes available for that pair. The method improves upon previous methods for this task. In addition, the weighting of the experts provides means to evaluate the prediction based on the high scoring feature experts.

Though above three approaches were all presented for the purpose of predicting PPIs, they are appropriate for different scenarios. The first combined framework aims to find

partners of the human membrane receptors and is proper for providing interesting hypotheses towards biological validations. The ranking approach fits well for the case that known positive PPI pairs are few and the ranking of the predictions is important. The multiple view learning method would be recommended for the situation with many heterogenous feature groups.

(IV) Complex formation is one typical group pattern in protein interaction networks. We present a novel algorithm for inferring protein complexes from weighted interaction graphs in a supervised style. By using graph topological patterns and biological properties as features, we model each complex subgraph by a probabilistic Bayesian Network (BN). We apply our method to protein interaction data in yeast. Our algorithm achieves a considerable improvement over clique based algorithms.

In summary, our proposed approaches provide strong computational tools for recovering and analyzing protein-protein interaction networks. They have been applied and generated promising results in multiple organisms.

8.2 Future Research Directions

Computational learning of protein interaction networks is still a relatively new research domain. Though several sub-problems have been studied by the community for a while, many important questions remain. Figure 8.1 describes five major challenges related to learning of protein-protein interaction networks. Three of them (a-c) have been covered in this dissertation. The remaining two challenges (d-e) and related extensions of the three covered challenges are discussed below.

Active PPI Predictions Lab experiments for validating protein-protein interactions are expensive and time-consuming. Currently the reliable PPIs are small in number, and this size limitation greatly reduces the predictive power of the computational algorithms [122]. Considering the urgent need to recover all the protein-interactions in networks (Figure 8.1a), it would be useful to design and validate new computational strategies for choosing the most informative sets of protein pairs for lab experiments. Computational active

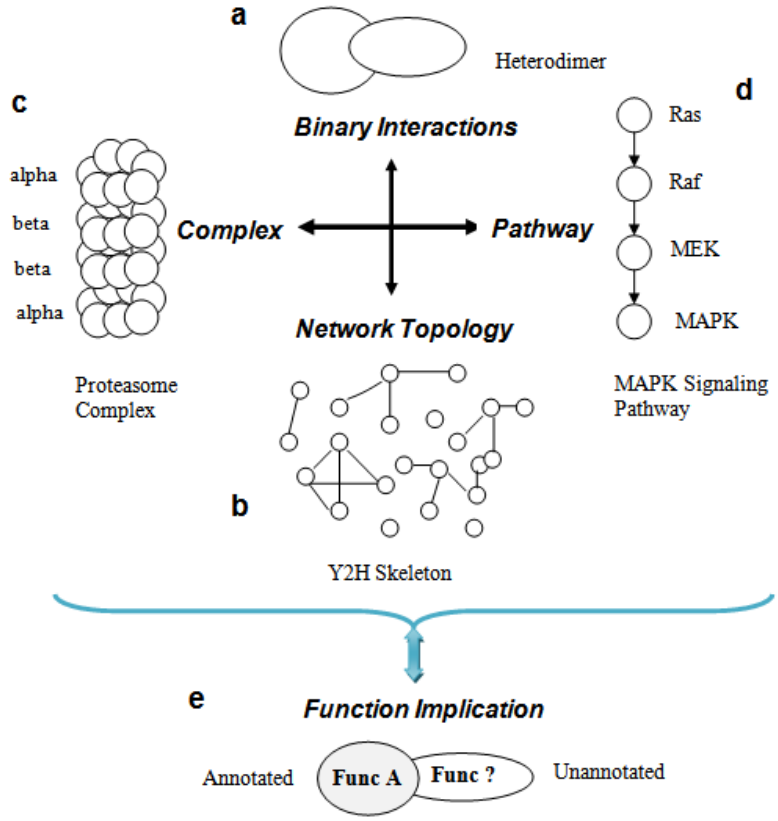

Figure 8.1: Overview of computational challenges in learning of protein interaction networks. (Modified from Figure 1 of [1], included as background information only.) (a) Recovering pairwise interaction edges in the network is the first problem to solve. (b) It is important to investigate global properties of PPI networks. (c) Identification of protein complexes is critical in understanding the cell. (d) PPI networks form a backbone of signaling pathways and metabolic pathways and underly all cellular processes required for normal cell functions. Defining and modeling theses pathway is essential, albeit challenging, to understand the cell. (e) Implications about protein functions can be made based on protein-protein interaction relationships. Making predictions regarding protein function is one of the most important tasks in current computational biology.

learning strategies typically make use of a few labeled instances and a large number of unlabeled examples, and must select particular instances for labels in order to optimize learning. This seems to fit well to the active PPI learning scenario. Unlike the general learning task, actively building an informative candidate set for PPI predictions is not just

a computational issue. Biological importance, experimental conditions and feature availability/importance should be considered in the design as well.

More PPI Predictions Currently it is of particular interests to study the protein interactions between organisms, like protein interactions between virus and host [7]. Extending the proposed framework to this type of interaction-prediction tasks would be interesting and significant. Moreover the proposed methods might be useful for PPI predictions in model organisms other than yeast and human. For these extended applications, computational difficulties, including the noisy/missing feature problem, the heterogeneous attribute properties and the lack of negative reference sets, all exist. How to develop robust learning methods towards reliable identifications of PPIs is still computationally challenging (Figure 8.1a).

Pathway Detection The PPI network forms a backbone of signaling pathways, metabolic pathways and cellular processes required for normal cell functions [1]. Pathways in cellular network graphs can represent a transformation path (or chain structure) from a nutrient to an end-product in a metabolic network, or a chain of post-translational modifications from the sensing of a signal to its intended target in a signal transduction network [145]. Complete knowledge of these pathways will help in the understanding of the cell's normal processes, as well as how diseases develop from mutation of individual pathway components [145, 1]. Computationally speaking, pathway structures (Figure 8.1d) correspond to linear paths or similar structures in the protein interaction networks [146]. However the relatively high degree of noise in the PPI graphs makes pathway modeling very challenging. The question of how to integrate prior biological knowledge or other data evidence to make the pathway inference more robust and efficient is an important and interesting computational task.

Domain/Motif Interaction Current protein interaction graphs do not directly reveal where two proteins interact [41] on their protein chains. Protein interactions occur through physical binding of small regions on the surface of proteins. Insights into the mechanisms with which different proteins fulfill their roles could be obtained by understanding the interaction sites where protein binding takes place. Moreover, a detailed understanding of the

binding sites at which an interaction takes place can provide both scientific insight into the causes of human disease as well as a starting point for drug design [41]. Computationally representing the binding site is a hard question (an extended problem of Figure 8.1a). Various sizes of local units have been previously proposed [147, 43], including contact patches, motifs and domains. Most proteins contain multiple possible positions (motifs or contact interfaces) for binding. Thus inferring interactions between domains or motifs from protein-protein interactions would be a challenging and interesting task [39, 42, 43, 148].

Protein Function Predictions Protein-protein interactions directly contribute to protein functions and function prediction is one of the major challenges in computational biology today (Figure 8.1e). The biological functions are still unknown for a large proportion of sequenced proteins [149]. Furthermore a given protein may have more than one function, so many proteins that are known to belong to some class may have as yet undiscovered functionalities. Implications about function can be made via protein-protein interactions based on the premise that the function of unknown proteins may be discovered if captured through their interaction with a known protein target having known function(s) [13, 14]. Besides protein interaction evidence, the function of an unannotated protein can be predicted based on various other data sets, including sequence homology, phylogenetic profile, gene expression and so on. Combining multiple data sources together for protein function prediction is an interesting computational problem [150, 151, 152, 153, 154, 155, 156]. We can also exploit the fact that protein function labels often exhibit a certain hierarchical structure. For instance, the Gene Ontology (GO) [99] representation entails a directed acyclic graph (DAG) taxonomy for functional annotations. To achieve a systematical identification of protein functions, the taxonomy of protein function labels and the correlations between different functions should be considered in the the computational design. Thus, how to effectively incorporate heterogeneous information sources, and at the same time to consider multiple function labels in the taxonomy structure, remains a challenging task.

Bibliography

[1] Ghavidel A, Cagney G, Emili A: A skeleton of the human protein interactome. *Cell* 2005, **122**(6):957–68.

[2] von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P: Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 2002, **417**(6887):399–403.

[3] Futschik ME, Chaurasia G, Herzel H: Comparison of human protein-protein interaction maps. *Bioinformatics* 2007, **23**:605–611.

[4] Qi Y, Bar-Joseph Z, Klein-Seetharaman J: Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *PROTEINS: Structure, Function, and Bioinformatics.* 2006, **63**(3):490–500.

[5] Qi Y, Dhiman H, et al, Bar-Joseph Z, Klein-Seetharaman J: The human membrane receptor interactome. *(In Review)* 2008.

[6] Nozawa H, et al, Qi Y, Klein-Seetharaman J, et al, Thomas S: Combined inhibition of plc-gamma-1 and c-src abrogates epidermal growth factor receptor-mediated head and neck squamous cell carcinoma invasion. *Clinical Cancer Research (In Press)* 2008.

[7] Tastan O, Qi Y, Carbonell J, Klein-Seetharaman J: Prediction of interactions between HIV-1 and human proteins. *(In Review)* 2008.

[8] Qi Y, Klein-Seetharaman J, Bar-Joseph Z: Random forest similarity for protein-protein interaction prediction from multiple sources. *Pacific Symposium on Biocomputing* 2005, **10**:531–542.

[9] Qi Y, Klein-Seetharaman J, Bar-Joseph Z: A mixture of experts approach for protein-protein interaction prediction. *Proceedings of Neural Information Processing Systems (NIPS): The workshop on Computational Biology and the Analysis of Heterogeneous Data.* 2005.

[10] Qi Y, Klein-Seetharaman J, Bar-Joseph Z: A mixture of feature experts approach for protein-protein interaction prediction. *BMC Bioinformatics* 2007, **8**(S10):S6.

[11] Qi Y, Balem F, Faloutsos C, Klein-Seetharaman J, Bar-Joseph Z: Protein complex identification by supervised graph clustering. *The 16th Annual International Conference Intelligent Systems for Molecular Biology (ISMB), (In Press)* 2008.

[12] Alberts, et al: *Molecular Biology of the Cell (4th Edition).*

[13] Phizicky E, Fields S: Protein-protein interactions: methods for detection and analysis. *Microbiol Rev.* 1995, **59**:94–123.

[14] Pierce biotechnology, http://www.piercenet.com/proteomics/, 2006.

[15] Royer C: Protein-protein interactions. *Outline of the thermodynamic and structural principles governing the ways that proteins interact with other proteins. Previously published in the Biophysics Textbook Online (BTOL).* 1999.

[16] Shoemaker BA, Panchenko AR: Deciphering protein-protein interactions. part i. experimental techniques and databases. *PLoS Comput Biol* 2007, **3**(3):e42.

[17] Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y: A comprehensive two-hybrid analysis to explore the yeast proteininteractome. *Proc. Natl. Acad. Sci. USA* 2001, **98**(8):4569–4574.

[18] Uetz P, Giot L, Cagney G, Mansfield TA, et al, Rothberg JM: A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* 2000, **403**:623–627.

[19] Rual JF, Venkatesan K, et al, Roth FP, Vidal M: Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 2005, **437**(7062):1173–8.

[20] Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck F, et al, Wanker E: A human protein-protein interaction network: A resource for annotating the proteome. *Cell* 2005, **122**(6):830–2.

[21] Rual JF, Venkatesan K, et al, Roth FP, Vidal M: Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 2005, **437**(7062):1173–8.

[22] Gavin AC, Bosche M, Krause R, et al, Superti-Furga G: Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 2002, **415**(6868):141–7.

[23] Gavin A, Aloy P, Grandi P, et al, Superti-Furga G: Proteome survey reveals modularity of the yeast cell machinery. *Nature* 2006, **440**(7084):631–6.

[24] Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, et al, Tyers M: Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* 2002, **415**(6868):180–3.

[25] Shoemaker BA, Panchenko AR: Deciphering protein-protein interactions. part ii. computational methods to predict protein and domain interaction partners. *PLoS Comput Biol* 2007, **3**(4):e43.

[26] Mewes H, Amid C, Arnold R, Frishman D, Guldener U, et al: MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res.* 2004, **32**(Database issue):D41–4.

[27] Bader GD, Donaldson I, Wolting C, Ouellette BF, Pawson T, Hogue CW: BIND -the biomolecular interaction network database. *Nucleic Acids Res* 2001, **29**:242–5.

[28] Reguly T, Breitkreutz A, Boucher L, et al, Tyers M: Comprehensive curation and analysis of global interaction networks in saccharomyces cerevisiae. *Journal of Biology* 2006, **5**(4):11.

[29] Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, et al, Prasad TS, Pandey A: Human protein reference database–2006 update. *Nucleic Acids Res* 2006, **34**(Database issue):D411–4.

[30] Peri S, Navarro JD, Kristiansen TZ, et al, Pandey A: Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res* 2004, **32**(Database issue):D497–501.

[31] Ramani AK, Bunescu RC, Mooney RJ, Marcotte EM: Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biol.* 2005, **6**(5):R40.

[32] Lehner B, Fraser AG: A first-draft human protein-interaction map. *Genome Biol.* 2004, **5**(R63):R63.

[33] Brown KR, Jurisica I: Online predicted human interaction database. *Bioinformatics.* 2005, **21**(9):2076–82.

[34] Stelzl U, Worm U, Lalowski M, et al, Wanker EE: A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 2005, **122**(6):957–68.

[35] Barabasi A, Oltvai Z: Network biology: understanding the cell's functional organization. *Nat Rev Genet.* 2004, **5**:101–113.

[36] Bar-Joseph Z, Gerber G, Lee T, Gifford D, et al: Computational discovery of gene modules and regulatory networks. *Nature Biotechnology* 2003, **21**(11):1337–42.

[37] Caffrey D, Somaroo S, Hughes J, Mintseris J, Huang E: Are protein protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Science* 2003, **13**:190–202.

[38] Gomez S, Noble W, Rzhetsky A: Learning to predict protein-protein interactions from protein sequences. *Bioinformatics* 2003, **19**(15):1875–81.

[39] Deng M, Mehta S, Sun F, Chen T: Inferring domain-domain interactions from protein-protein interactions. *Genome Res.* 2002, **12**(10):1540–8.

[40] Wang H, Segal E, Ben-Hur A, Koller D, Brutlag DL: Identifying protein-protein interaction sites on a genome-wide scale. *In Advances in Neural Information Processing Systems 17* 2005, **1**:1465–1472.

[41] Wang H, Segal E, Ben-Hur A, Li Q, Vidal M, Koller D: InSite: A computational method for identifying protein-protein interaction binding sites on a proteome-wide scale. *Genome Biology* 2007, **8**(9):R192.1–R192.18.

[42] Li MH, Lin L, Wang XL, Liu T: Protein-protein interaction site prediction based on conditional random fields. *Bioinformatics* 2007, **23**:597–604.

[43] Wu S, Zhang Y: A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. *Bioinformatics* 2008, **24**:924–931.

[44] Espadaler J, Romero-Isart O, Jackson R, Oliva B: Prediction of protein-protein interactions using distant conservation of sequence patterns and structure relationships. *Bioinformatics.* 2005, **21**(16):3360–8.

[45] Chia JM, Kolatkar PR: Implications for domain fusion protein-protein interactions based on structural information. *BMC Bioinformatics* 2004, **5**:161.

[46] Bader J, Chaudhuri A, Rothberg J, Chant J: Gaining confidence in high-throughput protein interaction networks. *Nature Biotechnology* 2004, **22**:78–85.

[47] Gilchrist M, Salter L, Wagner A: A statistical framework for combining and interpreting proteomic datasets. *Bioinformatics* 2004, **20**:689–700.

[48] Jansen R, Yu H, Dreenbaum D, Kluger Y, et al: A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* 2003, **302**:449–53.

[49] Lee I, Date S, Adai A, Marcotte E: A probabilistic functional network of yeast genes. *Science* 2004, **306**(5701):1555–8.

[50] Lin N, Wu B, Jansen R, Gerstein M, Zhao H: Information assessment on predicting protein-protein interactions. *BMC Bioinformatics* 2004, **5**:154.

[51] Yamanishi Y, Vert J, Kanehisa M: Protein network inference from multiple genomic data: a supervised approach. *Bioinformatics* 2004, **20**:363–370.

[52] Zhang L, Wong S, King O, Roth F: Predicting co-complexed protein pairs using genomic and proteomic data integration. *BMC Bioinformatics* 2004, **5**:38.

[53] Ben-Hur A, Noble W: Kernel methods for predicting protein-protein interactions. *Bioinformatics (Proceedings of the Intelligent Systems for Molecular Biology Conference)* 2005, **21**:i38–i46.

[54] Rhodes DR, Tomlins SA, Varambally S, Mahavisno V, Barrette T, Kalyana-Sundaram S, Ghosh D, Pandey A, Chinnaiyan AM: Probabilistic model of the human protein-protein interaction network. *Nat Biotechnol.* 2005, **8**:951–9.

[55] von Mering C, Jensen L, Snel B, Hooper S, Krupp M, Foglierini M, Jouffre N, Huynen M, Bork P: STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res.* 2005, **33**:D433–7.

[56] Jaimovich A, Elidan G, Margalit H, Friedman N: Towards an integrated protein-protein interaction network: a relational markov network approach. *J Comput Biol.* 2006, **13**(2):145–64.

[57] Guo X, Liu R, et al, Shriver C: Assessing semantic similarity measures for the characterization of human regulatory pathways. *Bioinformatics* 2006, **22**:967–73.

[58] Scott MS, Barton GJ: Probabilistic prediction and ranking of human protein-protein interactions. *BMC Bioinformatics* 2007, **8**:239.

[59] NCBI BLAST 2005, [http://www.ncbi.nlm.nih.gov/].

[60] Xenarios I, Salwinski L, Duan X, Eisenberg D, et al: DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* 2002, **30**:303–5.

[61] Sprinzak E, Sattath S, Margalit H: How reliable are experimental protein-protein interaction data. *J Mol Biol.* 2003, **327**(5):919–23.

[62] Kanehisa M, Goto S: KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 2000, **28**:27–30.

[63] King A, Przulj N, Jurisica I: Protein complex prediction via cost-based clustering. *Bioinformatics* 2004, **20**:3013–20.

[64] Dunn R, Dudbridge F, Sanderson C: The use of edge-betweenness clustering to investigate biological function in protein interaction networks. *BMC Bioinformatics* 2005, **6**:39.

[65] Farutin V, Robison K, Lightcap E: Edge-count probabilities for the identification of local protein communities and their organization. *Proteins* 2006, **62**:800–818.

[66] Pereira-Leal J, Enright A, Ouzounis C: Detection of functional modules from protein interaction networks. *Proteins* 2004, **54**:49–57.

[67] Bader G, Hogue C: An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 2003, **13**(4):2.

[68] Adamcsek B, Palla G, et al, Farkas I: Cfinder: locating cliques and overlapping modules in biological networks. *Bioinformatics* 2006, **22**(8):1021–3.

[69] Spirin V, Mirny L: Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci USA.* 2003, **100**:12123–8.

[70] Rives A, Galitski T: Modular organization of cellular networks. *Proc Natl Acad Sci USA* 2003, **100**:1128–33.

[71] Arnau V, Mars S, Marin I: Iterative cluster analysis of protein interaction data. *Bioinformatics* 2005, **21**:364–78.

[72] Sharan R, Ideker T, Kelley B, Shamir R, Karp R: Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. *J Comput Biol* 2005, **12**(6):835–846.

[73] Scholtens D, Vidal M, Gentleman R: Local modeling of global interactome networks. *Bioinformatics* 2005, **21**(17):3548–57.

[74] Chu W, Ghahramani Z, Krause R, Wild DL: Identifying protein complexes in high-throughput protein interaction screens using an infinite latent feature model. *Pacific Symposium on Biocomputing* 2006, **11**:231–242.

[75] Aittokallio T, Schwikowski B: Graph-based methods for analysing networks in cell biology. *Briefings in Bioinformatics* 2006, **7**(3):243–255.

[76] Zotenko E, Guimaraes KS, Jothi R, Przytycka TM: Decomposition of overlapping protein complexes: A graph theoretical method for analyzing static and dynamic protein associations. *Algorithms Mol Biol* 2006, **1**:7.

[77] Brohee S, van Helden J: Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics* 2006, **7**:488.

[78] Han JD, Dupuy D, Bertin N, Cusick ME, Vidal M: Effect of sampling on topology predictions of protein-protein interaction networks. *Nat Biotechnol* 2005, **23**(7):839–44.

[79] Faloutsos M, Faloutsos P, Faloutsos C: On power-law relationships of the internet topology. *ACM Press* 1999.

[80] Chakrabarti D: Tools for large graph mining (advisor: Dr. christos faloutsos). *PhD thesis*, School of Computer Science, Carnegie Mellon University 2005.

[81] Virtanen SE: Properties of nonuniform random graph models. Research Report A77, Helsinki University of Technology, Laboratory for Theoretical Computer Science, Espoo, Finland 2003.

[82] Shen-Orr SS, Milo R, Mangan S, Alon U: Network motifs in the transcriptional regulation network of escherichia coli. *Nature Genetics* 2002, **31**:64 – 68.

[83] Yeger-Lotem E, Sattath S, Kashtan N, Itzkovitz S, Milo R, Pinter RY, Alon U, Margalit H: Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *Proc Natl Acad Sci USA* 2004, **101**(16):5934–9.

[84] Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U: Network motifs: simple building blocks of complex networks. *Science* 2002, **298**(5594):824–7.

[85] Stumpf MP, Wiuf C, May RM: Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proc Natl Acad Sci U S A* 2005, **102**(12):4221–4.

[86] Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasii AL: The human disease network. *Proc Natl Acad Sci USA.* 2007, **21**:8685–90.

[87] Kim PM, Lu LJ, Xia Y, Gerstein MB: Relating three-dimensional structures to protein networks provides evolutionary insights. *Science.* 2006, **314**(5807):1882–3.

[88] Lage K, Karlberg EO, Storling ZM, Olason PI, Pedersen AG, Rigina O, Hinsby AM, Tumer Z, Pociot F, Tommerup N, Moreau Y, Brunak S: A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol* 2007, **25**(3):309–16.

[89] Linding R, Jensen L, Ostheimer G, van Vugt M, Jxrgensen C, Miron I, Diella F, Colwill K, Taylor L, Elder K: Systematic discovery of in vivo phosphorylation networks. *Cell* 2007, **129**(7):1415–1426.

[90] Getoor L, Taskar B: *Introduction to Statistical Relational Learning*. MIT Press 2007.

[91] Robins G, Pattison P, Kalish Y, Lusher D: A workshop on exponential random graph (p*) models for social networks. Social Networks working paper No 1/05, Psychology Department, University of Melbourne. 2005.

[92] Zhang Q, Bhola N, Lui V, Siwak D, Thomas S, Gubish C, Siegfried J, Mills G, Shin D, Grandis J: Antitumor mechanisms of combined gastrin-releasing peptide receptor and epidermal growth factor receptor targeting in head and neck cancer. *Mol Cancer Ther.* 2007, **6**:1414–1424.

[93] Cytoscape 2007, [http://www.cytoscape.org].

[94] Ben-Shlomo I, Yu Hsu S, Rauch R, Kowalski HW, Hsueh AJ: Signaling receptome: a genomic and evolutionary perspective of plasma membrane receptors involved in signal transduction. *Sci STKE.* 2003, **17**(RE9).

[95] Ewing RM, Chu P, Elisma F, Li H, Taylor P, Climie S, et al, Topaloglou T, Figeys D: Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol Syst Biol* 2007, **3**:89.

[96] Stagljar I, Fields S: Analysis of membrane protein interactions using yeast-based technologies. *Trends Biochem Sci* 2002, **27**(11):559–63.

[97] Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan N, Chung S, Emili A, Snyder M, Greenblatt J, Gerstein M: A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science.* 2003, **302**:449–453.

[98] Breiman L: Random forests. *Machine Learning* 2001, **45**:5–32.

[99] Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A, Dolinski K, et al, Sherlock G: Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet.* 2000, **25**:25–29.

[100] Pontius JU, Wagner L, , Schuler GD: Unigene: A unified view of the transcriptome. 2006, [ftp.ncbi.nih.gov/repository/UniGene/Homosapiens/].

[101] GEO: Gene expression omnibus 2005, [ftp://ftp.ncbi.nih.gov/pub/geo/data/gds].

[102] Hong E, Balakrishnan R, Christie K, Costanzo M, Dwight S, et al, Cherry J: Saccharomyces genome database 2005, [ftp://ftp.yeastgenome.org/yeast].

[103] Peri S, Navarro J, Amanchy R, Kristiansen T, Jonnalagadda C, et al, Pandey A: Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Research.* 2003, **13**:2363–2371.

[104] Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, et al, Tatusova L T A and-Wagner, Yaschenko E: Database resources of the national center for biotechnology information. *Nucleic Acids Res* 2008, **36**(Database issue):D13–21.

[105] Strobl C, Boulesteix AL, Zeileis A, Hothorn T: Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics* 2007, **8**:25.

[106] Hastie T, Tibshirani R, Friedman J: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York, NY 2001.

[107] Jansen R, Gerstein M: Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. *Curr Opin Microbiol.* 2004, **7**:535–45.

[108] Madeira SC, Oliveira AL: Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Trans. Comput. Biol. Bioinformatics.* 2004, **1**:24–45.

[109] Trible RP, Emert-Sedlak L, Smithgall TE: HIV-1 nef selectively activates src family kinases hck, lyn, and c-src through direct sh3 domain interaction. *J. Biol. Chem.* 2006, **281**:27029–27038.

[110] Pecquet C, Nyga R, Penard-Lacronique V, Smithgall T, Murakami H, Lassoued K, Gouilleux F: The src tyrosine kinase hck is required for tel-abl- but not for tel-jak2-induced cell transformation. *Oncogene.* 2007, **26,**:1577–1585.

[111] Damke H, Baba T, Warnock D, Schmid S: Induction of mutant dynamin specifically blocks endocytic coated vesicle formation. *J. Cell Biol.* 1994, **27**(4):915–934.

[112] Xi S, Zhang Q, Dyer K, Lerner E, Smithgall T, Gooding W, Kamens J, Grandis J: Src kinases mediate stat growth pathways in squamous cell carcinoma of the head and neck. *J Biol Chem* 2003, **278**:31574–31583.

[113] Cai K, Klein-Seetharaman J, Hwa J, Hubbell WL, Khorana HG: Structure and function in rhodopsin: effects of disulfide cross-links in the cytoplasmic face of rhodopsin on transducin activation and phosphorylation by rhodopsin kinase. *Biochemistry* 1999, **38,**:12893–12898.

[114] Booth V, Clark-Lewis I, Sykes BD: NMR structure of CXCR3 binding chemokine CXCL11 (ITAC). *Protein Sci.* 2004, **13,**:2022–2028.

[115] Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, Fox BA, TI L, Teller DC, Okada T, Stenkamp RE, Yamamoto M, Miyano M: Crystal structure of rhodopsin: A G protein-coupled receptor. *Science* 2000, **289**(5480):739–45.

[116] Comeau SR, Gatchell DW, Vajda S, Camacho CJ: ClusPro: a fully automated algorithm for protein-protein docking. *Nucleic Acids Res.* 2004, **32,**:W96–W99.

[117] Xia Y, Lu L, Gerstein M: Integrated prediction of the helical membrane protein interatome in yeast. *J. Mol. Biol.* 2006, **357**:339–349.

[118] Ekman D, Light S, Bjorklund AK, Elofsson A: What properties characterize the hub proteins of the protein-protein interaction network of saccharomyces cerevisiae? *Genome Biology* 2006, **7**(6):R45.

[119] Tong A, Lesage G, Bader G, Ding H, Berriz G, et al: Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* 2001, **294**:2364–2368.

[120] Xing E, Ng A, Jordan M, Russell S: Distance metric learning, with application to clustering with side-information. *Advances in NIPS* 2003, **15**.

[121] Blum A, Mitchell T: Combining labeled and unlabeled data with co-training. *In Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT-98)* 1998.

[122] Muslea I: Active learning with multiple views. *PhD thesis*, Department of Computer Science, University of Southern California 2002.

[123] Denis F, Laurent A, Gilleron R, Tommasi M: Text classification and co-training from positive and unlabeled examples. *ICML-2003 workshop: the Continuum from labeled data to unlabeled data in Machine Learning and Data Mining.* 2003.

[124] Muslea I, Minton S, Knoblock C: Active + semi-supervised learning = robust multi-view learning. *Proceedings of ICML-02, 19th International Conference on Machine Learning* 2002, :435–442.

[125] Waterhouse SR: Classification and regression using mixtures of experts. *PhD thesis*, Department of Engineering, Cambridge University. 1997.

[126] Jordan MI, Jacobs RA: Hierarchical mixtures of experts and the em algorithm. *Neural Computation* 1994, **6**(2):181–214.

[127] Mishra G, Suresh M, Kumaran K, Kannabiran N, Suresh S, et al, Pandey A: Human protein reference database–2006 update. *Nucleic Acids Res* 2006, **34(Database issue)**:D411–4.

[128] Joachims T: Learning to classify text using support vector machines. *Dissertation* 2002. [Phdthesis].

[129] Witten IH, Frank E: *Data Mining: Practical machine learning tools with Java implementations*. San Francisco: Morgan Kaufmann 2000.

[130] Probost F: Machine learning from imbalanced data sets 101. *Invited paper for the AAAI'2000 Workshop on Imbalanced Data Sets.* 2000.

[131] Elion E: Ste5: a meeting place for map kinases and their associates. *Trends Cell Biol.* 1995, **5**:322–7.

[132] Bader GD, Hogue CW: Analyzing yeast protein-protein interaction data obtained from different sources. *Nature Biotechnology* 2003, **20**(10):991–997.

[133] Borgwardt KM, Kriegel HP, Vishwanathan SVN, Schraudolph NN: Graph kernels for disease outcome prediction from protein-protein interaction networks. In *Proc. of Pacific Symposium on Biocomputing (PSB)*, *Volume 12*, Maui, Hawaii: World Scientific 2007:4–15.

[134] Prulj N: Biological network comparison using graphlet degree distribution. *Bioinformatics* 2007, **23**(2):e177–e183.

[135] Yan X, Han J: gspan: Graph-based substructure pattern mining. Technical report uiucdcs-r-2002-2296, Dept. of Computer Science, Univ. of Illinois at Urbana-Champaign. 2002.

[136] Barabasi A, Oltvai Z: Network biology: understanding the cell's functional organization. *Nat Rev Genet.* 2004, **5**(2):101–13.

[137] Dolinski K, Balakrishnan R, Christie K, Costanzo M, et al: Saccharomyces genome database (SGD). http://www.yeastgenome.org 2004.

[138] Zhu D, Qin Z: Structural comparison of metabolic networks in selected single cell organisms. *BMC Bioinformatics.* 2005, **6**:8.

[139] Hanneke S, Xing E: Discrete temporal models of social networks. *Statistical Network Analysis: Models, Issues, and New Directions, A Workshop at the 23rd International Conference on Machine Learning (ICML)* 2006.

[140] Manning, Schutze: *Foundations of Statistical Natural Language Processing*. MIT press 1999.

[141] Cormen, Leiserson, Rivest, Stein: *Introduction to Algorithms*. McGraw-Hill, second edition edition 2001.

[142] Ideker T, Ozier O, Schwikowski B: Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 2002, **18 Suppl 1**:S233–40.

[143] Jones KS: *Information retrieval experiment*. London Butterworths: Morgan Kaufmann Publishers 1981.

[144] Krogan N, et al, Greenblatt J: Global landscape of protein complexes in yeast saccharomyces cerevisiae. *Nature* 2006, **440**(7084):637–43.

[145] Albert R: Scale-free networks in cell biology. *J Cell Sci* 2005, **118**:4947–57.

[146] Scott J, Ideker T, Karp R, Sharan R: Effcient algorithms for detecting signaling pathways in protein interaction networks. *RECOMB* 2005, :1–13.

[147] Kim PM, Sboner A, Xia Y, Gerstein M: The role of disorder in interaction networks: a structural analysis. *Molecular Systems Biology* 2008, **4**:179.

[148] Zhou HX, Qin S: Interaction-site prediction for protein complexes: a critical assessment. *Bioinformatics* 2007, **23**(17):2203–2209.

[149] Sharan R, Ulitsky I, Shamir R: Network-based prediction of protein function. *Mol Syst Biol* 2007, **3**:88.

[150] Pena-Castillo L, et al, Qi Y, et al, Roth F: A critical assessment of m. musculus gene function prediction using integrated genomic evidence. *Genome Biology (In Press)* 2008.

[151] Deng M, Zhang K, Mehta S, Chen T, Sun F: Prediction of protein function using protein-protein interaction data. *J Comput Biol.* 2003, **10**(6):947–60.

[152] Carroll S, Pavlovic V: Protein classification using probabilistic chain graphs and the gene ontology structure. *Bioinformatics* 2006, **22**:1871–78.

[153] Engelhardt B, Jordan M, Muratore J, Brenner S: Protein function prediction by bayesian phylogenomics. *PLoS Computational Biology* 2005, **1**:432–45.

[154] Nabieva E, Jim K, Agarwal A, Chazelle B, Singh M: Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics* 2005, **21**(S1):i302–10.

[155] Deng M, Chen T, Sun F: An integrated probabilistic model for functional prediction of proteins. *J Comput Biol.* 2004, **11**(2-3):463–75.

[156] Deng M, Tu Z, Sun F, Chen T: Mapping gene ontology to proteins based on protein-protein interaction data. *Bioinformatics* 2003, **20**(6):895–902.

Glossary

AD Domain that activates transcription.

BD Domain that directs binding to a promoter DNA sequence.

BIND Biomolecular Interaction Network Database.

Cell The structural and functional unit of all known living organisms.

Classification Predict a discrete Y from X.

Clustering Put data into groups.

DIP Database of interacting proteins.

DNA A nucleic acid containing the genetic instructions.

EM Expectation Maximization.

Estimation Using data to estimate an unknown quantity.

Example Any item in the instance space.

Feature Vector Describe an example with a vector of various attributes' values.

Gene Expression Measurement of mRNA concentration.

Genome An organism's whole hereditary information and is encoded in the DNA.

In Vitro Experiment in a controlled environment outside of a living organism.

In Vivo Taking place inside an organism.

Instance Any item in the instance space, also called as example.

Instance Space The set of all possible examples.

Interactome Whole set of molecular interactions in cells.

kNN k-Nearest Neighbor.

Ligand An molecule that bonds to a central metal.

LR Logistic regression.

ME Mixture of Experts.

MFE Mixture of Feature Experts.

MFE-FM Mixture-of-Feature-Experts with missing values filled.

mRNA A molecule of RNA encoding a chemical blueprint for a protein product.

MS Mass spectroscopy.

NB Naive Bayes.

NN Nearest Neighbor.

PPI Protein-Protein Interaction.

Protein-Protein Interaction The association of protein molecules.

Proteome Entire complement of proteins expressed by a genome, cell, tissue or organism.

RF Random Forest.

SVM Support Vector Machine.

TAP Tandem affinity purification.

Tissue A set of interconnected cells that perform a similar function within an organism.

Y2H Yeast-Two-Hybrids.