PROTEIN-PROTEIN INTERACTION NETWORKS AND SOME RELATED PROBLEMS

by

NGUYEN THANH PHUONG

submitted to
Japan Advanced Institute of Science and Technology
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

Supervisor: Professor HO TU BAO

School of Knowledge Science
Japan Advanced Institute of Science and Technology

September 2008

Abstract

In all areas of biological and medical research, the role of the computer has been dramatically enhanced in the last five- to ten-year period. While the first wave of computational analysis did focus on sequence analysis, where many highly important unsolved problems still remain. Outstandingly, protein-protein interaction research looks into the association of proteins to discover the rules controlling their interactions, which are key parts of cell mechanism. Because many major biological processes are controlled by protein interaction networks, the comprehensive description of protein- protein interactions (PPI) is necessary to understand the genetic program of life. The research also aims at predicting a novel protein function given the relation with a well-characterized one.

In addition to the central problem of PPI prediction, two relevant problems have been raising and developing rapidly, i.e., signal transduction networks (STN) and disease-causing genes from PPI networks. STN play an important role in the control of most fundamental cellular processes, including cell proliferation, metabolism, differentiation, and survival. It is known that STN most likely dependent on PPI. Also, discovering human disease-causing genes (disease genes in short) is one of the most challenging problems in bioinformatics and biomedicine, because most of diseases are related in some way to our genes. There are several evidences of the relationship between PPI and disease genes. Keeping with the most attractive problems, our study targets three significant problems: (1) protein-protein interaction prediction, (2) signal transduction construction, and (3) disease-causing gene prediction.

In fact, three problems have long histories of their own in biology and medicine. However, the traditional wet-lab methods often required many efforts. Since the work itself much involves quantitative tasks, computer science came to the scene bringing about another approach, the computational one, to the standing issues. Thanks to life scientists, the amount of biological and medical data is growing exponentially. As a result, computational methods become more and more essential to mine the huge amount of data and discovery useful knowledge for life science.

In such context, our strategy is twofold. The first one is to take the full advantage of the biological nature of PPI, STN and disease-causing genes underlying a titanic amount of data. The second one is to develop appropriate and robust computational methods to integrate those complex biological and medical data, and then solve three targeted problems. These proposed methods fill the gaps of existing methods and achieve considerable contributions as follows.

1. *Protein-protein interaction prediction: We developed a novel integrative domain-based method to predict protein-protein interactions.*
   The previous works either used multiple data sources instead of a single source as in integrative methods or only protein domain features as in the domain-based methods. The key idea of our computational method is to integrate protein domain features and genomic and proteomic features from multiple data sources into PPI prediction using Inductive Logic Programming (ILP). Comparing with other methods, our method outperforms in terms of several evaluation measures. Moreover, represented in forms of ILP rules, the predictions were easy to interpret and useful for biologists.

2. *Signal transduction network construction: We developed an effective computational method to construct human signal transduction networks from protein-protein interaction networks.*
   Our method is better than the previous ones by exploiting three biological facts of STN applied to human, i.e., rich-information of protein-protein interaction networks, signaling features, and sharing components among STN. We firstly considered many levels of the signaling machinery in terms of signaling features. Secondly, soft-clustering effectively detected the sharing components among STN. The early work was done for yeast, and later we shifted to human STN, a currently significant challenge. To the best of our knowledge, this study is the first one that has taken effort to construct human STN computationally. Both the evaluations of STN construction for yeast and human were promising with high performance and gain some considerable findings.

3. *Disease-causing gene prediction: We developed a new method for discovering disease genes with the exploitation of semi-supervised learning, protein-protein interactions and multifarious disease-related features.*
   Differed from existing work, our method based on semi-supervised learning (i) solved imbalance between known disease genes and unknown disease genes, (ii) integrated multifarious data related to disease genes, and (iii) exploited both useful information of labeled and unlabeled data. The contributions of this work were not only the new and effective method for disease gene prediction but also new significant findings. The comparative results demonstrated that our method obtained the higher sensitivity, specificity, precision, accuracy, and balanced F-score. Testing with all interacting partners of disease proteins, we found 568 putative disease genes.

In conclusion, our effort in analyzing the protein interaction network data is to mine the coherent information, forecast unobserved interactions, and then detect relevant biological functions and processes, i.e., signal transduction networks and disease-causing

genes. The thesis focused on benefiting multiple data sources to effectively solve three biologically significant problems related to PPI in both theoretical and empirical aspects. The theoretical aspect of this thesis concerns about the design of new and effective methods for protein-protein interaction prediction, signal transduction network construction and disease-causing gene prediction. The other aspect is the application of these methods to produce lot of biological findings that can be useful sources for life scientists.

We further expect to combine three current works in our research to build up a complete decision support system for disease diagnostics and drug design.

Key words: Bioinformatics, Protein-Protein Interaction Networks, Protein Domain-Domain Interactions, Signal Transduction Networks, Disease-Causing Genes, Inductive Logic Programming, Soft Clustering, Semi-supervised Learning.

Acknowledgments

First of all, I wish to express my deepest respect and appreciation to my supervisor: Professor Tu Bao Ho, School of Knowledge Science, Japan Advanced Institute of Science and Technology (JAIST) for his kindly guidance, warm encouragement, and supports before and during my study. He has given me much invaluable knowledge not only how to formulate a research idea or to write a good paper, but also the vision, and much useful experience in the academic life. I am grateful for his patient supervision, and I am really lucky and proud to be one of his students. Besides academic life, Professor Ho also cared and supported about my daily life. I am really impressed by his kind consideration. Everything that I studied under his supervision will go along with me all of my life.

I wish to say sincere thanks to Professor Eiichiro Ichiishi, Professor Yoshiteru Nakamori, Hashimoto Takashi from JAIST, and Associate Professor Kenji Satou from Kanazawa University for being the members of my defense committee. I highly appreciate all of their worthy comments.

I wish to convey sincere thanks to Associate Professor Kenji Satou, for his valuable discussions and supports not only during my sub-theme research but also in the whole study period of mine in JAIST. He gave me the very interesting knowledge in his course Bioinfomatics and unconditionally discussed with me whenever I asked for help. In addition, he allowed me to access his computer systems and biological databases that are very important for my experiments.

I would also like to express my appreciation to Professor Nguyen Ngoc Binh, College of Technology, Vietnam National University, Hanoi (VNUH), Lecturer Do Van Uy, Lecturer Bui Thi Hoa, Professor Luong Chi Mai, Institute of Information Technology, National Center for Science and Technology of Vietnam for their kindly recommendations and constant encouragement before and during my research at JAIST.

I would like to thank the Lecture Mary Ann Mooradian, and the Technical Communication Reviewer Mark G. Elwell, Mrs. Judith A. Steeh for their help in proof-reading and correcting errors in my papers. I have learnt a lot from their corrections.

I have received a lot of help from colleagues and friends in Ho-Lab and Satou-Lab during last three years. Let me say special thanks to Assistant Professor Kawasaki Saori, and Mr. Oota Takato for all their helps for my life from the first day I came to Japan. I would like to thank my colleagues in Ho-Lab, Mr. Nguyen Canh Hao, Mr. Tran Dang Hung, Dr. Tran Tuan Nam and in Satou lab, Dr. Jose Carlos Clemente and others

for sharing their research ideas, useful experiences, as well as valuable discussions and comments.

I also wish to send my deep acknowledgements to JAIST staffs for their kind and convenient procedures and services.

I'd like to appreciate the authors of open source tools/packages such as BioPerl, Support Vector Machines (SVM^{light}), Aleph, SemiL, Mfuzz, Endevour. Without these packages, my experiments were hard to be completed.

Last but not least, my family is really the biggest motivation behind me. My husband, Nguyen Duy Cu, who always gives me love and encouragement. He contributes to all aspects of my life. My dear parents, my dear brother' family, my parent-in-law's family together with their unconditional sacrifices, love, and supports are always endless sources of inspiration for me to move forwards, so this thesis is dedicated to them.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this chapter, we briefly introduce the research on protein-protein interactions, the problem statement and the research context, and the contributions of the thesis. We first introduce the research on protein-protein interaction networks. Secondly, we state the research problems that the thesis targets to solve in the context of other work. The main contributions of the thesis are later described corresponding to each stated problem. Finally, we outline the structure of the thesis.

1.1 Research on Protein-Protein Interaction Networks

Bioinformatics has been emerging as one of the most prominent fields in the 2000s for its huge contribution to improve human life. Bioinformatics involves the use of techniques including applied mathematics, informatics, statistics, computer science, artificial intelligence, chemistry and biochemistry to solve biological problems usually on the molecular level. Major research efforts in the field include sequence alignment, gene finding, genome assembly, protein structure alignment, protein structure prediction, prediction of gene expression and protein-protein interactions, and the modeling of evolution. Among them, protein-protein interaction networks are taken much interest of many ongoing research projects. Biologically, in performing their functions, proteins rarely function in isolation. Interactions among proteins are intrinsic to almost all cellular functions and biological processes. Because of the importance of PPI, in addition to traditional experimental methods, there is a great need to develop computational methods to study protein-protein interactions and discover their biologically relevant effects in cells.

In the PPI research, there are three major subareas. The first one is to predict protein-protein interactions. The second one is study the characteristics of protein-protein interactions. The last one is to detect biological phenomena related to protein-protein

interactions.

The first subarea is to predict whether two arbitrary proteins interact or not. This subarea has attracted a lot of works and achieved many notable results. These works mainly fall into two categories: the homology (or similarity) methods and the non-homology methods. The homology methods generally assume that proteins with similar sequences or structures perform similar functions [Bock and Gough, 2001, Matthews et al., 2001]. However, proteins with similar sequences or structures could perform either similar functions (e.g., in the case of ortholog proteins) or different functions (e.g., in the case of paralog proteins), while two proteins with low sequence similarity could play similar roles (e.g. in the case of remote homology). The non-homology methods utilize properties of proteins other than sequence or structure similarity, such as gene neighborhood, domain composition and gene expression [Pellegrini et al., 1999]. As the data is more and more available and the computational techniques are much more improved as well, the research on PPI prediction is non-stop and requires more high-quality results [Ng and Tan, 2003].

The second subarea is to study insights of PPI characteristics. Besides the features of each interacting partner, PPIs have dynamic and complex characteristics of their own that differentiate one from another and later determine the functionality of PPI in cells. These characteristics are biological, physical , chemical, and also the combinations like physiological or chemico-physical characteristics, etc. Among numerous PPI characteristics, the recent work is mainly focusing on stability and transience, sites and interfaces of PPI. Starting from various points of view, the related work did not share a common approach, they were mostly based on a specific type of data and the popular methods have been the statistical methods [Mintseris and Weng, 2005, Ofran and Rost, 2007, Reichmann et al., 2005, Yan et al., 2004]. As the support data is limited, and much biologically concrete knowledge and wet-lab experiments are required, the second subarea still remains very challenging.

The third subarea is to discover the phenomena that are biologically close with PPI and affected to each other. For a long time, research groups working on detecting protein function detection from PPI have been very strong and have opened an important trend in protein function research [Marcotte et al., 1999, Chen and Yuan, 2006, Baudot et al., 2006]. In addition, lots of works have taken much effort to propose alternative ways in finding out other biological phenomena by investigating PPI networks. More specifically, many works have been interested in studying signal transduction networks (STN) and disease-causing genes (disease genes). In the signal transduction network study, it is known that STN most likely dependent on PPI. The strong association between PPI and signal transduction encourages the researchers to develop computational methods for constructing and modeling STN from PPI networks [Allen et al., 2006, Liu and Zhao, 2004, Steffen et al., 2002]. In the disease-causing genes study, the topology of PPI networks is widely confirmed to be

related to disease-causing genes. And several biological evidence have shown the effect of disruption of PPI in causing diseases. There is no doubt that disease genes can be discovered through PPI networks. Because of the significance of this problem, in some recent years, this topic has been raised as one of the hottest topics in the PPI research [Benjamin Schuster Bockler and Alex Bateman , 2008],
[Ideker and Sharan, 2008].

In just the past two to three years, large human protein-protein interaction networks have been increasing rapidly. This availability inspires many current studies to shift from attempting to understand networks encoded by model species to understanding the biomedical secrets that underlie human networks. The future of the the PPI research is going on brightly and splendidly.

1.2 Problem Statement and Research Context

We are highly motivated by the central roles of protein-protein interactions network study in bioinformatics. After investigating and analyzing the current research on PPI networks, our study targets to the first and the third subarea. Concretely, the thesis pursues three predominant problems: (1) protein-protein interaction prediction, (2) signal transduction network construction, and (3) disease-causing gene prediction.

The first problem investigated here is *the integrative approach to predict protein-protein interactions*. The motivation is the reliability and the comprehensive interpretation of PPI predictions.

At the early time, most of works tried to employ one protein feature to predict PPI. With the recent blooming of public proteomic and genomic databases, numerous computational approaches offer a chance to study protein-protein interactions more widely and deeply. Depending on the source of information used, computational approaches can be categorized into three groups: structure-based, sequence-based, and genome-based, as shown in [Bock and Gough, 2001, Matthews et al., 2001, Pellegrini et al., 1999] respectively. Besides the methods based on a single data source, many bioinformaticians attempted to use multiple data sources, *the integrative approach*, to better predict PPI. Jansen *et al.* used a Bayesian network approach to integrate weakly predictive genomic features into the predictions of protein-protein interactions [Jansen et al., 2003]. Several kernels for different data sources like protein sequences, Gene Ontology annotations, local properties of networks, etc. were combined [Ben-Hur and Noble, 2005]. Some other works were on probabilistic decision tree approach, [Zhang et al., 2004] inductive logic programming method [Tran et al., 2005], probabilistic model [Rhodes et al., 2005], and etc. From multiple data sources, it is possible to extract and combine various genomic and proteomic

features related to PPI. The obtained results showed the potential of multiple data source integration.

Protein domains, structural and/or functional units of proteins, are conserved through the evolution to represent protein structures or functions. They are the key regulators in protein-protein interactions. Interactions among domains act as stable channels of PPI. Recently, *the domain-based approach* to the prediction of PPI has received much attention in many ongoing studies. One of the pioneer works based on protein domains was association method [Sprinzak and Margalit, 2001]. Kim *et al.* improved this method by considering the number of domains in each protein [Kim et al., 2002]. Han *et al.* proposed a domain combination-based method by considering the possibility of domain combinations appearing in both interacting and non-interacting sets of protein pairs [Han et al., 2003]. A graph-oriented method, called the interacting domain profile pairs, was proposed by Wojcik and Schachter [Wojcik and Schachter, 2001]. Chen *et al.* used domain-based random forest framework [Chen and Liu, 2005]. Last but not least, Martin *et al.* used signatures generated from sequences [Martin et al., 2005].

The shortcoming of integrative methods is that they did not take protein domains into account while there are evidences that the biological mechanisms underlying protein-protein interactions are protein domains and their interactions [Pawson et al., 2002]. Furthermore, while domain-based methods all treasured the biological roles of protein domains in PPI prediction, most of them merely considered the co-occurrence of domains-domain pairs. To predict PPI comprehensively, it is reasonable to employ genomic and proteomic features in domain-based methods. Additionally, PPI predictions are much more useful if their interpretations are easy-understanding.

The second problem is to *construct signal transduction networks from PPI networks*. The motivation is how to take the best of combining biologically significant nature of signal transduction in terms of signaling features of components and sharing components among networks.

Constructing STN based on PPI is an area of much ongoing research. Using Markov chain Monte Carlo method, Gomez *et al.* modeled STN in terms of domains in upstream and downstream protein interactions [Gomez et al., 2001]. Steffen *et al.* developed a computational method for generating static utilized PPI maps produced from large-scale two-hybrid screens and expression profiles from DNA micro-arrays in STN construction [Steffen et al., 2002]. Liu *et al.* applied a score function that integrated PPI data and micro-array gene expression data to predict the order of signaling pathway components [Liu and Zhao, 2004]. Concerning protein modification time-course data, Allen *et al.* applied a method of computational algebra to modeling of signaling networks [Allen et al., 2006]. Fukuda *et al.* represented the model of signal transduction path-

ways based on a compound graph structure [Fukuda and Takagi, 2001]. A recent work has searched for the optimal subnetworks from PPI according to some cost functions [Zhao et al., 2008].

Although the previous work achieved many results, there are still some biological characteristics of STN that the previous works did not take much into account. First, it is known that the deeper level underlying the PPI to transmit signals in cells are functional domains, so-called signaling domains, and their interactions [Pawson et al., 2002], [Eungdamrong and Iyenga, 2004]. Data of those significant signaling features are structured, complexly relational, and sparse in different data sources. In order to construct STN effectively, those data is needed to be appropriately integrated. Second, STN indeed have many overlapping components including proteins and their interactions
[Neves and Iyengar, 2005]. As a result, the method that can join these characteristics of STN is promising to achieve good results in STN construction.

Finally yet importantly, the third problem is to *predict disease-causing genes using PPI networks*. The motivation is to benefit information-rich PPI networks, the useful data from various databases and lastly the effectiveness of machine learning techniques.

Research on protein-protein interaction networks and diseases has been rapidly increasing in the last two or three years. Many PPI-based methods have been proposed, each with a different way of exploiting the key assumption that "the network-neighbor of a disease gene is likely to cause the same or a similar disease", see [Goh et al., 2007], [Ideker and Sharan, 2008], and [Benjamin Schuster Bockler and Alex Bateman , 2008]. In an early work, disease genes were uncovered by topological features in human PPI networks using the k-nearest neighbor algorithm [Xu and Li, 2006]. Because of the sparseness of other proteomic/genomic data associated with certain diseases, several PPI-based methods have been required the integration of heterogeneous biomedical data in order to understand the complex interplay between genes/proteins and diseases [Kann, 2007]. A disease gene classification system has been proposed, to integrate the topological features of protein interaction networks with sequence and other features, and to analyze these features using support vector machines [Smalter et al., 2007]. Lage et *al.* (2007) used the phenomic ranking of protein complexes linked to human diseases to develop a Bayesian model for predicting new candidates for disorders. Borgwardt and Kriegel (2007) integrated graph kernels for gene expression and human PPI to predict disease genes. In another direction, some work has concentrated on using PPIs to discover disease genes for specific diseases, i.e., Alzheimer's disease, using heuristic score functions [Chen et al., 2006], [Krauthammer et al., 2004].

It can be seen that so far the ratio of known human disease genes to the total number of human genes is small. All previous work employed supervised learning schemes, which

exploited data regarding known disease genes to predict new disease gene candidates. The disease predictions are potentially better if we can benefit the hidden information of many unlabeled data. In addition, it is worth noting that information of proteins is not contained in a single database but resides in different databases. The integration of human protein-protein interactions with various biological data extracted from multiple proteomic/genomic databases can improve the performance of the disease-gene prediction.

For all above, the goal of this study is to fill the gaps of previous methods to better solve three mentioned problems in the PPI research. Our strategy is twofold. Firstly, various biologically significant data sources are investigated and then we take full advantage of these informative data that show the nature of phenomena. Secondly, we develop appropriate and robust computational methods for mining those complex biological and medical data.

1.3 Main Contributions

The thesis aims at studying protein-protein interactions and related problems in biomedicine. The ultimate goal is that we can effectively and computationally mine enormous biologically significant data to discover useful knowledge supporting to life scientists. The main contributions of this thesis are summarized as follows.

The first contribution is that we have proposed a novel integrative domain-based method to predict protein-protein interactions.

The key idea of this computational method is to integrate protein domain features and multiple genomic/proteomic features into PPI prediction using inductive logic programming.

While the integrative methods use multiple data sources instead of a single source, the domain-based methods often use only protein domain features. Integration of both protein domain features and genomic/proteomic features from multiple data sources can more effectively predict PPI. Moreover, it allows discovering the reciprocal relationships between PPI and biological features of their interacting partners. To integrate efficiently such two kinds of features, we specified two main tasks. The first is to extract as many as possible useful domain and genomic/proteomic features related to PPI. From seven popular databases, we extracted more than 278,000 ground facts of domain fusion, domain-domain interaction features and various biologically significant genomic/proteomic features. The second is to employ inductive logic programming (ILP) on the huge amount of background knowledge to infer PPI effectively.

To demonstrate the advantages of the above-mentioned integration, we conducted multiple 10-fold cross validations to compare our method with two other methods based on single domain features, as well as with the non domain-based approach using multiple genomic databases. The performance measures include Receiver Operating Characteristic (ROC) curves, sensitivity and specificity. In all tests, our method performed considerably better than the others did. Sharing the same ILP framework, domain-domain interactions were successfully inferred with high sensitivity and specificity. Lastly, analyzing various produced rules (of both PPI and DDI), many interesting relationships among PPI, DDI, and protein functions, biological processes, were found. Our proposed method can be tuned to predict PPI and DDI for diverse organisms and other genomic and proteomic data sources.

The publications related to PPI prediction are [Nguyen and Ho, 2008b],
[Nguyen et al., 2007, Nguyen and Ho, 2007b, Nguyen and Ho, 2007c, Nguyen and Ho, 2006].

The second contribution is that we have put forward an effective soft-clustering method to construct signal transduction networks from PPI networks.

The key idea of this computational method is to grasp the nature of signal transduction networks underlying the variety of signaling features and PPI networks using soft-clustering.

In case of previous methods, they did not much care deep levels of STN, i.e., signal domain, signaling domain interactions, and sharing components between STN. This work aims to solve those two intricate problems of STN to better construct STN from PPI networks. To this end, we introduced an effective computational method to construct STN that (1) exploited integrated multiple signaling features of STN from heterogenous sources, i.e., protein-protein interactions, signaling domains, domain-domain interactions, and protein functions, (2) detected overlapping components using soft-clustering. Additionally, in previous work clustered objects were often individual proteins, but our method handled clustered objects as the functional or physical protein interactions because these interactions are the means to transmit signals in cells. Not only limited to yeast STN, we currently shift to work on human STN. To the best of our knowledge, this work is the first one that computationally solves the STN problem for *Homo Sapiens*.

Experimental evaluation showed the high performance of our proposed method. We could reconstruct signal transduction networks with the small error and more importantly detect the overlapped parts among networks. The method is promising to discover new STN and solve other related problems in computational and systems biology from large-scale human protein interaction networks. Other early results related to yeast STN were also comparative. Signaling domain-domain interactions were well predicted and yeast

MAPK STN were reconstructed with considerable coverage.

The related publications related to STN construction are [Nguyen and Ho., 2006] and [Nguyen and Ho, 2008a].

The third contribution is that we have developed a new method for discovering disease genes by the exploitation of semi-supervised learning, protein-protein interactions and multifarious disease-related features.

The key idea of this work is, how best to utilize the wealth of existing data that may contain information about unknown genes. All previous work employed supervised learning schemes, which exploited data regarding known disease genes to predict new disease gene candidates. However, it has recently been shown by Oti *et al.* that genes associated with a particular phenotype or function are not randomly positioned in the PPI network, but tend to exhibit high connectivity; they cluster together and occur in central network locations [Oti et al., 2006]. That overriding property suggests that semi-supervised learning can be used in this prediction problem to exploit not only data concerning discovered disease genes but also data which may concern disease genes that are not yet known. In fact, this property solidifies the fundamental assumptions about the consistency of semi-supervised learning and provides more evidence for taking into account information regarding the unknown genes. Moreover, in addition to the protein topological features extracted from PPI databases, semi-supervised learning enables a systematic consideration of proteomic/genomic features related to diseases from various available data sources, which further enriches this computational scheme.

This work not only proposes a new and effective method for disease gene prediction, but also has generated significant new findings. We carefully carried out various experiments with disease gene information extracted from the OMIM (Online Mendelian Inheritance in Man) database (version 2007) [Hamosh et al., 2005]. Testing with all interacting partners of disease proteins, we found 568 putative disease genes. Some encouraging results were indirectly validated in various ways.

We performed two comparative experiments to evaluate the performance of the method. First, 10 times stratified 10-fold cross validations were conducted using our new Semi-Supervised Learning (SSL) method, the Support Vector Machines (SVMs) method [Smalter et al., 2007], and the k-nearest neighbor (k-NN) method [Xu and Li, 2006]. The results show that the SSL method outperforms the other two in terms of sensitivity, specificity, precision, accuracy, and a balanced F-score. Next, we compared our SSL method to the k-NN method with different sizes of labeled sets, and did twenty trials for each experiment to evaluate the accuracy. It turns out that the achieved accuracy of SSL is higher than that of k-NN.

The publications related to disease-casing gene discovery are [Nguyen and Ho, 2008c] and [Nguyen and Ho, 2007a].

All in all, this thesis focuses on benefiting multiple data sources for solving three biologically significant problems in the PPI research in both theoretical and empirical aspects. The theoretical aspect of this thesis concerns about the design of new and effective methods for protein-protein interaction prediction, signal transduction network construction and disease-causing gene prediction. The other aspect is the application of these methods to produce lot of new biological findings that can be useful sources for life scientists.

1.4 Dissertation Organization

The thesis is divided into six chapters, including the current one. The organization of the dissertation is as follows.

Chapter 2 introduces the background knowledge of molecular biology and protein-protein interactions. Firstly, some key points of molecular biology are summarized. Next, the PPI networks and theirs characteristics are presented. The last parts will discuss the relationships between PPI and STN, and between PPI and disease genes as well.

Chapter 3 presents the integrative domain-based method for PPI prediction. There are two main tasks in the method. Two main tasks are: (1) Constructing integrated background knowledge of domain features and multiple genomic/proteomic features, and (2) Learning PPI predictive rules by ILP from the constructed background knowledge. The experiments and evaluation with good results and new findings are shown in this chapter.

Chapter 4 describes the work of STN construction from PPI networks. First, the signaling features of STN are investigated and then weighted according to their significance. Soft-clustering method is proposed to detect the sharing components among STN and combine the signaling features. The experiments and evaluation are given with promising results. Some results related to yeast STN are also presented. At last, the outlook of this work is discussed.

Chapter 5 proposes to a semi-supervised learning method for disease-causing gene prediction problem. In succession, the chapter addresses three main issues to successfully predict disease genes. The first one is based on the similar phenotype and genotype of neighbors in PPI networks to reliably extend the known disease gene

set as the set of disease gene candidates. The second one is to extract from various databases the protein/gene traits that relate to disease. The third one is to integrate the whole rich data from (1) and (2) for achieving the best performance of semi-supervised learning in disease-causing prediction. The experimental results have demonstrated the robustness and accuracy of the proposed method. The new findings are initially validated.

Chapter 6 first summarizes the main tasks of the thesis, including the main achievements and contributions, and some shortcomings as well. Next, some interesting problems are opened and will be mentioned as the future research directions.

Chapter 2

Background

In this chapter, we first provide some introductions about molecular biology. Then, protein-protein interactions and their particular characteristics are presented. Finally, we would like to give the overall picture of the relationships between PPI and related problems, i.e., signal transduction networks and disease-causing genes.

2.1 Molecular biology

A cell, numbered to 100 billion of 320 different types in human body, is composed of nucleus, membrane, and the building blocks. Cellular basic activities include growth, division, and differentiation, which are completed by the functioning of the building blocks including macromolecules and small ones. Macromolecules are of three types: Proteins as basic constructional blocks hosting metabolism and maintaining cellular environment, Deoxyribonucleic acid (DNA) carrying inheritable information, and Ribonucleic acid (RNA) fostering protein synthesis. These are in turn built by small molecules, amino acids and nucleotides respectively. Generally, the flow of information in the cell is as below.

$$\text{DNAs} \Rightarrow \text{RNAs} \Rightarrow \text{Protein} \Rightarrow \text{Functions}$$

2.1.1 DNA

DNA stands for deoxyribonucleic acid. DNA is an extremely long molecule that forms a double-helix. The double-helix backbone of the molecule consists of sugars and phosphates, and there is one base attached to each sugar. There are four types of bases: Cytosine (C), Guanine (G), Adenine (A), and Thymine (T). The DNA consists of two strands, and each base attached to one strand forms a bond with a corresponding base on the other strand. A only links with T and C links with G. A triplet of bases encodes an amino acid. Protein is a sequence of amino acids, and the functional subunit of DNA that encodes a protein is called a gene [Alberts, 2002]. In the cell, DNA is the core of the chromosome. The human genome is stored on 23 chromosome pairs. Twenty-two of these are autosomal chromosome pairs, while the remaining pair is sex-determining. The *haploid* human genome occupies a total of just over 3 billion DNA base pairs and has a data size of approximately 750 Megabytes[1]. Figure 2.1 shows the flow from DNA to chromosomes and finally to cell.[2]

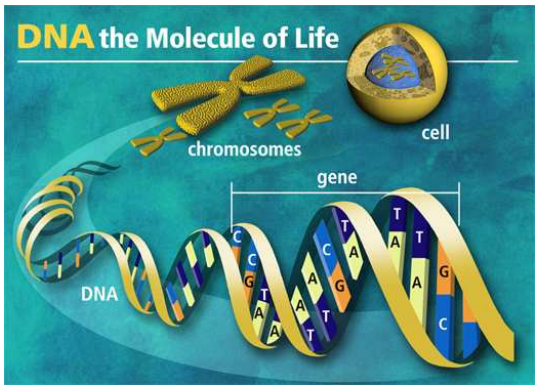

Figure 2.1: DNA - Molecule of life.

[1] http://en.wikipedia.org/wiki/Human_genome
[2] http://www.genetic-identity.com/Basic_Genetics/basic_genetics.html

2.1.2 Gene Expression

Gene expression is a two-step process in which DNA is converted into a protein it encodes. The first step is DNA transcription. In this step, the information from the archival copy of DNA is imprinted into short-lived mRNA. The structure of RNA is a little different, it contains ribose instead of deoxyrybose, and the four bases that bind to it are Cytosine (C), Guanine (G), Adenine (A) and Uracil (U). During the transcription, DNA unfolds, and mRNA is created by pairing mRNA bases with the bases of RNA. In this process, C in DNA translates to G, G to C, A to U, and T to A. After mRNA is translated, it is transported to the ribosome.

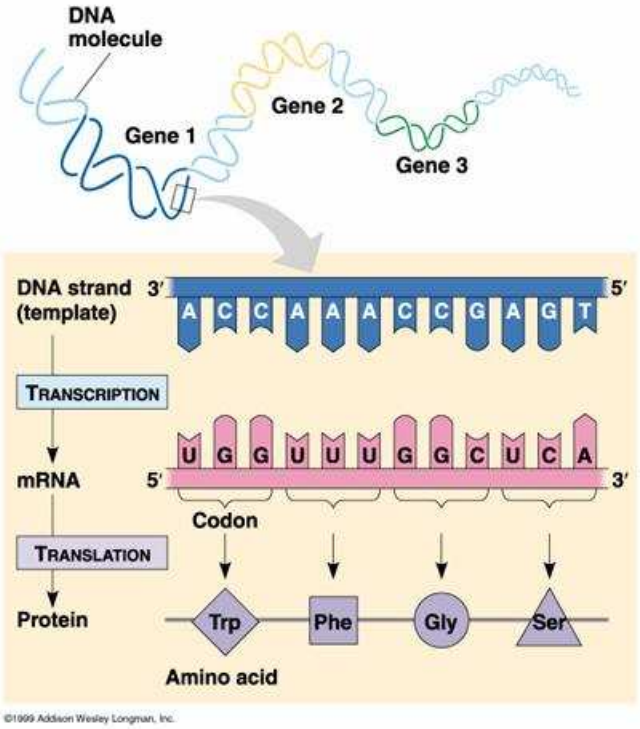

Figure 2.2: The process of gene expression.

The second step, protein translation occurs at the ribosome. During translation, the sequence of codons (triplets of bases) of mRNA is, with the help of tRNA, translated into a sequence of amino acids. Since many diseases result from complex changes on the molecular level, we need to observe and model these processes on the system level. Gene regulatory circuits are an example of machinery that allows us to depict gene expression graphically [Alberts, 2002]. Figure 2.2 illustrates the process of gene expression[3].

2.2 Molecular interactions

Proteins with appropriate structure can interact with other molecules to perform specific functions. Life is based on molecular interactions: underlying every biological process there is a multitude of proteins, nucleic acids, carbohydrates, hormones, lipids, and cofactors, binding to and modifying each other, forming complex frameworks and assemblies, and catalyzing reactions. Molecular interactions can be:

1. *protein-nucleic acid interactions:* proteins bind to DNA and RNA that mediate a number of processes, including regulation of gene expression, gene transcription, DNA replication, and mRNA intron splicing.

[3]http://fajerpc.magnet.fsu.edu/Education/2010/Lectures/26_DNA_Transcription.htm

2. *protein-ligand interactions:* proteins bind to some target molecule or a set of target molecules, and perform some action: enzymes bind to substrate molecules and then catalyze chemical reaction that would otherwise occur too slowly to be biologically useful; some proteins involved in cellular signaling bind to a signal molecule and undergo a conformational change leading to further signaling or changes in cellular processes.

3. *protein-protein interactions:* many proteins function by forming active complexes with each other. The RNA polymerase II complex is an example of such an assembly. Protein-protein interactions are also involved in antibody-antigen binding, large scale organismal motion, and cell adhesion.

2.3 Protein-Protein Interactions and Their Characteristics

Protein-protein interactions are specific interactions between two or more proteins. Figure 2.3 demonstrates the example of enzyme-inhibitor complex; antibody-antigen complex; receptor-ligand interactions, multi-protein complexes such as ribosomes or RNA polymerases [Cramer et al., 2001]. Part A is ribbon representation of the RNA Polymerase II complex structure and part B is schematic interaction diagram for the 10 subunits. The thickness of the connecting lines corresponds to the surface area buried in the corresponding subunit interface. Colors of subunits are identical in (A) and (B).

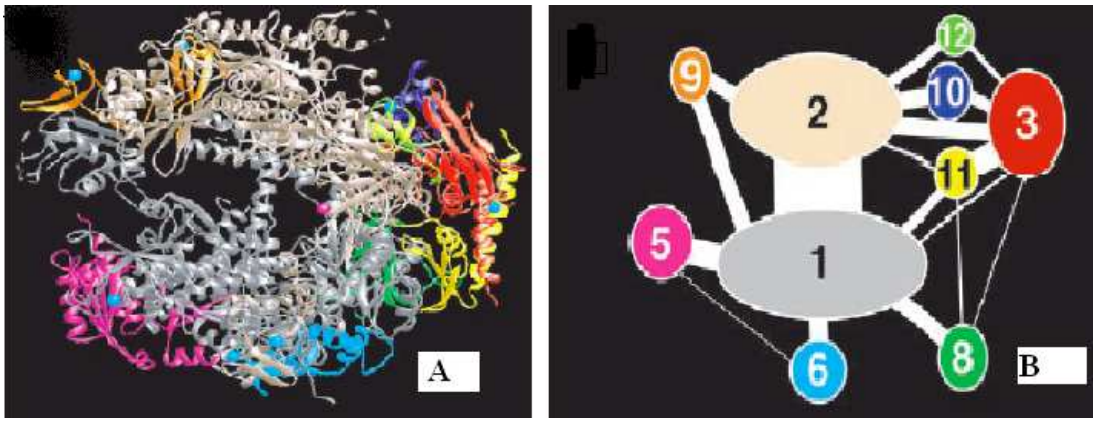

Figure 2.3: Large protein complex and its protein-protein interactions.

2.3.1 Biological Characteristics of Protein-Protein Interactions

The followings are the summary of general characteristic of protein-protein interactions [Uetz and Vollert, 2006].

Classification: Protein-protein interactions can be arbitrarily classified based on the proteins involved (structural or functional groups) or based on their physical properties (weak and transient, non-obligate vs. strong and permanent). Protein interactions are usually mediated by defined domains, hence interactions can also be classified based on the underlying domains.

Universality: All of molecular biology is about protein-protein interactions [Alberts, 2002]. Protein-protein interactions affect all processes in a cell: structural proteins need to interact in order to shape organelles and the whole cell, molecular machines such as ribosomes or RNA polymerases are hold together by protein-protein interactions, and the same is true for multi-subunit channels or receptors in membranes.

Specificity: distinguishes such interactions from random collisions that happen by Brownian motion in the aqeous solutions inside and outside of cells. Note that many proteins are known to interact although it remains unclear whether certain interactions have any physiological relevance. Number of interactions: It is estimated that even simple single-celled organisms such as yeast have their roughly 6000 proteins interact by at least 3 interactions per protein, i.e. a total of 20,000 interactions or more. By extrapolation, there may be on the order of 100,000 interactions in the human body.

Protein-protein interactions and protein complexes: Most protein-protein interactions are detected as interacting pairs or as components of protein complexes. Such complexes may contain dozens or even hundreds of protein subunits (ribosomes, spliceosomes etc.). It has even been proposed that all proteins in a given cell are connected in a huge network in which certain protein interactions are forming and dissociating constantly.

In Figure 2.4, complex networks showing the interactions among proteins help scientists understand how a drug affecting one protein will affect overall cell functioning. This protein network for Brewer's yeast shows which proteins are critical for survival (red), which are important for growth but not critical to survival (orange), which can be removed without slowing growth or killing the cells (green), and which are of unknown importance (yellow)[4].

[4]http://www.sciencenews.org/view/access/id/31301

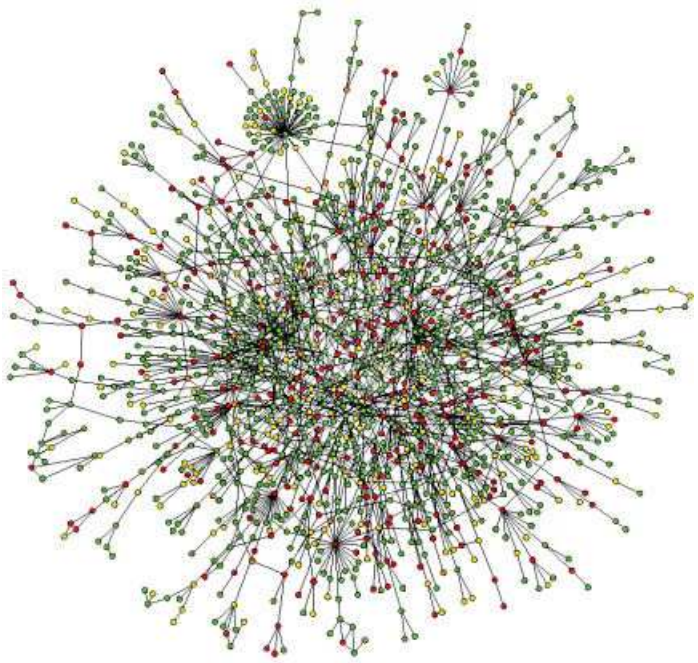

Figure 2.4: Protein network for Brewer's yeast.

2.3.2 Topological Characteristics of Protein-Protein Interaction Networks

The followings are some topological characteristics of protein-protein interaction networks [Lin et al., 2006].

Scale-free network: Protein-protein interactions have the features of a scale-free network, meaning that their degree distribution approximates a power law, $P(k) \sim k^{\gamma}$. In scale-free networks, most proteins participate in only a few interactions, while a few (termed "hubs") participate in dozens of interactions.

Small-world effect: Protein-protein interaction networks have a characteristic property known as the "small world effect", which states that any two nodes can be connected via a short path of a few links. Although the small-world effect is a property of random networks, the path length in scale-free networks is much shorter than that predicted by the small-world effect. Therefore, scale-free networks are "ultra-small". This short path length indicates that local perturbations in metabolite concentrations could permeate an entire network very quickly.

Disassortativity: In protein-protein interaction networks, highly-connected nodes (hubs) seldom directly link to each other. This differs from the assortative nature of social networks, in which well-connected people tend to have direct connections to each other. By contrast, all biological and technological networks have the property of disassortativity, in which highly-connected nodes are infrequently linked each other.

2.4 Protein-Protein Interactions Networks and Signal Transduction Networks

2.4.1 Signal Transduction Networks

Signal transduction is the primary means by which eukaryotic cells respond to external signals from their environment and coordinate complex cellular changes. It plays an important role in the control of most fundamental cellular processes including cell proliferation, metabolism, differentiation, and survival. Extracellular signal is transduced into the cell through ligand-receptor binding, followed by the activation of intracellular signaling pathways that involve a series of protein phosphorylation and dephosphorylation, protein-protein interaction, and protein-small molecules interaction [Liu and Zhao, 2004].

Figure 2.5 shows big number of published papers related to signal transduction in recent years. The total number of papers published in each year since 1977 containing the specific phrase signal transduction in either their title or abstract section, are plotted.[5]

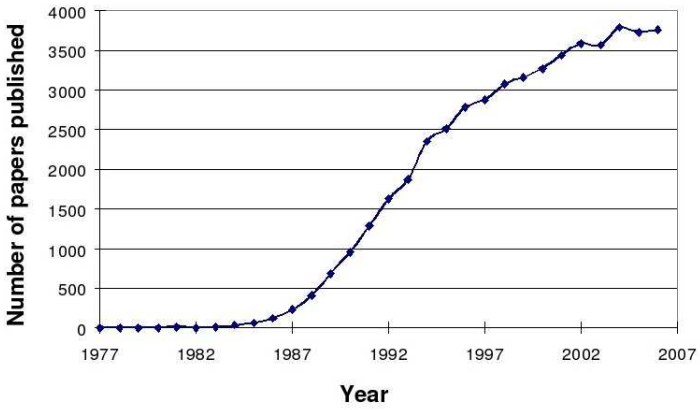

Figure 2.5: Occurrence of the term signal transduction.

The molecular components involved in cellular signaling form signal transduction pathways. A signal transduction pathway affecting a cell is composed of the following events:

- A signaling molecule arrives outside the cell
- A receptor on the extracellular surface of the cell membrane interacts with the signaling molecule
- The receptor interacts with intracellular pathway components, starting a cascade of protein interactions that propagates the signal inside the cell
- A signaling molecule arrives outside the cell.

[5]http://en.wikipedia.org/wiki/signal_transduction

Additionally, an intracellular signaling cascade can no longer be viewed as a linear pathway that relays and amplifies information. It is known that the cell uses these pathways as a way of integrating multiple inputs to shape a uniquely defined output. Hence the interactions of different pathways and the dynamic modulation of the activities of the components within signaling pathways can create a multitude of biological outputs. The cell appears to use these complex networks of interacting pathways and regulatory feedback mechanisms to co-coordinately regulate multiple functions. These outputs allow the cell to respond to and adapt to an ever-changing environment. Due to this increasing complexity, it is often not possible to understand intuitively the systems behavior of signaling networks.

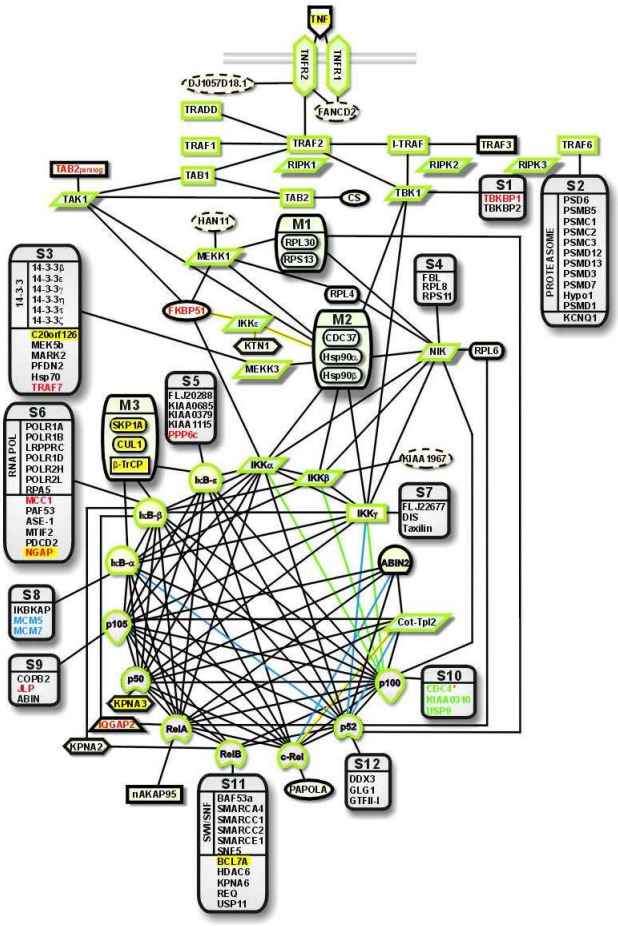

Figure 2.6: A physical map of the human TNFα/NF-κB signal transduction pathway.

Because of their size and complexity, these networks are often too complicated for the human mind to organize and analyze. Therefore, it has become necessary to develop mathematical models to understand the system behavior of signaling networks, and to predict higher order functions that can be validated by experiments [Neves and Iyengar, 2005].

2.4.2 From Protein-Protein Interactions Networks To Signal Transduction Networks

Protein kinases and various transcription factors are known to be crucial in controlling signal transduction processes that are important in mediating cellular functions such as proliferation, differentiation, and apoptosis. Deciphering the complex cascades of binding events in a whole cell will help define signal transduction and metabolic pathways or enzymatic complexes activated by a variety of stimuli. Signal transduction is most likely dependent on specific protein-protein interactions and the identification of specific binding partners could have significant implications in the treatment of cancer and other diseases.

The Figure 2.6 shows the interaction network which is based on a comprehensive Tandem Affinity Purification-Mass Spectrometry (TAP-MS) analysis of proteins implicated in Tα/NF-κB signal transduction pathways.[6]

The first level of complexity in cellular signaling derives from the large number of molecules and multiple types of interactions between them. Considering the read functionality alone, detecting extracellular cues requires several classes of sensor stimulus interactions, typically involving receptor-ligand binding. While some receptors, such as those for growth factors and cytokines, are found on the cell surface, intracellular receptors bind to small molecules such as steroids capable of passing through the cell membrane. Also extracellular matrix (ECM) proteins and proteins found on adjacent cell surfaces also regulate cell functions through their interaction with adhesion and cell-cell contact receptors such as integrins and cadherins, respectively. In addition to the size of the signaling machinery, a second layer of complexity inter-connectivity of signaling biochemistry is apparent from the fact that signaling proteins often contain multiple functional domains, thus enabling each to interact with numerous downstream targets [Eungdamrong and Iyenga, 2004].

2.5 Protein-Protein Interactions Networks and Disease-Causing Genes

2.5.1 Disease Causing-Genes

Most diseases are related in some way to our genes. The information contained in our genes is so critical that simple changes can lead to a severe inherited disease, make us more inclined to develop a chronic disease, or make us more vulnerable to an infectious disease. In Figure 2.7, the rules of governing monogenetic diseases and complex diseases are showed.[7]. While monogenetic diseases are caused by a single gene and complex diseases are complicated combinations of many genes.

Scientists currently believe that single gene mutations cause approximately 6,000 inherited diseases. These diseases are called single gene or monogenic diseases because a change in only one gene causes the disease. These diseases include a number of lung and blood disorders, such as cystic fibrosis, sickle cell anemia, and hemophilia. Although these conditions are not popular however they still affect millions of people worldwide.

[6] http://tnf.cellzome.com/

[7] http://www.genetics.gsk.com/link.htm.

The rules that underlie the inheritance of major common diseases are not as straightforward. These diseases include heart disease, diabetes, Alzheimer disease, psychiatric disorders, and osteoarthritis. These common diseases result not just from a change in one or a few genes, but from a combination of the effects of the environment and a number of susceptibility genes.

Susceptibility genes contribute to an individual's risk of developing a specific disease, but usually are not enough to cause the disease. Susceptibility genes may influence the age of onset of a disease, contribute to its rate of progression, or help to protect against it. Understanding the rules of their inheritance and their roles in disease is not a simple task. Different alleles may be associated with different degrees of susceptibility, or risk. The *APOE* gene on chromosome 19 is one example of a disease susceptibility gene. An individual who has two copies of one variant allele of *APOE* is more likely to develop Alzheimer disease at an earlier age than an individual with a different *APOE* genotype.

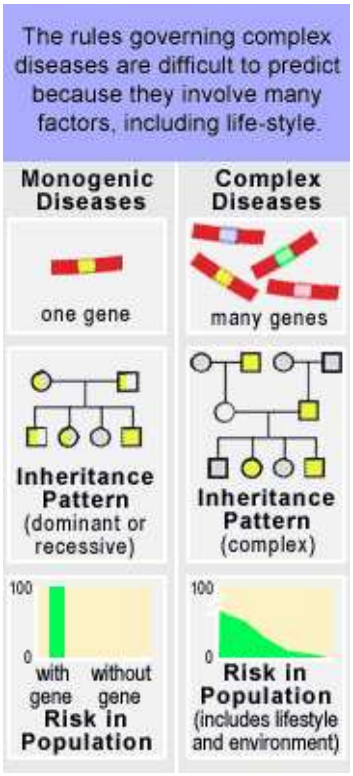

Figure 2.7: Monogenetic diseases and complex diseases.

Even when the genetic basis of a disease is well understood, not much is known about the molecular mechanisms leading to the disorders.

2.5.2 From Protein-Protein Interactions Networks To Disease-Causing Genes

Interactions between specific pairs or groups of proteins are essential to all stages of development and homeostasis. Not surprisingly, many human diseases can be traced to aberrant proteinprotein interactions, either through the loss of an essential interaction or through the formation of a protein complex at an inappropriate time or location. There are some relationships between diseases and protein interactions such as a hitchhiker's guide to pathogen host interactions, normal protein-protein interactions gone wrong, protein protein interaction inhibitors [Ryan and Matthews, 2005].

For oligogenic diseases, synergistic contribution of genes from several loci could explain disruptions in their products, in particular when these proteins are directly or indirectly interacting [Kann, 2007, Ideker and Sharan, 2008]. Two models, namely the dosage and the poison model, have been used to explain the molecular mechanisms of the disruption.

The dosage model explains disruptions of two proteins within a complex. Mutations in one protein alone weaken the interaction but do not affect the phenotype. Only when the two proteins are mutated, the complex is not formed and the phenotype is affected. For instance, mutations that affect ligand-receptor interactions could be explained with such a model. In the poison model, mutations in one of the proteins disrupt the complex but enough of the unchanged complexes are still available to maintain the function. Addition of another mutated subunit will further decrease the already reduced number of normal complexes, resulting in phenotype changes. The molecular models described earlier could be also used to explain indirect interactions between proteins (i.e. proteins that do not physically interact but participate in the same functional pathway).

One example of protein interaction disruption causes is Huntington's disease. In Huntington's disease, an N-terminal region of the protein *huntingtin* (htt) is expanded to contain at least 37 glutamines (*polyQ-htt*). *htt* is a ubiquitously expressed protein that has many known protein interaction partners and a range of functions, including anti-apoptotic effects, transcription regulation, cellular trafficking and neuronal development. It has recently come to light that specific interactions between *polyQ-htt* and other proteins may contribute directly to the neuropathology of Huntington's disease.

Given the highly interlinked internal organization of the cell, it should be possible to improve the single gene single disorder approach by developing a conceptual framework to link systematically all genetic disorders (the human disease phenome) with the complete list of disease genes (the disease genome), resulting in a global view of the *diseasome*, the combined set of all known disorder/disease gene associations [Goh et al., 2007].

In Figure 2.8, Goh *et al.* present the construction of the diseasome bipartite network to show the association between disease networks and gene networks (on the way, protein network as the the product of genes networks). The figure in the center is a mall subset of OMIM-based disorder disease gene associations, where circles and rectangles correspond to disorders and disease genes, respectively. A link is placed between a disorder and a disease gene if mutations in that gene lead to the specific disorder. The size of acircle is proportional to the number of genes participating in the corresponding disorder,andthe color corresponds to the disorder class to which the disease belongs. The figure on the left is the HDN projection of the diseasome bipartite graph, in which two disorders are connected if there is a gene that is implicated in both. The width of a link is proportional to the number of genes that are implicated in both diseases. For example, three genes are implicated in both breast cancer and prostate cancer, resulting in a link of weight three between them. The figure on the right is the DGN projection where two genes are connected if they are involved in the same disorder. The width of a link is proportional to the number of diseases with which the two genes are commonly associated [Goh et al., 2007].

Many human genetic diseases can be caused by multiple genes. Since they lead to

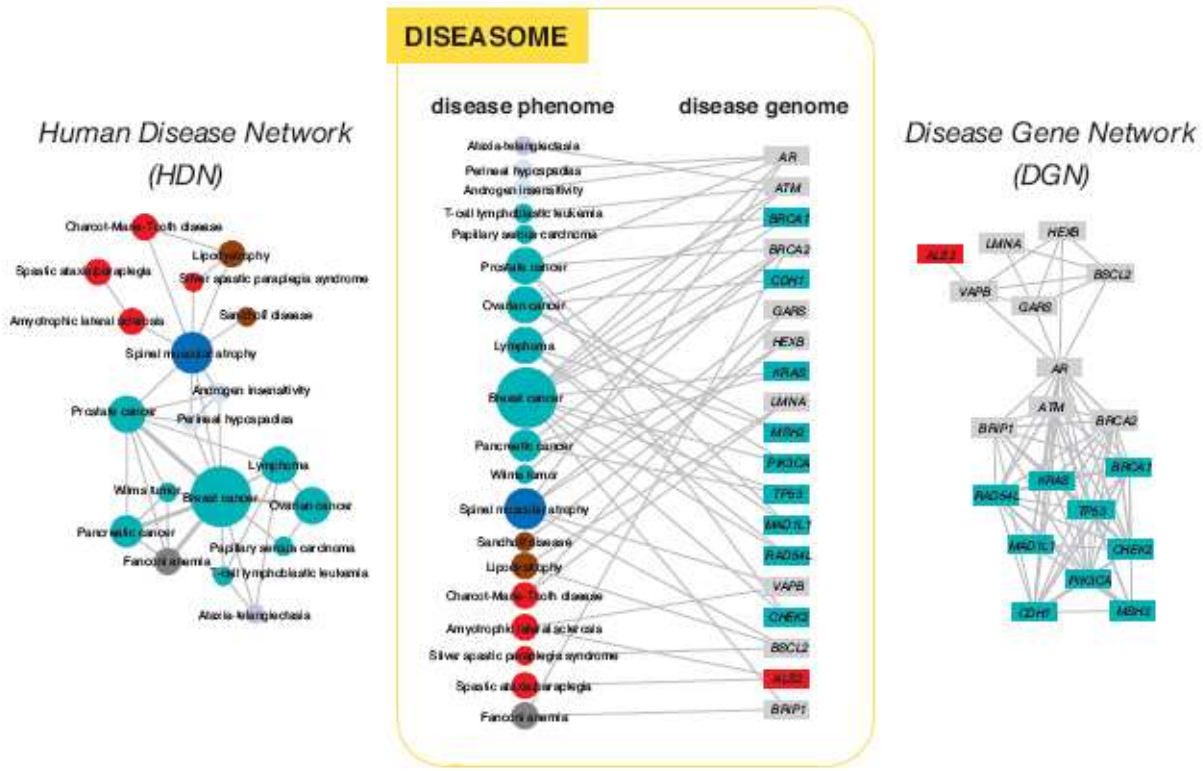

Figure 2.8: Construction of the diseasome bipartite network.

the same or similar disease phenotypes, the underlying genes are likely to be functionally related. Such functional relatedness can be exploited to aid in the finding of novel disease genes. Direct protein.protein interactions are one of the strongest manifestations of a functional relation between genes, so interacting proteins may lead to the same disease phenotype when mutated. Indeed, several genetically heterogeneous hereditary diseases are known to be caused by mutations in different interacting proteins, such as *Hermansky-Pudlak* syndrome and *Fanconi anaemia* [Oti et al., 2006].

2.6 Summary

In this chapter, we first provide some introduction about molecular biology. Then, protein-protein interactions and their particular characteristics are presented. Finally, we would like to give the pictures of PPI and related problems, i.e., signal transduction networks and disease-causing genes. With the key roles of PPI in cells, it is promising to use machine learning and data mining techniques to discover useful knowledge from protein-protein interaction data produced by high-throughput biological techniques for further studies in life science.

Chapter 3

An Integrative Domain-Based Approach to Predicting Protein-Protein Interactions

In this chapter, we present a novel integrative domain-based method for predicting protein-protein interactions using inductive logic programming (ILP). Two principal domain features used were domain fusions and domain-domain interactions. Various relevant features of proteins were exploited from five popular genomic and proteomic databases. By integrating these features, we constructed biologically significant ILP background knowledge of more than 278,000 ground facts. The experimental results through multiple 10-fold cross-validations demonstrated that our method predict protein-protein interactions better than other computational methods in terms of typical performance measures. The proposed ILP framework can be applied to predict domain-domain interactions with high sensitivity and specificity. The induced ILP rules gave us many interesting biologically reciprocal relationships among PPI, protein domains, and PPI related genomic/proteomic features.

3.1 Introduction

Proteins are macro molecules made of twenty amino acids arranged in a linear chain, which participate in every process within cells. Many proteins play a key role in bio-chemical reactions, and structural or mechanical functions, and thus understanding functions of proteins is a main task in molecular biology. Early work has focused on finding protein functions via prediction of protein structures [Bock and Gough, 2001, Matthews et al., 2001, Pellegrini et al., 1999]. Recently, detecting protein functions via protein-protein interactions (PPI) has emerged as a new trend in computational biology [Marcotte et al., 1999, Chen and Yuan, 2006, Baudot et al., 2006]. Protein-protein interaction study is not only crucial in finding protein functions, but also is a significant task, as protein interactions are one of the most important regulatory mechanisms in cells, and most of the cellular processes are coordinated by specific protein interactions. For example, form the physical association between a novel protein and a well-characterized protein, we can infer the functions of the former.

Discovering protein-protein interactions has been a key problem in molecular biology and bioinformatics. Some good surveys about protein-protein interaction research have been available [Ng and Tan, 2003], [Uetz and Vollert, 2006]. Generally, there are experimental and computational methods for prediction of protein interactions. The experimental methods are divided into two groups, the traditional and the high-throughput ones. Traditional experimental methods typically include co-immunoprecipitation and synthetic lethal screening. Although the high-throughput experimental detection methods for PPI (typically, yeast two-hybrid [Ito et al., 2001, Uetz et al., 2000], phage display [Smith, 1985], affinity purification and mass spectrometry[Bauer and Kuster, 2003], and protein micro-arrays)present many advantages over traditional experimental methods, but they are still tedious, labor-intensive and usually have high false positive and high false negative rates.

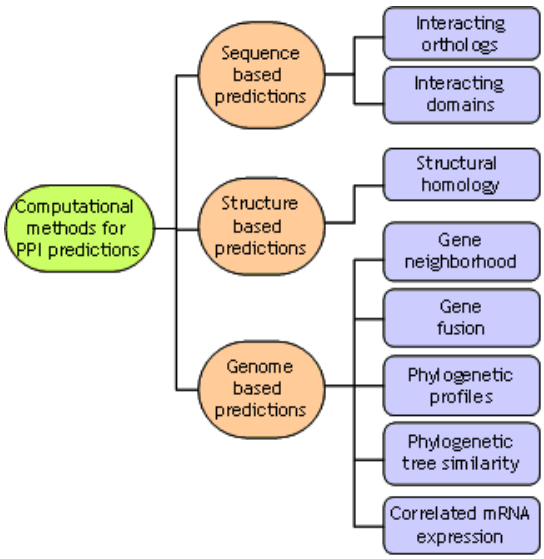

Figure 3.1: An overview of computational methods for PPI prediction.

Computational methods for detecting protein interactions, recently developed with various machine learning techniques and various types of available biological data, al-

low a chance to study more widely and deeply about protein-protein interactions. These works mainly fall into two categories: the homology (or similarity) methods and the non-homology methods. Sharing the same view but dividing more concretely, Ng and Tan grouped these works into three groups, based on information of protein sequences, structures or genomes [Ng and Tan, 2003]. Figure 3.1 synthesizes these three groups. Below is a short description of the methods (the more details are referred to [Ng and Tan, 2003]).

The first group consists of methods that exploit information in amino acid sequences of proteins. As the information of amino acid sequences of most proteins is available, these sequence-based prediction methods currently are widely applicable. There are mainly two approaches of sequence-based methods: one is based on interactions of orthologs, i.e., sequence homology across various species (interacting orthologs) and the other is based on interactions of protein domains (interacting domains).

The second group consists of methods that exploit information on protein structures. This is based on the fact that the three-dimensional shapes of proteins play a major role in their interactions. The key idea is to exploit structure homology, i.e., if two protein A and B interact, and other two proteins A' and B' have similar structures to A and B, respectively, then A' and B' are likely to interact. Though being a powerful approach to protein interaction pre-diction, the structure-based methods have to deal with two problems. One is the lack of docking algorithms for predicting large protein molecules, and the other is the difficulty of protein structure determination, especially the tertiary structures [Ng and Tan, 2003]. These difficulties currently limit the usage of this approach, but also encourage further research to solve them.

The third group consists of methods that exploit genomic data, especially gene locality context (gene neighbourhood or fusion), phylogenetic context (profiles or tree similarity) and gene expression, to study protein interactions. As various genomes have already been sequenced, the genome-based methods are widely applicable.

Based on the same assumption mentioned in three above groups, some early work was based on a single data source [Bock and Gough, 2001, Matthews et al., 2001], [Pellegrini et al., 1999]. Also, recently many bioinformaticians attempted to use multiple data sources, the *integrative approach*, to better predict PPI, such as Bayesian network approach [Jansen et al., 2003], kernel methods [Ben-Hur and Noble, 2005], probabilistic decision tree approach [Zhang et al., 2004], inductive logic programming method [Tran et al., 2005], probabilistic model [Rhodes et al., 2005]. With the rapidly increasing data sources, integrative approaches showed up many advantages in PPI prediction.

Along with the development of integrative methods, *domain-based approach* to the prediction of PPI has received much attention in many ongoing studies. As proteins are assumed to interact through their domains, which are considered to be the building blocks of proteins, a domain-based approach for inferring interactions is adopted, e.g., association

method[Sprinzak and Margalit, 2001, Kim et al., 2002], probabilistic combination-based method [Han et al., 2003], graph-oriented method [Wojcik and Schachter, 2001], random forest framework [Chen and Liu, 2005].

The shortcoming of integrative methods is that they do not take protein domains into account while there are evidences that the biological mechanisms underlying protein-protein interactions are protein domains and their interactions.[Pawson et al., 2002] Furthermore, while domain-based methods all treasured the biological roles of protein domains in PPI prediction, most of them merely considered the co-occurrence of domains/ domain pairs. To predict PPI comprehensively, it is reasonable to employ genomic and proteomic features in domain-based methods.

This work presents a novel integrative domain-based method using inductive logic programming to predict protein-protein interactions. The key idea of this computational method is to integrate protein domain features and multiple genomic/proteomic features into PPI prediction. To integrate efficiently such two kinds of features, we specified two main tasks. The first is to extract as many as possible useful domain and genomic/proteomic features related to PPI. From seven popular databases, we extracted more than 278,000 ground facts of domain fusion, domain-domain interaction features and various biologically significant genomic/proteomic features. The second is to employ inductive logic programming (ILP) on the huge amount of background knowledge to infer PPI effectively.

To demonstrate the advantages of the above mentioned integration, we conducted multiple 10-fold cross validations to compare our method with two other methods based on single domain features, as well as with the non domain-based approach using multiple genomic databases. The performance measures include Receiver Operating Characteristic (ROC) curves, sensitivity and specificity. In all cases, our method performed considerably better than the others did. Sharing the same ILP framework, domain-domain interactions were successfully inferred with high sensitivity and specificity. Lastly, analyzing various produced rules (of both PPI and DDI), many interesting relationships among PPI, DDI, and protein functions, biological processes, were found. Our proposed method can be tuned to predict PPI and DDI for diverse organisms and other genomic and proteomic data sources.

The rest of this chapter is organized as follows. In Section 3.2, we present our proposed method to predict PPI based on domains using ILP and multiple genomic and proteomic databases. The comparative evaluation of the experiments is showed in Section 3.3. Predictive rules of PPI and DDI, as well as discussion, are presented in Section 3.4. Section 3.5 gives some concluding remarks.

3.2 Materials and Methods

In this section, we present our proposed method to predict protein-protein interactions based on domain and multiple genomic/proteomic data using ILP. Two main tasks are: (1) Constructing integrated background knowledge[1] of domain features and multiple genomic/proteomic features, and (2) Learning PPI predictive rules by ILP from the constructed background knowledge. Constructing ILP background knowledge requires two steps. The first one is defining ILP predicates. The second one is extracting ground facts to define predicates extensionally. When choosing a feature, we concentrated on two points: (i) the biological role of that feature in protein-protein interactions or domain-domain interactions, and (ii) the availability of data for that feature. Consulting results of experimental and computational researches on PPI, twenty-two features of protein domains, genes, and proteins were chosen and formulated using ILP predicates. A large database of more than 278,000 ground facts of twenty-two predicates is sufficient for the accurate PPI prediction.

First, we give a brief introduction about Inductive Logic Programming and some bioinformatics applications of ILP in Section 3.2.1. Then, the first task in the proposed method is presented in Subsections 3.2.2, 3.2.3, and 3.2.4. Subsection 3.2.5 describes the second task.

3.2.1 Inductive Logic Programming

Inductive logic programming is an intersection of machine learning and logic programming [Muggleton, 1992]. Inductive logic programming uses logic programming as a uniform representation for examples, background knowledge and hypotheses. Given an encoding of the known background knowledge and a set of examples (positive and negative examples) represented as a logical database of ground facts, an ILP system will derive hypotheses in forms of logical rules that entail the entire positive and none of the negative examples. The schema of ILP is the following

$$\textit{Positive examples} + \textit{Negative examples} + \textit{Background knowledge} \Rightarrow \textit{Hypotheses}$$

From inductive machine learning ILP inherits its goal to develop tools and techniques to induce hypotheses from observations examples and to synthesise new knowledge from experience By using computational logic as the representational mechanism for hypotheses and observations inductive logic programming can overcome the two main limitations of classical machine learning techniques such as the Top-Down-Induction of Decision Tree (TDIDT) family:

[1] the terms 'background knowledge' and 'ground facts' (the second task) are used in terms of the language of inductive logic programming.

1. the use of a limited knowledge representation formalism essentially a propositional logic and

2. difficulties in using substantial background knowledge in the learning processes [Muggleton and Raedt, 1994].

Distinguishing features of ILP are its ability to represent the background (domain) knowledge in the form of logic programs and the expressive power of discovered pattern's language [Dzeroski and Lavrac, 2001].

There have been many ILP systems applied to various problems in bioinformatics. ILP is particularly suitable for bioinformatics tasks because of its ability to take into account background knowledge and work directly with structured data [Page and Craven, 2003]. The ILP system GOLEM was applied to find the predictive theory about the relationship between chemical structure and activity [King et al., 1992]. The training data consisted of 44 trimethoprim analogues and their observed inhibition of E.coli dihydrofolate reductase. Eleven additional compounds were used as unseen test data. GOLEM obtained rules that were statistically more accurate on the training data and on the test data than a previously published linear regression model. Other central concerns of bioinformatics were convincingly solved by ILP, such as protein secondary structure prediction [Muggleton et al., 1993], and protein fold recognition [Turcotte et al., 1998], etc.

3.2.2 Extracting Domain Fusion and Domain-Domain Interaction Data

Protein domains form the structural or functional units of proteins that partake in the intermolecular interactions. The existence of certain domains in proteins can, therefore, suggest the propensity of the proteins to interact or form a stable complex bringing about certain biological functions. Owing to their important biological roles in PPI prediction [Pawson et al., 2002, Marcotte et al., 1999], domain fusion and domain-domain interaction features were used .

Let P denote the set of considered proteins p_i. Let denote D the set of all protein domains d_k that belong to proteins p_i. A pair of interacting proteins (p_i, p_j) is denoted by p_{ij}, and a protein pair that does not interact with each other by $\neg p_{ij}$.

Domains of interacting proteins have more chance to fuse together than domains of non-interacting proteins do. Therefore, once finding a pair of proteins, which have fused domains, we can predict an interaction between them [Enright et al., 1999]. Domain fusion data was extracted from Domain Fusion Database [Truong and Ikura, 2003]. Truong *et al.* employed relational algebra to find domain fusions in protein sequence databases.

We extracted domain fusion data for all protein pairs (p_i, p_j), $p_i, p_j \in P$. The following predicate represents the domain fusion between two proteins

```
domain_fusion(+protein, +protein, #FUSION).
```

Note that in the ILP system used – the learning engine Aleph for proposing hypothesis [2] – there are some *mode declarations* to build the bottom clauses, and a simple mode type is one of the following: (1) *the input variable* (+), (2) *the output variable* (−), or (3) *the constant term* (#). Predicate `domain_fusion` means whether two input proteins, A and B, have fused domains or not. This predicate is supported by a set of ground facts $\mathrm{G}_{domain\_fusion}$, e.g., domain_fusion (ap3m_yeast, ap3b_yeast, yes). After preprocessing step, the set G_{domain_fusion} consists of 2,761 ground facts.

Let d_{kl} and $\neg d_{kl}$ denote a domain-domain interaction and a non-interacting pair respectively. The assumption that proteins interact with each other through interactions of their domains is widely accepted and already validated. To predict PPI more reliably, we extracted DDI data from **iPfam** database [3] which is a resource describing domain-domain interactions observed in PDB entries. When two or more domains occur in a single structure, the domains are analyzed to see if they form an interaction. If the domains are close enough to form an interaction, the bonds forming the interaction are calculated and reported.

We considered two features of DDI. The first feature is whether a protein pair (p_i, p_j) has a domain interaction d_{kl}, and if yes, how many d_{kl} it has. This information is formulated by predicate

```
hasddi(+protein, +protein, #DDI).
```

where the `#DDI` value is the number of DDI mediating the same PPI p_{ij}. The set of ground facts for this predicate G_{ddi} includes 657 ground facts, e.g. hasddi(jsn1_yeast,yip1_yeast,2), and hasddi(msh4_yeast,msh5_yeast,5), etc.

The interacting possibility of one protein may depend on the number of domain-domain interactions occurring on it. Therefore, we considered the relationship between PPI and the number of DDI of each interacting partner. This relationship is presented in the following predicate

```
num_ddi(+protein, #NUM_DDI).
```

Denoted by G_{num_ddi} the set of ground facts of the above predicate contains 505 ground facts. We found that there were proteins having many DDI, for example, num_ddi(did4_yeast, 20) or num_ddi(bud27_yeast, 39), and these proteins potentially interact with many other proteins.

[2] http://web.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph/
[3] http://www.sanger.ac.uk/Software/Pfam/iPfam/

3.2.3 Extracting Proteomic and Genomic Data From Multiple Databases

In addition to domain fusion and domain-domain interaction features, we mined genomic and proteomic data from UniProt database, CYGD database, InterPro database, Gene Ontology database, and Gene Expression database to detect useful genomic and proteomic features for PPI prediction. Table 3.1 shows 19 predicates corresponding to genomic/proteomic data extracted from multiple databases.

As the world's most comprehensive catalog of information on proteins **UniProt database** [Bairoch et al., 2005], provides various protein data, e.g. functions, structures (in Keyword - KW line); regions or sites of interest in the sequences (in Feature Table - FT lines); Enzyme Commission (EC) numbers. Others are pointers to different data collections such as GO, PIR, PROSITE, Pfam, and Interpro database (in Database cross-Reference -DR line). Three predicates present general protein's features that should effect their interactions. Other predicates give references to other databases. Data from different databases related to PPI are bound by these predicates. Some examples of these predicates are keyword(ace1_yeast, transcription regulation), feature(ldb7_yeast, chain chromatin structure remodeling complex), coded_enzyme(uqcr1_yeast, ec1.10.2), and dr_go(twoa5d_yeast,go0005935), etc.

The MIPS Comprehensive Yeast Genome Database (**CYGD**) [4] presents information on the molecular structure and functional network of the entirely sequenced, well-studied model eukaryote, and the budding yeast *Saccharomyces cerevisiae*. Among various information provided by CYGD, the following should be mined to discover the relationship between CYGD's categories and protein-protein interactions, i.e. category of functions, category of subcellular locations, category of phenotypes, category of complexes, and category of proteins. A protein has more chance to interact with proteins in the same category than with proteins in different categories. Here are some examples: subcell_cat (ahc1 yeast, cytoplasm), phenotype_cat(cyk2 yeast, cell cycle defects), etc.

InterPro database [5] is a database of protein families, domains and functional sites. We considered the association between InterPro annotations and GO terms. For example, interpro_go(ipr000009,go0007165), interpro_go(ipr000009,go0000159).

Gene Ontology database [6] has three organizing principles: molecular function, biological process and cellular component. The terms in an ontology are linked by two relationships, *is_a* and *part_of*. The GO relationship between interacting partners may effect their interaction. Some examples are is_a (go0000002, go0007005), part_of (go0000032, go0007047).

[4] http://mips.gsf.de /genre/proj/yeast/
[5] http://www.ebi.ac.uk/interpro/
[6] http://www.geneontology.org/

Table 3.1: Predicates used as background knowledge in various genomic/proteomic data sources

Database	Background knowledge predicates	#Ground fact
UniProt	`keyword(+protein,#Keyword)` A protein has a proteins keyword	43,539
	`feature(+protein,#Feature)` A protein has a protein feature	
	`coded_enzyme(+protein,#EC)` A protein has a enzyme commission number	
	`dr_prosite(+protein, -PROTSITE_ID)` A protein has a PROSITE annotation number	
	`dr_interpro(+protein, -INTERPRO_ID)` A protein has an InterPro annotation number	
	`dr_go(+protein,-GO_TERM)` A protein has a GO term	
	`dr_pfam(+protein, -PFAM_ID)` A protein has an Pfam annotation number	
	`dr_pir(+protein, -PIR_ID)` A protein has a Pir annotation number	
CYGD	`subcell_cat(+protein, #SUBCELLCAT)` A protein has a subcellular structure in which it is found	11,909
	`function_cat(+protein, #FUNCAT)` A protein has a certain function category	
	`protein_cat(+protein, #PROTEINCAT)` A protein has a certain protein category	
	`phenotype_cat(+protein, #FENCAT)` A protein has a certain phenotype category	
	`complex_cat(+protein, #COMPLEXCAT)` A protein has a certain complex category	
InterPro	`interpro_go(+INTERPRO_ID, -GO_TERM)` Relation of InterPro annotations and GO terms	4,965
GO	`is_a(+GO_TERM,-GO_TERM)` `is_a` relation between two GO terms	1,142
	`part_of(+GO_TERM,-GO_TERM)` `part_of` relation between two GO terms	
Gene Expression	`expression(+protein, +protein, #COEFFICIENT)` Gene expression correlation coefficient of two proteins	200,000
DIP	`num_ppi(+protein, +protein, #NUM_PPI)` A protein has a number of protein-protein interactions	13,376
	`ig(+protein, +protein, #IG)` Interaction generality of two proteins is the number of protein that interact with just two considered proteins	

Interacting proteins are often co-expressed, hence gene expression coefficients between two proteins are useful to predict PPI. The **Gene Expression** coefficients between two proteins are referred to Jansen *et al.*'s work [Jansen et al., 2003] that contains 25,000,000 pairwise coefficients for about 18,773,128 protein pairs. In our work, we randomly extracted 200,000 gene expression coefficients in terms of ground facts represented by predicates `expression(+protein, +protein, #COEFFICIENT)` for about 11,000 positives and negatives in the training data sets.

Two last predicates represent information about the number of protein-protein interactions and interaction generality of two interacting partners. Interaction generality is the number of proteins that interact with both interacting partners in an interaction. The interacting pairs are extracted corresponding to these predicates from DIP core data set (see more in Section 3.3.1).

3.2.4 Constructing Background Knowledge for Predicting Protein-Protein Interactions

After defining twenty-two predicates, we exploited data in terms of ground facts for these predicates from seven databases (two databases for domain features and five for genomic/proteomic features).

Algorithm 1 Extracting domain feature data and genomic /proteomic feature data from multiple sources.

Input:
Set of proteins $\{p_i\} \subseteq P$.

Output:
Sets of ground facts $G_L = \{G_l\}, G_l \in \{G_{domain\_fusion}, G_{ddi}, G_{num\_ddi}, G_{UniProt}, G_{CYGD}, G_{InterPro}, G_{GO}, G_{expression}, G_{ig}, G_{num\_ppi}\}$.

Initialize all sets of ground facts $G_l := \emptyset$; $D := \emptyset$.
Extract all domains d_k belonging to proteins p_i; $D := D \cup \{d_k\}$.
for each protein pair (p_i, p_j)
  for all $d_k \in p_i$ and $d_l \in p_j$
    if $fused(d_k, d_l) =$ `true` **then**
      $G_{domain\_fusion} := G_{domain\_fusion} \cup \{(p_i,\ p_j)\}$.
    if $\exists\ d_{kl}$ **then**
      $G_{ddi} := G_{ddi} \cup \{(p_i,\ p_j)\}$
      Count the number of DDI for proteins p_i and p_j for G_{num_ddi}, respectively.
for each protein $p_i \in P$
  Extract $G_{UniProt}$ and G_{CYGD} from UniProt and CYGD database, respectively.
  Extract mapping data between GO terms g_i and Interpro identifiers t_i related to p_i from InterPro database for $G_{Interpro}$; $G_{InterPro} = G_{InterPro} \cup \{t_i, g_i\}$.
for each protein $p_i \in P$
  for each protein $p_j \in P$
    Extract the relationship r_{ij} between GO terms $(g_i,\ g_j)$ related to $(p_i,\ p_j)$ from GO database; $G_{GO} = G_{GO} \cup \{r_{ij}(g_i, g_j)\}$.
    Extract the expression correlation coefficients e_{ij} of (p_i, p_j); $G_{expression} = G_{expression} \cup \{p_i, p_j, e_{ij}\}$.
    Extract the interaction generality of PPI n_{ij} of $(p_i,\ p_j)$; $G_{ig} = G_{ig} \cup \{p_i, p_j, n_{ij}\}$.
    if $\exists\ p_{ij}$ **then**
      $num\_ppi_i := num\_ppi_i\ +\ 1$;
  $G_{num\_ppi} := G_{num\_ppi} \cup \{(p_i, num\_ppi_i)\}$.
return G_L.

In succession, we denote the sets of ground facts extracted from UniProt database, CYGD database, InterPro database, Gene Ontology database, and Gene Expression database by $G_{UniProt}$, G_{CYGD}, $G_{InterPro}$, G_{GO}, and $G_{expression}$, respectively. Algorithm 1 presents the procedure to extract data from multiple databases to construct background knowledge.

3.2.5 Predicting Protein-Protein Interaction With Integrative Domain-Based ILP Framework

Algorithm 2 describes the integrative domain-based ILP framework for predicting PPI from multiple genomic/proteomic databases. After initializing the set of rule R in Step 1, Step 2 and Step 3 are for generating positive and negative example sets $S_{interact}$ and $S_{\neg interact}$, respectively (see more in Subsection 3.3.1). In Step 4, we constructed background knowledge $S_{background}$ with sets of ground facts of twenty-two predicates. In Step 5, of our experiments, Aleph was applied to induce rules.

Algorithm 2 An integrative domain-based ILP framework for PPI prediction

Input:

Set of protein-protein interactions $S_{interact} = \{p_{ij}\}$
Number of negative examples ($\neg p_{ij}$) N
Sets of ground facts $G_L = \{G_l\}, G_l \in \{G_{domain\_fusion}, G_{ddi}, G_{num\_ddi}, G_{UniProt}, G_{CYGD}, G_{InterPro}, G_{GO}, G_{expression}, G_{ig}, G_{num\_ppi}\}$.

Output:

Set of rules R for protein-protein interaction prediction.

1: $R := \emptyset$.
2: Extract positive examples for the set $S_{interact}$.
3: Generate N negative examples $\neg p_{ij}$;
$S_{\neg interact} = \{\neg p_{ij}\}$.
4: **call** Algorithm 1 to generate sets of ground facts G_l;
$S_{background} = G_L = \{G_l\}$.
5: Run an ILP program with $S_{interact}$, $S_{\neg interact}$ and $S_{background}$ to induce the set of rules R.
6: **return** R.

Aleph is an advanced ILP system that uses a top-down ILP covering algorithm. All predicates appearing in hypothesized clauses have to be declared, and amongst them, the target predicate is learned to induce rules. The target predicate in our work is `has_int(+protein, +protein)`, meaning that two arbitrary proteins interact. Aleph learns three inputs (positive examples, negative examples and background knowledge) and induces rules (hypothesized clauses) in terms of the relationships between the target predicate and other predicates declared in background knowledge.

3.3 Evaluation

We concentrated on predicting PPI for *Saccharomyces cerevisiae*, the budding yeast. We did experimental comparative evaluations, two experiments for protein-protein interaction prediction (in Section 3.3.1) and domain-domain interaction prediction (in Section 3.3.2).

3.3.1 Predicting Protein-Protein Interactions

Experiment Design of Protein-Protein Interaction Prediction

To assess the performance of PPI prediction, we firstly did two comparative tests to demonstrate: (1) the advantages of the integration of multiple proteomic and genomic features and (2) the advantages of using protein domain features. The 10-fold cross validation was conducted 10 times with each of two negative data sets to compare our proposed method with other domain-based methods, particularly Association method (AM) and Support vector machines (SVMs) method. Secondly, we conducted 10-fold cross validation tests for ILP method with multiple genomic databases, but not using domain features,[Tran et al., 2005] and then compared those results with our method in terms of sensitivity and specificity.

In two comparative tests with AM and SVMs method, we used the core data of DIP data set [7]. This is a large and reliable set of interactions each of which was observed by at least three different methods. Each interaction in DIP database is originally presented by ORF name (Open Reading Frame). We excluded all the interactions in which bait ORF or/and prey ORF is not found in UniProt database. The final positive data set has 5,512 interacting pairs out of the original 5,963 pairs. We generated two data sets of negatives (5,512 examples for each one) according to two popular methods [Ben-Hur and Noble, 2006]. The first one is the set of random protein pairs that do not belong to the positive data set $S_{interact}$. The second one is the set of protein pairs of which two proteins are located in different subcellular compartments. In the test with the negatives generated by the second method, we excluded predicate `subcell_cat(+protein, #SUBCELLCAT)`. Then, the negative data set of the second test was assured to be independent of the background knowledge.

Result of Protein-Protein Interaction Prediction

With the same training data sets and the same set of extracted protein domains, we conducted 10-fold cross validation tests for our method, AM and SVMs method. AM calculated the probability of protein pairs based on protein domains [Sprinzak and Margalit, 2001]. In our experiment, the probability threshold was set to 0.05. For SVMs method, we used SVM^{light} [Joachims, 1998]. The linear kernel with default values of the parameters was used.

For Aleph, we selected $minpos = 2$ and $noise = 0$, i.e. the lower bound of the positive example range to be covered by an acceptable clause is 2, and there are no negative examples allowed to be covered by an acceptable clause. Other parameters in Aleph were

[7] http://dip.doe-mbi.ucla.edu/

defaulted to have fair comparative comparisons with AM and SVMs method.

The ROC curves of ILP, AM and SVMs methods with 5,512 randomly selected negative examples are shown in Figure 3.2. ROC curve (Receiver Operating Characteristic curve) shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity). The sensitivity of a test is described as the proportion of true positives it detects out all the positives, measuring how accurately it identifies positives. The specificity of a test is the proportion of true negatives it detects of all the negatives, thus is a measure of how accurately it identifies negatives.

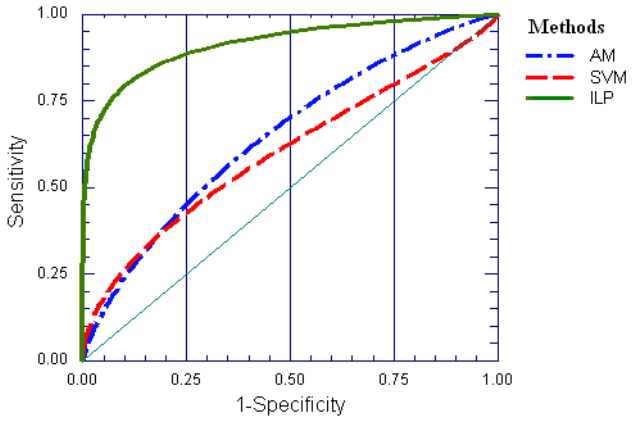

Figure 3.2: Comparative ROC curves of ILP, SVMs and AM with 5,512 random negative examples.

The ROC curve of our method is close to the left-hand border and then the top border of the ROC space, while the ROC curves of AM and SVMs method are close to the 45-degree diagonal of the ROC space. The ROC curve demonstrates that our method performs considerably better than AM and SVMs method do.

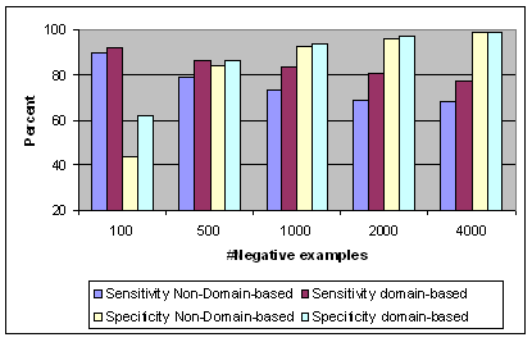

Figure 3.3: Comparison of sensitivity and specificity of non-domain based method and our proposed method with various sets of negative examples by 10 times 10-fold cross-validation.

In the test with negative examples chosen in separate sub-cellular compartments, we carried out 10 trials of 10-folds cross validation, then calculated the average sensitivity (SS) and specificity (SP) of these 10 trials for each of our ILP method, AM, and SVMs method. Our method outperformed with SS 84% and SP 90% compared to AM with SS 82% and SP 34%, and SVMs method with SS 47% and SP 75%.

Reproducing the same experiments to non domain-based approach using ILP [Tran et al., 2005] with the same training negatives (with different numbers of negatives)

and positives (the data set of Ito *et al.* with at least 3 hit interactions), the results of 10 times 10-fold cross-validation are demonstrated in Figure 3.3. They showed that our integrative domain-based method achieved higher sensitivity, and higher or equal specificity, than the non-domain based approach.

Furthermore, the number of unknown interacting protein pairs is, in fact, much larger than the known ones. We also did comparative experiments with imbalanced training sets. According to [Ben-Hur and Noble, 2006], the negative example set should be 4 times larger than the positive example set, thus we randomly selected 2,500 positives from DIP core data set and random 10,000 negatives. Sensitivity and specificity of gained method are 78% and 95% (in this case, sensitivity and specificity of AM are 75% and 30% respectively, and sensitivity and specificity of SVMs methods are 30% and 94%, respectively). As a result, even in testing with imbalanced training data sets, our method effectively predicted PPI.

3.3.2 Predicting Domain-Domain Interactions

Experiment Design of Domain-Domain Interactions Prediction

Domain-domain interaction (DDI) prediction is biologically significant to understand protein-protein interactions in depth. Inheriting the ILP framework for PPI prediction, we applied ILP framework to infer domain-domain interactions. Different from previous works on DDI prediction which exploit only a single protein database, we exploited and combined various domain and protein data. The experimental results of DDI prediction are promising.

To assess the performance for DDI prediction, sensitivity and specificity were evaluated through the 10-fold cross validation tests. We used about 3,000 interactions in InterDom database as positive examples [Ng et al., 2003]. Positive examples are domain-domain interactions in InterDom database that have score thresholds over 100 and are not false positives. Because there is currently no experimental and computational method for detecting non-interacting domain pairs, the negative examples were randomly generated. A domain pair is considered a negative example if the pair does not exist in the interaction set. Various numbers of negatives, 500, 1,000, 2,000 and 3,000 negatives, were chosen. We also implemented the AM and SVMs method to compare sensitivity and sensitivity. We input to AM and SVMs the same databases employed in ILP method. The probability threshold is set to 0.05 for the simplicity of comparison. For SVM method, we used SVM^{light} [Joachims, 1998]. The linear kernel with default values of the parameters was used.

Result of domain-domain interactions prediction

In fact, the interaction of two domains depends on: (i) domain features of interacting partners themselves, and (ii) protein features of host proteins consisting of those domains.

We modeled twenty predicates from seven databases (see more in Supplementary materials [8]). Among those, there are thirteen predicates as protein features extracted from three genomic/proteomic databases of UniProt database, CYGD database, and GO database and seven predicates for domain features corresponding to four domain databases of Pfam[9], PRINT[10], PROSITE [11], and Interpro. In case of domain-domain interaction prediction, we did not use domain-domain interaction and domain fusion data in ILP background knowledge. The target predicate of DDI prediction is `interact_domain(+domain, +domain)`. With more than 100,000 ground facts, we effectively predicted domain-domain interactions by ILP.

Results conducted from 10 times 10-fold cross-validation show that our method obtained higher sensitivity and specificity in the comparison with AM and SVMs. The performance in terms of specificity and sensitivity is also statistically tested by confidence intervals. To estimate 95% confidence interval for each calculated specificity and sensitivity, we used t distribution. Table 2 shows the tested specificity and sensitivity.

Table 3.2: The sensitivity and specificity are obtained for each randomly chosen set of negative examples by 10 times 10-fold cross-validation.

# Neg	Sensitivity			Specificity		
	AM	SVMs	ILP	AM	SVMs	ILP
500	0.49±.027	**0.86**±.010	0.83±.016	0.54±.074	0.24±.004	**0.61**±.075
1000	0.57±.018	0.63±.074	**0.78**±.042	0.44±.033	0.49±.009	**0.68**±.042
2000	0.50±.015	0.32±.014	**0.69**±.027	0.50±.021	0.73±.015	**0.80**±.018
3000	0.49±.021	0.22±.017	**0.62**±.027	0.53±.022	0.81±.013	**0.84**±.010
Avg.	0.51±.020	0.51±.029	**0.73**±.028	0.50±.038	0.57±.010	**0.73**±.036

Besides calculating cross-validated sensitivity and specificity, cross-validated accuracy and precision were considered. All of our experiment results obtained high accuracy and precision. The average accuracy and precision were 0.76 and 0.82, respectively.

[8] http://www.jaist.ac.jp/s0560205/PPIandDDI/
[9] http://www.sanger.ac.uk/Software/Pfam/
[10] http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/
[11] http:// au.expasy.org/prosite/

3.4 Discussion

The experimental results have shown that our method potentially predicts PPI and DDI with high sensitivity and specificity. Furthermore, the induced predictive rules encouraged us to discover many interesting biologically reciprocal relationships among protein-protein interactions and protein domains, and other genomic/proteomic features related to protein-protein interactions. Compare our results to information in biological literatures, we found that ILP induced rules could be further applied to related studies in biology. The full list of both PPI predictive rules and DDI predictive rules and the covered positives are available as Supplemental materials[12] . Figure 3.4 shows some induced rules for PPI prediction.

Studying the rules of PPI prediction that are related to domain-domain interaction, we found many interesting rules (Rule 1, 2, and 3). In Rule 1, if a protein pair has at least one DDI, in which protein A belongs to the subcellular compartment *'cytoplasm'* and protein B has the *'coiled-coil'* domain(s), they have a chance to interact. Similar to Rule 1, Rule 2 shows that two proteins mediated by at least one DDI, A and B, will interact if one of them is located in the compartment *'nucleus'* and the other's functions are *'cell cycle and DNA progressing'*. The evidences for this rules are interactions (kar4_yeast,ime4_yeast), (rsc6_yeast,rsc8_yeast), (cdc23_yeast,leur_yeast), (cdc23_yeast,nip29_yeast).

Considering the group of proteins, which may be required for the production of *pyridoxine* (vitamin B6), i.e., sno1_yeast, snz3_yeast snz1_yeast, and snz2_yeast, we found that each pair in this group has an interaction which satisfies Rule 3 and Rule 4. Rule 3 means interaction of protein A and protein B may occur if the proteins satisfy three conditions. First is that they interact with the same protein. Second is that they have at least one DDI. Third is that one of them is categorized to function catalogue *'cell rescue defense and virulence'*. We know that PPI plays an important role in drug design, so such rules and their evidences are expected to be instrumental in discovering helpful relationships between PPI, DDI and protein functions in pharmaceuticals.

Two rules with large number of covered positives prove that if two proteins, A and B, are located in the same subcellular compartment, protein A potentially interacts with protein B. There are 216 covered positives for *'nucleus'* compartment, 284 ones for *'cytoplasm'* compartment, and 15 ones for *'mitochondria'* compartment. However, among induced rules, we surprisingly found Rule 6 with 37 positives that showed the phenomenon of two proteins being in different subcellular locations but interacting. This phenomenon can occur when there is a certain translocation or post-translation modification of proteins in different subcellular compartments. Some evidences are interactions (lsm2_yeast, pat1_yeast), (ntc20_yeast, syf1_yeast), (yb89_yeast, yp33_yeast), (kar4_yeast, ime4_yeast),

[12]http://www.jaist.ac.jp/s0560205/PPIandDDI/

Rule 1 [Pos cover = 26 Neg cover = 0]
$has\_int(A, B) :- \quad hasddi(A, B, C), gteq(C, 1), subcell\_cat(A, cytoplasm),$
$keyword(A, coiled\ coil).$

Rule 2 [Pos cover = 8 Neg cover = 0]
$has_int(A, B) :- \quad hasddi(A, B, C), gteq(C, 1), subcell\_cat(B, nucleus),$
$function\_cat(A, cell\ cycle\ and\ DNA\ processing).$

Rule 3 [Pos cover = 8 Neg cover = 0]
$has\_int(A, B) :- \quad ig(A, B, C), C = 1, hasddi(A, B, D), gteq(D, 1),$
$function\_cat(B, cell\ rescue\ defense\ and\ virulence).$

Rule 4 [Pos cover = 8 Neg cover = 0]
$has\_int(A, B) :- \quad domain\_fusion(A, B, yes),$
$feature(B, chain\ probable\ pyridoxin\ biosynthesis\ protein).$

Rule 5 [Pos cover = 199 Neg cover = 0]
$has\_int(A, B) :- \quad dr\_go(B, C), part\_of(C, D),$
$domain\_fusion(A, B, yes).$

Rule 6 [Pos cover = 37 Neg cover = 0]
$has\_int(A, B) :- \quad subcell\_cat(B, nucleus), subcell\_cat(A, cytoplasm),$
$function\_cat(A, transcription).$

Rule 7 [Pos cover = 14 Neg cover = 0]
$has\_int(A, B) :- \quad num\_ppi(A, C), C = 4, subcell\_cat(A, cytoplasm),$
$function\_cat(A, protein\ fate).$

Rule 8 [Pos cover = 5 Neg cover = 0]
$has\_int(A, B) : -function\_cat(B, protein\ fate), expression(A, B, C),$
$gteq(C, 0.688153).$

Rule 9 [Pos cover = 4 Neg cover = 0]
$has\_int(A, B) :- \quad ig(A, B, C), C = 2, protein\_cat(A, polymerases),$
$phenotype\_cat(A, conditional\ phenotypes).$

Rule 10 [Pos cover = 92 Neg cover = 0]
$has\_int(A, B) :- ig(A, B, C), C = 1, hasddi(A, B, D), D = 1.$

Figure 3.4: Some induced rules obtained with $minpos = 3$.

(sas10_yeast, yb9x_yeast).

Rule 8 means that if there is a high correlation between two proteins and one of them has a function as *'protein fate'*, they have more chance to interact with each other than random protein pairs. This rule reconfirms the theory that the expression of two interacting proteins is highly correlated. Some of evidences of this rules are interactions (uqcr1_yeast,uqcr2_yeast), (rsc6_yeast,grpe_yeast), (ady3_yeast,atg17_yeast), (va0d_yeast,vate_yeast), (yn97_yeast,bsc5_yeast)

From the analysis of DDI predictive rules, some interesting associations between DDI and other domain and protein features are discovered.

Related to the domain feature *'motif compound'*, we found that the more motifs a domain has, the more interactions the domain has with other domains. This means that domains, which have many conserved motifs, tend to interact with others. The interactions having these domains play an important role in forming stable domain-domain interactions in particular, and protein-protein interactions in general.[Moon et al., 2005] If two domains, A' and B', the domain A' has a PRINTS annotation C, and C is with eight motifs and the rest domain B' belongs to proteins categorized in function category *'protein synthesis'*, they interact. This rule covers 23 positives

interact_domain (A',B') :- prints (A', C), motif_compound
(C, compound(8)), function_category (B', protein synthesis).

The combination of inductive rules of ILP will be very useful to understand not only PPI and DDI, but also protein functions, and biological processes.

3.5 Summary

In this chapter, we have presented an integrative domain-based method using ILP and multiple genome databases to predict protein-protein interactions. The ILP framework was extended for domain-domain interaction prediction. The experimental results demonstrated that our proposed method could produce more comprehensible rules and outperformed other methods in protein-protein interaction prediction as well as domain-domain interaction prediction.

In future, we would like to do more comparative evaluation with other methods. We would like to investigate further the biological significance of novel protein-protein interactions obtained by our method, especially the induced rules. Other work is applying the ILP framework to other important tasks in biomedicine, such as determining protein functions, determining the sites, interfaces of PPI, etc..

Chapter 4

Constructing Signal Transduction Networks Using Multiple Signaling Feature Data

The essence of STN is underlain in some signaling features scattered in various data sources and biological components overlapping among STN. The integration of those signaling features presents a challenge. Most of previous works based on PPIs for STN did not much take the signaling properties of signaling molecules and components overlapping among STN into account. This chapter describes an effective computational method that can exploit three biological facts of STN applied to human: rich-information of protein-protein interaction networks, signaling features and sharing components. To this end, we introduce a soft-clustering method for doing the task by exploiting integrated multiple data, especially signaling features, i.e., protein-protein interactions, signaling domains, domain-domain interactions, and protein functions. The gained results are considerable showing that the method is promising to discover new STN and solve other related problems in computational and systems biology from large-scale protein interaction networks. Other interesting results of early work on yeast STN are presented to show the advantages of using signaling domain-domain interactions.

4.1 Introduction

Signal transduction networks are the primary means by which eukaryotic cells respond to external signals from their environment as well as coordinate complex cellular changes, and are crucial for inter- and intra-cellular signaling [Allen et al., 2006]. These networks are also important in the correct functioning of the cell and can produce appropriate outcomes, such as cell division, apoptosis, or differentiation in response to a variety of biological signals.

Because of the biologically significant roles of STN in cells, both biologists and bioinformaticians have taken much interest in finding out molecular components and/or the relations among these molecular components in STN. Experimental methods have been effective in generating detailed descriptions of specific linear signaling pathways; however our knowledge of complex signaling networks and their interactions remains incomplete [Asthagiri and Lauffenburger, 2000]. Recently, the enormous amount of high-throughput protein-protein interaction (PPI) data, one of important signaling features, has been generated and provided invaluable resources for STN study [Ito et al., 2001], [Liu and Zhao, 2004]. Consequently, there is a great need for developing computational methods to take advantage of information-rich protein interaction data to study complex signaling mechanisms inside STN.

Computational modeling has emerged as a powerful descriptive and predictive tool that allows the study of complex systems. This approach is becoming increasingly useful in many areas of biology, including in the study of signaling pathways given the identification of a growing number interactions within and between signaling pathways in the cell. The explicit modeling approach should allow the monitoring of the effects of multiple signal inputs that may arrive simultaneously and/or sequentially and the subsequent processing and integration of these signals. Such analysis would lead to understanding of the complexity underlying the higher order functions of signaling networks, and may even help identify novel properties that would not be observable by the study of isolated signaling pathways [Neves and Iyengar, 2005].

Recently, the enormous amount of protein-protein interaction data [Liu and Zhao, 2004], [Ito et al., 2001] has been generated and provides invaluable resources for STN study [Ng and Tan, 2003]. Consequently, there is a great need for developing computational methods to take advantage of information-rich protein interaction data for understanding complex signaling mechanisms inside signal transduction networks.

Constructing STN based on PPI is an area of much ongoing research. A statistical model, based on representing proteins as collections of domains or motifs, which predicts unknown molecular interactions within these biological networks was proposed by Gomez *et al.* [Gomez et al., 2001]. Using Markov chain Monte Carlo method, they then modeled

the signal transduction networks (STN) in terms of domains in upstream and downstream protein interactions. Steffen *et al.* developed a computational method for generating static models of STN which utilizes PPI maps generated from large-scale two-hybrid screens and expression profiles from DNA microarrays [Steffen et al., 2002]. Liu *et al.* applied a score function that integrated protein-protein interaction data and microarray gene expression data to predict the order of signaling pathway components [Liu and Zhao, 2004]. Concerning protein modification time-course data, Allen *et al.* applied a method of computational algebra to modeling of signaling networks [Allen et al., 2006]. Another work by Fukuda *et al.* is to represent the model of signal transduction pathways based on a compound graph structure. Their method is designed to capture directly the structure of pathways that biologists bear in mind or that are described in articles [Fukuda and Takagi, 2001]. Based on the signaling domain-domain interactions, Nguyen and Ho proposed a method that takes advantage of singling features of molecules to discover STN [Nguyen and Ho., 2006]. One of the most recent works is to search for the optimal subnetworks from PPI according to some cost functions [Zhao et al., 2008].

Although the previous work achieved many results, there are still some biological characteristics of STN that the previous works did not take much into account. First, it is known that the deeper level underlying the PPI to transmit signals in cells are functional domains, so-called signaling domains, and their interactions [Pawson et al., 2002], [Eungdamrong and Iyenga, 2004]. Data of those significant signaling features are structured, complexly relational, and sparse in different data sources. In order to construct STN effectively, those data is needed to be appropriately integrated. Second, STN indeed have many overlapping components including proteins and their interactions
[Neves and Iyengar, 2005]. This work aims to solve those two intricate problems of STN to better construct STN from PPI networks. To this end, we introduce an effective computational method to construct STN that (1) exploits integrated multiple signaling features of STN from heterogenous sources, i.e., protein-protein interactions, signaling domains, domain-domain interactions, and protein functions, (2) detects overlapping components using soft-clustering. Additionally, in previous work clustered objects were often individual proteins, but our method handled clustered objects as the functional or physical protein interactions because these interactions are the means to transmit signals in cells.

We evaluated the proposed method using human protein interaction network published in the database Reactome. Five complex biological processes were tested to demonstrate the performance. The clustered results are well-matched with these five processes. To the best of our knowledge, this work is the first one that computationally solves the STN problem for *Homo Sapiens*. The preliminary results open a prospect to study other problems related to complex biological systems in *Homo Sapiens*.

The rest of the chapter is organized as follows. In Section 4.2, we present our pro-

posed method to construct STN from human PPI networks and multiple databases using soft-clustering. The evaluation of the experiments is showed in Section 4.3. Experimental results, as well as discussion, are presented in Section 4.3.2. In Section 4.3.2, we first present current results for human STN, then some interesting results for yeast is summarized. Section 4.5 gives some concluding remarks.

4.2 Materials and Methods

The method does two main tasks. The first one is to extract and preprocess signaling feature data from various data sources. Those relational data in heterogenous types are then weighted and normalized by the proposed functions. Based on data extracted in the first task, the second is to combine weighted data and then cluster protein-protein interactions into STN using soft-clustering. In this section, Subsection 4.2.2 and Subsection 3.2.5 describe two mentioned tasks in succession.

4.2.1 Soft-clustering and PPI Networks

Clustering methods can be divided into hierarchical and partitioning ones. In partitioning clustering, there are two categories of hard-clustering and soft-clustering. On the one hand, hard-clustering is based on classical set theory and assigns an instance to exactly one cluster, e.g., k-means, SOMs, etc. On the other hand, soft-clustering can assign an instance to several cluster and differentiate grade of representation (cluster membership), e.g., fuzzy c-means, HMMs, etc. [Futschik and Carlisle, 2005].

In PPI networks, many proteins are believed to exhibit multiple functionalities in several STN, interacting with different groups of proteins for different functions. As a result, soft-clustering is appropriate to generate STN in terms of overlapping clusters which share common interactions. Some soft clustering methods are well applied to PPI networks.

The line graph generation is one of soft clustering techniques and has a number of attractive features [Lin et al., 2006]. It does not sacrifice informational content, because the original bidirectional network can be recovered at the end of the process. Furthermore, it takes into account the higher-order local neighborhood of interactions. Additionally, the graph it generates is more highly structured than the original graph. Finally, it produces an overlapping graph partitioning of the interaction network, implying that proteins may be present in multiple functional modules.

Ucar *et al.*'s work proposed a soft clustering method using hub-induced subgraphs [Ucar et al., 2006]. Their approach consists of two stages. In the first stage, they refine the PPI graph to improve functional modularity, using hub-induced subgraphs. They

employ the Edge betweenness measure to identify dense regions within the neighborhoods. In the second stage, they cluster the refined graph using traditional algorithms. Their end goal is to isolate components with high degree of overlap with known functional modules. An additional advantage of the refinement process is its ability to perform soft clustering of hub proteins. Owing to this approach, they improved functional modularity in PPI network.

Other soft clustering for PPI is an ensemble framework [Asur et al., 2007]. They construct a variant of the PCA-agglo consensus algorithm to perform soft clustering of proteins, which allows proteins to belong to multiple clusters. The hard agglomerative algorithm places each protein into the most likely cluster to satisfy a clustering criterion. However, it is possible for a protein to belong to many clusters with varying degrees. The probability of a protein belonging to an alternate cluster can be expressed as a factor of its distance from the nodes in the cluster. If a protein has sufficiently strong interactions with the proteins that belong to a particular cluster, then it can be considered amenable to multiple memberships.

4.2.2 Extracting signaling feature data from multi-data sources

STN have a complex two-level signaling machinery. The first level of complexity in cellsular signaling constructs from the large number of molecules and multiple types of interactions between them. The second layer of complexity of signaling biochemistry is apparent from the fact that signaling proteins often contain multiple functional domains, thus enabling each to interact with numerous downstream targets [Eungdamrong and Iyenga, 2004]. Considering these complexities, we extracted the following structured data of signaling features.

1. Protein-protein interactions (PPI): the upper level consists of the components as interfaces to transmit signals. PPI data were extracted from Reactome database[1].

2. Domain-domain interactions (DDI): the deeper level consists of the functional domains that perform as the basic elements in signal transduction. DDI data were extracted from iPfam database[2].

3. Signaling domain-domain interactions: the functional level consists of signaling domains (specific functional domains) that act as key factors to transduce signals inside STN. Signaling DDI data were extracted from SMART database[3] and referred in [Pawson et al., 2002].

[1] www.reactome.org/
[2] www.sanger.ac.uk/Software/Pfam/iPfam/
[3] smart.embl-heidelberg.de/

Table 4.1: List of signaling features and their corresponding data sources.

Feature	Database	Description of database
Protein-protein interactions	Reactome database	An online bioinformatics database of biology described in molecular terms. The largest set of entries refers to human biology.
Domain-domain interactions	iPfam database	A resource describing domain-domain interactions observed in PDB entries.
Signaling domains	SMART database and [Pawson et al., 2002]	SMART allows the identification and annotation of genetically mobile domains and the analysis of domain architectures.
Function of protein	Uniprot database	The world's most comprehensive catalog of information on proteins.

Functions of proteins in STN were also extracted from Uniprot database[4] in terms of keywords.

The extracted data are in different types, e.g., the numerical type for number of PPI, interaction generality, number of signaling DDI or categorical type for protein functions. Those data have complex relations, such as a protein may have many interactions and then each interaction may have many DDI. In a domain interaction, interacting partners may be a signaling domain or not. To exploit these relations, after extracting data from multi-data sources, we weighted and normalized these relational data by weight functions. Table 4.2 shows these proposed weight functions and the corresponding explanations.

- PPI weight function (w_{ppi}): The topological relation of proteins in the PPI network was extracted in terms of the numbers of interactions of each partner and the interaction generality.

- Signaling DDI weight function (w_{Sddi}): The relation between a PPI and their domains was exploited to study more deeply STN in terms the number of DDI and signaling DDI mediating this interactions.

- Keyword weight function (w_{func}): The relation of a PPI and protein functions was considered in terms of the keywords tagged in each partner and the keywords shared between them.

[4]www.uniprot.org/

Table 4.2: Signaling features and their weight functions.

Weight functions	Notations and explanation
$w_{ppi}(p_{ij}) = \frac{g_{ij}^2}{n_i * n_j}$	g_{ij}: Interaction generality, the number of proteins that interact with just two interacting partners, p_i and p_j. n_i: The number of protein-protein interactions of the protein p_i.
$w_{Sddi}(p_{ij}) = \frac{n_{Sddi}+1}{n_{ddi}+1}$	n_{Sddi}: The number of signaling domain-domain interactions shared between two interacting proteins. n_{Sddi}: The number of domain-domain interactions shared between two interacting proteins.
$w_{func}(p_{ij}) = \frac{k_{ij}^2}{k_i * k_j}$	k_{ij}: The number of sharing keywords k_{ij} of two interacting partners, p_i and p_j. k_i: The number of keywords of the protein p_i.

4.2.3 Combining signaling feature data to construct STN using soft-clustering

After weighing signaling features, it is necessary to combine them all in a unified computational scheme to take advantage of those data. We integrated these data and represented them in forms of feature vectors. Each interaction has its own feature vector that has three elements corresponding to three features, $v_{ij} = \{w_{ppi}, w_{Sddi}, w_{func}\}$. Subsequently, we employed a soft-clustering algorithm to cluster the interactions based on their features vectors. Soft-clustering can construct STN and detect the overlapping components that can not be found by traditional hard-clustering. Note that we used Mfuzz software package [Kumar and Futschik, 2007] to implement fuzzy c-means (FCM) clustering algorithm in our experiments. Fuzzy c-means (FCM) clustering algorithm is a popular soft-clustering algorithm.

Figure 3 summarizes the key idea of our method that does (1) extracting and weighing signaling features and (2) integrating and soft-clustering them into STN. Given a large protein-protein interaction network $\mathfrak{N}$, the outputs of our method are STN, which are the subgraphs of edges as protein interactions and node as proteins. Step 1 is to obtain the binary interactions from the protein-protein interaction network $\mathfrak{N}$. From Step 2 to Step 5 is to do the first task, extracting and then weighing signaling data features by functions shown in Table 4.2. These steps were done for all binary interactions to exploit the relations between PPI and signaling features. Step 6 and Step 7 are to perform the second task, combining weighted feature data, representing them in forms of feature vectors $v_{ij} = \{w_{ppi}, w_{Sddi}, w_{func}\}$ and lastly doing soft-clustering into STN $\mathcal{S}$. STN $\mathcal{S}$ are returned in Step 8.

Algorithm 3 The proposed method to construct STN from PPI networks using soft-clustering and multi-signaling feature data.

Input:

Protein-protein network $\mathfrak{N}$.

Set of features $\mathcal{F} \subset \{f_{ppi}, f_{Sddi}, f_{func}\}$.

Output:

Set of signal transduction networks $\mathcal{S}$.

1: Extract binary interactions $\{p_{ij}\}$ from the protein-protein network $\mathfrak{N}$. $\mathcal{P} := \{p_{ij}\}$.

2: For each interaction $p_{ij} \subset \mathcal{P}$

3: Extract and formalize data for the PPI data feature f_{ppi}

Calculate the number of interactions n_i, n_j of each interacting partner p_i and p_j, respectively.

Calculate the interacting generality g_{ij} of interaction p_{ij}.

Weigh the feature f_{ppi} by the numbers n_i, n_j, and g_{ij}.

4: Extract and formalize data for the signal DDI feature f_{Sddi}

Calculate the number of sharing domain-domain interactions n_{ddi} of two interacting partners, p_i and p_j.

Calculate the number of sharing signaling domain-domain interactions n_{Sddi} of two interacting partners, p_i and p_j.

Weigh the feature f_{Sddi} by the numbers n_{ddi}, n_{Sddi}.

5: Extract and formalize data for the function data feature f_{func}

Calculate the number of keywords k_i, k_j of each interacting partner p_i and p_j, respectively.

Calculate the number of sharing keywords k_{ij} of two interacting partners, p_i and p_j.

Weigh the feature f_{func} by the numbers k_i, k_j, and k_{ij}.

6: Combine and represent the all features in the feature vectors $v_{ij} = \{f_{ppi}, f_{Sddi}, f_{func}\}$.

7: Apply a soft-clustering algorithm with the set of feature vectors $\{v_{ij}\}$ to cluster interactions p_{ij} into signal transduction networks $\mathcal{S}$.

8: **return** $\mathcal{S}$.

4.3 Evaluation

To evaluate the performance of the method, we consider a complex PPI network to detect STN out of other biological processes. The tested PPI network does not contain only signaling processes but also other biological processes functioned inside the network as the nature in cells. The clustered results should reflect this complicated phenomena, well construct signaling processes and find overlapping components. We extracted five heterogeneous processes in Reactome database and the experimental results demonstrated that our method effectively constructed signaling processes from the PPI network. At

the end of this section, we also shortly present some results achieved for yeast STN included (1) signaling domain-domain interaction prediction, (2) yeast MAPK pathways reconstruction.

4.3.1 Experiments for Human STN construction

The Reactome database consists of 68 *Homo sapiens* biological processes of 2,461 proteins. They also published 6,188 protein interactions, among those there are 6,162 interactions participating in biological processes. Investigating known biological processes in Reactome database, there are 636 proteins partaking in at least 2 different processes, 400 proteins in at least 3 processes, 119 proteins in 5 processes. These phenomenon prove that there exists lot of proteins and their interactions overlapping among these processes.

In our experiments, we extracted a group of five biological processes which have from 30 to 50 proteins and include signaling networks. Table 4.3 shows some information related to these five processes. Totally, this group consists of 145 distinct interactions of 140 distinct proteins. Among these processes, there are overlapping interactions and proteins. Figure 4.1 illustrates the interaction network of five processes.

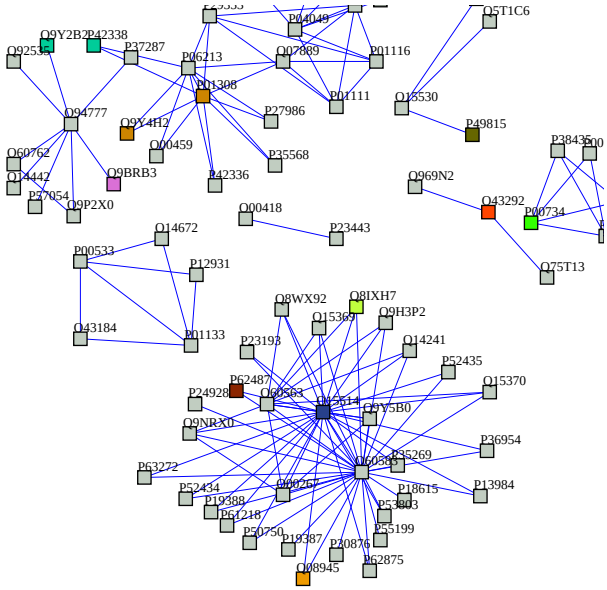

Figure 4.1: Protein interaction networks of the five testing processes.

Proteins partaking in these processes are extracted and looked for their interactions

Table 4.3: Five tested biological processes and some related information.

Reactome annotation	Description	#Proteins	#Interactions
REACT_1069	Post-translational protein modification	40	23
REACT_1892	Elongation arrest and recovery	31	68
REACT_498	Signaling by Insulin receptor	39	44
REACT_769	Pausing and recovery of elongation	31	68
REACT_9417	Signaling by EGFR	40	25

in the Reactome interactions set. We strictly extracted only the interactions that have both interacting partners joining in processes because the method considers the proteins but more importantly their interactions. The extracted interactions and their signaling features were then input in the soft-clustering algorithm.

In this work, we applied Mfuzz software package to run fuzzy c-means (FCM) clustering algorithm. It is based on the iterative optimization of an objective function to minimize the variation of objects within clusters [Kumar and Futschik, 2007]. As a result, fuzzy c-means produces gradual membership values μ_{ij} of an interaction i between 0 and 1 indicating the degree of membership of this interaction for cluster j. This strongly contrasts with hard-clustering, e.g., the commonly used k-means clustering that generates only membership values μ_{ij} of either 0 or 1. Mfuzz is constructed as an R package implementing soft clustering tools. The additional package Mfuzzgui provides a convenient TclTk-based graphical user interface.

Concerning the parameters of Mfuzz, the number of clusters was 5 (because we are considering 5 processes) and the so-called fuzzification parameter μ_{ij} was chosen 0.035 (because the testing data is not noisy).

4.3.2 Experimental Results and Discussion for Human STN construction

Actually, two processes REACT_1892 and REACT_498 share the same set of proteins and the same interactions as well. Also, two signaling processes, REACT_9417 and REACT_498 have 16 common interactions. Nevertheless, the process 'post-translational protein modification' is separated with the rest processes. In such complex case, the method should construct STN effectively and detect the overlaps among STN.

The threshold to output clusters is 0.1. The threshold means that if the membership of an interaction i with a cluster j $\mu_{ij} \geq 0.1$, this interaction highly correlates with the

cluster j and it will be clustered to cluster j. Five clusters are outcomes and then matched with 5 processes. The results are shown in Table 4.4.

Table 4.4 shows that we can construct signal transduction networks with the small error and can detect the nearly exact number of overlapping interactions. The combination of signaling feature data distinguished signaling processed from other biological processes and soft-clustering detected the overlapping components. When we checked the overlapping interactions among the clusters, there were exact 16 interactions that are shared in two signaling processes 'signaling by Insulin receptor' and 'signaling by EGFR'. Also, the same interaction set of the process 'elongation arrest recovery' and the process 'pausing and recovery of elongation' are found in their clusters. In fact, REACT_1069 does not overlap other processes but the results return three overlapping interactions, i.e., one with REACT_1892 and REACT_769 and two with REACT_498 and REACT_9417.

Table 4.4: Clustered results for five tested biological processes.

Process	True positive[1]	False negative[2]	False positive[3]	#Overlap_Int[4]
REACT_1069	0.565	0.174	0.435	3/0
REACT_1892	1.000	0.103	0.000	70/68
REACT_498	0.818	0.068	0.182	17/16
REACT_769	1.000	0.103	0.000	70/68
REACT_9417	0.960	0.120	0.040	17/16

1 True positive: the number of true interactions clustered/the number of interactions of the fact process.

2 False negative: the number of interactions missed in fact processes/the number of interactions of the fact process.

3 False positive : the number of false interactions clustered/the number of interactions of the fact process.

4 #Overlap_Int: the number of overlapping interactions among the clusters/the number of overlapping interactions among the fact processes.

Analyzing the case of interaction (P00734, P00734) shared among REACT_1069, REACT_498 and REACT_9417, we found some interesting findings. Protein P00734 (Prothrombin) functions in blood homeostasis, inflammation and wound healing and joins in biological process as cell surface receptor linked signal transduction (have GO term GO:0007166). In Reactome database, interaction(P00734, P00734) does not happen in the processes REACT_498 and REACT_9417, however according to the function of P00734, it probably partakes in one or two signaling processes REACT_498 and REACT_9417.

Although, the experiment carried out a case study of five biological processes; the proposed method is flexible to be applied to the larger scale of human interaction networks. In the intricate relations of many biological processes, the proposed method can well construct signal transduction networks.

In this chapter, we proposed a general framework to construct STN from mutilple signaling feature data using soft-clustering. The experiments with various parameters and other soft-clustering algorithms (not only FCM algorithm in Mfuzz) should be tested.

4.3.3 Some Results of Yeast STN Reconstruction

In addition to the work on human STN, we also carried out the work on yeast STN. This work consist of two parts: (1) signaling DDI prediction using ILP and (2) MARK yeast reconstruction.

This work concentrates on study STN for *Saccharomyces cerevisiae* – a budding yeast. The objective of this work is twofold. One objective is to present a method of predicting signaling domain-domain interactions (signaling DDI) using inductive logic programming (ILP), and the other is to present a method of discovering signal transduction networks (STN) using signaling DDI.

For signaling DDI prediction, we first examine five most informative genome databases, and extract more than twenty four thousand possible and necessary ground facts on signaling protein domains. We then employ inductive logic programming (ILP) to infer efficiently signaling DDI. Sensitivity (88%) and accuracy (83%) obtained from 10-fold cross validation show that our method is useful for predicting signaling domain interactions. Studying yeast MAPK pathways, we predicted some new signaling DDI that do not exist in the well-known InterDom database. Assuming all proteins in STN are known, we preliminarily build up signal transduction networks between these proteins based on their signaling domain interaction networks. We can mostly reconstruct the STN of yeast MAPK pathways from the inferred signaling domain interactions with coverage of 85%.

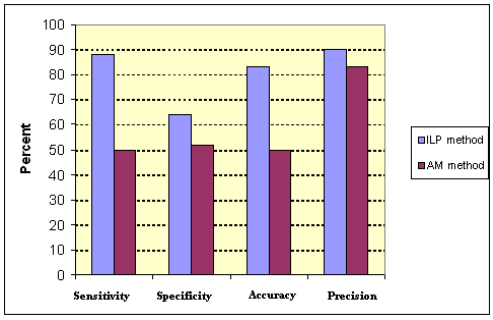

Figure 4.2: Performance of ILP method ($minpos = 3$ and $noise = 0$) compared with AM methods for signaling DDI prediction.

Figure 4.2 shows the results for signaling domain-domain interactions. Our experimental results obtained higher sensitivity, specificity, accuracy and precision compared with AM method [Sprinzak and Margalit, 2001].

From predicted (signaling) domain interaction networks, we raise the question of how

completely they cover the STN, and how to reconstruct STN using signaling DDI. Our motivation was to propose a computational approach to discover more reliable and stable STN using signaling DDI. When studying yeast MAPK pathways, the results of our work are considerable.

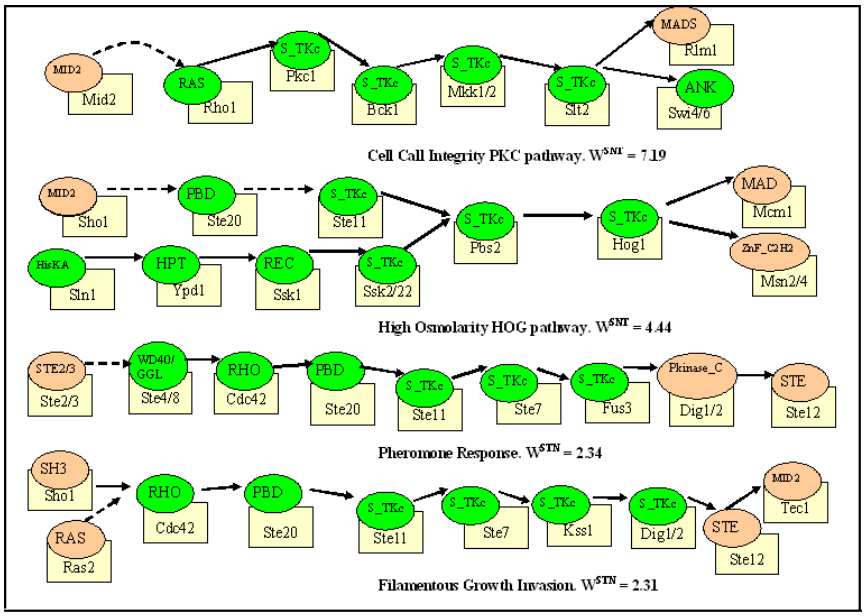

Figure 4.3: MAPK signal transduction pathways in yeast covered by signaling DDI networks. The rectangles denote proteins, the ellipses illustrate their domains and the signaling domains are depicted in dark. The signaling DDI are the lines with arrows, the missing interactions are dashed lines with arrows.

All extracted domains of proteins in MAPK pathways are inputs (testing examples) in our proposed predictor using ILP method [Nguyen and Ho., 2006]. With 32 proteins appearing in MAPK pathways, we extracted 29 different protein domains, and some of them are shared among proteins. Some domains are determined to be signaling domains, such as domain *pf00069* belonging to many proteins, for example, *ste11_yeast, fus3_yeast* or *pbs_2*, etc., and some of them are not signaling domains, such as *TEA* or *MID2*. Figure 4.3 shows yeast MAPK (mitogen-activated protein kinase) covered by signaling domain interactions. MAPK pathways involve pheromone response, filamentous growth, and maintenance of cell wall integrity pathways. Table 4.5 shows the results of predicted signaling DDI when reconstructing STN for the yeast MAPK pathways. Moreover, among predicted signaling DDI for yeast MAPK pathways, there are some DDI which are newly discovered, when compared with the InterDom database. For example, our predicted DDI *(pf00071,pf00768), (pf00768,pf00069), (pf00433,pf02200)* do not exist in the InterDom database.

Evaluating signaling domain interactions predicted from the testing set of MAPK do-

mains, 88% of protein relations in the Cell Wall Integrity PKC pathway, the Pheromone Response pathway, and the Filamentous Growth pathway are covered, and the Invasion High Osmolarity HOG pathway has coverage of 80%. Outstandingly, lots of domain interactions are found in which their corresponding proteins interacted in DIP (Database of Interacting Proteins) [5] and/or in CYGD (Comprehensive Yeast Genome Database)footnotehttp://mips.gsf.de /genre/proj/yeast/, for example, seven signaling domain interactions in the Cell Wall Integrity PKC pathway belong to 39 protein-protein interactions in CYGD database, and also belong to 47 protein-protein interactions in DIP. For estimating the reliability of STN, the reliability score W^{STN} (see in [Nguyen and Ho., 2006]) was calculated for yeast MAPK pathways. The reliability score of the Cell Wall Integrity PKC pathway is the highest with $W^{STN} = 7.19$.

Table 4.5: Results of predicted signaling DDI in the yeast MAPK pathways

The yeast MARK pathways pathways	Percentage of signaling DDI predicted	#CYGD PPI covered	#DIP PPI covered
Cell Wall Integrity PKC	88%	39	47
Pheromone Response	88%	41	42
Filamentous Growth	88%	40	38
Invasion High Osmolarity HOG	80%	40	53

The work is the first work that took effort to predict signaling DDI. The results on yeast STN confirmed the role of signaling domain-domain interactions and it

4.4 Outlook

The Section 4.3.1 presents a small scale of interaction networks for five biological processes, however, the method is easy to be applied to the larger scale of human interaction networks. In the intricate relationships with various processes, the proposed method can well detect signal transduction networks. The preliminary results encourage the further studies on biological complex systems.

1. Consider the whole interaction networks or some functional subnetworks, we definitely can construct not only the known signal transduction networks but also new ones. The components (proteins and their interactions) are shared among these networks to perform various functions in different biological processes.

[5] http://dip.doe-mbi.ucla.edu/

2. Given starting nodes (e.g., membrane proteins) and ending nodes (e.g, transcription factors), the proposed method can specify the signal transduction networks and then discover complete signaling pathways.

3. In human disease study, human interaction networks, signal transduction pathways and diseases have very close associations. Signaling network dysfunction can result in abnormal cellular transformation or differentiation, often producing a physiological disease outcome. The further work on identification of disease-related subnetworks are significant and can be investigated through the constructd signal transduction networks.

4. Our proposed soft-clustering method is simple to integrate other useful biological features and apply to other organisms.

We think that this presented work is feasible and potential to get more considerable results with many extensions in biological complex systems research.

4.5 Summary

In this chapter, we have presented a soft-clustering method to construct signal transduction networks from protein-protein networks. Many structured data of signaling features were extracted, integrated and exploited by soft-clustering to build STN. The experimental results demonstrated that our proposed method can construct STN effectively. The overlapping parts among STN were well detected. In future work, we would like to further investigate signaling features of proteins and protein interactions. Some other methods in relational learning and statistical learning will be consider to improve the work in some ways. It is also promising to discover the novel signal transduction networks from large interaction networks.

As proposing the general framework to construct signal transduction networks from protein interaction networks using soft-clustering, the method should be more carefully tested with various parameters and other algorithms (not only FCM algorithm in Mfuzz). Other computational measures also need calculated to better demonstrate efficiency of the method. Yet, the experimental results show that the proposed method is promising to construct signal transduction networks from protein-protein interaction networks.

In future work, we would like to further investigate signaling features of proteins and protein interactions. It is also promising to discover the novel signal transduction networks from large interaction networks. Not limited to signal transduction networks, the entire signal transduction pathways in particular and other complex biological systems in general, are able to be found when adding more information.

Chapter 5

A Semi-Supervised Learning Approach to Disease Gene Prediction

Discovering the human genes that cause disease (or "disease genes") is one of the emerging tasks in bioinformatics and biomedicine. In many ongoing research projects, protein-protein interaction networks (PPI) are being exploited in the discovery process, because there is a complex interplay between disease genes and PPI. Most current PPI-based methods only employ data regarding well-known disease genes, using supervised learning. However, there is a lot of valuable data containing information about unknown genes which could potentially enhance disease gene predictions. Combining multiple data sources for both known disease genes and unknown genes is expected to better predict which genes are likely to be disease genes. We have developed a novel method to effectively predict disease genes, by taking advantage of the wealth of existing data which may contain information about unknown genes. To this end, our method makes the best of semi-supervised learning, integrating data of human protein-protein interactions and various biological data extracted from multiple proteomic/genomic data sources. An experimental evaluation demonstrated that our proposed method outperformed other methods in terms of several measures including sensitivity, specificity, precision, accuracy, and a balanced F-score. A considerable number of potential disease genes were discovered and initially validated.

5.1 Introduction

One of the ultimate goals of life science is to improve our understanding of the processes and events related to disease. It is known that genetic diseases result from gene mutations caused by several factors, such as environment. Genes that have been identified as causing some diseases are called "disease-causing genes" or "disease genes" [NCBI, 2007]. Much biomedical work is focusing on monogenic diseases, investigating the position of a single gene in a chromosome through use of the positional cloning technique or linkage analysis, and on polygenic diseases using association analysis. However, we are still far from uncovering the molecular mechanisms of most diseases, which remain an important challenge to researchers.

The availability of important biological databases now allows research groups to develop computational methods for predicting disease genes from various data sources. Early work on disease gene prediction, typically methods based on sequences [Adie et al., 2005] or annotations [Turner et al., 2003], investigated disease genes as separate and independent entities.

However, it is well-known that biological processes are not the work of single molecules, but rather are the product of complex molecular networks, especially protein-protein interaction networks. Thus, there has been a shift from attempting to understand the molecular networks of other species to understanding the networks that underlie human diseases [Ideker and Sharan, 2008]. In particular, inspired by the findings for yeast PPI networks, several research groups are now focusing on the exploitation of human PPI networks to predict human disease genes via their corresponding proteins (intuitively, those are called disease proteins).

Research on protein-protein interaction networks and diseases has been rapidly increasing in the last two or three years. Many PPI-based methods have been proposed, each with a different way of exploiting the key assumption that "the network-neighbor of a disease gene is likely to cause the same or a similar disease", see [Goh et al., 2007], [Ideker and Sharan, 2008], and [Benjamin Schuster Bockler and Alex Bateman , 2008]. In an early work, disease genes were uncovered by topological features in human PPI networks using the k-nearest neighbor algorithm [Xu and Li, 2006]. Because of the sparseness of other proteomic/genomic data associated with certain diseases, several PPI-based methods require the integration of heterogeneous biomedical data in order to understand the complex interplay between genes/proteins and diseases [Kann, 2007]. A disease gene classification system has been proposed, to integrate the topological features of protein interaction networks with sequence and other features, and to analyze these features using support vector machines [Smalter et al., 2007]. Lage et al. (2007) used the phenomic ranking of protein complexes linked to human diseases to develop a Bayesian model

for predicting new candidates for disorders. Borgwardt and Kriegel (2007) integrated graph kernels for gene expression and human PPI to predict disease genes. In another direction, some work has concentrated on using PPIs to discover disease genes for specific diseases, i.e., Alzheimer's disease, using heuristic score functions [Chen et al., 2006], [Krauthammer et al., 2004].

The aim of the work presented in this chapter is to develop a novel and effective computational method for discovering disease genes, which takes into account recent biological results on protein networks and diseases.

The starting point of this work is, how best to utilize the wealth of existing data that may contain information about unknown genes. All previous work employed supervised learning schemes, which exploited data regarding known disease genes to predict new disease gene candidates. However, it has recently been shown by Oti *et al.* that genes associated with a particular phenotype or function are not randomly positioned in the PPI network, but tend to exhibit high connectivity; they cluster together and occur in central network locations [Oti et al., 2006]. That overriding property suggests that semi-supervised learning can be used in this prediction problem to exploit not only data concerning discovered disease genes but also data which may concern disease genes that are not yet known. In fact, this property solidifies the fundamental assumptions about the consistency of semi-supervised learning and provides more evidence for taking into account information regarding the unknown genes. Moreover, in addition to the protein topological features extracted from PPI databases, semi-supervised learning enables a systematic consideration of proteomic/genomic features related to diseases from various available data sources, which further enriches this computational scheme.

This work not only proposes a new and effective method for disease gene prediction, but also has generated significant new findings. We carefully carried out various experiments with disease gene information extracted from the OMIM (Online Mendelian Inheritance in Man) database (version 2007) [Hamosh et al., 2005]. Testing with all interacting partners of disease proteins, we found 568 putative disease genes. Some encouraging results were indirectly validated in various ways.

We performed two comparative experiments to evaluate the performance of the method. First, 10 times stratified 10-fold cross validations were conducted using our new Semi-Supervised Learning (SSL) method, the k-nearest neighbor (k-NN) method [Xu and Li, 2006], and the Support Vector Machines (SVMs) method [Smalter et al., 2007]. The results show that the SSL method outperforms the other two in terms of sensitivity, specificity, precision, accuracy, and a balanced F-score. Next, we compared our SSL method to the k-NN method with different sizes of labeled sets, and did twenty trials for each experiment to evaluate the accuracy. It turns out that the achieved accuracy of SSL is higher than that of k-NN.

There are about 25-30,000 genes in the human body. As reported in [Hamosh et al., 2005], some 3,053 of them are known to cause disease, and for these we use the term "known disease genes" or "disease genes". The more that 20,000 remaining genes, which may or may not cause disease, are called "unknown genes" in this chapter. Finally, genes that are assumed not to cause any disease are referred to as "known non-disease genes" or simply "non-disease genes".

The rest of the chapter is organized as follows. In Section 5.2, we present our proposed method to discover disease genes based on human PPI networks and multiple databases using semi-supervised learning. The comparative evaluation of the experiments is showed in Section 5.3.2. Predictive rules of PPI and DDI, as well as discussion, are presented in Section 5.4. Section 5.5 gives some concluding remarks.

5.2 Materials and Methods

In this section, we will describe our method to predict disease genes using semi-supervised learning. First, we give a brief introduction to semi-supervised learning, and explain why it is suitable for predicting disease genes with PPI networks. Then, we intuitively describe the proposed method for disease gene prediction. Finally, we present the score function for estimating the biological significance of extracted features for disease gene prediction.

5.2.1 Semi-Supervised Learning and Disease Gene Prediction

Semi-supervised learning (SSL) is halfway between supervised and unsupervised learning. It exploits both labeled data and unlabeled data to do either supervised learning or unsupervised learning. A given data set $\mathcal{X} = \{x_1, ..., x_l, x_{l+1}, ..., x_n\}$ can always be divided into two parts. The first one is the set of l data points $\mathcal{X}_l = \{x_1, ..., x_l\}$ which are labeled by the label set $\mathcal{Y}_l = \{y_1, ..., y_l\}$, and the other one is the data set of u data points $\mathcal{X}_u = \{x_{l+1}, ..., x_n\}$, the labels of which are not known. The goal is to predict labels of unlabeled data. Figure 5.1 shows the improvement of SSL in learning both labeled and unlabeled data [Zhu, 2005].

Semi-supervised learning methods use unlabeled data to either modify or reprioritize hypotheses obtained from labeled data alone. Although not all methods are probabilistic, it is easier to look at methods that represent hypotheses by p(y—x), and unlabeled data by p(x). Generative models have common parameters for the joint distribution p(x, y). It is easy to see that p(x) influences p(y—x). Mixture models with EM is in this category, and to some extent self-training. Many other methods are discriminative, including transductive SVM, Gaussian processes, information regularization, and graph-based methods. Original discriminative training cannot be used for semi-supervised learning,

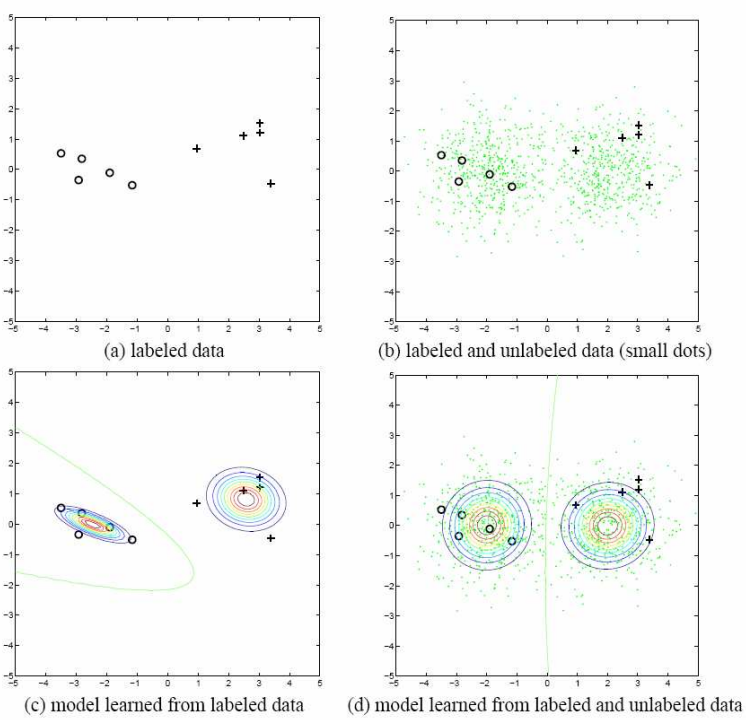

Figure 5.1: Semi-supervised learning.

since p(y—x) is estimated ignoring p(x). To solve the problem, p(x) dependent terms are often brought into the objective function, which amounts to assuming p(y—x) and p(x) share parameters [Zhu, 2005].

Some often-used semi-supervised learning methods include EM with generative mixture models, self-training, co-training, transductive support vector machines, and graph-based methods [Chapelle et al., 2006]. SSL is very useful in many real-word problems and has recently attracted an increasing number of researchers since labeling often requires much human labor, whereas unlabeled data is far easier to obtain [Chapelle et al., 2006]. In bioinformatics, SSL is also applied to solve many problems and has achieved considerable results, for example, in the study of protein classification [Weston et al., 2005] and in the functional genomics [Mark-A and Scheffer, 2004], etc.

The necessary question to ask before using SSL is whether the method is appropriate to the topology of PPI networks and the combination of multiple data features. The topology of PPI networks satisfies the fundamental assumptions about the consistency of SSL. These assumptions of consistency are: (1) nearby points are likely to have the same label, and (2) points on the same structure (typically referred to as a cluster or a manifold) are likely to have the same label [Zhou et al., 2004, Chapelle et al., 2006]. Considering the characteristics of disease gene distribution in PPI networks, we know that genes associated with a particular phenotype or function, including the progression

of disease, are not randomly positioned in the network. Rather, they tend to exhibit high connectivity, cluster together, and occur in central network locations [Oti et al., 2006]. As a result, SSL is appropriate for exploiting PPI networks in disease gene prediction.

Moreover, considering the topological properties of human protein-protein interaction networks, graph-based semi-supervised learning is suitable for the task of disease gene prediction. Another strength is that SSL is capable of combining various data not only of disease genes but also of their information-rich neighbors. The details of our proposed method are presented in Section 5.2.

5.2.2 The Proposed Method for Predicting Disease Genes

The key premise is to enrich the disease gene classifier by (1) making use of both known disease genes and as yet unidentified disease genes (unknown genes), and (2) integrating multiple data sources in a semi-supervised learning scheme. The method addresses three main tasks to successfully predict disease genes.

1. Extend the initial known disease protein set to be the set of both labeled data (as known disease proteins), and unlabeled data (as newly extracted candidate proteins). Based on the assumption of the phenotype and genotype similarity of neighbors in PPI networks, the interacting partners of known disease proteins were extracted to build an extensive candidate protein set. The interacting partners of disease proteins are considered potentially reliable candidate proteins. This task is to extract candidate proteins, rich data of which will be later exploited to complement known disease protein data.

2. Compile multiple disease-related feature data of all proteins in the extended set. Going through the literature, we detected several features that may affect diseases. However, data of these features is stored in various forms and scattered in different data sources. This task is to extract, preprocess, collect and represent those feature data in a unified form.

3. Integrate all of the rich compiled protein data in the extended set using SSL, to predict disease genes. This task is to achieve the best performance of semi-supervised learning in disease prediction by learning combined multiple features data of both labeled and unlabeled data.

Corresponding to these three main tasks, we carried out three steps. Figure 5.2 illustrates these steps: (1) Identify disease proteins, non-disease proteins, and candidate proteins, (2) Extract the topological features and disease-related proteomic/genomic features of proteins, and (3) Use semi-supervised learning to predict disease genes. Detailed

procedures for these steps are also shown in Table 5.1. Given the known disease protein set, Step 1 extends this set by adding more potential candidate proteins and also prepares the negative example data set (see more in Subsection 5.3.1). In Step 2, we investigated the seven data sources: OPHID [Brown and Jurisica, 2005], Uniprot[1], Gene Ontology[2] Pfam[3], InterDom[4], Reactome[5], and Gene expression data in [Mariadason et al., 2002] to extract nine features. The extracted data were preprocessed and weighted by our proposed score functions (see more in Subsection 5.2.3). Finally, in Step 3 we combined the feature data of all proteins in the extended set. Step 3 then builds a classifier using a semi-supervised learning algorithm. Specifically, a graph-based semi-supervised learning algorithm, the so-called Harmonic Gaussian method [Zhou et al., 2004], is appropriately applied.

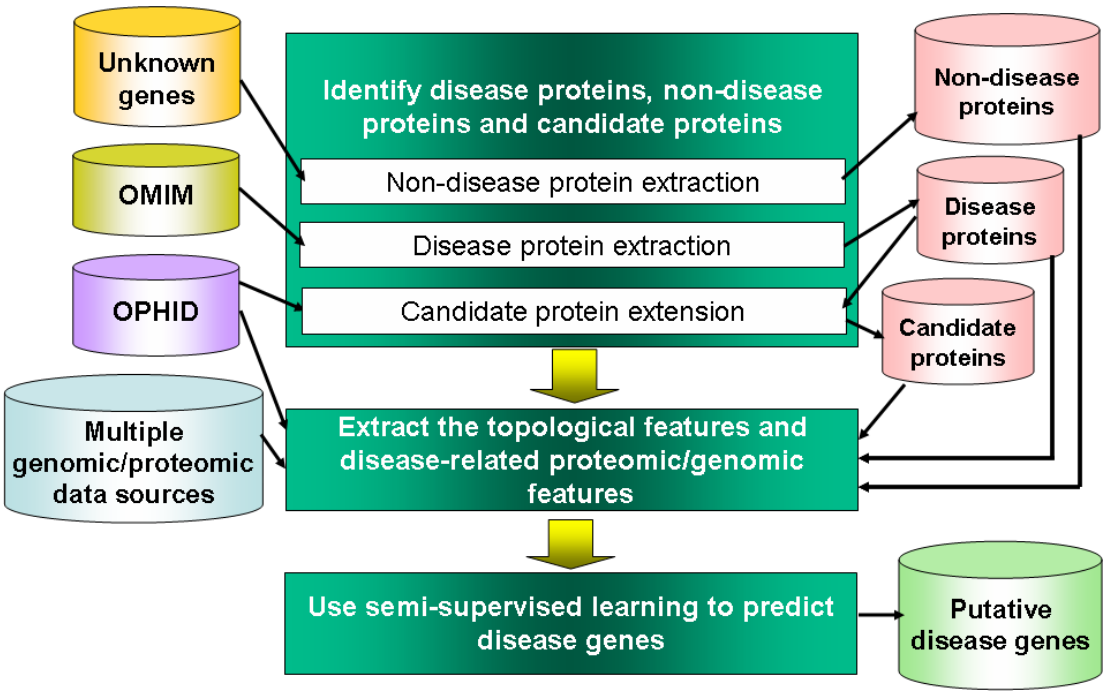

Figure 5.2: Three-step semi-supervised learning method for disease gene prediction.

5.2.3 Scores of Proteomic/Genomic Features

Several features are known to affect and cause diseases. However, data concerning those features are scattered in a wide range of data sources. These data should be integrated into

[1]http://www.uniprot.org/
[2]http://www.geneontology.org/
[3]http://www.sanger.ac.uk/Software/Pfam/
[4]http://interdom.i2r.a-star.edu.sg/
[5]http://www.reactome.org/

Table 5.1: Three main steps of the proposed semi-supervised learning method for disease gene prediction.

Step 1: *Identify disease proteins, non-disease proteins and disease protein candidates.* First, take available disease genes, e.g., from the OMIM database, as positive examples (labeled data) and the unknown genes (unlabeled data), and map them to the corresponding proteins identified by the Uniprot accessions. The genes which belong to the UEGH set are excluded, because they are essential genes having features that differ significantly, both from disease genes and from other genes [Tu et al., 2006]. As the result, we obtain the set $\mathcal{P}^+$ of disease proteins and the set $\mathcal{P}$ of unknown proteins. Second, starting from a protein interaction network, e.g., from the OPHID database, extract from $\mathcal{P}$ the set $\mathcal{P}^c$ of interacting partners of disease proteins, and consider them as candidate proteins. Third, randomly choose non-disease proteins (as negative examples) from the set $\mathcal{P}$ excluded proteins in the set $\mathcal{P}^c$. The number of proteins in the non-disease protein set $\mathcal{P}^-$ is equal to the number of proteins in the disease protein set $\mathcal{P}^+$. Denote by $\mathcal{P}^*$ the set obtained by the union of disease proteins, the candidate proteins, and the non-disease proteins, i.e., $\mathcal{P}^* = \mathcal{P}^+ \cup \mathcal{P}^- \cup \mathcal{P}^c$.

Step 2: *Extract and represent the topological features and disease-related proteomic/genomic features.* For each protein in $\mathcal{P}^*$, extract a number of topological features and disease-related proteomic/genomic features in terms of their numerical scores. In fact, each feature f^k corresponds to a score $score_k$ estimated from one of the seven data sources: OPHID, Uniprot, GO, Pfam, InterDom, Reactome, and Gene Expression.

Step 3: *Use semi-supervised learning to predict disease genes.* Due to the network nature of protein-protein interactions, graph-based methods of semi-supervised learning are appropriate to the prediction task, using the data obtained from Steps 1 and 2. The output is a set of new putative disease genes.

---

one computational scheme to better predict disease genes. Our proposed method makes use of semi-supervised learning to combine information about various features from both labeled and unlabeled data (as candidate proteins).

In addition to topological features of PPI extracted from the OPHID database, we extracted eight other proteomic/genomic features that have comprehensive associations with diseases from the six data sources: Uniprot (three features, sequence length, tagged keyword, and codded enzyme), GO (one feature, GO term), Pfam (one feature, protein domain), InterDom (one feature, domain-domain interaction), Reactome (one feature, pathway), and Gene Expression (one feature, gene expression). Table 5.2 shows the statistics of the extracted proteomic/genomic features from each data source. Columns 3 and 4 are the numbers of records extracted according to their respective features, and the last two columns are the numbers of feature categories.

Among the 5,557 proteins in $\mathcal{P}^*$, 31,465 data records were extracted for the keyword

Table 5.2: Statistics of two sets $\mathcal{P}^+$ and $\mathcal{P}^*$ with the eight extracted proteomic/genomic features.

Database	Feature f^k	#Record		#Category	
		in $\mathcal{P}^*$	in $\mathcal{P}^+$	in $\mathcal{P}^*$	in $\mathcal{P}^+$
Uniprot	f^{length}	4412	1496		
	f^{KW}	31465	13597	564	504
	f^{EC}	1123	451	133	106
Gene Otology	f^{GO}	17241	6404	2911	1817
Pfam	f^{Pfam}	6817	2426	1796	1413
Reactome	$f^{Pathway}$	1167	540	68	62
InterDom	f^{DDI}	3854	1322		
Gene Expression	$f^{expression}$	696	52		

features, and 1,123 for the enzyme features. These proteins share the same 564 keywords and 133 enzymes, as shown in Table 5.2. The keyword and enzyme data are categorical; for example, (P05067, alzheimer disease) and (P01011, disease mutation) where P05067, P01011 are the Uniprot names, and "alzheimer disease", "disease mutation" are their keywords, or (O75688, ec3.1.3) where O75688 is the Uniprot name and ec3.1.3 is the coded enzyme.

The features data are stored in different data types, i.e., numerical type such as sequence length and number of domain-domain interactions, or categorical type such as keywords, pathways and coded enzymes. Accordingly, we defined the score functions to weight and formalize the extracted features, and then represented them as feature vectors in order to integrate the features into to a unified computational scheme.

The score functions of selected features are introduced below summarized in Table 5.3.

- **The topological score**: This score is computed based on protein-protein interactions, and shows the topological association between a given protein and disease proteins. We can assume that if one protein has many interactions with disease proteins, and joins in the group of disease proteins, it is likely to be a disease protein. Therefore, the higher PPI score a protein has, the more probable that it causes a disease. Human protein-protein interactions are extracted from the OPHID database [Brown and Jurisica, 2005]. The score $score_{ppi}(p_i)$ for the feature f^{ppi} is defined as in Table 5.3.

- **The keyword score**: Disease proteins may have the same keywords, and these common keywords are tagged more frequently in the set of disease proteins than other proteins. Keywords are scored by their frequency and assigned to each protein p_i by $score_{kw}(p_i)$ shown in Table 5.3.

Table 5.3: Topological feature, genomic/proteomic features and their score functions

Score functions	Notations and explanations
$score_{ppi}(p_i) = \frac{\sum_{p_j \in \mathcal{P}^+} Int(p_i,p_j)}{\sum_{p_j \in \mathcal{P}^*} Int(p_i,p_j)} \times \frac{\sum_{p_j \in \mathcal{P}^+} Int(p_i,p_j)}{Avg_{ppi}}$	$Int(p_i,p_j) = \begin{cases} 1 & \text{if there is an interaction between proteins } p_i \text{ and } p_j, \\ 0 & \text{otherwise.} \end{cases}$ Avg_{ppi}: average of number of protein interactions belonging to disease proteins.
$score_{kw}(p_i) = \frac{1}{\sum_{\forall kw_i \in p_i} w_i^{kw}}$	$w_i^{kw} = freq^+(kw_i) \times freq^*(kw_i)$ $freq^+(kw_i)$: the frequency count of kw_i observed in $\mathcal{P}^+$. $freq^*(kw_i)$: the frequency count of kw_i observed in $\mathcal{P}^*$.
$score_{ec}(p_i) = freq^+(ec_i) \times freq^*(ec_i)$	$freq^+(ec_i)$: the frequency count of ec_i observed in $\mathcal{P}^+$. $freq^*(ec_i)$: the frequency count of ec_i observed in $\mathcal{P}^*$.
$score_{go}(p_i) = \frac{1}{\sum_{\forall go_i \in p_i} w_i^{go}}$	$w_i^{go} = (\#go_i^+ + 1)/(\#go_i^* + 1)$ $\#go_i^+$: the count of go_i observed in $\mathcal{P}^+$. $\#go_i^*$: the frequency counts of go_i observed in $\mathcal{P}^*$.
$score_{pfam}(p_i) = \frac{\#pfam_i^+ + 1}{\#pfam_i^* + 1}$	$\#pfam_i^+$: the number of domains d_j of a protein p_i observed in $\mathcal{P}^+$ $\#pfam_i^*$: the number of domains d_j of the protein p_i observed in $\mathcal{P}^*$.
$score_{length}(p_i) = \frac{length(p_i)}{Avg_{length}}$	$length(p_i)$: the sequence length of a protein p_i. Avg_{length}: the average sequence length of disease proteins in $\mathcal{P}^+$.
$score_{pathway}(p_i) = \sum_{\forall pathway_i \in p_i} w_i^{pathway}$	$w_i^{pathway} = freq^+(pathway_i) \times freq^*(pathway_i)$ $freq^+(pathway_i)$: the frequency count of $pathway_i$ observed in $\mathcal{P}^+$. $freq^*(pathway_i)$: the frequency count of $pathway_i$ observed in $\mathcal{P}^*$.
$score_{ddi}(p_i) = \frac{1}{\sum_{\forall ddi_i \in p_i} w_i^{ddi}}$	$w_i^{ddi} = \frac{Avg_{ddi}}{\#ddi(p_i)}$ $\#ddi(p_i)$: the number of DDI observed in $\mathcal{P}^+$ of protein p_i. Avg_{ddi}: the average number of DDI of disease proteins in $\mathcal{P}^+$.

- **The enzyme score**: Enzymes perform a wide variety of functions inside living organisms. The relationship between enzymes and diseases has been studied and proved in many works. Like the keyword feature, some enzymes are shared among the group of disease proteins. Score $score_{ec}(p_i)$ shows how probable it is that a protein is a disease protein, in terms of coded enzymes (shown in Table 5.3).

- **The sequence length score**: We investigated the protein sequence length feature to study how the sequence length of a protein relates to disease-causing mechanisms. Score $score_{length}(p_i)$ is the ratio of sequence length of a protein over the average length of disease proteins (shown in Table 5.3).

- **The GO term score**: GO terms are divided into three groups: molecular function, biological process, and cellular component. These terms present the general information about the proteins, and the terms of disease proteins might focus on some specific groups. The score for GO term feature is shown as $score_{go}(p_i)$ in Table 5.3.

- **The protein domain score**: Protein domains are the building blocks of proteins. Disease proteins may structurally or functionally depend on their domains. If a protein has many domains related to disease, it is more likely to be a disease protein. Pfam domains d_j of all considered proteins are extracted and scored by $score_{pfam}(p_i)$ (shown in Table 5.3).

- **The DDI score**: Domain-domain interactions (DDI) underlie the interactions of proteins, and themselves perform specific functions in cells. DDI may play an important role in the regulation of PPI in causing diseases. We extracted the DDI data from the InterDom database and weighted them by $score_{ddi}(p_i)$ based on the number of their DDI shared with disease proteins (shown in Table 5.3).

- **The biological pathway score**: Many disease processes arise from defects in biological pathways. In the Reactome database, all proteins in the extended set take part in 68 pathways. Among those pathways, 62 contain at least one disease protein. The $score_{pathway}(p_i)$ of feature $f^{pathway}$ is based on the frequency of the pathways observed in both $\mathcal{P}^*$ and $\mathcal{P}^+$ (shown in Table 5.3).

  There is no doubt about the association between disease and gene expression. The gene expression data used is a gene expression profile that defines colon cell maturation [Mariadason et al., 2002]. The gene expression feature was not scored because its original data was formalized. The above score functions demonstrate the similarity of candidate proteins and disease proteins. Combining all the features produces a biologically significant data set for disease gene prediction.

5.3 Evaluation

Performance of the proposed semi-supervised learning method was compared with that of two closely related works using k-nearest neighbors (k-NN) [Xu and Li, 2006] and support vector machines (SVMs) [Smalter et al., 2007] in terms of several measures.

5.3.1 Experiment Design

We carried out two comparative experiments. The first employed a 10 times stratified 10-fold cross validation to evaluate SSL, k-NN, and SVMs in terms of sensitivity, specificity, precision, accuracy, and a balanced F-score. In the second one, we varied the labeled set size l and performed twenty trials to compare the accuracy of the proposed method and the k-NN method [Xu and Li, 2006].

We prepared three data sets to carry out the experiments: (i) a set of disease genes, (ii) a set of non-disease genes, and (iii) a set of protein-protein interactions.

The OMIM database (version 2007) is a catalog of human genes and genetic disorders. In OMIM, the list of hereditary disease genes is described in the OMIM morbid map. As reported in [Hamosh et al., 2005], there are 4,512 records with 3,053 unique OMIM identifiers in the catalog. A total of 3,053 human disease genes were mapped, to look for their disease proteins identified by Uniprot names. The results showed 3,590 corresponding disease proteins. Among them, 1,502 proteins have published interactions in the OPHID database.

Compiling a list of non-disease genes is difficult. A recent study [Tu et al., 2006] showed that the human genome may contain thousands of essential genes having features that differ significantly, both from disease genes and from other genes. In the absence of a set of well-defined essential human genes, they considered the set of ubiquitously expressed human genes (UEHG), also known as housekeeping genes, as an approximation. Tu et *al.* proposed to classify them as a unique group for comparisons of disease genes with non-disease genes. Mapping to Uniprot names, there are 723 proteins corresponding to UEHG. From the set of unknown genes, we randomly chose negative examples, which do not belong to both the OMIM morbid map and the UEHG set. The number of negative examples is equal to the number of positive examples.

Human protein-protein interactions were extracted from the OPHID database [Brown and Jurisica, 2005]. Among 51,934 human protein-protein interactions stored in OPHID, there are 13,368 interactions which have at least one interacting partner belonging to the set of disease proteins. From 13,368 interactions, the initial set of 1,502 disease proteins was extended to 5,775 proteins.

We used Weka [Witten and Frank, 2005] to run k-NN and SVMs (Sequential Minimal

Optimization (SMO) algorithm). For k-NN, the different parameters k were chosen exactly as in [Xu and Li, 2006]. For SVMs, the kernels were RBF and linear kernel functions, and other parameters were default.

In the SSL implementation, SemiL software that developed by Huang and Kecman [Huang and Kecman, 2004] was employed to run the Harmonic Gaussian method. SemiL software is efficient for solving large-scale semi-supervised learning problems using graph kernels. The Harmonic Gaussian method is suitable for the topological characteristics of PPI networks and the combination of multiple feature scores. In the algorithm, data is represented as a graph $G = (V, E)$ with V as the set of nodes corresponding to both l labeled data points and u unlabeled ones. An $n \times n$ symmetric weight matrix W on the edges of the graph is given. W was defined as $W_{ij} = exp(- \parallel x_i - x_j \parallel)^2/2\sigma^2$ x_i if $x_i \neq x_j$, otherwise $w_{ij} = 0$, where x_i, x_j are data points. The nearby points in Euclidean space are assigned large edge weights. Intuitively, unlabeled points that are nearby in the graph have similar labels. In our experiment, the weight matrices W were calculated with two different distance functions, i.e., Euclidean distance and Cosine distance, and the degree of graph was 20.

5.3.2 Experiment Results

Five measures of prediction quality are as follows.

$$Sensitivity = \frac{TP}{TP + FN}, \quad Precision = \frac{TP}{TP + FP}, \quad F = 2 * \frac{SS * P}{SS + P},$$
$$Specificity = \frac{TN}{TN + FP}, \quad Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

where TP, FN, TN, FP, SS, P denote true positive, false negative, true negative, false positive, sensitivity and precision, respectively.

In the first experiment, performance of the methods was also statistically tested in terms of confidence intervals, to give us an estimate of the amount of error involved in our data. To estimate a 95% confidence interval for each calculated specificity and sensitivity, we used t distribution. The 95% confidence intervals are shown in Table 5.4. The experimental results demonstrated that the SSL methods outperformed other methods in disease gene prediction.

In the second experiment, from the training data set we randomly selected l data points as labeled data, while the rest *(n-l)* were unlabeled data. Then, accuracy was estimated by comparing the predicted labels and true labels. For each labeled set size l, we performed 20 trials. The result is the average accuracy of those 20 trials. These procedures were

Table 5.4: The 10 time 10-folds cross validation performance of SSL methods (SSL1 with Cosine distance and SSL2 with Euclidean distance) compared to two methods SVM and k-NN .

	Precision	Accuracy	Sensitivity	Specificity	F-measure
SSL1	**0.812±.042**	**0.823±.019**	**0.852±.031**	**0.794±.041**	**0.829±.013**
SSL2	0.806±.039	0.820±.019	0.850±.026	0.789±.036	0.825±.013
SVMs	0.713 ±.032	0.741±.023	0.804±.035	0.677±.038	0.756 ±.019
1-NN	0.779±.033	0.786±.032	0.798±.025	0.774±.042	0.789 ±.032
3-NN	0.768±.037	0.782±.020	0.806±.030	0.757±.037	0.787±.027
5-NN	0.771±.031	0.771±.017	0.819±.029	0.744±.037	0.789±.027
7-NN	0.761±.042	0.761±.024	0.822±.030	0.720±.022	0.782±.019
9-NN	0.776±.030	0.540±.025	0.770±.027	0.752 ±.034	0.763±.026

carried out for both the SSL method and the k-NN method [Xu and Li, 2006] on the same test data sets.

Figure 5.3 shows the accuracy of our method and the k-NN method with various parameters k. When the labeled set size is small (10% of the data set), semi-supervised learning obtained non trivial accuracy, 78%. When the amount of labeled data is at least half of the total data set, the accuracy of the SSL method is over 80%. By comparison, SSL method obtained higher accuracy than k-NN method with all the labeled set sizes. Moreover, the results also demonstrated that when little labeled data is available, semi-supervised learning can predict disease genes with high accuracy, and performs better than supervised learning.

5.4 Discussion

In addition to computational evaluation, we endeavored to look for biological evidence to support our method. We found some interesting evidence when verifying the new putative disease genes. Testing the whole network of protein interactions, we predicted 572 putative proteins corresponding to 568 putative genes.

For evaluating the findings, some indirect methods were used to demonstrate the potential of putative disease genes. These strategies were: (i) validating the putative disease gene's keywords and pathways shared with known disease genes, (ii) checking their functional categories and gene similarity via DAVID tools [Dennis et al., 2003]; (iii) testing them with Endeavour – Computer Program For Identifying Disease Genes [Aerts et al., 2006], and (iv) looking their disease-related information up in the literature. This section discusses some interesting findings.

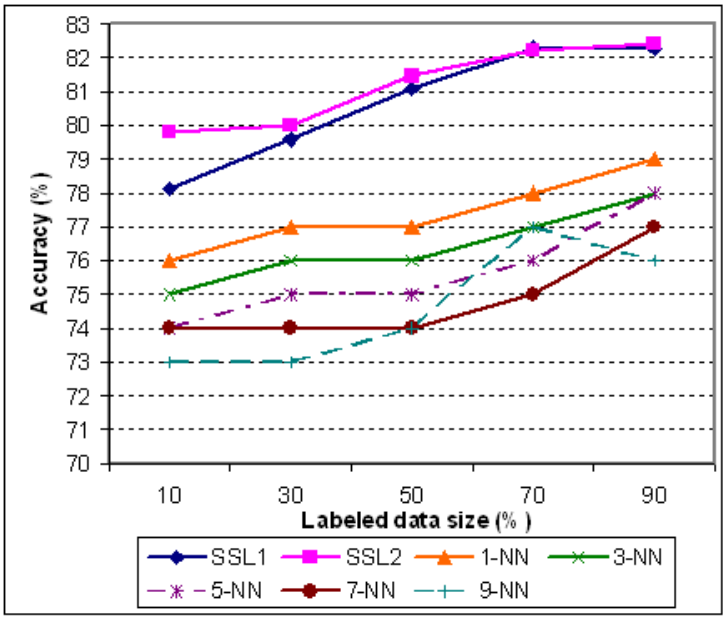

Figure 5.3: Accuracy of the proposed method with different sizes of labeled data for the Euclidean and Cosine distance compared to the k-NN method.

First, we checked whether the putative disease proteins have keywords and pathways of known disease proteins. Among 47 Reactome pathways shared with known disease proteins, we found that the set of putative proteins belonged to many pathways related to disease traits, such as *'Signaling in Immune system'* (29 putative proteins), e.g., O00459, P01112, P04439; *'Hemostasis process'* (25 putative proteins), e.g., O00459, P01112, P04085; and *'Gene expression process'* (21 putative proteins), e.g., O60563. Similarly, there are 270 Uniprot keywords that are tagged for known disease proteins. Among them, many putative disease proteins share the same keywords, e.g., '*alternative splicing*' with 212 proteins, '*polymorphism*' with 195 proteins, and '*glycoprotein*' with 187 proteins.

The second validation is to check functional categories, and gene similarity of the putative disease genes via DAVID tools. Interestingly, 29 genes were found in 67 records in GAO – Genetic Association Database (version 2008)[6]. For example, *IGFBP2* (insulin-like growth factor binding protein 2, 36kda), and *TNFSF8* (tumor necrosis factor (ligand) superfamily, member 8) are related to the term *'diabetes, type 1'*; IFNAR1 (interferon alpha, beta and omega receptor 1) is related to the term *'Hepatitis B, Chronic'*, and *ITGA3* (Integrin, alpha 3 (antigen CD49C, alpha 3 subunit of VLA-3 receptor)) is related to the term *'breast cancer'*. Checking the putative disease genes in OMIM, 2 genes are related to 8 records found in database OMIM with the term *'Colorectal cancer'*, e.g., *BAX* (bcl2-associated x protein) and *HRAS* (v-Ha-ras harvey rat sarcoma viral oncogene

[6]http://geneticassociationdb.nih.gov

homolog). The Figure 5.4 show some of genes related to term 'immune' in database GAO.

IL12RB2	interleukin 12 receptor, beta 2	Related Genes	Homo sapiens
GENETIC_ASSOCIATION_DB	IMMUNE, TYPE 1 DIABETES,		
KCNJ9	potassium inwardly-rectifying channel, subfamily j, member 9	Related Genes	Homo sapiens
GENETIC_ASSOCIATION_DB	METABOLIC, TYPE 2 DIABETES,		
CSF3R	colony stimulating factor 3 receptor (granulocyte)	Related Genes	Homo sapiens
GENETIC_ASSOCIATION_DB	IMMUNE, SEVERE CHRONIC NEUTROPENIA,		
BAX	bcl2-associated x protein	Related Genes	Homo sapiens
GENETIC_ASSOCIATION_DB	B CELL CHRONIC LYMPHOCYTIC LEUKAEMIA., LYMPHOCYTIC LEUKEMIA,		
NPPB	natriuretic peptide precursor b	Related Genes	Homo sapiens
GENETIC_ASSOCIATION_DB	CARDIOVASCULAR, IDIOPATHIC DILATED CARDIOMYOPATHY,		
TNFSF8	tumor necrosis factor (ligand) superfamily, member 8	Related Genes	Homo sapiens
GENETIC_ASSOCIATION_DB	IMMUNE, TYPE 1 DIABETES,		
HLA-A	major histocompatibility complex, class i, a	Related Genes	Homo sapiens
GENETIC_ASSOCIATION_DB	ASTHMA, CANCER, IMMUNE, INFECTION, LEPROSY, RENAL CELL CARCINOMA,		
ITGA3	integrin, alpha 3 (antigen cd49c, alpha 3 subunit of vla-3 receptor)	Related Genes	Homo sapiens
GENETIC_ASSOCIATION_DB	BREAST CANCER, CANCER,		
TIMP1	timp metallopeptidase inhibitor 1	Related Genes	Homo sapiens
GENETIC_ASSOCIATION_DB	CANCER, CARDIOVASCULAR, INTRACRANIAL ANEURYSMS, RECTAL CANCER,		
SHC1	shc (src homology 2 domain containing) transforming protein 1	Related Genes	Homo sapiens
GENETIC_ASSOCIATION_DB	AGING, LONGEVITY, METABOLIC, TYPE 2 DIABETES,		
IGFBP2	insulin-like growth factor binding protein 2, 36kda	Related Genes	Homo sapiens
GENETIC_ASSOCIATION_DB	IMMUNE, TYPE 1 DIABETES,		
RFC1	replication factor c (activator 1) 1, 145kda	Related Genes	Homo sapiens
GENETIC_ASSOCIATION_DB	NEURAL TUBE DEFECTS,		
CD86	cd86 antigen (cd28 antigen ligand 2, b7-2 antigen)	Related Genes	Homo sapiens
GENETIC_ASSOCIATION_DB	CELIAC DISEASE, IMMUNE, RHEUMATOID ARTHRITIS, SYSTEMIC LUPUS ERYTHEMATOSUS, TYPE 1 DIABETES,		
KLRC2	killer cell lectin-like receptor subfamily c, member 2	Related Genes	Homo sapiens
GENETIC_ASSOCIATION_DB	IMMUNE, RHEUMATIC DISEASES, RHEUMATOID ARTHRITIS, SYSTEMIC LUPUS ERYTHEMATOSUS,		
IL8	interleukin 8	Related Genes	Homo sapiens
GENETIC_ASSOCIATION_DB	CANCER, COLORECTAL CANCER, ENTEROAGGREGATIVE ESCHERICHIA COLI DIARRHEA, IMMUNE, INFECTION, LUPUS, MICROSATELLITE POLYMORPHISM OF THE INTERLEUKIN 8 (IL-8) GENE, SEVERE RSV BRONCHIOLITIS, TUBERCULOSIS DISEASE,		

Figure 5.4: The testing putative disease genes with database GAO related to the term 'immune'.

In the third evaluation, we used the Endeavour system to rank the putative disease genes. Endeavour identifies disease genes by ranking them from the rank of each individual data source. We did two tests with the Endeavour system. First, all the test data were input into Endeavour and ranked with all data sources. There are 42 genes with *p-value* ≤ 0.05 which are found in the set of predicted disease genes. Some of them obtained a very high rank with a statistically significant *p-value*. Table 5.5 shows top 10 putative genes as ranked by Endeavour system.

To study how the candidate genes related to specific diseases, we then applied Endeavour to test 568 candidate genes for three diseases, i.e., cancer, diabetes, and Alzheimer's. In the Endeavour system, the score of each gene was precalculated by the Ouzounis and Prospectr systems to investigate the similarity between the genes being tested and the genes related to the three diseases in these two systems. It is interesting that the Endeavour system returned high ranked genes with *p-value* ≤ 0.01, for example, genes *MYH10, DYNC1H1, CACNA1D, LAMA4, GIPC1, NCOR1.*

Finally, we looked to the biomedical literature to find the evidence for the putative disease genes. As in [Tu et al., 2006], ubiquitously expressed human genes (UEGH) should be regarded as the most severe disease genes. Among 568 newly predicted disease proteins, there are 6 proteins which correspond to UEHG genes: *nherf_human, ddx3x_human,*

Table 5.5: List of some putative disease proteins and the corresponding disease genes.

#Rank	Gene ID	Q-int	P-value
2	ENSG00000133026	1.16E-08	0.00026408
4	ENSG00000010810	3.70E-07	0.001516515
7	ENSG00000171094	2.47E-06	0.003953699
10	ENSG00000065361	4.90E-06	0.005591587
11	ENSG00000178568	5.64E-06	0.006002614
13	ENSG00000112769	6.97E-06	0.006679875
14	ENSG00000162434	7.68E-06	0.00701467
16	ENSG00000115306	9.49E-06	0.007801282
17	ENSG00000123384	9.80E-06	0.007931698
22	ENSG00000077522	1.26E-05	0.00899866

tyy1_human, 1433t_human, ctbp1_human, spta2_human.

The hepatitis C virus (HCV) core protein influences the expression of host genes [Owsianka and Patel, 1999]. *Ddx3x_human* (ATP-dependent RNA helicase DDX3X) acts as a cofactor for XPO1-mediated nuclear export of incompletely spliced HIV-1 Rev RNAs, and is also involved in HIV-1 replication. This protein interacts specifically with the HCV core protein, resulting in a change in intracellular location.

Protein *tyy1_human* acts as a repressor in the absence of adenovirus E1A protein, but as an activator in its presence. A adenoviruses, a group of viruses that infect the membranes (tissue linings) of the respiratory tract, the eyes, the intestines, and the urinary tract, account for about 10% of acute respiratory infections in children, and are a frequent cause of diarrhea.

Protein *trrap_human* is the isolation of highly conserved 434 kDa protein; and it interacts specifically with the c-Myc N terminus, and has homology to the ATM/PI3-kinase family. *Trrap_human* (related to gene *trrap*) also interacts specifically with the E2F-1 transactivation domain. Expression of transdominant mutants of the protein *trrap_human* or antisense RNA blocks c-Myc- and E1A-mediated oncogenic transformation. Then, *trrap* was suggested as an essential cofactor for both the c-Myc and E1A/E2F oncogenic transcription factor pathways [McMahon et al., 1998].

5.5 Summary

in this chapter, we have introduced a method based on semi- supervised learning, integrating multiple data features, for disease gene prediction. The method proposed here is not restricted to any particular disease or particular group of data features. We investigated and chose several features that are considered relevant to diseases. Later, when there

are other, better features, the method is flexible enough to combine them as well. The experimental results demonstrated that our proposed method performed well with high accuracy, and at the same time, predicted some new disease genes. Moreover, the experimental results with small amounts of labeled data demonstrated an improved ability to study specific diseases when the known disease genes (as labeled data) are very limited.

In future work, we would like to validate the predicted disease genes in a wet-lab. Other work will involve applying and comparing the performance of the Harmonic Gaussian algorithm with other semi-supervised learning algorithms for disease genes prediction. Various protein-protein interaction databases should be combined to widen knowledge of the interaction networks of disease genes.

Chapter 6

Conclusions and Future Work

6.1 Summary of the Dissertation

In this thesis, we have presented a study of protein-protein interactions and two related problems in medicine, i.e., (1) protein-protein interaction prediction, (2) signal transduction network construction, and (3) disease-causing gene prediction. The study considered the research problems in both theoretical and practical views. The theoretical view concerned about the proposal of new methods for protein-protein interaction prediction, signal transduction network construction, and disease-causing gene prediction. The practical views came from the effective application of these works in biomedicine. Among the six chapters of the thesis, the main chapters are chapters 3, 4, and 5. The main contributions of the thesis can be summarized as follows.

1. *The first contribution is that we developed novel integrative domain-based method to predict protein-protein interactions.*

   Our method was based on protein domains, the basic functional and structural parts of proteins. In addition, we investigated and combined various informative genomic and proteomic data from multiple data sources using inductive logical programming. The advantages of the method demonstrated in both the computational and biological aspects.

   We took biological usability expertise along with ILP method to predict protein-protein interactions in an efficient way. ILP is appropriate to unify different types of data that are acquired from the great deal of expert knowledge. By integrating a large amount of data from seven databases, 278,000 ground facts of domain fusion, domain-domain interaction features and various biologically significant genomic/proteomic features, were extracted and represented in forms of ILP predicates. After obtaining the large set of ground facts, we applied the Aleph system to

learn these ground facts (as background knowledge), negative and positive examples. The Aleph system then induced predictive rules.

Through 10-fold cross validations, the performance measures, including Receiver Operating Characteristic (ROC) curves, sensitivity and specificity, showed that our method achieved better performance than other methods, such as support vector machines method and association method. Moreover, thanks to ILP rules, the predictions were more interpretable and useful for biologists. Analyzing produced rules (of both PPI and DDI), many interesting relationships among PPI, DDI, and protein functions, biological processes, were found. Our proposed method can be tuned to predict PPI and DDI for diverse organisms and other genomic and proteomic data sources.

This work was presented in Chapter 3.

2. *The second contribution is that we developed an efficient soft-clustering method to construct signal transduction networks from PPI networks.*

   Unlike previous method, the proposed method firstly considered different levels of signaling machinery, particularly protein-protein interaction networks, domain-domain interactions, signaling domains, protein functions, which are useful for STN construction. Secondly, the sharing components among STN were detected by soft-clustering. Our method did not separate the networks into individual proteins, but carries out them in associations with other proteins in terms of their functional or physical interactions. In addition, differed from existing methods, our work shifted from yeast STN to human STN, a currently significant challenge.

   For human STN, experimental evaluation showed the high performance of our proposed method. The method is promising to discover new STN and build up the complete pathways. For Yeast STN, the results of signaling domain-domain interaction prediction were comparative with other methods. To discover the roles of signaling domains in STN, the signaling DDI occurring in yeast MAPK STN were predicted and then matched with well-known MARK pathway.

   This work was presented in Chapter 4.

3. *The third contribution is that we developed a new effective method for discovering disease genes by the exploitation of semi-supervised learning, protein-protein interactions and multifarious disease-related features.*

   The key premise is to enrich the disease gene classifier by (1) making use of both known disease genes and as yet unidentified disease genes (unknown genes), and (2) integrating multiple data sources in a semi-supervised learning scheme.

In order to utilize unlabeled data which are available with large volumes and cheap to collect, we proposed a semi-supervised learning method to improve the performance of disease gene predictions. Supportingly, many useful data were extracted for both labeled and unlabeled data. These attractive advantages made our method better than other existing works.

We performed two comparative experiments to evaluate the performance of the method. First, 10 times stratified 10-fold cross validations were conducted using our new semi-supervised learning method, the k-nearest neighbor method, and the Support Vector Machines method . The results show that the SSL method outperforms the other two in terms of sensitivity, specificity, precision, accuracy, and a balanced F-score. Next, we compared our SSL method to the k-NN method with different sizes of labeled sets, and did twenty trials for each experiment to evaluate the accuracy. It turns out that the achieved accuracy of SSL is higher than that of k-NN.

The contributions of this work are not only high computational performance for disease gene prediction but also new significant findings. Considering the whole networks of disease proteins, we found out 568 putative disease genes. Some encouraging results were indirectly validated in various ways, including (i) validating the putative disease gene's keywords and pathways shared with known disease genes, (ii) checking their functional categories and gene similarity via DAVID tools [Dennis et al., 2003]; (iii) testing them with Endeavour – Computer Program For Identifying Disease Genes [Aerts et al., 2006], and (iv) looking their disease-related information up in the literature.

This work was presented in Chapter 5.

6.2 Future Directions

Three proposed methods obtained good results in both computational and biological aspects. Because of the objective and subjective reasons, they still remain some extension to solve the problems completely. As one of the target of this thesis is to detect and combine biologically significant features, some feature selection methods should be applied to find the best features or the best groups of features (for all three problems 1, 2, and 3). However, the feature selection problem itself is another difficult problem in data mining and require much effort. The other problem is the proposed score functions in both problems 2 and 3. The better the score can reflect the significance of the features, the better the integrations of the methods are. We also expecte that the investigation and the development of other algorithms for ILP, soft-clustering, and semi-supervised learning

can improve the results. One of important issues is how to validate the new findings biologically. The cooperation with biologists and doctors will push up the findings to real life. The concrete discussions about shortcomings and improvement can be found in each chapter for each problem.

For long-term plan, we would like to pursue the following promising works.

For protein-protein interaction prediction. The future work is firstly to complete and clean two created databases of ground facts on proteins and protein domains from multiple genomic and proteomic databases. The other work is to choose good positive and negative examples (i.e., labeled and unlabeled examples, respectively) of proteins and protein domains. For the first plan, we will exploit other genome databases. One major requirement for the methods is the databases of ground facts on proteins and protein domains should be updated regularly according to updates of the above-mentioned databases. For the second plan, we will develop a semi-supervised transduction method that uses a proposed similarity measure between proteins to choose and increase the number of good negative examples from multiple protein interaction databases, typically BIND, MINT, Yeast PPI, DIP, and etc. Others are finding PPI network motifs, detecting stable and transient PPI.

For signal transduction network construction. We first would like to consider the whole interaction networks or some functional subnetworks to discover new signal transduction networks. Given starting nodes (e.g., membrane proteins) and ending nodes (e.g, transcription factors), the proposed method can be improved to specify the signal transduction networks and then discover complete signaling pathways. In human disease study, human interaction networks, signal transduction pathways and diseases have very close associations. Signaling network dysfunction can result in abnormal cellular transformation or differentiation, often producing a physiological disease outcome. The further works on identification of disease-related subnetworks are significant and can be investigated through signal transduction networks.

For disease-causing gene prediction. One of the ultimate goals of biological sciences, and certainly one with a high impact on society, is to improve our understanding of the processes and events related to diseases. We approach diseases in terms of disease pathogenic mechanisms by using knowledge from protein-protein interactions, as discovery of the reciprocal relationship between protein-protein interaction networks, signal transduction networks and disease-causing genes. Other clinical data are considered to understand deeply disease pathogenic mechanisms. Other related work is to study the disease pathogenic mechanisms based on the host PPI networks (human PPI networks) and the pathogen PPI. The further expectation is to build up a complete decision support system for for disease diagnostics and drug design.

Bibliography

[Adie et al., 2005] Adie, E., Adams, R. R., Evans, K. L., Porteous, D. J., and Pickard, B. (2005). Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics*, 6(55).

[Aerts et al., 2006] Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., Tranchevent, L.-C., De Moor, B., Marynen, P., Hassan, B., Carmeliet, P., and Moreau, Y. (2006). Gene prioritization through genomic data fusion. *Nature Biotechnology*, 24(5):537–544.

[Alberts, 2002] Alberts, B. (2002). *Molecular biology of the cell*. Garland Science.

[Allen et al., 2006] Allen, E., Fetrow, J., Daniel, L. W., Thomas, S., and John, D. (2006). Algebraic dependency models of protein signal transduction networks from time-series data. *Journal of Theoretical Biology*, 238(2):317–330.

[Asthagiri and Lauffenburger, 2000] Asthagiri, A. R. and Lauffenburger, D. A. (2000). Bioengineering models of cell signaling. *Annual Review of Biomedical Engineering*, 2(1):31–53.

[Asur et al., 2007] Asur, S., Ucar, D., and Parthasarathy, S. (2007). An ensemble framework for clustering protein protein interaction networks. *Bioinformatics*, 23(13):i29–40.

[Bairoch et al., 2005] Bairoch, A., Apweiler, R., Wu, C., Barker, W., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M., Natale, D., O'Donovan, C., Redaschi, N., and Yeh, L. (2005). The universal protein resource (uniprot). *Nucleic Acids Research*, 33:D154–D159.

[Baudot et al., 2006] Baudot, A., Martin, D., and Mouren, P. (2006). PRODISTIN Web Site: a tool for the functional classification of proteins from interaction networks. *Bioinformatics*, 22(2):248–250.

[Bauer and Kuster, 2003] Bauer, A. and Kuster, B. (2003). Affinity purification-mass spectrometry: Powerful tools for the characterization of protein complexes. *Eur. J. Biochem.*, 270(4):570–578.

[Ben-Hur and Noble, 2005] Ben-Hur, A. and Noble, W. S. (2005). Kernel methods for predicting protein-protein interactions. *Bioinformatics*, 21(suppl1):i38–46.

[Ben-Hur and Noble, 2006] Ben-Hur, A. and Noble, W. S. (2006). Choosing negative examples for the prediction of protein-protein interactions. *BMC Bioinformatics*, 7(suppl 1):–.

[Benjamin Schuster Bockler and Alex Bateman , 2008] Benjamin Schuster Bockler and Alex Bateman (2008). Protein interactions in human genetic diseases. *Genome Biology*, 9(1):R9.1–R9.12.

[Bock and Gough, 2001] Bock, J. R. and Gough, D. A. (2001). Predicting protein-protein interactions from primary structure. *Bioinformatics*, 17(5):455–460.

[Brown and Jurisica, 2005] Brown, K. and Jurisica, I. (2005). Online predicted human interaction database. *Bioinformatics*, 21(9):2076–2082.

[Chapelle et al., 2006] Chapelle, O., Scholkopf, B., and Zien, A. (2006). *Semi-Supervised Learning*. The MIT Press.

[Chen et al., 2006] Chen, J., Shen, C., and Sivachenko, A. (2006). Mining Alzheimer Disease Relevant Proteins from Integrated Protein Interactome Data. In *Pacific Symposium on Biocomputing*, volume 11, pages 367–378.

[Chen and Yuan, 2006] Chen, J. and Yuan, B. (2006). Detecting functional modules in the yeast protein-protein interaction network. *Bioinformatics*, 22(18):2283–2290.

[Chen and Liu, 2005] Chen, X. and Liu, M. (2005). Prediction of protein-protein interactions using random decision forest framework. *Bioinformatics*, 21(24):4394–4400.

[Cramer et al., 2001] Cramer, P., Bushnell, D. A., and Kornberg, R. D. (2001). Structural Basis of Transcription: RNA Polymerase II at 2.8 Angstrom Resolution. *Science*, 292(5523):1863–1876.

[Dennis et al., 2003] Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., and Lempicki, R. A. (2003). David: Database for annotation, visualization, and integrated discovery. *Genome Biol*, 4(5).

[Dzeroski and Lavrac, 2001] Dzeroski, S. and Lavrac, N., editors (2001). *Relational Data Mining*. Springer.

[Enright et al., 1999] Enright, A., Iliopoulos, I., Kyrpides, N., and Ouzounis, C. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402:86–90.

[Eungdamrong and Iyenga, 2004] Eungdamrong, N. J. and Iyenga, R. (2004). Modeling cell signaling networks. *Biology of the Cell*, 96(5):355–362.

[Fukuda and Takagi, 2001] Fukuda, K. and Takagi, T. (2001). Knowledge representation of signal transduction pathways. *Bioinformatics*, 17(9):829–837.

[Futschik and Carlisle, 2005] Futschik, M. and Carlisle, B. (2005). Noise-robust soft clustering of gene expression time-course data. *J Bioinform Comput Biol.*, 3(4):965–88.

[Goh et al., 2007] Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabasi, A.-L. (2007). The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690.

[Gomez et al., 2001] Gomez, S. M., Lo, S., and Rzhetsky, A. (2001). Probabilistic Prediction of Unknown Metabolic and Signal-Transduction Networks. *Genetics*, 159(3):1291–1298.

[Hamosh et al., 2005] Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*, 33 Database Issue.

[Han et al., 2003] Han, D., H.S.Kim, J.Seo, and W.Jang (2003). A domain combination based probabilistic framework for protein protein interaction prediction. In *Genome Inform. Ser. Workshop Genome Inform*, pages 250–259.

[Huang and Kecman, 2004] Huang, T. M. and Kecman, V. (2004). SemiL, Software for solving semi-supervised learning problems.

[Ideker and Sharan, 2008] Ideker, T. and Sharan, R. (2008). Protein networks in disease. *Genome Res.*, 18(4):644–652.

[Ito et al., 2001] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. In *Proc. Natl. Acad. Sci. USA 98*, pages 4569–4574.

[Jansen et al., 2003] Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F., and Gerstein, M. (2003). A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data. *Science*, 302(5644):449–453.

[Joachims, 1998] Joachims, T. (1998). Making large-scale support vector machine learning practical. In Scholköpf, B., Burges, C., and Smola, A., editors, *Advances in Kernel Methods: Support Vector Machines*. MIT Press, Cambridge, MA.

[Kann, 2007] Kann, M. G. (2007). Protein interactions and disease: computational approaches to uncover the etiology of diseases. *Brief Bioinform*, 8(5):333–346.

[Kim et al., 2002] Kim, R., Park, J., and Suh, J. (2002). Large scale statistical prediction of protein - protein interaction by potentially interacting domain (PID) pair. In *Genome Inform. Ser. Workshop Genome Inform*, pages 48–50.

[King et al., 1992] King, R., Muggleton, S., Lewis, R., and Sternberg, M. (1992). Drug design by machine learning: The use of inductive logic programming to model the structure activity relationships of trimethoprim analogues binding to dihydrofolate reductase. *Proc. of the National Academy of Sciences of the USA*, 89(23):11322–11326.

[Krauthammer et al., 2004] Krauthammer, M., Kaufmann, C. A., Gilliam, T. C., and Rzhetsky, A. (2004). Molecular triangulation: Bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. *PNAS*, 101(42):15148–15153.

[Kumar and Futschik, 2007] Kumar, L. and Futschik, M. (2007). Mfuzz: A software package for soft clustering of microarray data. *Bioinformation*, 2(1):5–7.

[Lin et al., 2006] Lin, C., Cho, Y., Hwang, W., Pei, P., and Zhang, A. (2006). Clustering methods in protein-protein interaction network. *Knowledge Discovery in Bioinformatics: Techniques, Methods and Application*.

[Liu and Zhao, 2004] Liu, Y. and Zhao, H. (2004). A computational approach for ordering signal transduction pathway components from genomics and proteomics data. *BMC Bioinformatics*, 5(158).

[Marcotte et al., 1999] Marcotte, E. M., Pellegrini, M., and et al., H. L. N. (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428):751–753.

[Mariadason et al., 2002] Mariadason, J. M., Arango, D., Corner, G. A., Aranes, M. J., Hotchkiss, K. A., Yang, W., and Augenlicht, L. H. (2002). A Gene Expression Profile That Defines Colon Cell Maturation in Vitro. *Cancer Res*, 62(16):4791–4804.

[Mark-A and Scheffer, 2004] Mark-A, M. and Scheffer, T. (2004). Multi-relational learning, text mining, and semi-supervised learning for functional genomics: Special issue: Data mining lessons learned. *Machine Learning*, 57(1-2):61+.

[Martin et al., 2005] Martin, S., Roe, D., and Faulon, J. L. (2005). Predicting protein-protein interactions using signature products. *Bioinformatics*, 21:218–226.

[Matthews et al., 2001] Matthews, L. R., Vaglio, P., and et al., J. R. (2001). Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or 'interologs'. *Genome Res.*, 11(12):2120–2126.

[McMahon et al., 1998] McMahon, S., Van Buskirk, H. A., Dugan, K., Copeland, T. D., and Cole, M. D. (1998). The novel atm-related protein trrap is an essential cofactor for the c-myc and e2f oncoproteins. *Cell*, 94:363–374.

[Mintseris and Weng, 2005] Mintseris, J. and Weng, Z. (2005). Structure, function, and evolution of transient and obligate protein-protein interactions. *Proceedings of the National Academy of Sciences*, 102(31):10930–10935.

[Moon et al., 2005] Moon, H., Bhak, J., Lee, K., and Lee, D. (2005). Architecture of basic building blocks in protein and domain structural interaction networks. *Bioinformatics*, 21(8):1479–1486.

[Muggleton, 1992] Muggleton, S. (1992). *Inductive Logic Programming*. Academic Press.

[Muggleton et al., 1993] Muggleton, S., King, R., and Sternberg, M. (1993). Protein secondary structure prediction using logic-based machine learning. *Protein Eng.*, 6(5).

[Muggleton and Raedt, 1994] Muggleton, S. and Raedt, L. D. (1994). Inductive logic programming: Theory and methods. *Journal of Logic Programming*, 19/20:629–679.

[NCBI, 2007] NCBI (2007). *Genes and disease*. National Library of Medicine (US), NCBI.

[Neves and Iyengar, 2005] Neves, S. R. and Iyengar, R. (2005). Modeling Signaling Networks. *Sci. STKE*, 2005(281):tw157–.

[Ng et al., 2003] Ng, S., Zhang, Z., Tan, S., and Lin, K. (2003). InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes. *Nucleic Acids Res*, 31(1):251–254.

[Ng and Tan, 2003] Ng, S. K. and Tan, S. H. (2003). Discovering protein-protein interactions. *Journal of Bioinformatics and Computational Biology*, 1(4):711–741.

[Nguyen and Ho., 2006] Nguyen, T. and Ho., T. (2006). Discovering signal transduction networks using signaling domain-domain interactions. *Genome Informatics*, 17(2):35–45.

[Nguyen and Ho, 2006] Nguyen, T. and Ho, T. (2006). Prediction of Domain-Domain Interactions Using Inductive Logic Programming from Multiple Genome Databases. In *Proceedings of The 9th International Conference on Discovery Science (DS'06)*, pages 185–196.

[Nguyen and Ho, 2007a] Nguyen, T. and Ho, T. (2007a). A Semi-Supervised Learning Approach to Disease Gene Prediction. In *Proceedings of 2007 IEEE International Conference on BioInformatics and BioMedicine (BIBM 2007)*, pages 423–428.

[Nguyen and Ho, 2007b] Nguyen, T. and Ho, T. (2007b). Combining Domain Fusions And Domain-Domain Interactions To Predict Protein-Protein Interactions. In *Proceedings of The 7th International Workshop on Data Mining in Bioinformatics (BIOKDD '07), ACM SIGKDD '07*, pages 27–34.

[Nguyen and Ho, 2007c] Nguyen, T. and Ho, T. (2007c). Prediction of protein-protein interactions and disease genes by machine learning. In *Proceedings of international Workshop on Data Mining and Statistical Science (DMSS2007)*, pages 51–70.

[Nguyen and Ho, 2008a] Nguyen, T. and Ho, T. (2008a). A soft-clustering method for deriving signal transduction networks from protein-protein interaction networks. In *Proceedings of The 7th International Workshop on Data Mining in Bioinformatics (BIOKDD '08), ACM SIGKDD '08.* (submitted).

[Nguyen and Ho, 2008b] Nguyen, T. and Ho, T. (2008b). An Integrative Domain-Based Approach to Predicting Protein-Protein Interactions. *Journal of Bioinformatics and Computational Biology.* (in press).

[Nguyen and Ho, 2008c] Nguyen, T. and Ho, T. (2008c). Discovering Disease Genes Using Protein-Protein Interaction Networks and Semi-Supervised Learning. *Bioinformatics.* (submitted).

[Nguyen et al., 2007] Nguyen, T., Nguyen, N., and Ho, T. (2007). Prediction of Protein-Protein Interactions Using Bayesian Networks. In *Proceedings of The 2nd International Conference on Knowledge, Information and Creativity Support Systems (KICSS 2007)*, pages 207–214.

[Ofran and Rost, 2007] Ofran, Y. and Rost, B. (2007). ISIS: interaction sites identified from sequence. *Bioinformatics*, 23(2):e13–16.

[Oti et al., 2006] Oti, M., Snel, B., Huynen, M. A., and Brunner, H. G. (2006). Predicting disease genes using protein-protein interactions. *J Med Genet*, 43(8):691–698.

[Owsianka and Patel, 1999] Owsianka, A. and Patel, A. H. (1999). Hepatitis c virus core protein interacts with a human dead box protein ddx3. *Hepatitis C Virus Core Protein Interacts with a Human DEAD Box Protein DDX3*, 257:330 – 340.

[Page and Craven, 2003] Page, D. and Craven, M. (2003). Biological applications of multi-relational data mining. In *SIGKDD Explorations*, volume 5, pages 69–79.

[Pawson et al., 2002] Pawson, T., Raina, M., and Nash, N. (2002). Interaction domains: from simple binding events to complex cellular behavior. *FEBS Letters*, 513(1):2–10.

[Pellegrini et al., 1999] Pellegrini, M., Marcotte, E. M., and et al., M. J. T. (1999). Assining protein functions by comparative genome analysis: Protein phylogenetic profiles. In *Proc. Natl. Acad. Sci. USA 96(8)*, pages 4285–4288.

[Reichmann et al., 2005] Reichmann, D., Rahat, O., Albeck, S., Meged, R., Dym, O., and Schreiber, G. (2005). From The Cover: The modular architecture of protein-protein binding interfaces. *PNAS*, 102(1):57–62.

[Rhodes et al., 2005] Rhodes, D. R., Tomlins, S. A., Varambally, S., Mahavisno, V., Barrette, T., Kalyana-Sundaram, S., Ghosh, D., Pandey, A., and Chinnaiyan, A. M. (2005). Probabilistic model of the human protein-protein interaction network. *Nat Biotech*, 23(8):1087–0156. http://www.nature.com/nbt/journal/v23/n8/suppinfo/nbt1103_S1.html.

[Ryan and Matthews, 2005] Ryan, D. and Matthews, J. (2005). Proteinprotein interactions in human disease. *Current Opinion in Structural Biology*, 15(4):441–446.

[Smalter et al., 2007] Smalter, A., Lei, S., and Chen, X.-w. (2007). Human disease-gene classification with integrative sequence-based and topological features of protein-protein interaction networks. In *IEEE BIBM 2007*, pages 209–216.

[Smith, 1985] Smith, G. P. (1985). Filamentous fusion phage: Novel expression vectors that display cloned antigens on the virion surface. *Science*, 228(4705):1315–1317.

[Sprinzak and Margalit, 2001] Sprinzak, E. and Margalit, H. (2001). Correlated sequence-signatures as markers of protein-protein interaction. *Journal of Molecular Biology*, 311(4):681–692.

[Steffen et al., 2002] Steffen, M., Petti, A., Aach, J., D'haeseleer, P., and Church, G. (2002). Automated modelling of signal transduction networks. *BMC Bioinformatics*, 3(34).

[Tran et al., 2005] Tran, T., K.Satou, and T.B.Ho (2005). Using inductive logic programming for predicting protein-protein interactions from multiple genomic data. In *PKDD*, pages 321–330.

[Truong and Ikura, 2003] Truong, K. and Ikura, M. (2003). Domain fusion analysis by applying relational algebra to protein sequence and domain databases. *BMC Bioinformatics*, 4(16):1–10.

[Tu et al., 2006] Tu, Z., Wang, L., Xu, M., Zhou, X., Chen, T., and Sun, F. (2006). Further understanding human disease genes by comparing with housekeeping genes and other genes. *BMC Genomics*, 7(31).

[Turcotte et al., 1998] Turcotte, M., Muggleton, S., and Sternberg, M. (1998). Protein fold recognition. In *Proc. of the 8th International Workshop on Inductive Logic Programming (ILP-98)*, pages 53–64.

[Turner et al., 2003] Turner, F. S., Clutterbuck, D. R., and Semple, C. (2003). Pocus: mining genomic sequence annotation to predict disease genes. *Genome Biology*, 4(R75).

[Ucar et al., 2006] Ucar, D., Asur, S., Catalyurek, U., and Parthasarathy, S. (2006). Improving Functional Modularity in Protein-Protein Interactions Graphs using Hub-induced Subgraphs. *PKDD*, page i371382.

[Uetz et al., 2000] Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J. M. (2000). A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, 403(6770):623–627.

[Uetz and Vollert, 2006] Uetz, P. and Vollert, C. (2006). Protein-Protein Interactions. *Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine*, 17:–.

[Weston et al., 2005] Weston, J., Leslie, C., Ie, E., Zhou, D., Elisseeff, A., and Noble, W. S. (2005). Semi-supervised protein classification using cluster kernels. *Bioinformatics*, 21(15):3241–3247.

[Witten and Frank, 2005] Witten, I. H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

[Wojcik and Schachter, 2001] Wojcik, J. and Schachter, V. (2001). Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics*, 17(suppl1):S296–305.

[Xu and Li, 2006] Xu, J. and Li, Y. (2006). Discovering disease-genes by topological features in human protein-protein interaction network. *Bioinformatics*, 22(22):2800–2805.

[Yan et al., 2004] Yan, C., Dobbs, D., and Honavar, V. (2004). A two-stage classifier for identification of protein-protein interface residues. *Bioinformatics*, 20(suppl1):i371–378.

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

[Zhao et al., 2008] Zhao, X., Wang, R., Chen, L., and Aihara, K. (2008). Automatic modeling of signal pathways from protein-protein interaction networks. In *The Sixth Asia Pacific Bioinformatics Conference*, pages 287–296.

[Zhou et al., 2004] Zhou, D., Bousquet, O., Lal, T. N., Weston, J., and Schlkopf, B. (2004). Learning with local and global consistency. *Advances in Neural Information Processing Systems*, 16:321–328.

[Zhu, 2005] Zhu, X. (2005). Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison. http://www.cs.wisc.edu/∼jerryzhu/pub/ssl_survey.pdf.

Publications

[1] Nguyen, T. P. and Ho, T.B., (2008). An Integrative Domain-Based Approach to Predicting Protein-Protein Interactions. *Journal of Bioinformatics and Computational Biology*, Imperial College Press. (in press)

[2] Nguyen, T. P. and Ho, T.B., (2008). Discovering Disease Genes Using Semi-Supervised Learning and Protein-Protein Interaction Networks. *Bioinformatics*, Oxford Press. (submitted)

[3] Nguyen, T. P., Satou, K. and Ho, T.B., (2008). Constructing Signal Transduction Networks Using Multiple Signaling Feature Data. *Statistical and Relational Learning in Bioinformatics 2008 Workshp, PKDD/ECML 2008.* (accepted)

[4] Nguyen, T.P. and Ho, T.B. (2007). Prediction of protein-protein interactions and disease genes by machine learning, *International Workshop on Data Mining and Statistical Science (DMSS2007)*, October 5-6, Tokyo, Japan, 51-70, The Institute of Statistical Mathematics Press.

[5] Nguyen, T.P. and Ho, T.B., Nguyen, N.B. (2007). Prediction of Protein-Protein Interactions Using Bayesian Networks, *The 2nd International Conference on Knowledge, Information and Creativity Support Systems (KICSS 2007)*, November 5-7, Ishikawa, Japan, 207-214, JAIST Press.

[6] Nguyen, T.P. and Ho, T.B. (2007), A Semi-Supervised Learning Approach to Disease Gene Prediction, *2007 IEEE International Conference on BioInformatics and BioMedicine (BIBM 2007)*, November 2-4, San Jose, CA, USA, 423-428, IEEE Society Press.

[7] Nguyen, T.P. and Ho, T.B. (2007), Combining Domain Fusions And Domain-Domain Interactions To Predict Protein-Protein Interactions, *The 7th International Workshop on Data Mining in Bioinformatics (BIOKDD '07), ACM SIGKDD '07*, August 12, San Jose, CA, USA, 27-34, ACM Digital Library.

[8] Ho, T.B. and Nguyen, T.P. and Tran, T.N. (2007), Study of protein-protein interactions from multiple data sources. *Data Mining and Knowledge Discovery Technologies*, David Taniar (Ed.), IGI Global Publishers (in press).

[9] Nguyen, T. P. and Ho, T.B., (2006). Discovering Signal Transduction Networks Using Signaling Domain-Domain Interactions. *Genome Informatics*, Vol. 17, No. 2, 35-45, Universal Academic Press.

[10] Nguyen, T.P. and Ho, T.B. (2006), Prediction of Domain-Domain Interactions Using Inductive Logic Programming from Multiple Genome Databases, *The 9th International Conference on Discovery Science (DS'06)*, October 7-10, Barcelona. Lecture Notes in Artificial Intelligence LNAI 4265, 185-196, Springer.

[11] Nguyen, T.P. and Ho,T.B., Nguyen, N.B. (2005), Discovering Reliable Protein Interactions Using Bayesian Networks, T*he 16th International Conference on Genome Informatics (GIW'05)*. December 19-21, Yokohama. P038.1-2.