DISCOVERING DOMAIN-DOMAIN INTERACTIONS TOWARD GENOME-WIDE PROTEIN INTERACTION AND FUNCTION PREDICTIONS

BY

Mei Liu

Submitted to the graduate degree program in Electrical Engineering & Computer Science and the Graduate Faculty of the University of Kansas School of Engineering in partial fulfillment of the requirements for the degree of Doctor of Philosophy

_____
Chairperson

Committee members* _____*

_____*

_____*

_____*

Date Defended: _____

The Dissertation Committee for Mei Liu certifies that this is the approved version of the following dissertation:

DISCOVERING DOMAIN-DOMAIN INTERACTIONS TOWARD GENOME-WIDE PROTEIN INTERACTION AND FUNCTION PREDICTIONS

Committee:

_____
Chairperson*

_____

_____

_____

_____

Date Approved: _____

Acknowledgements

I would like to express my deepest gratitude to my advisor, Dr. Xue-wen Chen, for his effective guidance and support over the years that helped me in completing my dissertation work. I am also grateful to Dr. Robert Ward and Dr. Jun (Luke) Huan for their expert advices in biology during different stages of my research and Dr. Agah and Dr. Grzymala-Busse for serving on my committee. Thanks also go to Mike Wasikowski for proofreading my dissertation.

Most importantly, none of this would have been possible without the love and support of my family. My parents, Qiwu Liu and Zhuye Jin, have provided me with all their love, support, and strength my entire life. My husband, Yong Hu, has comforted and encouraged me throughout this endeavor. I would like to express my heart-felt gratitude to them.

Abstract

To fully understand the underlying mechanisms of living cells, it is essential to delineate the intricate interactions between the cell proteins at a genome scale. Insights into the protein functions will enrich our understanding in human diseases and contribute to future drug developments. My dissertation focuses on the development and optimization of machine learning algorithms to study protein-protein interactions and protein function annotations through discovery of domain-domain interactions. First of all, I developed a novel domain-based random decision forest framework (RDFF) that explored all possible domain module pairs in mediating protein interactions. RDFF achieved higher sensitivity (79.78%) and specificity (64.38%) in interaction predictions of *S. cerevisiae* proteins compared to the popular Maximum Likelihood Estimation (MLE) approach. RDFF can also infer interactions for both single-domain pairs and domain module pairs. Secondly, I proposed cross-species interacting domain patterns (CSIDOP) approach that not only increased fidelity of existing functional annotations, but also proposed novel annotations for unknown proteins. CSIDOP accurately determined functions for 95.42% of proteins in *H. sapiens* using 2,972 GO 'molecular function' terms. In contrast, most existing methods can only achieve accuracies of 50% to 75% using much smaller number of categories. Additionally, we were able to assign novel annotations to 181 unknown *H. sapiens* proteins. Finally, I implemented a web-based system, called PINFUN, which enables users to make online protein-protein interaction and protein function predictions based on a large-scale collection of known and putative domain interactions.

Keywords: Systems Biology, Bioinformatics, Computer Science, Data Mining, Machine Learning, Protein Interaction Network, Domain Interaction Network, Protein Function

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1. Introduction

The science of biology was established in the early 19th century as scientists discovered that living organisms on earth shared fundamental characteristics, and ever since, people have become increasingly fascinated by the origin, evolution, classification, structure, function and growth of all living organisms. We are genuinely intrigued by the seemingly simple yet intricate question, "What is life?"

Over the last half century, molecular biologists have adopted reductionist thinking in understanding life from a scientific point of view by breaking living organisms down to their fundamental components, the individual genes and molecules. The principle behind reductionist thinking is that we can figure out what happens at the higher-levels from what we observe at the lower-levels. As Denis Noble interpreted in his book, the reductionist causal chain for living organisms runs entirely 'one-way' upward, from genes to the organism [1]. It is commonly known that genes produce proteins, the proteins constitute cells, and the cells organize into tissues such as skin, bone, and muscle. Then the tissues form organs such as heart and kidney, and finally, all the organs together with immune and hormonal systems form the organism (Figure 1.1).

Throughout the years, reductionist thinking and methods have produced tremendous amount of knowledge about the molecular static properties. It has answered many questions but raised many more at the same time. The main challenge now is to work out how to extend this lower-level knowledge up to entire living systems. This is not an easy problem. A living organism is a complex system featuring a large number of simple and identical components (e.g. genes and proteins) interacting with each other whose collective activity is nonlinear. When we move from proteins to their interactions, the problem becomes seriously complicated. Yet understanding the complexities is only the beginning to answer the larger

question “What is life?” From genes to organisms (Figure 1.1), assorted components at each level are embedded in an integrated system where each has its own logic. Carlson and Doyle described an interesting analogy of biological cells to central processing units (CPU) in computers [2]. Each cell or CPU is considered to be a complex system composed of thousands of components i.e. genes and proteins or transistors. But these systems themselves are also components embedded in larger systems such as organs or control systems of machines. This encapsulation continues upwards to even larger networks that make up organisms, ecosystems, and computer networks [2]. Thus, it is not possible to learn the true logic of each system by only examining properties of its individual components. In fact, we are confronted by a crucial challenge today: will we be able to reassemble the little pieces back together to their original form again? This is where systems biology comes into play.

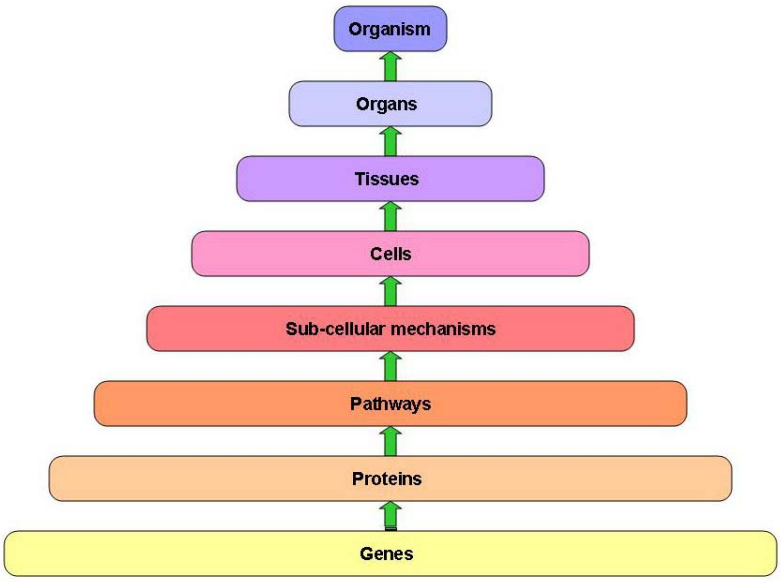

Figure 1.1 Tree of life – the reductionist causal chain of living organisms

Systems biology is a relatively new field in biological study that focuses on the systematic understandings of complex interactions in biological systems. Different people

bear different perspectives on the field. There are those who see it as a logical continuation of functional genomics where genome-scale experiments are carried out to understand better the entirety of processes that occur in a biological system [3, 4]. Their aim is to understand how the whole is greater than the sum of its parts. On the other hand, others see systems biology as a branch of mathematical biology which involves the study of small systems [5, 6]. For these small systems, it is assumed that sufficient parameters can be measured to allow simulations of how the individual molecules function together to achieve a particular outcome. Aloy and Russell [7] view systems biology as both of these things. Indeed, studying single macromolecules no longer dominates the current trend of molecular biology. Now, it is the norm to analyze how groups of molecules behave together in complexes, pathways, or even complete organisms.

Over the years, genome-sequencing projects have provided an encouraging domain for the studies in systems biology by producing a near complete list of the components that are present in various organisms. Post-genomic efforts have aspired to determine relationships between those catalogued components. Essentially, systems biology is about understanding these complex relationships when all components are considered together in a biological network. For example, full comprehension of the metabolic and signaling pathways or gene-regulatory networks relies on the detailed discernment of the interactions between proteins, metabolites, and nucleic acids.

In this dissertation, I will address two specific problems: (1) elucidating the complex interplays between proteins and (2) inferring functions of uncharacterized proteins. Section 1.1 discusses the problems and their significance in detail. Section 1.2 and 1.3 provides in-depth reviews of related works.

1.1 Significance of Protein Interaction & Function Predictions

Proteins are the major components of living organisms and the fundamental units of life. The entire complement of proteins expressed at any given time under defined conditions by a genome, cell, tissue, or organism is called proteome. Compared to an organism's genome, the proteome is much more dynamic because it changes during development due to external stimuli, and the proteins can form large interaction networks to regulate and support each other. Proteins are fascinating molecular devices, in which they exhibit a wide variety of roles in cellular processes such as composing cellular structure, promoting chemical reactions, carrying messages from one cell to another and acting as antibodies. For example, structural proteins define the physical shape of cells and other parts of our body. Enzymatic proteins are the most varied and the most highly specialized proteins that catalyze different kinds of chemical reactions. Moreover, there are transport proteins, storage proteins, receptor proteins involved in cell's response to chemical stimuli, hormonal proteins in coordinating bodily activities, contractile proteins important in movement, and defensive proteins for our immune system.

For an increasing number of organisms, near-complete lists of genes and encoded proteins are made available as a result of the genome sequencing projects. Sequence data is deposited at an exponential rate. In the human genome alone, it is estimated that there are approximately 20,000 to 25,000 protein coding genes [8]. Nowadays, amino acid sequences are available for millions of proteins. However, in order to fully understand the cellular machinery, simply cataloging the protein sequences is not enough. The multiplicity of functions that proteins execute in most cellular processes and biochemical events is attributed to their interactions with other proteins. Thus, it is necessary to delineate the intricate interplays between proteins in the post-genomic era.

Fundamentally, protein-protein interactions can be studied from two different perspectives. In the traditional view, the goal has been to study each interaction individually to understand the physical interaction mechanism between two proteins. These individual interactions can be determined through experimental methods such as genetic, biochemical, and biophysical techniques. To expedite the interaction discovery process, the more recent 'high-throughput' view of protein interactions emerged. It aims to understand the system of interactions as a whole by treating proteins as logical entities in which their interactions can be visualized as a network. The high-throughput experimental technologies include yeast two-hybrid system, tandem affinity purification, and mass spectrometry.

However, results from two proteome-wide screens in the *Saccharomyces cerevisiae* [9, 10] yielded very little overlap. In two later studies by mass spectrometry of purified protein complexes [11, 12], the results again exhibited little overlap with each other as well as with the interactions from yeast two-hybrid analysis. Such unexpected results have prompted speculations that the little overlap between studies may be caused by two factors. First, the large-scale interaction screens may be associated with high error rates. Second, the number of interactions in yeast is probably much larger than what was expected.

As a matter of fact, in the most extensively studied organism, *Saccharomyces cerevisiae*, there are approximately 6,000 proteins, and more than 80,000 interaction pairs are recorded up to date. *Drosophila melanogaster* is much bigger compared to *Saccharomyces cerevisiae*, which consists of roughly 13,600 genes. In yeast, it was estimated that each protein may be involved in 3-10 interactions [13]. Assuming that this also holds true for the *Drosophila* proteins, the total number of interaction pairs would undoubtedly exceed 100,000. The protein interaction statistics in BioGRID [14] indicates that there are only 32,852 interaction pairs currently observed in *D. melanogaster*. Moreover, the human proteome contains close

to 100,000 proteins, which would lead to a conservative estimate of at least 300,000 interactions, and there are only ~35,000 pairs currently available in Human Protein Reference Database (HPRD) [15]. Hence, we are still far away from full understanding of the protein interactomes. Within one cell, the number of possible interactions is enormous, and this presents a potentially limiting factor for experimental analyses. It is simply infeasible to experimentally study each and every protein pair in all proteomes. Therefore, we need fast, cost-effective and reliable *in silico* approaches to extrapolate the accumulated knowledge of characterized proteins to the uncharacterized proteins.

It is these interaction linkages between proteins that provide the basis for a precise understanding of cellular pathways. Since proteins play a central role in an organism's life, they may guide the discovery of biomarkers that are indicative of a particular disease. There are many proteins specific to pathogens that we may want to deactivate. A complete protein interaction network may ultimately help us to understand disease mechanisms and facilitate the development of therapies optimized for efficacy.

Although most amino acid sequences of proteins encoded by the genome may be known, only a fraction of the protein functions have been annotated. For instance, among the current list of *Drosophila* genes downloaded from FlyBase in November 2006 [16], only 54% are annotated with "molecular function" terms in Gene Ontology (GO) [17]. Additionally, many proteins are modular, consisting of multiple functional domains, so the existing annotations may still be incomplete. While experimental methods such as loss of function mutational analysis, RNAi, or targeted misexpression approaches have been very successful in identifying protein functions, they are labor-intensive and time-consuming. As a result, much of the genome-wide functional annotations are based upon *in silico* methods. Most function prediction algorithms can predict protein functions with 50% - 75% accuracy, and this

performance may not be of practical use for biologists. Moreover, some methods use only several tens to hundreds of functional categories in the prediction process which results in more generic rather than specific functional assignments. Therefore, developing more effective *in silico* methods to increase the fidelity of these functional annotations and to propose novel functions for currently uncharacterized proteins presents a major challenge to the life science community. The research will eminently aid the biological community as higher quality functional annotations are often used by scientists to generate new hypotheses and direct their research focus.

Our understanding of protein functions and disease mechanisms is still under-developed relative to the proportion of available supporting biological data. Post-genomic biological discoveries have confirmed that proteins execute multiple biochemical functions in extended networks. In particular, many proteins must physically bind to other proteins, either stably or transiently, to perform their functions. Hence, the functions proteins exhibit are inseparable from their interactions. It is only possible for us to study the function of a living cell when we understand the structure and dynamics of the complex web of protein interactions. The goal of my dissertation research is not only to study the protein interaction network, but also to elucidate the unknown protein functions.

1.2 Methods for Protein Interaction Prediction

There exist numerous computational approaches for protein interaction discovery. Some earlier methodologies focused on estimating the interaction sites by recognizing specific residue motifs [18] or by using features and properties related to interface topology, solvent accessible surface area and hydrophobicity [19, 20]. Many computational techniques are based on other interesting features and combination of features. Table 1.1 lists different

prediction methods and for which purpose each is designed for. In this dissertation, these methods are classified into two main categories: comparative genomic and classification approaches. Details are discussed in the following sections.

Table 1.1 Different protein interaction prediction methods

Method Name	Protein (P) / Domain (D) Interaction	Physical Binding (P) / Functional Association (F)
Gene co-expression	P	F
Gene neighborhood / cluster	P	F
Rosetta Stone	P, D	F
Phylogenetic profile	P, D	F
Sequence co-evolution	P, D	F
Protein structure	P, D	P
Interaction network topology	P	P, F
Interlog	P, D	P, F
Amino acid physiochemical property	P	P
Kmers	P	P
Motifs	P	P
Domain association	D	P
Domain profile pairs (IDPP)	P	P
Set cover	D	P
Probabilistic model	P, D	P, F
Domain pair exclusion	D	P
Parsimonious explanation	D	P
Domain combination	D	P
Data Integration	P, D	P, F

First column lists different prediction methods, while the second column shows if each method is designed to predict protein (P) or domain (D) interactions. The third column shows if the method is intended to infer direct physical interaction (P) of indirect function associations (F). Note that methods based on protein domain can subsequently be used to predict protein interactions. The predicted protein interactions can be used to identify potential protein functional annotations.

1.2.1 Comparative Genomic Approaches

In the late 1990s, a class of computational methods to decipher protein-protein interactions emerged on the basis of genomic context. Genomic context is described as any statistical, physical or biological properties of genes that can be measured. The unifying theme is to propose protein interactions for which there is evidence of an association. The association can be similarities either in gene expression patterns or how the genes are placed relative to each

other in known genomes (e.g. chromosomal location). Interactions derived from a genomic context do not necessarily imply a direct physical binding, but instead suggest functional associations between proteins. For example, proteins at opposite ends of a single pathway or complex may yield the same signal as those proteins in tight, direct, physical contact. In addition, errors in the underlying genomic context data may lead to false positives and false negatives. Caution should be taken when applying the context based methods in eukaryotic genomes because some are tuned primarily in prokaryotes. The following sections describe each genomic context in detail.

Gene Co-expression

The functionality of a protein complex is often determined by the functionality of all its subunits, and these subunits are often observed to be co-expressed or correlated (Figure 1.2). The gene expression profile data can be obtained from cell cycle experiments and expression flunctuations of a gene under different conditions. Similarity between expression profiles is defined as the correlation coefficient between relative expression levels of two genes or encoded proteins [21-24]. Researchers have observed that interacting proteins are much more likely to have their genes co-expressed than the non-interacting ones [21, 25-31]. Moreover, expression levels of protein pairs that physically interact tend to co-evolve, and this co-evolution of gene expressions can be a better predictor than the co-evolution of sequences [31].

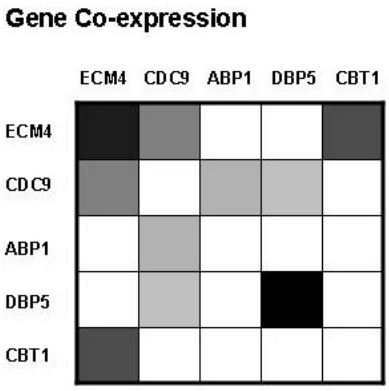

Figure 1.2 PPI prediction based on gene co-expression

Local Context of Genes

It is a well-known fact that bacterial genomes are organized into regions that tend to code for functionally related proteins called operons. These co-transcribed genes, or gene clusters, often fill related function roles, bind to one another, or act in the same metabolic pathway. Researchers have developed different methods to predict operons based on intergenic distances [32-36] (Figure 1.3). This gene clusters relationship sometimes is conserved in different species which forms a gene neighborhood (Figure 1.3). The co-regulated genes determined by the gene neighborhood based methods can provide additional evidence about functional linkages between the genes [37-40]. Research groups analyzed the gene order conservation in three bacterial and archaeal genomes and found that 63% to 75% of the co-regulated genes interact physically [37, 41]. Similar results were observed in some eukaryotes [27]. Nevertheless, this method has a major limitation: it may only be directly applicable to bacteria genomes where the genome order is a consistent property. In addition, prediction of unknown operons is a difficult and error-prone procedure.

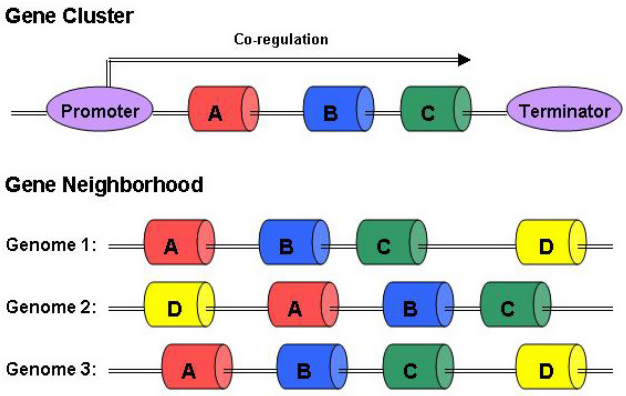

Figure 1.3 PPI prediction based on local context of genes – gene cluster and gene neighborhood. Different boxes represent different genes.

"Rosetta Stone"

Many researchers deduced interactions between proteins from the same protein domain occurrence in different genomes. This is called the gene fusion method or "Rosetta stone" (Figure 1.4) [42-46]. Research groups have detected that certain protein families in a given species consist of fused domains, and such proteins are referred to as composite proteins. Each domain of the composite proteins sometimes corresponds to single and full-length proteins in other species, and these are called component proteins. Based on this gene fusion event, one can conclude that the two component proteins may interact with each other. Analysis found that more than half of the interacting protein pairs proposed by the Rosetta stone approach were functionally related [42].

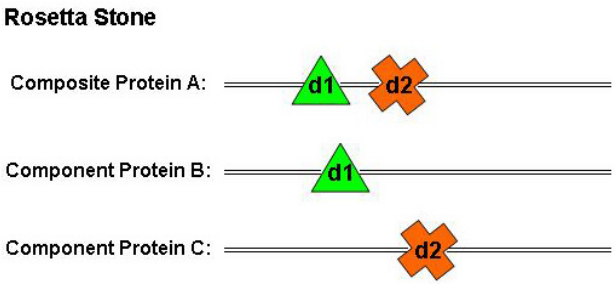

Figure 1.4 PPI prediction based on Rosetta Stone

Phylogenetic Profiles

The fundamental idea of phylogenetic profile (PP) based methods is to discover patterns of the presence or absence of a given gene in a set of genomes. The basic assumption is that under evolutionary pressure, it is necessary for an organism to encode both or either of the proteins within its genome because encoding just one lowers its fitness. Therefore, genes with similar phylogenetic profiles are more likely to be functionally connected than those with different profiles [47-51].

A phylogenetic profile of a protein is constructed as a vector of *N* elements where *N* is the number of genomes (Figure 1.5). Each element of a profile can be either "0" or "1", which indicates the absence or presence of a given gene, respectively. Profile similarity can be calculated using distance measures including Hamming distance and Pearson correlation coefficient. Proteins can then be clustered based on these distance measures, and proteins from the same cluster may be considered as functionally related. Some researchers have even taken higher-order relationships between proteins into consideration [52, 53], and some have incorporated phylogenetic tree to analyze the correlated gains and losses of protein pairs [54].

Despite the promising results, the PP method suffers from high computational cost and its dependence on high information profiles. Moreover, the PP method is more suited for interaction prediction in prokaryotic genomes. In prokaryotic genomes, functionally related genes are often transferred as a unit between organisms. Thus, the group of genes can maintain their association directly in various genomes. Secondly, those linked genes are often located near each other; they have a higher probability of being transferred or lost at the same time. Eukaryotic genomes, on the other hand, may not have these properties, which may result in less informative profiles. In order for PP to perform well, one must select the optimal reference organism by picking organisms that are evolutionarily farthest from the rest [50].

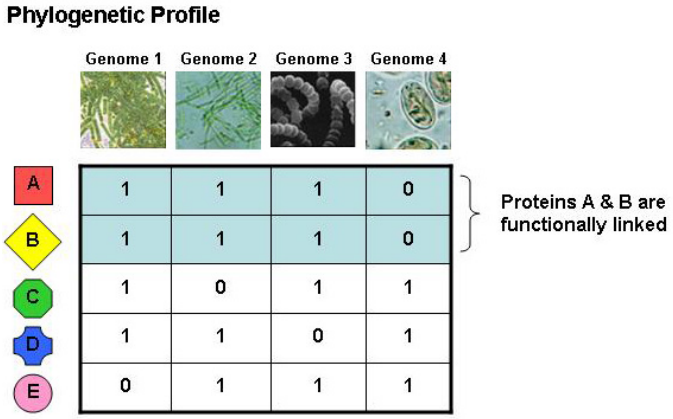

Figure 1.5 PPI prediction based on phylogenetic profile – showing absence (0) or presence (1) of four proteins in four genomes.

Sequence Co-evolution

Many groups observed interacting proteins to evolve at a similar rate. The property of co-evolution is realized when changes in one protein lead to the loss of function or compensated interaction by correlated changes in the other protein. Different types of computational techniques exist based on genomic sequence analysis, including analyzing correlated mutations in amino acid sequences between interacting proteins [55, 56] and exploring the similarity of phylogenetic trees (Figure 1.6) [57-59].

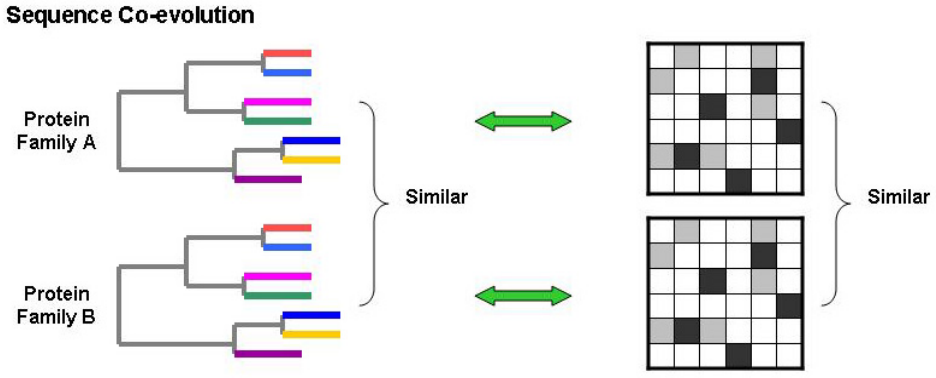

Figure 1.6 PPI prediction based on sequence co-evolution. It looks for similarity between distance matrices of two phylogenetic trees.

The aim of the correlated mutation based method is to quantify the degree of co-variation between two residues from pairs of proteins. On the other hand, the phylogenetic tree based method studies co-evolution in terms of similarity between phylogenetic trees of two non-homologous interacting protein families. It is believed that phylogenetic trees of interacting proteins show a greater degree of similarity or symmetry than expected in non-interacting proteins. The similarity between two phylogenetic trees can be quantified by the correlation coefficient between distances matrices employed in constructing the trees. Each element in a distance matrix corresponds to a tree branch. Agreements between branches of two trees are required in computing correlation coefficients; however, such information is not always available. Several groups have tried to address this issue by identifying specific interaction partners between two interacting families [60-62]. It was found later that similarity between two phylogenetic trees may be influenced by the speciation process which implies that there is always a "background" similarity between trees of any proteins. In order to account for the "background" similarity, groups have developed different statistical techniques for "phylogenetic subtraction" [63-65].

1.2.2 Classification Approaches

Researchers have proposed different classification methods to infer protein interaction partners using various data sources. The basic idea is to train a classifier to find patterns that can distinguish between interacting and non-interacting protein pairs. The following sections discuss the different information sources employed.

Protein Structure

The three-dimensional (3D) structure of a protein is determined by its amino acid sequence. In a given environment and physiological conditions, a protein can only assume one 3D

structure. The biochemical function of a protein predominantly depends upon this unique 3D structure. Consequently, researchers have adopted the structural information of proteins in protein interaction prediction [66-69]. The 3D structure information can enhance our knowledge in how protein molecules interact because they provide crucial atomic details about binding.

Unfortunately, determining protein 3D structures is difficult. Structural information can be experimentally determined by X-ray crystallography or nuclear magnetic resonance (NMR) experiments. It took decades to solve the first X-ray structure. Nowadays, individual protein structures can be learned in a matter of weeks if sufficient material is available, but obtaining sufficient material can be an enormous problem when we have large complexes of proteins. Complex assembly requires precise control and timing in the cell which is not easy to reproduce *in vitro*. In addition, these techniques are labor intensive and time consuming, and not all proteins' structures in the universe can be determined experimentally.

Several groups have proposed *in silico* methods to predict atomic details involved in a pair of interacting proteins. The classic docking approach attempts to find the docked complex based on the shape or electrostatic complementarity between interfaces. Instead of distinguishing pairs of proteins that interact from those that do not, the current methods only search for the optimal fit between two proteins. In the past five years, a new class of techniques has emerged. It models interacting structures by homology. Basically, it uses protein-protein complexes with known structural data to model interactions between their homologues by assessing how well a homologous sequence pair 'fit' onto a previously determined structure of a complex. These approaches suffer when the interactions undergo conformational changes at the interface. Even though it is increasingly rare to not find a protein with structural information from existing database or through homology, the current

3D structure based approaches still struggle with what it can deliver. Large protein complexes and entire systems would require years of study in order to reach a detailed understanding.

Orthology and Interaction Network Topology

Many researchers think that protein interaction pairs co-evolve across multiple organisms, and these conserved interactions are referred to as interlogs. Walhout et al. [58] showed that many interactions in signal transduction pathways or molecular machines are conserved across different species. Matthews et al. [70] applied BLAST [71] to search for potential orthologs of known interacting pairs and attempted to identify possible conserved interactions.

Several research groups strived to determine topological structure of a protein interaction network in order to infer protein interactions [72, 73]. A protein interaction network is usually created using available high-throughput protein interaction data in which network nodes represent the proteins and connected edges depict interaction. Sharan et al. [74] performed multiple comparisons between interaction networks of model organisms to discover conserved network topological patterns. More specifically, they searched for two types of conserved subnetwork structures: short linear paths of interacting proteins, which may model signal transduction pathways, and dense clusters of interactions, which may model protein complexes.

Protein Sequence

This group of methods intends to explore characteristic sequences, or structural motifs that distinguish interacting proteins from non-interacting ones based on the protein primary structure, namely their amino acid sequences [75]. Bock and Gough [76] introduced a method to predict protein interactions based on physiochemical properties of the associated residues in protein sequences. Hydrophobicity profiles have been demonstrated to be sensitive

descriptors of local interaction sites, and Bock and Gough extended it by including more properties such as charge and surface tension. Martin et al. [77] later proposed a method strictly based on the protein sequences to identify signature patterns. In their study, a signature is similar to a k-tuple (subsequences) except it is represented in a tree structure instead of linear string. Fundamentally, their goal is to detect subsequence pairs more likely to occur together when two proteins interact.

Integrative Method

Recently, numerous independent research groups have been exploiting the idea of combining various interaction evidences from different data sources such as gene expression, gene fusion, correlated mutations, etc. for interaction prediction. von Mering et al. [78] is among the first to explore the idea of data integration in assessing the reliability of inferred protein interactions. They validated protein-protein interactions (PPIs) based on overlapping interaction data from multiple sources including yeast two-hybrid system, mass spectrometry of purified protein complexes, correlated mRNA expression profiles, genetic interaction data, and computationally predicted interaction data. At the same time, numerous other groups have tried to combine a variety of data sources through intersection [79], association rule discovery to uncover PPI related knowledge [80], and greedy algorithms in which a dataset with the lowest error rate is added successively until a good compromise between the error rate and coverage is reached [81].

Although promising, the aforementioned integrative methods do not consider error rates in individual data sources. To address the issue, researchers combined heterogeneous data sources to infer protein interactions *de novo* according to reliabilities of each independent evidence source using Bayesian statistics [82-86]. The reliability of each source is analyzed

by comparison against the known true positives and true negatives which are referred to as gold-standards. Gilchrist et al. [87] employed a statistical model to analyze tandem affinity purification (TAP) and high-throughput mass spectrometric protein complex identification (HMS-PCI) datasets. They considered the number of experimental trials performed between bait and prey proteins and the number of associations observed. One shortfall of their method is that they limit the datasets to provide information on the same level of biological organization. Several groups developed kernel methods to infer protein interaction networks from multiple genomic data types [88, 89]. Other research groups attempted to combine heterogeneous data sources through probabilistic decision trees [90, 91], logistic regression [92], and Markov Random Fields (MRF) [93]. In another study, Random Forest (RF) was applied to measure similarities between protein pairs where each node in the tree corresponds to a feature or data source, and a k-nearest neighbor algorithm was then adopted to classify protein pairs according to the calculated similarities [94].

All of the above mentioned integrative methods for protein interaction predictions were applied in *S. cerevisiae*. In 2005, Rhodes et al. [95] employed a naïve Bayesian network to integrate different evidences across model organisms to infer interactions in human. They suspected interactions in three model organisms, *S. cerevisiae*, *D. melanogaster*, and *C. elegans*, may suggest interactions among orthologous proteins in human. Other features considered were: similar gene expression profiles across human tissue samples, enrichment of protein domain pairs among human protein interactions, and shared biological functions. Their preliminary results seemed promising with nearly 40,000 predicted protein interactions in human. However the naïve Bayesian network assumes conditional independence between data sources. High dependence between the genomic features may exist, and it would become more prominent as more features are integrated.

More recently, Zhong and Sternberg [96] combined multiple data sources from the three model organisms to predict genetic interactions in *C. elegans* using logistic regression. The orthologous information was exploited in a different way. For each *C. elegans* gene pair as well as its orthologous pairs in *D. melanogaster* and *S. cerevisiae*, they attempted to identify five features: identical anatomical expression, phenotype, function annotation, microarray coexpression, and presence of interlogs. Finally, interaction predictions are made on the basis of all features from the *C. elegans* gene pair and its orthologs. The algorithm employed in their study, logistic regression, is a statistical regression model for binary dependent variables. It is a generalized linear model that utilizes the logit as its link function. The logarithm of the likelihood odds is modeled as a linear function of the explanatory variables; therefore, it is not adept at modeling nonlinear complex systems such as interaction networks.

1.3 Methods for Protein Function Annotation

The most established computational approaches to function detection primarily depend on homology matching to genes with known functions utilizing programs such as FASTA [97] and PSI-BLAST [71]. However, assuming functional annotations by sequence similarity poses some critical questions: at what level of sequence similarity can we feel assured that the two proteins carry out the same function, and even if the function is conserved, at what level of detail is the conservation? Over the years, numerous non-homology based computational techniques have been developed to derive protein functions from additional sources of biological data such as gene fusion events [42, 43], phylogenetic profiles of proteins in multiple genomes [47], gene expression and mutant phenotype data [98], and heterogeneous data sources including gene expression, physical interactions, motif information and transcription factor binding sites data [99-102].

With the ever-increasing accumulation of high-throughput protein-protein interaction data, a number of computational approaches emerged to take advantage of these data for gene function prediction [103-110]. In general, these approaches are based upon the premise that proteins often physically interact to achieve a common objective. Hence, it may be possible to infer functions for a protein based on its interaction partners. The concept is also known as 'guilt-by-association', which assumes that interacting proteins are more likely to carry out similar functions. Schwikowski et al. [103] introduced a neighbor counting method where unknown proteins were assigned functions with the most occurrences among their interaction partners. Thereafter, several research groups attempted to improve the neighbor counting method through application of χ^2 statistics [104], Bayesian analysis [105], and Markov Random Field analysis [106, 107]. Moreover, several researchers have introduced protein interaction network based methods [108, 109], and Brun et al. and other colleagues [110-113] clustered the *Saccaromyces cerevisiae* proteome into several groups to predict cellular functions using protein interaction data.

1.4 My Research Contribution

Despite various progresses computational methods have made toward protein interaction predictions and their fair share of successes, the existing methods still have a limited range of applicability: the specificity and sensitivity are normally low. To this day, high-throughput experimental data describing protein interactions still provide the most coverage. There is still a strong need for more reliable *in silico* models for the inference of protein interactions that can cover a larger spectrum of the interactome. The available protein interaction information is extremely valuable, but it does not provide any insights into how the molecules are associated or interacting. In fact, proteins physically bind to each other only through small

regions on their surfaces. The ability to pinpoint such binding sites is critical in comprehending cellular roles that different proteins fulfill; in particular, mutations at the binding sites may disrupt existing protein interactions or create new undesirable interactions that could lead to many human diseases [114].

Essentially, the protein-protein interaction prediction problem is a two-fold problem. (1) *Given two proteins A and B, do they interact or not?* (2) *If proteins A and B are known to interact, how do the two molecules interact?* In other words, what is the three-dimensional (3D) structure of the protein complex? Solving the latter question will provide crucial atomic details about protein binding, and these details may permit more rational design of biological experiments to disrupt an interaction. Before studying the atomic details, we need to predict who interacts with whom. In this dissertation, one of my two main goals is to answer the first question, whether two proteins interact or not, and provide a confined search space for the second question, or approximately where the two proteins physically bind.

It is sometimes feasible to locate finer details such as domains or segments of proteins that may mediate the interaction so that interacting sites of larger proteins are narrowed down. Protein domains are considered to be the building blocks of proteins that are conserved through evolution to represent protein functions or structures. It is a generally-acknowledged assumption that protein interactions involve bindings of two or more specific domains. Hence, understanding domain interactions can not only predict protein-protein interactions, but it also can provide additional information on how two molecules are interacting.

The other main goal of my dissertation is to develop computational models to study protein functions. Improving the quality of current functional annotations and proposing new annotations to the numerous uncharacterized proteins still presents a major challenge, and this research can eminently assist the research community. Protein functions are inseparable from

its interactions. Proteins do not perform functions in isolation but as part of a complex network of physical complexes and pathways. Since domain-domain interactions are the foundation of protein-protein interactions, domain interactions can undoubtedly be used to learn protein functions.

It is obvious that domain-domain interaction is the key here. Therefore, my approach is to first discern domain-domain interactions and then use the knowledge to extrapolate protein-protein interactions and protein functions. In the elucidation of protein-protein interactions, I have made several contributions. First, protein interaction prediction is formulated as a binary classification problem with novel domain-based feature representation. Due to the features' unique characteristics, the standard random forest algorithm cannot be directly applied to the protein-interaction inference problem. Instead, a new framework based on random forest, RDFF, is proposed. Second, unlike most of the existing domain-based computational approaches, RDFF does not assume domain pairs to be independent of each other. Third, rather than considering single-domain pairs as the basic unit of protein interactions, contributions of all the possible domain combinations to protein interactions are explored. As a result, the RDFF method not only inferred domain-domain interactions but also predicted protein-protein interactions with better performance compared to the popular Maximum Likelihood Estimation (MLE) method (in terms of the specificity and sensitivity).

For the protein function determination problem, I proposed a novel approach CSIDOP, Cross-Species Interaction Domain Patterns, to take advantage of the ever-increasing accumulation of high-throughput protein interaction data. CSIDOP is fundamentally different from other protein interaction based function detection algorithms where the function of a target protein is determined strictly by the annotations of its interaction partners. Compared with existing methods, CSIDOP is distinctive in the following aspects: (i) protein functions

are detected through the shared interacting domain patterns, (ii) the patterns are mined from cross-species protein-protein interaction data, and (iii) unknown proteins can be assigned to various functional categories in GO, in contrast to most other methods where proteins are assigned with a limited number of functional categories such as MIPS [115] that are less specific than GO.

Finally, to tie everything up, I integrated our discovered domain-domain interactions with other known and putative domain interactions into an online system called PINFUN for protein interaction and function predictions.

In summary, my dissertation consists of the following three major modules. Details about each module are respectively discussed in Chapter 3, 4, and 5.

1. A novel domain-based framework for protein interaction prediction.
2. A novel model to construct the domain-domain interaction network for protein function prediction.
3. An online system tool PINFUN for people to use in protein interaction prediction and protein function annotations.

Chapter 2. Research Background

Domain-domain interaction is the foundation of our approach to protein interaction and function predictions. This chapter provides basic background information on protein domains and different types of domain interactions in sections 2.1 and 2.2, respectively. Section 2.3 discusses the existing domain-domain interaction prediction methods based on protein-protein interaction data.

2.1 Protein Domain

Protein domains are defined as evolutionary, functional, and structural units of proteins that can fold independently of other such units. In order for a domain to fold into stable and unique structures when excised from a complete protein, the cohesion between side chains is required for its organization, and it needs to be of a certain size before recognized. Domains do vary in size, but they typically have 100 to 250 residues [116].

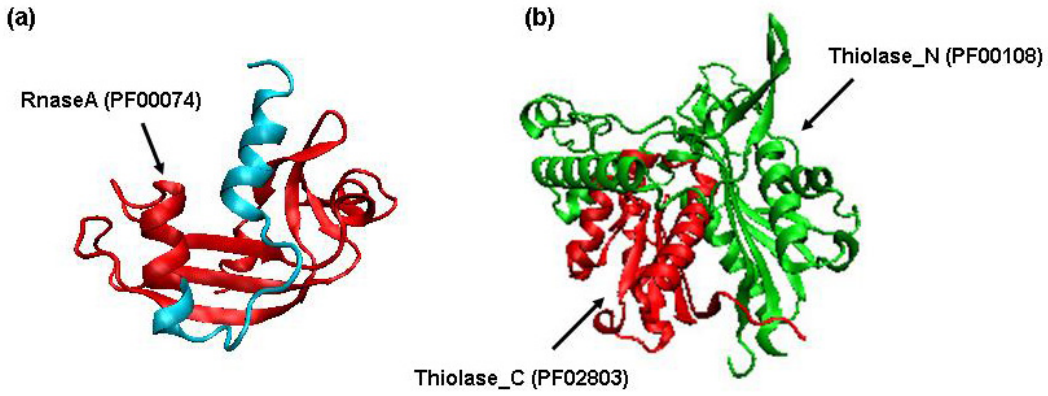

Figure 2.1 Examples of single-domain and multi-domain proteins. PDB structures of **(a)** Single-domain protein ANGI_HUMAN (P03950) in *H. sapiens*. The 3D structure is the PDB entry 1b1e. Red colored residues belong to the domain RnaseA (PF00074). **(b)** Multi-domain protein THIC_HUMAN (Q9BWD1) in *H. sapiens* (PDB entry 1wl5). Red and Green colored residues represent two different domains: Thiolase_C (PF02803) and Thiolase_N (PF00108), respectively.

Domains come in different types, and proteins are created from this pool of limited types of domain architectures [117, 118]. These compact and stable domain units alone can form single-domain proteins or undergo duplication and recombination with others to form multi-domain proteins (Figure 2.1). The majority of proteins, especially in higher organisms, are built from the fusion and shuffling of domains [116, 119-122]. Small proteins usually contain a single domain whereas larger proteins (i.e. having more complex architectures) are formed by combinations of domains. For instance, some human proteins can contain up to 130 domains.

It is believed that once a set of domains with sufficient functions to support the basic life form, it would be much easier and faster for the genome to produce various new proteins by duplication, divergence, and recombination [116]. Koonin et al. observed that there is a propensity for eukaryotic proteins to have more domains than their prokaryotic homologs, this is termed domain accretion [123]. A rough estimate reveals that approximately two-thirds of proteins in prokaryotes and four-fifth of proteins in eukaryotes are multi-domain proteins [124]. This observation suggests a connection between increased recombination of domain architectures and organism complexity. Indeed, Koonin et al. observed that the likelihood of domain combination increases in the order of archea, bacteria, and eukaryotes [120]. It is now clear that the complexity of an organism is not determined by its number of genes; for instance, fruit flies have fewer genes than nematodes and humans have fewer genes than rice. However, the organism complexity does seem to be related to the extent of their domain duplications and recombinations. The reason for this is that organism complexity mainly originates from the complex networks of protein interactions, and a modest increase in the number of domains in interacting partners may directly translate into numerous new interactions. This probably explains why complex organisms have fewer genes [125].

2.2 Domain-Domain Interaction

While examining protein-protein interactions, it is often feasible to locate finer details such as domains or segments of proteins that mediate the interaction such that interacting sites of larger proteins are narrowed down. Protein interactions involve binding between two or more specific domains is a generally acknowledged notion. In addition to the determination of protein interactions, analysis of the domain architectures can be extremely beneficial for function predictions of uncharacterized proteins. A domain can either exhibit an independent function or cooperate with other domains to execute a certain function of a multi-domain protein. In brief, domains and their interactions determine the functions of proteins [121]. Thus, it is important to understand the principles of domain-domain interactions.

There are four basic types of domain-domain interactions that can explain protein-protein interactions, and these are illustrated in Figure 2.2.

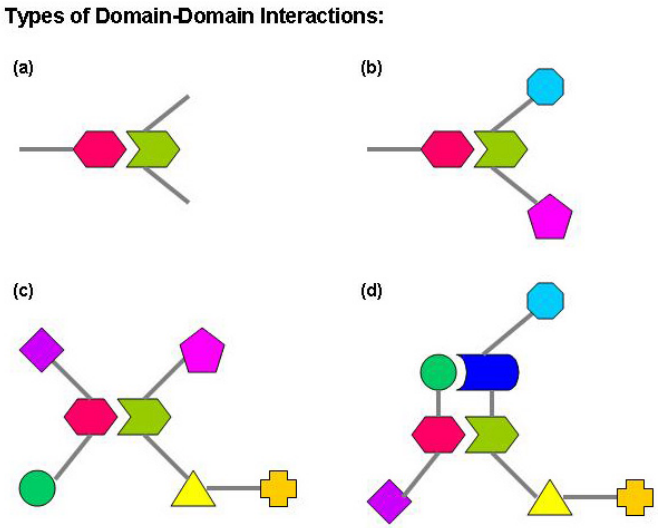

Figure 2.2 Types of domain-domain interactions. Various shapes and colors representing different domains **(a)** Interactions between two single-domain proteins. **(b)** Interactions between a single-domain protein and a multi-domain protein. **(c)** Multi-domain proteins bind through a single domain. **(d)** Multi-domain proteins bind through multiple domains.

- *Singlet-to-Singlet*: Interactions between two single-domain proteins can be explained by the binding of their respective single domains (Figure 2.2a).
- *Singlet-to-Multiplet*: Interaction is often observed between single-domain proteins and multi-domain proteins where the single domain binds to one of the multiple domains (Figure 2.2b).
- *Multiplet-to-Multiplet via Single Binding*: If two multi-domain proteins interact, the interaction may be explained by a single binding between one of many domains from each protein (Figure 2.2c).
- *Multiplet-to-Multiplet via Multiple Bindings*: If two multi-domain proteins interact, the interaction may also be explained by multiple bindings between two or more domains from each protein (Figure 2.2d).

The analysis of domain-domain interactions is an extremely important task for protein interaction and function annotation as mutations at the binding sites can disrupt existing protein interactions or create new undesirable interactions, and this may, as a consequence, disturb normal functions of proteins in cells and cause various diseases for the organism. Hence, understanding interactions at the domain level is not only a critical step towards thorough understanding of protein interaction networks and their evolution, but it is also one step closer to acquiring insights into the functions of proteins and causes of human diseases. The research will inevitably contribute to protein and drug design in the future.

2.3 Existing Methods for Domain Interaction Prediction

In the past decade, domain-domain interaction discovery has been the subject of intensive study. Some methods focused on interactions involving specific mediating domains such as SH2, SH3, and PDZ domains [126, 127]. Some attempted to understand domain-domain

interactions (DDI) from genetic variations among multiple genomes [128-130]. Some analyzed architecture of the domain interaction networks that revealed high scoring domain interactions indeed give rise to reliable protein interactions [131-135]. Some approaches utilized heterogeneous data [136, 137]. Many other DDI prediction methods rely on the protein-protein interaction (PPI) network. The following sections give an overview of these prediction methods for domain-domain interactions utilizing PPI data (Table 1.1).

2.2.1 Domain Interactions as Predictor of Protein Interactions

The original studies in domain-domain interaction view it as a predictor of protein-protein interactions. The Association Method is one of the pioneering works that seeks domain pairs co-occurring more often in interacting protein pairs than expected by chance. Available experimental protein interaction data is utilized to compute the probability of two domains interacting. This is the fraction of interacting protein pairs among all protein pairs containing the domain pair (Eq. 2.1) [138].

$$P(d_m, d_n) = \frac{I_{mn}}{N_{mn}} \tag{2.1}$$

I_{mn} refers to the number of interacting protein pairs that contain domain pair (d_m, d_n), and N_{mn} is the total number of protein pairs that contain the domain pair (d_m, d_n). However, Equation 2.1 may assign high association scores to domain pairs with low frequency which may not correspond well to the interaction probability. Other colleagues tried to address the issue by considering the number of domains in each protein, but this correction may preferentially identify promiscuous domain interactions because they screen for pairs that occur with the highest frequency [139]. Generally, the association-based method considers interacting domain pairs to be independent which ignores other domains in a given protein pair. Moreover, they do not explicitly consider the errors in interaction dataset.

Taking these errors into account, an optimization approach, Maximum Likelihood Estimation (MLE), was proposed to infer domain interactions by maximizing the likelihood of the observed protein interaction data [140, 141]. The likelihood function is a function of parameters θ (λ_{mn}, *fp*, *fn*), where λ_{mn} is the probability that domain *m* interact with domain *n*, *fp* refers to false-positive rate, and *fn* refers to false-negative rate. The domain interaction probability λ_{mn} was optimized using the Expectation-Maximization (EM) algorithm. Other groups proposed Bayesian network based models to capture the protein interaction probabilities based on domain pair frequencies in a domain attraction-repulsion model [142-144]. Later, Nye et al. described the p-value method which tests the null hypothesis that the presence of a domain pair in a protein pair do not affect whether the two proteins interact or not [145]. To test the hypothesis, p-value statistics are calculated considering fractions of false positives and false negatives. The domain pair with the lowest p-value is the most likely to interact. Results suggest that the p-value method performs reasonably well when there are nine or more domains in a protein pair.

Furthermore, an interesting graph-oriented approach called the Interacting Domain Profile Pairs (IDPP) was proposed, which employed a combination of sequence similarity searches and clustering based on interaction patterns and domain information across multiple species [146]. First, they clustered domains of different proteins that interact with a common region of another protein into interaction clusters (IC). Then the domains within an IC are regrouped by sequence similarity, and the regrouped domains are later clustered into n-SICs (Similarity & Interaction Cliques). Each n-SIC consists of domains that are similar in sequence and interact with common domains. The Interacting Domain Profile Pairs are generated from the interactions between SICs. The use of domain profile pairs has resulted in better predictions than methods solely based on sequence information. Nevertheless, the main

goal of the IDPP method is to infer the protein interaction map of a target organism from a large-scale interaction map of a source organism which can be very expensive to obtain.

2.2.2 Domain Interactions as Explanation of Protein Interactions

More recently, a unique class of methods for domain-domain interaction prediction emerged, and they consider DDIs as putative explanations of protein-protein interactions rather than their predictors. For instance, the domain pair exclusion analysis (DPEA) method [147] introduced a new measure for each potentially interacting domain pair, called E-score, that measures degree of reduction in likelihood of observing the given protein interaction network when excluding a domain pair. Different from previous methods where the most probable domain interactions identified tend to be the most promiscuous, or least specific, DPEA can detect specific interacting domain pairs. It extends the MLE method by adding a likelihood ratio test to assess the contribution of each potential interacting domain pair to the likelihood of a set of observed protein interactions. This is achieved by estimating the E_{ij} score which is defined as the logarithm of two probabilities. The numerator probability represents the probability of two proteins interacting given that domains i and j interact. The denominator probability corresponds to the probability of two proteins interacting given that the domains do not interact. For a given domain pair, the numerator probability is computed with the EM procedure to maximize θ. Higher E-scores indicate a higher tendency for the two domains to interact. In order to identify specific domain interaction pairs, one can simply screen for low θ and high E-score values. A variation of the method was proposed later [148].

To generalize the complex problem of interactions among proteins and their corresponding domain architectures, Huang et al. [149] conceptualized a maximum-specificity set cover procedure (MSSC). Formally, they represented a protein interaction

network as a set cover problem by defining $Y = \{all\ protein\ pairs\ (P_i, P_j) \mid P_i, P_j \in P\}$ as the set of all protein pairs, $X = \{protein\ pairs\ (P_i, P_j) \mid P_i\ interacts\ with\ P_j\}$ as the set of interacting protein pairs, and F as the set of all domain pairs (d_m, d_n). Essentially, it strives to find a set C of domain pairs that can "cover" the given protein-protein interactions to the largest extent.

On the other hand, Guimaraes et al. explain protein interactions as evolving in parsimonious ways [150, 151]. Parsimony is a 'less is better' concept that displays preference for the least complex explanation of an observation. In general, mathematical models with the smallest number of parameters are preferred because each parameter introduced into the model inserts more uncertainty to it. The Parsimonious Explanation (PE) approach hypothesized that interactions between proteins evolve in parsimonious way. That means the set of true domain-domain interacting pairs should be well approximated by the minimal set of domain pairs necessary to explain a given protein interaction data.

Chapter 3. Domain Interaction Based PPI Prediction

Ever since the beginning of the 21st century, there has been a growing interest in the inference of protein-protein interactions from their corresponding domain-domain interactions. One of the pioneering works is the Association Method by Sprinzak and Margalit [138]. Numerous methods followed, and the preliminary results have demonstrated their feasibility. However, a majority of the approaches do not consider the fact that multiple domains in a protein can collaborate with each other as a single module to interact with another domain module in the other protein. Moreover, most approaches assume independence of domain-domain interactions which means that they ignore other possible domain interaction information between a pair of proteins. To overcome the inherent flaws of existing methods, I propose a novel domain-based random forest framework (RDFF) to learn protein-protein interactions. Details on the methodology and experimental evaluation are provided in the upcoming sections.

3.1 Methodology

The following sections describe our domain based approach to protein-protein interaction prediction. Section 3.1.1 describes the feature representation, section 3.1.2 explains the learning model selection, and section 3.1.3 discusses our proposed model.

3.1.1 Feature Representation

The protein-protein interaction (PPI) prediction problem can be formulated as a two-class classification problem where each pair of proteins is considered to be a sample belonging to either the 'interaction' class (i.e. two proteins interact with each other) or 'non-interaction' class (i.e. two proteins do not interact). Since the goal in this study is to investigate protein

interactions at the domain level, it is necessary to characterize a protein pair by their respective domains. Thus, each protein pair is represented by a vector of features where each feature corresponds to a domain. Let $D = [X_1, X_2, ..., X_n]$ represent n data samples and $X_i = [x_1^{(i)}, x_2^{(i)}, ..., x_M^{(i)}, y_i]$ represent the i-th sample with M feature attributes x_j belonging to the class y_i. In the problem formulation, $y_i = 1$ denotes the 'interaction' class and 0 refers to the 'non-interaction' class. The feature vector size M is the total number of unique domains in the dataset, and each feature attribute x_j has a discrete value of 0, 1, or 2. If both proteins in a sample pair do not contain the domain attribute x_j, the associated feature value will be 0. On the other hand, if one of the proteins has the domain, then its value is 1. Lastly, if both proteins contain the domain, its assigned value is 2. This ternary-value feature representation is distinct compared to the binary representations in other domain-based methods. It is necessary as domains often self-interact. Moreover, it allows us to distinguish between protein pairs with a domain existing in one protein and those in both proteins.

3.1.2 Model Selection

The unique characteristic of our data feature prompts a challenge for the well known machine learning algorithms like Bayesian Network, Neural Network, and Decision Tree. Here, the size of the feature space is equivalent to the total number of unique domains which is extremely large, in the range of thousands. An extraordinarily large feature space can be destructive to the learning process of classification algorithms with reduced accuracy and longer learning time. In order to tackle the challenge, I propose to adopt the random forest framework.

The 'random forest' is an ensemble classifier. Its main principle is that when the input space is extraordinarily large, random subspace (RS) feature selection can potentially

improve classifier diversity. The algorithm for inducing a random forest was developed and popularized by Leo Breiman [152]. It involves construction of an ensemble of decision trees from randomly sampled subspaces of the input features, and final classification is obtained by combining results from the trees via voting. It has been shown that combination of multiple trees produced from randomly selected subspaces can improve the generalization accuracy [153, 154]. When using the combined power of multiple trees to increase accuracy, it is particularly important to produce a large number of sufficiently different trees. The application of randomization in feature selection is a way to explore various possibilities of subspaces. While most classification methods suffer from the curse of dimensionality, the RS feature selection method can take advantage of the high dimensionality. In contrast to Occam's Razor, the method improves accuracy as it grows in complexity [155].

Generally, the random decision forest constructs many decision trees, and each is grown from a different set of training data. To construct individual decision trees, training samples are randomly selected with replacement from the original training dataset. More precisely, if the number of samples in the original training set is N, then N samples are randomly drawn with replacement. At each splitting or decision node, it determines the best splitting feature from a randomly selected subspace of m features where m is much smaller than the M total number of features. Each tree in the forest is then grown to the largest extent possible without pruning. To classify a new object, each tree in the forest outputs a classification which is interpreted as the tree 'voting' for that specific class. The final classification of the object is determined by a majority vote among the classes decided by the forest of trees.

3.1.3 Domain-based Random Decision Forest Framework

Because of the distinctive characteristics of our domain-based features, a traditional random forest cannot be directly applied here. Similar to the standard random forest algorithm, our individual decision trees in the forest are still built from different sets of training data. For each tree, positive and negative samples from the original training dataset are selected randomly with replacement. To preserve the same proportion of positive samples among all samples, we drew samples from positive and negative sets separately.

While building a decision tree, the standard random forest randomly selects a subspace of features to focus on at each splitting node. However, our application is unique, and this randomness introduced may not work as well as in other applications. In other applications, all features contain information for classification no matter what the values are. In our application, a feature with a value 0 does not give us any information about the interaction status of a pair of proteins. Consider the following example. A protein pair (P1, P2) has domains {a, b, c} and {d, e}, respectively. Assume that the true domain interacting pair is (a, d) and it has appeared frequently in many different protein pairs. With random selection, domain {a} or {d} may not be selected properly even though they are the domains that appear in many proteins. We could randomly select {a} and {e} as the splitting attributes to classify the protein pair (P1, P2) as interacting. Although the classification is correct, we have the wrong conclusion on the interacting domain pair.

In order to address this issue, we introduce probability selection for the feature subspace. Each feature in the entire feature space is assigned with a selection probability. The probability is calculated based on the number of protein interaction pairs in the original training dataset that include such domain feature (i.e. at least one protein in a pair contains the domain). A Roulette wheel representing the feature attributes is then created. Each

domain feature is assigned with a real number in range [0.0, 1.0] to represent a section on the wheel where the range is calculated based on the probabilities. Thus, if a domain is common among large number of proteins, it will have a higher probability of being selected.

Each decision tree is built level by level from a bootstrapped training dataset starting at the root. At each splitting node of a decision tree, in order to form the feature subspace, we spin the Roulette wheel by generating random numbers. If the generated random number falls in between a feature's range, then the feature is added to the subset unless it is already used as a splitting attribute by one of the parent nodes in the same branch up to the root. This process continues until $log_2M + 2$ (M is the total number of features) features are selected for the feature subspace. For each feature in the subspace, information gain splitting criteria by Quinlan [156, 157] is calculated as the 'goodness of split' measure which is based on the classic formula from information theory. The information gain measures theoretical information content of a code by $\sum_i p_i \log(p_i)$, where p_i is the probability of the i-th message. Assume that the number of samples in 'interaction' class and 'non-interaction' class are n_1 and n_2, respectively. The information required to classify samples given only the decision class totals as a whole is

$$H(C) = -(P(y=0)\log P(y=0) + P(y=1)\log P(y=1)) \tag{3.1}$$

where $P(y)$ is the class probability among all samples (i.e. $P(y=1) = n_1/n$ and $P(y=0) = n_2/n$). The information needed to classify samples given knowledge of the attribute x_j is defined as

$$H(C \mid x) = \sum_{j=1}^{3} P(x = x_j) H(C \mid x = x_j) \tag{3.2}$$

where $P(x = x_j)$ is the probability of the attribute x taking the value x_j. In our ternary-value representation, x can take three discrete values: 0, 1, and 2. The information needed given each attribute value $H(C \mid x = x_j)$ is then defined by

$$H(C \mid x = x_j) = -(P(y = 0 \mid x = x_j) \log P(y = 0 \mid x = x_j) - P(y = 1 \mid x = x_j) \log P(y = 1 \mid x = x_j)) \quad (3.3)$$

where $P(y = y_i \mid x = x_j)$ is the conditional probability of i-th class given attribute value x_j. Finally, the information gain (IG) measure for an attribute x can be calculated with Eq. 3.1 and Eq. 3.2 as follows.

$$IG(x) = H(C) - H(C \mid x) \quad (3.4)$$

The attribute with the largest information gain is selected from the subspace of features.

Individual trees of a traditional random forest are normally built completely without pruning. In an ideal situation, a decision tree stops growing only when all samples are well classified. By well classified, it means that all samples at a leaf node must belong to the same class. Unfortunately, the world is never perfect. In addition, due to the high dimensionality of our data, each tree is expected to be extremely large when entirely grown. Thus it is necessary to impose some stopping conditions. Several early stopping criterions are employed as a forward pruning technique that stops pursuing branches with little statistical significance. A node in a decision tree stops splitting when any of the following conditions is met: a node is at the maximum level, the node impurity is smaller or equal to a certain threshold, or there are minimum number of samples left to be classified. Node impurity is defined as the proportion of samples that are in the minority class.

Finally, a forest forms when multiple decision trees are grown. To classify a protein pair, instead of collecting votes from all trees in the forest as in the standard random forest algorithm, votes are obtained only from the trees that contain at least one domain feature from each protein in the pair as a splitting attribute. This is essential because domain features of the input protein pair not appearing in a decision tree as nodes do not imply non-interaction for the proteins. In that situation, we consider the decision tree as incompetent to

come up with an adequate verdict on this particular protein pair; thus, we take away its voting rights. Certainly, this action reduces the number of voting trees as a consequence. In order to have an appropriate number of voters, a sufficient number of decision trees need to be built to cover all domains found in the training samples. In other words, we need to make sure that each domain feature in the training samples is covered by at least a certain number of trees. In the worst case scenario, if no tree is able to vote for a protein pair, we assume the protein pair to be non-interacting. Otherwise, protein pairs are classified by majority votes. A tree casts a vote of value 1 for interaction and a vote of value 0 for non-interaction.

The benefit of our domain-based random decision forest framework (RDFF) is two-fold. First of all, RDFF can predict whether two proteins interact or not. Secondly, it can infer domain-domain interactions. Since each splitting attribute represents a single domain, if an occurrence of two or more such attributes in different proteins of a pair leads to an interaction classification, then we can interpret the domains as forming an interacting domain pair or domain combination pair (Figure 3.1). However, a potential problem associated with our ternary-value feature representation is that it cannot tell whether two domains are from the same protein or from two different proteins. To handle this problem, in decision-making procedures, we consider the domains as an interacting domain pair only if these domains (two or more) are from different proteins and they lead to an interaction classification. For example, assuming that (a) and (b) in Figure 3.1 are decision trees in the forest, we can infer interacting domain pairs from a path of decision nodes (green colored nodes) that induce a classification result of 1 (represented as pink dashed boxes) for a particular protein pair if and only if domains from both proteins appear in the path. At each node, it trails down the leftmost branch if it does not contain the domain feature and the feature value is 0 (details in Section 3.1.1). The middle branch is pursued if the feature value is 1, and otherwise the

rightmost branch is chosen. The tree (a) in Figure 3.1 classifies a protein pair to be interacting if one protein in the pair contains domain 25 and the other protein contains domain 15. One can also conclude that domains 15 and 25 may interact with each other because they contribute to an interaction prediction. Similarly in tree (b), domains 38, 269 or 848 may form interacting domain combination pairs.

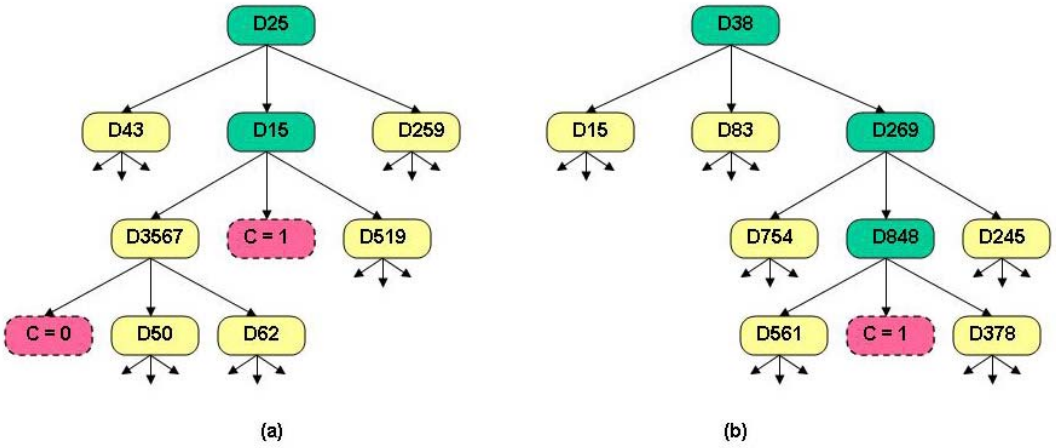

Figure 3.1 Domain-domain interaction inference using RDFF. Assume that **(a)** and **(b)** are two decision trees in a forest. The leftmost branch is taken if the domain attribute value is 0, middle branch is followed if it is 1, and lastly rightmost branch is chosen if domain feature value is 2. Boxes with solid borderlines represent decision nodes with selected domain features among which the green boxes represent domains that form interacting domain or domain combination pairs and the yellow boxes are the regular domain feature nodes. Pink boxes with dashed borders are the leaf nodes that display classification results.

3.2 Experimental Results

Here, we discuss the experiments performed in evaluation of our RDFF method. Details about the experimental data, evaluation metrics, and model parameters are presented in sections 3.2.1, 3.2.2 and 3.2.3. The performance of RDFF in both domain-domain interaction and protein-protein interaction predictions are demonstrated in sections 3.2.4 and 3.2.5.

3.2.1 Data Source

For protein-protein interaction (PPI) predictions, our domain-based random forest framework is evaluated over the most studied model organism *Saccharomyces cerevisiae*. Protein interaction data in yeast is collected from the Database of Interacting Proteins (DIP) [158, 159], Deng et al. [140], and Schwikowski et al. [103]. The PPIs used in Deng et al. is a combined interaction dataset from experimental two-hybrid assays by Uetz et al. [9] and Ito et al. [10]. Schwikowski et al gathered their data from yeast two-hybrid, biochemical and genetic data.

Initially, we obtained 15,409 interacting protein pairs in the yeast organism from DIP, 5,719 pairs from Deng et al. and 2,238 pairs from Schwikowski et al. The datasets are then merged by removing overlapping interaction pairs. Also, because domains are the basic units of protein interactions, proteins without domain information cannot provide any useful information for our prediction. Therefore, we only keep the pairs where both proteins have domain information. 9,834 protein interaction pairs remained among 3,713 proteins, and these pairs are evenly separated (4917 pairs each) into training and testing datasets. Since non-interacting protein information is not available, the negative samples are randomly generated. A protein pair is considered to be a negative sample if the pair does not exist in the interaction set. A total of 8,000 negative samples were generated and also separated into two halves. Both final training and testing datasets contain 8917 samples, 4917 positive and 4000 negative samples.

The protein domain information is gathered from Pfam [160] which is a protein domain family database that contains multiple sequence alignments of common domain families. In Pfam, hidden Markov model profiles were used to find domains in new proteins. The Pfam database consists of two parts: Pfam-A and Pfam-B. Pfam-A is manually curated, and Pfam-

B is automatically generated. Both Pfam-A and Pfam-B families are used here. In total, there are 4293 Pfam domains defined by the set of proteins.

3.2.2 Evaluation Metrics

After building our domain-based random forest PPI prediction system, it is important to estimate how accurately the model will perform in practice. Ordinarily, a predictive model has one or more unknown parameters and the model can be fitted as well as possible on a training dataset through parameter optimization. Nonetheless, the model may not perform as well when fitted on an independent test dataset. This is commonly known as overfitting. Overfitting is most likely to occur when the training dataset is small or number of parameters is large.

In order to assess the fit of our domain-based random forest model on test data, we calculated two statistical measures: sensitivity (SN) and specificity (SP). The sensitivity (also known as recall in other fields) measures the proportion of true positives (TP) which are correctly identified. The specificity measures the proportion of true negatives (TN) which are correctly predicted. In our case, true positives and true negatives are the observed protein interaction and non-interaction pairs, respectively. False negatives are the observed interaction pairs wrongly identified as non-interactions. Similarly, false positives are the non-interaction samples wrongly classified as interactions. Sensitivity and specificity are defined in the following formulas (Eq. 3.5 & 3.6).

$$SN = \frac{TP}{TP + FN} \tag{3.5}$$

$$SP = \frac{TN}{TN + FP} \tag{3.6}$$

3.2.3 Model Parameter Selection

The predictive accuracy of a random forest depends on the strength of the individual tree classifiers, and these may be affected by tree size. In our implementation, we have set three stopping criteria to limit the tree size, and they are maximum tree level, impurity and minimum node size thresholds. Minimum node size defines the minimum number of samples to be classified by each node. In our forest, each decision tree is constructed with node impurity threshold of 0.01, and the minimum number of samples at a node is 3. Among those early stopped nodes, less than 10% reached impurity and minimum node size thresholds. This implies that the maximum tree level criterion has the most impact in restricting the tree size.

In order to make an appropriate parameter choice for the maximum tree level threshold, we grew multiple forests with trees having different heights and analyzed their generalization accuracies via cross-validation. Keeping in mind the computational time, K-fold cross-validation seems to be more suitable. In K-fold cross-validation, the original set of samples is partitioned into *K* subsets of samples. Among the *K* subsets of samples, one subset is drawn to be the validation dataset for testing the model, and the remaining (*k*-1) subsets are used as training data. The cross-validation process is repeated for *K* times (folds), in which each of the K subsets are used exactly once as the validation set. To produce the final accuracy estimation, results from the *K* folds are averaged. Here, we used 5-fold cross-validation and found that classification error rates over the validation sets decrease first as the tree levels increase. This is due to the increased performance of each individual tree. Because of majority voting, when each individual tree in a random forest performs better, the entire forest also performs better. As shown in Figure 3.2, the forest classification error rate reaches the minimum for the heights of 350 and 450 and increases slightly after 450. Therefore, we select the maximum tree size at 450 levels.

Another parameter to be considered is how many trees should be grown. To determine an appropriate number of trees in a forest, we set a limit on minimum coverage of each domain feature at 30 trees. In other words, it makes sure that each domain feature appeared in the training dataset is one of splitting attributes in at least 30 trees. In this way, we guaranteed that at least a certain number of trees will vote to classify each protein pair. It is estimated from experiment that with 100 trees in a forest, each domain feature would be covered by at least 74 trees. To be on the safe side, we grew 150 trees.

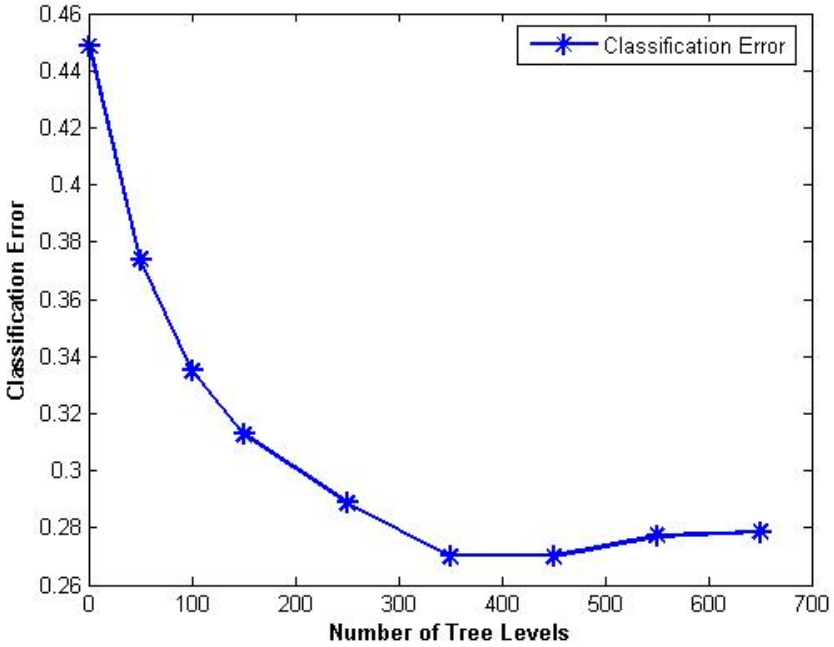

Figure 3.2 RDFF maximum tree height parameter selection – classification error comparisons of different tree sizes using 5-fold cross-validation

3.2.4 Domain-Domain Interaction Prediction by RDFF

After training our domain-based random forest model, for each true protein interaction pair, we can derive domain-domain interactions from the domain nodes that contributed to the interaction classification by tracing the branch or path the protein pair followed to reach such

classification. In total, RDFF predicted 4,366 single-domain interaction pairs (i.e. one domain from one protein). Among them, 1,891 pairs are found with Pfam-B domains for which interaction information is not available from Pfam or other sources, and remaining 2,475 pairs are composed of Pfam-A domains. Out of the 2,475 single Pfam-A domain interaction pairs, 95 of them are also reported by the iPfam database [161] and 2,239 of them are found in the InterDom database [162]. In iPfam, two domains are defined as interacting if and only if they are close enough in at least one PDB complex to form an interaction. The domain interactions reported by crystal structures of protein complexes are generally regarded by colleagues as true positives. On the other hand, the InterDom contains putative interacting domains from heterogeneous data sources; thus, they can also be treated as high confident domain interactions.

Table 3.1 Examples of inferred single-domain interaction pairs confirmed by iPfam

Domain A Pfam_ID	Domain A Name	Domain B Pfam ID	Domain B Name
PF00069	Pkinase	PF00023	Ank
PF00069	Pkinase	PF02984	Cyclin_C
PF00069	Pkinase	PF00134	Cyclin_N
PF00364	Biotin_lipoyl	PF02852	Pyr_redox_dim
PF00117	GATase	PF02786	CPSase_L_D2
PF00117	GATase	PF02787	CPSase_L_D3
PF00117	GATase	PF00289	CPSase_L_Chain
PF00071	Ras	PF00996	GDI
PF00560	LRR_1	PF00076	RRM_1
PF00183	HSP90	PF00515	TPR_1

Table 3.1 lists some of the single-domain interacting pairs identified by our method and also confirmed by the iPfam. For example, the domain biotin_lipoyl (PF00364) is annotated as biotin-requiring enzyme and it has a conserved lysine residue that binds to biotin or lipoicacid. Biotin performs catalysis in some carboxyl transfer reactions and is covalently attached to a lysine residue via an amide bond. The pyr_redox_dim (PF02852) domain is

annotated as pyridine nucleotide-disulphide oxidoreductase, dimerization domain and determined to involve in oxidation–reduction reaction.

Table 3.2 lists some identified single-domain interaction pairs that are not found in iPfam, but are found as interacting with a high confidence by the InterDom [162]. For example, SH3 (PF00018) and Pkinase (PF00069) in Table 3.2 are derived from a PPI only involving single-domain proteins by InterDom. In InterDom, a protein is considered as a single-domain protein if it has only one domain and the domain accounts for at least 50% of the protein length [162]. Domain interactions derived from such single-domain protein interactions are usually considered to be highly likely. The SH3 domain is also found to interact with Pkinase_Tyr (PF07714) by iPfam [161]. Pkinase and Pkinase_Tyr are both members of the protein kinase superfamily clan. A complete list of single domain interaction pairs identified is available in our publication [163].

Table 3.2 Examples of inferred single-domain interaction pairs confirmed by InterDom

Domain A Pfam ID	Domain A Name	Domain B Pfam ID	Domain B Name
PF00153	Mito_carr	PF01423	LSM
PF00248	Aldo_ket_red	PF00106	adh_short
PF00155	Aminotran_1_2	PF00735	GTP_CDC
PF00018	SH3_1	PF00069	Pkinase
PF00241	Cofilin_ADF	PF00400	WD40
PF00694	Aconitase_C	PF01028	Topoisom_I
PF00330	Aconitase	PF01336	tRNA_anti
PF00501	AMP-binding	PF01253	SUI1
PF00022	Actin	PF01853	MOZ_SAS
PF00249	Myb_DNA-binding	PF00098	zf-CCHC

While most of the existing domain-based methods can only infer interactions for single-domain pairs, RDFF is capable of retrieving two or more domains for each protein in a pair from a tree branch that leads to an interaction classification. This is attractive as in some PPIs, it is highly probable that two or more domains in a protein cooperate with each other to form a module that interacts with another domain or domain module in the other protein. Here, a

domain module is defined as two or more domains functioning as a whole during interaction binding. Some of our identified domain module interactions are listed in Table 3.3.

Table 3.3 Examples of domain module interactions discovered

Domain Module in Protein A	Domain Module in Protein B
PF00083	PF00397; PF00168
PF00676	PF02779; PF02780
PF00036	PF00612; PF02736; PF00063
PF00009	PF02798; PF00043; PF00647
PF00459	PF00627; PF00442; PF00443
PF00026	PF00176; PF00271; PF00097; PB019909
PF01412	PF02826; PF00389; PF01842; PB042699
PF00076; PF00806	PF00248
PF00249; PF00569	PF00628
PF00004; PB030344; PF01426	PF02178
PF00006; PF02874; PF00306	PF00231
PF00169; PF00620; PF00617	PF00252

It is observed that domains in a module demonstrate strong association. For example, domains in the module {PF00006, PF02874, PF00306} listed in Table 3.3 (row #11) are annotated in Pfam as ATP synthase alpha/beta family, nucleotidebinding domain; ATP synthase alpha/beta family, beta-barrel domain; and ATP synthase alpha/beta chain, C-terminal domain, respectively. Actually, the three domains were identified by iPfam to cooperate with each other in binding to the ATP synthase (PF00231). Moreover, another domain module {PF00612, PF02736, PF00063} in Table 3.3 (row #3) is annotated as IQ calmodulin-binding motif; Myosin N-terminal SH3-like domain; and Myosin head (motor domain), respectively. The iPfam reported that the domains work together to form bonds with the EF hand (PF00036). Furthermore, the domains of the module {PF02779, PF02780} in the second row are Transketolase, pyridine binding domain, and Transketolase, C-terminal domain, respectively. The two domains are identified by iPfam to bind together in proteins to interact with the dehydrogenase E1 component (PF00676). In this study, total 867 interactions between domain modules were identified. A complete list can be found in our

publication [163]. Verifying those predictions is a challenging task because currently there are not enough resources available on domain module interaction pairs.

3.2.5 Protein-Protein Interaction Prediction by RDFF

With the putative domain–domain interactions, we can predict PPIs. To exemplify this, we select some predicted domain–domain interactions and then find proteins that contain these domains to see if these proteins interact with each other or not. For instance, as identified by Pfam [160], cell division control protein 7 (CDC7) contains protein kinase domain (PF00069). Both our model and Pfam identify the domain to be interacting with the ankyrin repeat (PF00023) domain. Regulatory protein SWI6 is known to contain the ankyrin repeat. Our model predicts the proteins CDC7 and SWI6 to be interacting. Indeed, the protein CDC7 is a conserved Dbf4-dependent protein kinase (DDK). Bailis et al. [164] has demonstrated that *Schizosaccharomyces prombe* Hsk1 (CDC7) regulates replication initiation, interacts and phosphorylates the heterochromatin protein 1 (HP1) which is the equivalent of SWI6. For another example, cell cycle protein kinase DBF2 contains the protein kinase domain (PF00069) and protein G2/mitotic-specific cyclin 2 (CLB2) contains Cyclin, N-terminal domain (PF00134). The PF00069 and PF00134 domain pair is inferred by our model and verified by iPfam as an interacting domain pair. The cell cycle protein kinase DBF2 and G2/mitotoicspecific cyclin 2 protein is predicted by our model to be an interacting protein pair. The DBF2 protein kinase is found to control the inactivation of the CLB2 (G2/mitotic-specific cyclin 2) kinase in late mitosis [165]. This clearly demonstrates the potential of the domain-based random forest framework in protein-protein interaction prediction.

We assessed the performance of RDFF for protein-protein interactions prediction in *Saccharomyces cerevisiae* and compared its results to the MLE method by Deng et al. [140].

The MLE method requires two input parameters: false positive (*fp*) and false negative (*fn*) rates. Deng et al. analyzed various values of the two parameters but did not observe any significant change in accuracy among the tested various values. Therefore, we choose to use one of their tested values, *fp* = 1.0E-5 and *fn* = 0.85, to train the MLE method over our training dataset. For RDFF, we chose a forest of 150 trees and each with maximum height threshold equal to 450. Impurity and minimum node size are set to 0.01 and 3, respectively. Details regarding the parameter selection are discussed in previous Section 3.2.3. To compare the MLE method and our method RDFF, sensitivity vs. (1 - specificity) was calculated on the test dataset and plotted in a receiver operating characteristic (ROC) in Figure 3.3.

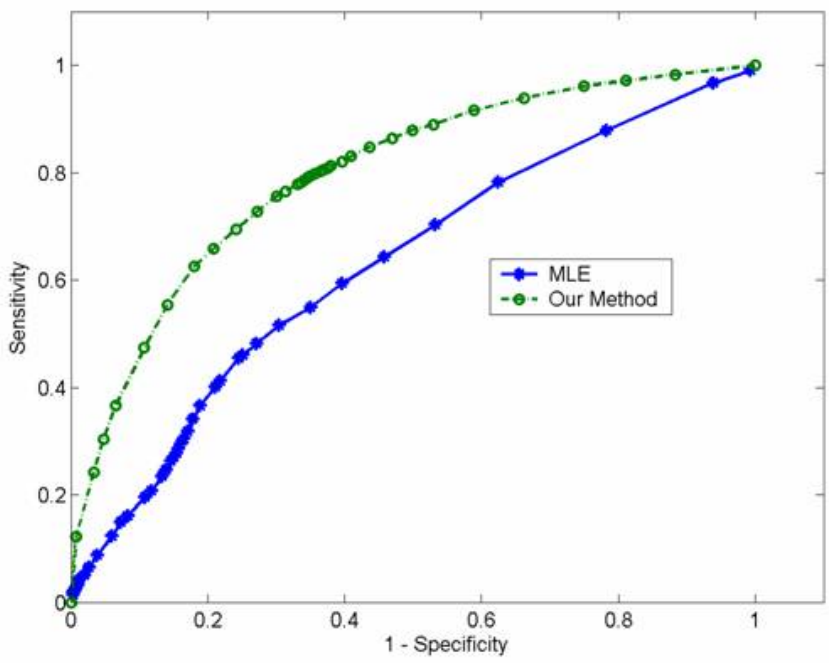

Figure 3.3 ROC performance comparison of RDFF and MLE

The ROC curve of our model is constructed by varying a classification threshold placed on the number of extra 'interaction' votes required for the final interaction prediction. Typically, majority votes win if the threshold equals to 0; however, the threshold can be

changed. For instance, a threshold at 5 implies that at least five more 'interaction' votes than 'non-interaction' votes must be obtained to classify a protein pair as interacting. Otherwise, the pair is classified as non-interacting. Therefore, with different thresholds, our model would perform differently in terms of specificity and sensitivity. There has always existed a trade-off between specificity and sensitivity. Regardless of the trade-off between the measures, it is clearly shown in Figure 3.3 that our method outperforms the MLE method in all ranges of the classification threshold. Table 3.4 compares the results of our method and the MLE over the test dataset. With comparable sensitivities fixed at approximately the same level 79%, RDFF can achieve 64.38% in specificity and the MLE can only reach 37.53% in specificity.

Table 3.4 Accuracy comparisons of RDFF and MLE

Method Name	**RDFF**	**MLE**
True Positives (TP)	3923	3850
False Positives (FP)	1425	2499
True Negatives (TN)	2575	1501
False Negatives (FN)	994	1067
Sensitivity (SN)	79.78%	78.30%
Specificity (SP)	64.38%	37.53%

Chapter 4. Domain Interaction Network for Function Prediction

As we move into the post genome-sequencing era, an immediate challenge is how to make the best use of the large amount of available high-throughput experimental data to assign functions to currently uncharacterized proteins. While experimental methods have been successful in identifying protein functions, they are extremely labor intensive and time consuming. Thus, genome-wide functional annotations must rely on *in silico* methods. Over the years, many computational methods have been developed and shown great promises; however, they still suffer from two major limitations: (1) low accuracy – the prediction accuracy of most methods are under 75%, which may not be of practical use for biologists; and (2) low specificity – predictions are normally generic functional categories.

In this research, I first explore a novel model to construct a weighted network of domain-domain interactions from cross-species protein interaction data where nodes are the domains and edges correspond to their interactions (Section 4.2). The proposed method is named CSIDOP for Cross-Species Interacting Domain Patterns. After building the DDI network with CSIDOP, we then statistically analyze the network with metrics that combine the static network topology and weights of the underlying interactions (Section 4.3). Finally, domain-domain interactions and protein-protein interactions predictions by the CSIDOP method are statistically evaluated in Section 4.4 and 4.5.

4.1 Principle of CSIDOP

It is well known that protein domains are the structural and/or functional units of proteins that are conserved through evolution. Some protein domains serve specific functions such as tyrosine kinase domains that covalently attach phosphate groups to select tyrosine residues in target proteins. Other protein domains may be more generic; for example, they may

participate in protein-protein binding and thereby are associated with numerous biological activities. A protein may encompass only one domain or multiple domains. In some cases, multiple domains may work together as a unit in a protein to direct physical bindings and executions of protein functions [166]. Pereira-Leal and Teichmann [167] suggested that protein interactions often evolve through duplication of the proteins involved in the interaction. In their work, Pereira-Leal and Teichmann defined partial duplicates as any two interaction pairs with one protein in common and homology between the other proteins. They defined complete duplicates as any two interactions where both proteins are homologous. Their results indicated that the duplicated modules in proteins tend to retain similar general functions. This suggests that interacting modular domains may be conserved over time and between organisms. Moreover, a shared pattern between two interacting protein pairs may signify that both protein pairs interact through the shared modular domains; thus they may exhibit similar functions.

Under this hypothesis, if two PPI pairs contain a common interacting domain pattern, then proteins in the two pairs with similar modular domains are more likely to be associated with similar functions. For example, assume that there exist two PPI pairs: protein A interacts with protein B and protein C interacts with protein D. If proteins A and C contain the same modular domain X that interact with the modular domain Y in proteins B and D, then we conclude that the two PPI pairs share a common interaction domain pattern. Therefore, we extrapolate that proteins A and C are more likely to have similar functions, and the same applies to proteins B and D (Figure 4.1).

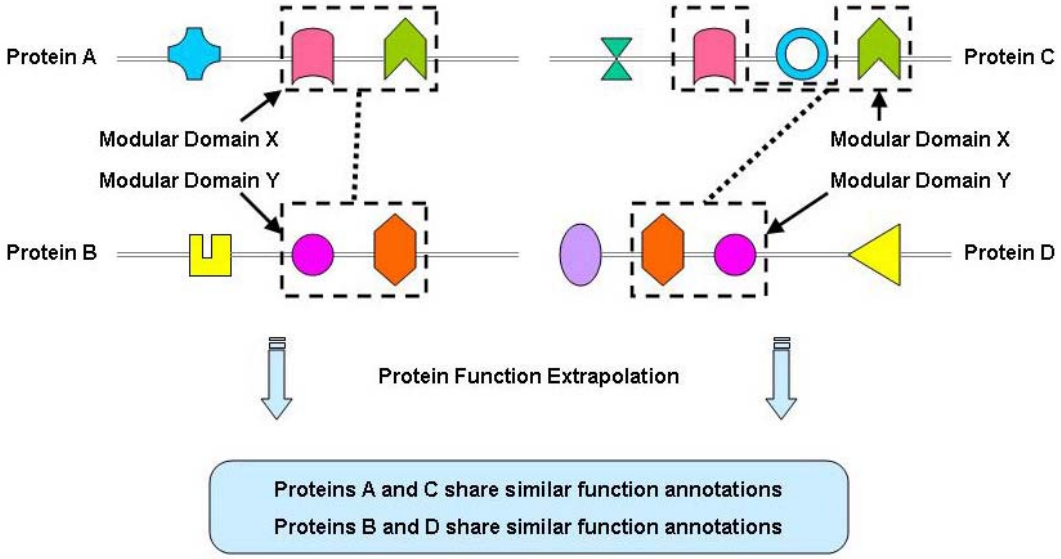

Figure 4.1 Function annotation scheme based on interacting domain patterns. This also illustrates how domain interaction can contribute to protein interactions. One or more domains in a protein may form modular domains and interact with other modular domains in other proteins. Dashed rectangles represent modules. In each module, one or more domains may exist and form a unit during interaction. The dashed lines represent interactions between proteins. Since the protein-protein interaction pairs A-B and C-D share common domain interaction patterns, and proteins A and C and B and D share the same interacting modular domains, we may deduce that the proteins are associated with similar functional annotations.

To further analyze the hypothesis in real data, we looked into protein interaction pairs from different species and observed evidence of this conservation of function between the pairs. For example, in *C. elegans*, nhr-67 [Swiss-Prot: Q9XVV3] and daf-21 [Swiss-Prot: Q18688] have been shown to interact [168], whereas in *H. sapiens*, ESR1 [Swiss-Prot: P03372] and HSP90AA1 [Swiss-Prot: P07900] are also known to interact [169]. Both protein interaction pairs contain a common domain interaction pattern, (PF00105)-(PF02518, PF00183), where '-' denotes interaction and the parentheses denote modular domains. PF00105 is described by Pfam [160] as the zinc finger, C4 type domain, and PF02518 and PF00183 refer to HATPase_c and HSP90 domains, respectively. The proteins nhr_67 and ESR1 contain the PF00105 domain, whereas daf-21 and HSP90AA1 contain the modular

domain (PF02518, PF00183). In the Gene Ontology database [17], the proteins nhr-67 in *C. elegans* and ESR1 in human are annotated to the same GO terms including regulation of transcription, DNA dependent (GO:0006355) and DNA binding (GO:0003677). Analogously, daf-21 and HSP90AA1 were also found to be annotated with the same GO terms: ATP binding (GO:0005524) and protein folding (GO:0005515). Hence, we can explore this property of interaction conservation as means to build a domain interaction network and assign protein functions by concentrating on protein-protein interaction (PPI) pairs with similar interacting modular domain patterns.

4.2 Domain-Domain Interaction Network

Most DDI prediction methods are based on protein interaction data. As much as we appreciate the available data generated from high-throughput protein interaction experiments, several independent studies have indicated their false positive rates to be in the order of 50% [22, 29, 78, 92]. In order to construct a reliable domain-domain interaction network from the noisy protein interaction network, we employ the CSIDOP principle to extract conserved interacting domain patterns buried in the enormously noisy assembly of protein interaction pairs across diverse species.

4.2.1 Data Source

To build a DDI network, we utilized a large scale collection of protein-protein interaction data in four species: *S. cerevisiae*, *C. elegans*, *D. melanogaster*, and *H. sapiens*, from the DIP January 2008 release [159], BioGRID 2.0.38 release [14], and HPRD September 2007 release [15]. Specifically, this dataset contains 54,987, 3,085, 5,375, and 30,223 protein interaction pairs among 3,794, 1,609, 2,059, and 7,167 proteins in *S. cerevisiae*, *C. elegans*, *D. melanogaster*, and *H. sapiens*, respectively.

Protein domain information is extracted from Pfam 22.0 [170]. For each protein, both pfam-A and pfam-B domains are considered. Among our protein interaction datasets, there are 4,542, 2,346, 3,715, and 12,082 unique Pfam domains identified in *S. cerevisiae*, *C. elegans*, *D. melanogaster*, and *H. sapiens*, respectively. Some domains are shared across species. In total, there are 429 Pfam domains in common between all four species. Complete information regarding domain distribution across the four organisms is available in Figure 4.2. For protein annotation information, we obtained 'molecular function' terms from the GO January 2008 release [17]. Within our interaction dataset, there are a total of 2,972 unique GO annotated molecular function terms.

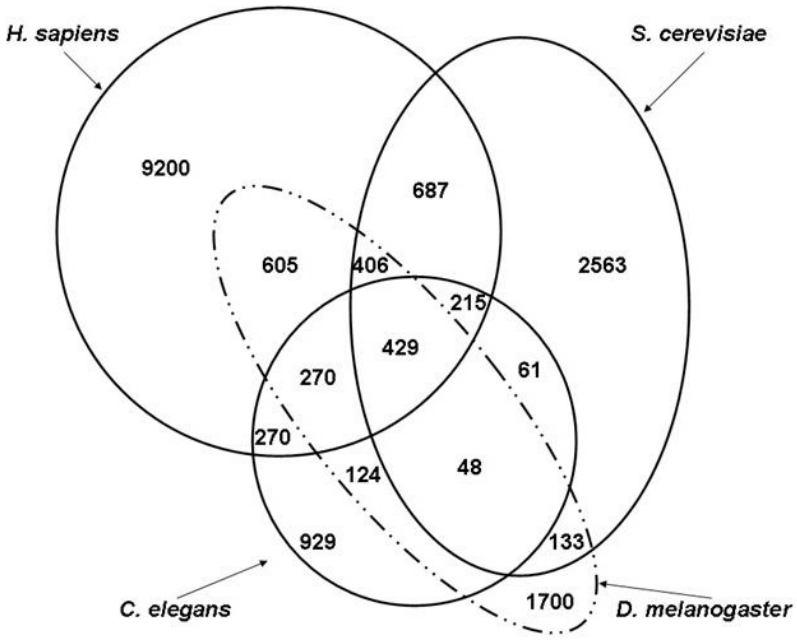

Figure 4.2 Domain distribution of different organisms: *S. cerevisiae*, *C. elegans*, *D. melanogaster*, and *H. sapiens*. Each organism can share different number of Pfam domains with other organisms. Among the protein-protein interactions we have collected for the four organisms, there are 429 Pfam domains in common between all four species as shown. There are total 753, 1,016, and 1,737 common domains between *S. cerevisiae* and the remaining three organisms: *C. elegans*, *D. melanogaster*, and *H. sapiens*. On the other hand, *C. elegans* has 808 and 1,194 domains in common with *D. melanogaster* and *H. sapiens*. Finally, *D. melanogaster* and *H. sapiens* share 1,710 Pfam domains.

4.2.2 Weighted DDI Network Construction by CSIDOP

In an effort to construct a reliable domain-domain interaction network, a new algorithm is devised to gather functionally related protein-protein interaction pairs into groups and then apply χ^2-statistics to acquire significantly conserved interacting domain patterns. Figure 4.3 is a flowchart of the CSIDOP approach. In essence, for each pair of interacting proteins in the protein interaction network, we strive to identify its close neighbors based on functional distances between individual proteins. The distance between proteins is defined as the closest GO-graph-node distance among their annotated GO molecular function terms. Since GO is designed as a directed acyclic graph where each node represents a GO term, the GO-graph-node distance is defined here as the least number of nodes separating two GO terms. More precisely, each protein interaction pair in the PPI network serves as a mean point, or centroid, and functional distances between the centroid pair and all remaining pairs are computed. An incoming PPI pair is accepted to join the group if and only if the distances among individual proteins in the centroid pair and the pair under consideration are below a certain threshold *t*. For instance, assume that there are two PPI pairs, A-B and C-D, where '-' denotes interaction. The two pairs are said to be functionally related if and only if the following condition is satisfied: the closest GO-graph-node distance between either (A, C) and (B, D) or (A, D) and (B, C) are less than or equal to *t*. The threshold *t* is empirically set to 3 in this study. In the end, PPI pairs in the same cluster are assuredly more likely to share the same or similar functions.

After forming a group of functionally related PPI pairs, we attempt to derive the most representative interacting domain patterns that are uniquely conserved across multiple organisms with the same or similar functions (i.e. in the same group). Since proteins often contain multiple domains, and one or more domains may form a functional unit during

interaction (i.e. modular domain), both single and modular domains are considered. Therefore, all possible combinations of modular domains in a protein are considered in generating the potential interacting domain patterns. Due to the existence of some big proteins with more than 15 domains, measures must be taken to trim down the set of all possible combinations by restricting the modular domain size to 4. This assumption is reasonable because it is unlikely for a large number of domains to come together and form a single unit during interaction. Additionally, the same set of a large number of domains is unlikely to occur repeatedly in other proteins across organisms.

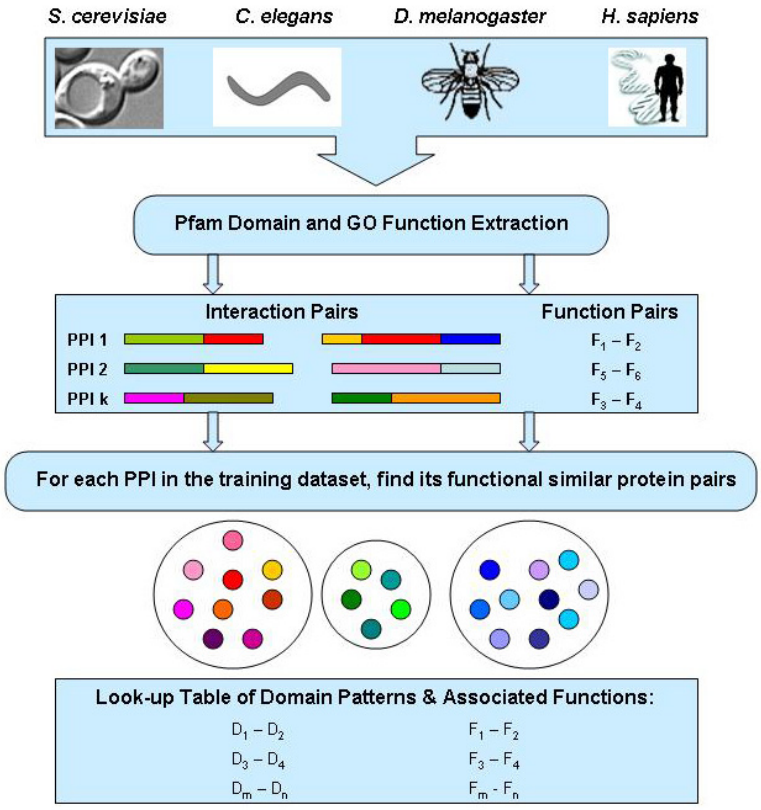

Figure 4.3 Flowchart of the CSIDOP approach. The model begins with a collection of protein interaction pairs across various species and extraction of their domain and function information. Then, for each PPI pair, it searches for its close neighbors based on their GO-graph-node functional distances. Finally, from each group, significant interacting domain patterns can be derived and in turn form a lookup table of patterns and associated functional assignments.

As a result, for each PPI pair in an individual group, a list of possible interacting domain patterns is enumerated. Each domain pattern will be associated with a list of function terms from their corresponding PPI pairs. In order to select the most significant interacting domain patterns, the χ^2 statistic is calculated for each potential pattern using the following formula (Eq. 4.1):

$$\chi^2 = \frac{N \times (AD - CB)^2}{(A + C)(B + D)(A + B)(C + D)} \tag{4.1}$$

N is the total number of PPI pairs in the protein interaction network. Variable A is the number of PPI pairs in the group that contain the particular 'pattern', and B is the number of remaining PPI pairs outside the group that contain the 'pattern'. Variables C and D are the number of PPI pairs that do not contain the 'pattern' in the group and in the remaining samples outside the group, respectively. An interacting domain pattern occurring more frequently in PPI pairs inside the group than outside the group is expected to have a higher χ^2 value; hence it is more significant. Finally, the highest scoring interacting domain patterns are reported to create the DDI network where nodes denote domains and edges represent interactions with strength equals to its χ^2-value and a lookup table for protein function annotations (Figure 4.3).

4.2.3 Weighted DDI Network Analysis

Network structures are observed in many natural and man-made complex systems such as social phenomena (e.g. scientific collaborations), communication systems (e.g. Internet), transportation infrastructures (e.g. airline routes), and biological systems (e.g. gene and/or protein interaction network). These highly interconnected systems have been extensively studied, which highlighted a number of topological features. One remarkable finding is the presence of scale-free nature in these networks [171], or a degree distribution showing

power-law behavior, which is extremely relevant to networks' robustness or vulnerability. In scale-free networks, there are a small number of highly connected nodes, often referred to as hubs, which secure the network integrity. The prevalence of this property in biological systems may indicate an evolutionary advantage because scale-free networks are more robust to random perturbations than other network architectures [172]. However, these networks are vulnerable to targeted attacks of their hubs. For instance, viruses may interfere with activities of the hub proteins to induce massive changes in cellular behavior. In past studies, these topological features have mainly been considered in networks with links between nodes that represent binary states. However, biological networks are generally not determined only by their layouts. Many expect heterogeneity to exist in the capacity and intensity of the connections. Recently, Barrat et al. [173] studied weighted scientific collaboration and world-wide transportation networks considering interplays between links and their weights.

Here, we present a statistical analysis of the weighted Pfam domain interaction network constructed in Section 4.2.2 using the CSIDOP method where the nodes denote Pfam domains and edges represent interactions. The DDI network is composed of 5,582 domains and 20,837 interactions. Each link between domains is weighted by a score – expectation value – to reflect the strength of the interaction. The score is calculated through χ^2-statistic which are used to derive significant domain interacting patterns in CSIDOP. In this network analysis, we desire to investigate the interplay between the DDI network topology and strength of the underlying interactions.

As a first insight into the role of weights, we examine the interplays between vertex degree products and edge weights. In real world networks, interrelation of the degree product $k_i k_j$ and weight w_{ij} often follows a power-law curve. This possibly signifies dependency between the network layout and weights of links. For an undirected graph, the degree of a

vertex $k_i = \sum_j a_{ij}$ is the number of edges incident to the vertex i where $a_{ij} = 1$ if a link exists between vertex i and j. In our DDI network, we hardly notice any such dependence where the mean expectation value is almost constant throughout a wide range of degree products (Figure 4.4). This phenomenon generally implies a lack of correlation between number of domain interaction partners and their interaction strengths [173].

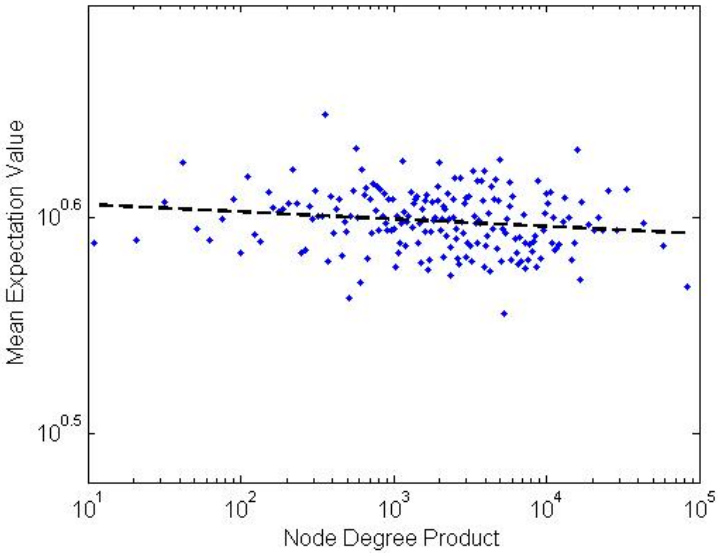

Figure 4.4 Node degree product vs. mean expectation value. The dependence of mean interaction expectation value from the domain degree product k_ik_j shows a weak correlation $\sim k_ik_j^{-0.003}$.

To further investigate the dependence between the DDI network topology and their interaction strengths, we apply a series of measures introduced by Barrat et al. [173] that combines both network topology and weights to assess the impact of weights. In a weighted DDI network, the initial definition of vertex degree $k_i = \sum_j a_{ij}$ is extended in terms of domain strength s_i, which is defined as

$$s_i = \sum_j a_{ij} E_{ij} \tag{4.2}$$

where E_{ij} is the expectation values or strengths of interactions of domain i. When comparing cumulative frequency distributions of the domain degree k and its strength s, we observe that both distributions have power-law tails $P(x) = x^{-\alpha}$ (Figure 4.5). The observation of power-law tails in the degree distribution implies the existence of the scale-free property [174]. In other words, it suggests that the DDI network integrity depends on a small subset of highly interactive domains. Similarly, it signifies that a majority of domains have low strength while only a minority of domains can reach high strength.

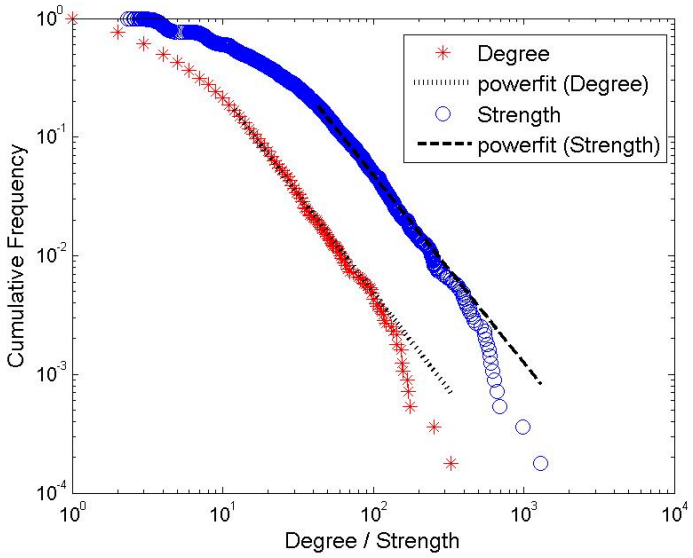

Figure 4.5 Cumulative frequency distributions of node degree and strength. For measures that consider single domain such as the node degree k and interaction strength s, we observe a power-law tail in cumulative frequency distribution with $P(k) = k^{-2.64}$ and $P(s) = s^{-2.5822}$.

Another typical network analysis measure is called local network cohesiveness. The clustering coefficient C_i (Eq. 4.3) of a vertex in an unweighted network is calculated to quantify how close its neighbors are to being a clique (complete graph). For additional information about the structure of the underlying DDI network, we examine the average clustering coefficient $C(k)$ that is restricted to a class of domains with degree k.

$$C_i = \frac{2a_{jh}}{k_i(k_i-1)} : v_j, v_h \in neighborhood\ N_i \qquad C(k) = \frac{1}{n}\sum_{i=1}^{n} C_i \tag{4.3}$$

In most real world networks, *C(k)* exhibits a highly nontrivial behavior with a power-law decay as degree *k* increases. High *C(k)* values indicate that low degree vertices belong generally to well interconnected communities, and small *C(k)* values mean that hubs serve as bridges for many vertices. Here, we examine the average clustering coefficients of domains with a certain degree *k* in both weighted and unweighted networks. By considering interaction strengths, the weighted clustering coefficient can be defined as in Eq. 4.4.

$$C_i^w = \frac{1}{s_i(k_i-1)} \sum_{j,h} \frac{E_{ij}+E_{ih}}{2} a_{ij} a_{ih} a_{jh} \tag{4.4}$$

As expected, power-law dependence is observed in both networks which demonstrates that the weighted clustering coefficient preserves its dependence from the degree *k* (Figure 4.6).

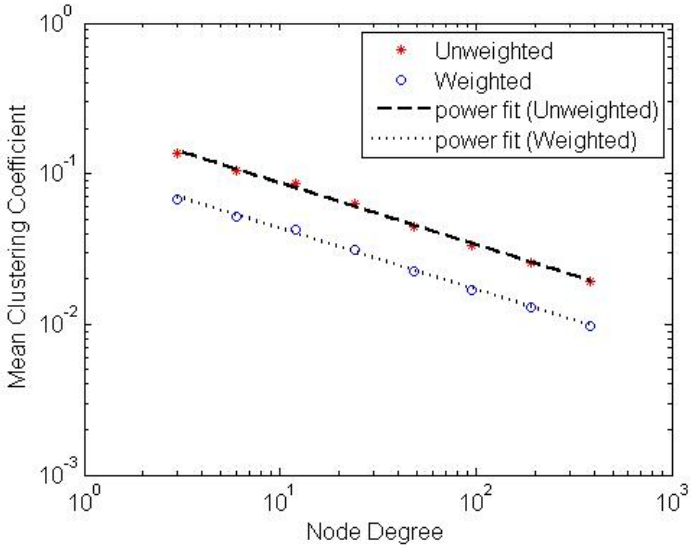

Figure 4.6 Mean clustering coefficient. For clustering coefficient that measures local group cohesiveness or network modularity, we notice the dependence of unweighted clustering coefficient *C* decays as a power-law, $C(k) \sim 0.223k^{-0.41}$. The same trend is observed for the weighted clustering coefficient $C^w(k) \sim 0.111k^{-0.405}$. We logarithmically binned the data points first and calculated mean values in each bin.

Furthermore, to gain a deeper understanding of the relationship between network layout and weights, degree-degree correlations are evaluated by the average degree of nearest neighbors $k_{nn,i}$. In unweighted and weighted networks, $k_{nn,i}$ is defined respectively as

$$k_{nn,i} = \frac{1}{k_i}\sum_j a_{ij}k_j \qquad k^w_{nn,i} = \frac{1}{s_i}\sum_j a_{ij}E_{ij}k_j \tag{4.5}$$

The average nearest-neighbor degree with connectivity k is referred to as $k_{nn,i}(k)$. This can be used to identify two types of networks. If $k_{nn,i}(k)$ is an increasing function of k, then higher degree vertices are more likely to connect with large-degree vertices, a property that is known as assortative mixing. In contrary, if $k_{nn,i}(k)$ is an decreasing function of k, then it means that majority of the higher degree vertices are connected to ones with low degree, which is called disassortative mixing. In our case, we have found both unweighted and weighted DDI networks have a trend toward disassortative mixing (Fig. 4.7) which was also recognized in other biological networks [133, 175].

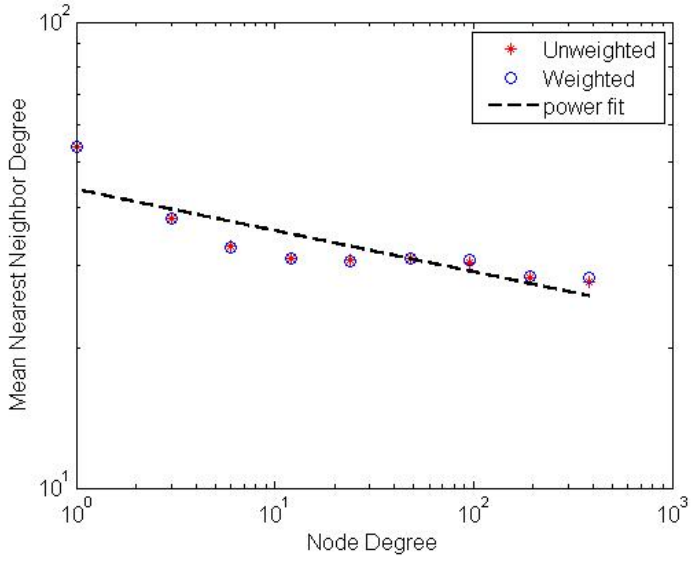

Figure 4.7 Average nearest neighbor degree. We reached roughly the same result for unweighted and weighted average nearest neighbor degree, which slightly decays as increasing degree. The weak dependency can be approximated by a power-law ~ $43.686k^{-0.088}$. We logarithmically binned the data points first and calculated mean values in each bin.

In summary, our constructed DDI network exhibits the scale-free property, and while investigating interplays between the network topology and strengths of the underlying interactions, we observe that the unweighted measures and their weighted counterparts generally follow the same trends. The same phenomena were also identified by Wuchty [133]. Thus, these observations further confirm the effectiveness of our CSIDOP method for DDI predictions.

4.3 Domain-Domain Interaction Predictions

In the previous section, our network analysis of the DDI network built by CSIDOP showed that the network as a whole exhibits the scale-free property which is consistent with the observations of other studies in domain interaction network. In this section, we would like to assess the quality of individual domain-domain interactions in the network.

4.3.1 Statistical Evaluations

Due to limited availability of a gold standard dataset for domain-domain interactions, evaluation of domain interaction prediction methods becomes a challenging problem. As a standard solution, domain pairs reported to interact in crystal structures of protein complexes are often used as the benchmark of true positives. To verify the DDIs predicted by our algorithm CSIDOP, we compare them to domain pairs reported in iPfam [161] and 3DID [176], which are referred to here as known DDIs. Both databases regard two domains as interacting if and only if they are close enough in at least one PDB complex to form an interaction. However, one must keep in mind that iPfam and 3DID only embody a small fraction of all true interacting domain pairs. According to a recent study conducted by Itzhaki et al. [177], domain interaction pairs stored in iPfam and 3DID databases account for no more than 20% of the protein-protein interactions for any of the *E. coli*, *S. cerevisiae*, *C. elegans*, *D.*

melanogaster, and *H. sapiens* organisms examined. Hence, it is expected that the number of predicted DDIs to be verified by the two databases is low. For instance, only 11.35% of the predictions by the domain co-evolution based method, RCDP, are confirmed by iPfam [129]. Those predictions without any PDB evidences are not necessarily false positives. It is highly probable that domain interaction prediction methods will retrieve many true interacting domain pairs that are just not part of the high-confidence databases yet.

To evaluate the reliability of the predicted DDIs, we adopt a statistical approach described by Deng et al. [140]. If CSIDOP is reasonable, a real domain interaction pair should be much more likely to be verified by the known interacting domain pairs than random pairs. To measure the excess, we calculate a quantity called Fold or ratio of the fraction of matched DDIs in predicted domain pairs with those in all pairs (Eq. 4.6).

$$Fold = \frac{k_0 / K}{n / L} \tag{4.6}$$

L denotes the total number of domain pairs, n is the number of predicted DDIs, K is the total number of known DDIs, and k_0 is the number of matching domain pairs between the predicted and known DDIs.

When the DDI network was built, a χ^2-value or expectation value is assigned to each derived domain pair. One would assume that the larger the score, the more likely the interaction is real. Thus, to assess performance of our method in detail, we set up a threshold by which two domains are considered to be interacting if and only if the expectation value is greater than this threshold. Comparing to the known DDIs, experiments are performed over various combinations of the threshold. Since individual databases cover different parts of the protein domain space, we only consider those domain pairs that exist in each. Fold calculation of the predicted DDIs to the known ones in iPfam and 3DID databases are shown in Table 4.1 and 4.2, respectively.

Table 4.1 Evaluation of the predicted domain-domain interactions vs. known interactions in iPfam

Top % scoring DDIs considered	Total # of predictions	# of overlaps with iPfam	Fold
10%	298	48	103.38
20%	630	84	85.58
30%	1036	104	64.43
40%	1442	122	54.30
50%	1745	141	51.86
60%	2125	162	48.93
70%	2548	182	45.84
80%	3169	201	40.71
90%	4572	266	37.34
100%	5796	319	35.32
Random	2377290	3704	1

Table 4.2 Evaluation of the predicted domain-domain interactions vs. known interactions in 3DID

Top % scoring DDIs considered	Total # of predictions	# of overlaps with 3DID	Fold
10%	324	52	195.68
20%	730	96	160.34
30%	1183	119	122.65
40%	1677	138	100.33
50%	2071	161	94.79
60%	2545	182	87.19
70%	3058	207	82.53
80%	3813	229	73.23
90%	5513	310	68.56
100%	6934	373	65.59
Random	6126750	5025	1

As shown in Table 4.1 and 4.2, the domain-domain interaction predictions by our algorithm CSIDOP are significantly better than random. The comparison results against iPfam and 3DID databases demonstrate similar trends. The Fold decreases as we lower the threshold in determining whether two domains interact or not. This is expected as the higher the threshold, the smaller and more reliable the resulting DDI network is.

4.3.2 Comparison to Other Methods in DDI Prediction

We compared the domain interaction predictions by CSIDOP with those of our previous method RDFF [163], RCDP by Jothi et al. [129] and DPEA by Riley et al. [147]. The

objective of this comparison is to figure out how the percentage of the CSIDOP predictions confirmed by iPfam compares against other methods. It must be emphasized here that this is only an indirect comparison because different datasets were utilized in each study, and it would be extremely difficult to test these methods on the same dataset as some of the methods impose unique set of constraints on the input dataset. For example, RCDP [129] considers only those PPIs with both proteins having orthologous hits in 10 or more genomes. As shown in Table 4.3, when the top 10% and 20% of the scoring DDIs are utilized in CSIDOP, the percentage of overlap with iPfam is 16.11% and 13.33%, which are slightly lower than the best performing RCDP with a test set of SLA ≥ 75% (17.26%).

Table 4.3 Indirect comparison of CSIDOP to RDFF, DPEA, and RCDP

	# of Pfam DDI predictions	**# of predictions confirmed by iPfam**	**% of predictions confirmed by iPfam**
RDFF	2475	104	4.20%
DPEA	1812	185	10.21%
RCDP_SLA50	960	109	11.35%
CSIDOP_20	630	84	13.33%
CSIDOP_10	298	48	16.11%
RCDP_SLA75	336	58	17.26%

Numbers listed in the table for RDFF, DPEA, and RCDP are obtained from the study by Jothi et al. [129]. For our method CSIDOP, top 10% and 20% scoring DDIs are utilized in the comparisons.

Furthermore, we want to find what percentage of our predictions is confirmed by other methods. We compared the CSIDOP prediction results with DDIs listed in the DOMINE database [178]. DOMINE collects DDIs inferred from PDB entries and those by eight different computational approaches. For each domain pair, it assigns a label HCP, MCP, LCP or PDB to represent high-, medium-, or low-confidence and PDB inferred interactions. High-confidence pairs (HCP) are those predicted using multiple sources of information or by at least two sufficiently different computational methods. Medium-confidence pairs (MCP) are predicted by just one approach in which both domains are a part of the same GO biological

process. Low-confidence pairs (LCP) are the ones predicted simply by one computational approach. The comparison summary is illustrated in Table 4.4. For different choices of parameters, 15.06% to 28.17% of our predictions are confirmed by the DOMINE database, and among them 6.65% to 19.21% are verified either by PDB determined interactions or predictions from heterogeneous data sources and multiple approaches.

Table 4.4 Fraction of CSIDOP domain-domain interaction predictions confirmed by DOMINE

Top % scoring DDIs considered	% of DDI predictions confirmed by DOMINE	% of DDI predictions confirmed by PDB + HCP
10%	28.17%	19.21%
20%	27.74%	17.60%
30%	23.10%	13.77%
40%	19.77%	11.53%
50%	19.14%	10.56%
60%	18.07%	9.53%
70%	17.77%	8.98%
80%	16.49%	7.96%
90%	15.52%	6.91%
100%	15.06%	6.65%

For those DDI predictions not in DOMINE, we investigate them further for supporting evidences. As we know, two domains often interact to achieve a common objective; therefore, two interacting domains are more likely to share similar GO annotations. By examining the closest GO-graph-node distance between each pair of domains predicted, many pairs are found to share the same GO annotation or have the closest GO-graph-node distance of 1 indicating a direct parent-and-child relationship where the parent is a more general description and the child is more specific. For example, ARID (PF01388) and SAM_PNT (PF02198) domains are predicted to interact. The ARID domain is known to participate in DNA binding (GO:0003677). The SAM_PNT domain is known to execute sequence-specific DNA binding (GO:0043565), which is a children term of the DNA binding (GO:0003677). For another example, our method predicted the ERM (PF00769) and WW (PF00397)

domains to interact. The ERM domain participates in protein binding (GO:0005515), and the WW domain is annotated to cytoskeletal protein binding (GO:0008092). Cytoskeletal protein binding (GO:0008092) is a more specific term of the protein binding (GO:0005515).

Under a fixed parameter (i.e. top 10% scoring DDIs), a total of 329 DDIs were predicted but not found in DOMINE. Among those, 176 pairs contain both domains with GO annotation from Pfam. Out of the 176 pairs, we found 47 domain pairs (26.70%) with constituent domains having GO-graph-node distances less than and equal to 2. To assess the significance of the percentage, we computed the GO-graph-node distance for all possible domain pairs (2,377,290 in total) to see how many of the random domain pairs would have a distance less than or equal to 2. Out of all 2,377,290 domain pairs, 1,408,681 pairs had GO annotations available for both domains. Among the 1,408,681 random domain pairs, 167,309 (11.88%) had GO-graph-node distance of 2 or less, which is more than a two-fold reduction compared to what is observed in predicted DDIs (26.70%). Table 4.5 summarizes the results for various GO-graph-node distances as thresholds.

Table 4.5 Comparison between fractions of our DDI predictions (176 pairs) and random domain pairs (1,408,681 pairs) having certain GO-graph-node distance

Shortest GO-graph-node distance	# Random domain pairs	% Random domain pairs*	# Predicted DDIs	% Predicted DDIs¶
= 0	47,150	3.35%	21	11.93%
≤ 1	76,648	5.44%	32	18.18%
≤ 2	167,309	11.88%	47	26.70%
≤ 3	330,705	23.48%	62	35.23%
≤ 4	559,503	39.72%	83	47.16%

*****Calculated based on a total of 1,408,681 random domain pairs
¶Calculated based on a total of 176 predicted DDIs

Annotations of some predicted DDIs with GO distance of 1, 2, and 3 are listed in Table 4.6. Examples of some predicted domain pairs with shared annotations (i.e. distance = 0) are illustrated in Table 4.7. A complete list of the 329 domain pairs can be found in appendix A.

Table 4.6 Examples of domain-domain interactions predicted with the closest GO-graph-node distance of 1, 2, and 3 (not found in DOMINE)

Domain A	Domain B	Annotation A	Annotation B	Dist
RNase_PH (PF01138)	DEAD (PF00270)	RNA binding (GO:0003723)	Nucleic acid binding (GO:0003676)	1
RNA_pol_Rpb1_7 (PF04990)	RRM_1 (PF00076)	DNA binding (GO:0003677)	Nucleic acid binding (GO:0003676)	1
bZIP_1 (PF00170)	SH2 (PF00017)	Protein dimerization activity (GO:0046983)	Protein binding (GO:0005515)	1
Ets (PF00178)	ARID (PF01388)	Sequence-specific DNA binding (GO:0043565) Transcription factor activity (GO:0003700)	DNA binding (GO:0003677)	1
RBD (PF02196)	Hint (PF01079)	Signal transduction (GO:0007165)	Cell communication (GO:0007154)	1
Zf-C2H2 (PF00096)	KIX (PF02172)	Nucleic acid binding (GO:0003676)	Protein binding (GO:0005515)	2
C1_1 (PF00130)	Hint (PF01079)	Intracellular signaling cascade (GO:0007242)	Cell communication (GO:0007154)	2
RBD (PF02196)	HH_signal (PF01085)	Signal transduction (GO:0007165)	Cell-cell signaling (GO:0007267)	2
TFIIA (PF03153)	Homeobox (PF00046)	RNA polymerase II transcription factor activity (GO:0003702)	Transcription factor activity (GO:0003700)	2
SAM_PNT (PF02198)	TSC22 (PF01166)	Sequence-specific DNA binding (GO:0043565)	Transcription factor activity (GO:0003700)	2
HLH (PF00010)	Wnt (PF00110)	Transcription regulator activity (GO:0030528)	Signal transducer activity (GO:0004871)	3
RHD (PF00554)	Ribosomal_S5_C (PF03719)	Regulation of transcription (GO:0045449)	Translation (GO:0006412)	3
Cadherin (PF00028)	ZZ (PF00569)	Calcium ion binding (GO:0005509)	Zinc ion binding (GO:0008270)	3
SQS_PSY (PF00494)	TIP49 (PF06068)	Transferase activity (GO:0016740)	DNA helicase activity (GO:0003678)	3
C1_1 (PF00130)	HH_signal (PF01085)	Intracellular signaling cascade (GO:0007242)	Cell-cell signaling (GO:0007267)	3

Table 4.7 Examples of domain interaction pairs predicted with shared GO annotations (not found in DOMINE)

Domain A	Domain B	Shared GO Annotation
Pkinase_Tyr (PF07714)	Furin-like (PF00757)	ATP binding (GO:0005524) Protein amino acid phosphorylation (GO:0006468)
Ets (PF00178)	TSC22 (PF01166)	Regulation of transcription, DNA-dependent (GO:0006355) Transcription factor activity (GO:0003700)
PHD (PF00628)	Zf-C4 (PF00105)	Zinc ion binding (GO:0008270)
Ku (PF02735)	Histone (PF00125)	DNA binding (GO:0003677)
wnt (PF00110)	Hint (PF01079)	Multicellular organismal development (GO:0007275)

The above results suggest that our algorithm CSIDOP is able to discover biologically relevant novel interacting domain pairs. Our method can indeed predict true interacting domain pairs overlooked by other methods. It can be used along with other method to detect unrecognized domain interactions; thus it provides a wider coverage of the entire domain interaction space.

4.4 Protein Function Predictions

An essential issue concerning the protein function prediction problem is the assessment of method reliability. To evaluate the CSIDOP method, the collected protein interaction data is partitioned into two groups: training and testing. The training data is used to extract interacting domain patterns, and it only contains PPI pairs where both proteins are annotated in the GO. The test dataset, on the other hand, consists of interaction pairs that have either one of the proteins uncharacterized or both unknown. Thus, we can assess the reliability of the CSIDOP method by determining how well it works in function prediction for those GO-

characterized proteins and infer novel functions for proteins that are currently not characterized in GO in the test dataset.

For the function prediction evaluation, we decided to analyze the CSIDOP method over proteins in *H. sapiens*. The collected human protein interaction data are separated accordingly into training and test datasets. Interaction pairs in the two datasets are exclusively different, so a protein pair can only belong to one set. To train the CSIDOP method, we integrated PPI data from the organisms *S. cerevisiae*, *C. elegans*, and *D. melanogaster,* in addition to the large dataset from *H. sapiens*. In order to assess the relative performance of our method, inferred functions of the *H. sapiens* proteins (by CSIDOP) were then compared to the known functions in the GO database which we designate as the 'true' terms. Throughout this paper, the 'true' function terms of a protein refer to the known function terms of this protein listed in the GO. One match between the predicted terms and the corresponding true GO terms for a protein indicates a correct prediction; it is otherwise a wrong prediction.

4.4.1 Comparison to Other Methods in Function Prediction

After training, CSIDOP produces a lookup table of significant interacting modular domain patterns from protein interaction pairs in the training dataset where each pattern is associated with a number of function terms (please refer to Section 4.2.2 for details). Function annotations can be assigned to a PPI pair in the test dataset if it contains at least one interacting modular domain pattern listed in the table. Overall, we assigned GO function terms to 618 *H. sapiens* proteins from PPIs with common domain patterns in the lookup table. Among the 618 predicted proteins, 437 had existing annotations in the GO database and could be used to evaluate the CSIDOP method. Among the 437 proteins, 417 were assigned with correct functions by the CSIDOP (assigned functions have an exact match with the

'true' terms), i.e., the CSIDOP method had an accuracy of 95.42% (Table 4.8) using 2,972 GO functional terms, which is higher than most of the existing *in silico* methods. For comparison, we also tested the Majority Rule (MR) method by Schwikowski et al. [103], a simple domain based method, and an orthology based method.

Table 4.8 Accuracy comparison for different function prediction methods

Method	Accuracy
CSIDOP	95.42%
Orthology based method	83.86%
Pfam domain based method	61.98%
Majority Rule (MR)	59.50%

Accuracies of the listed methods are compared in protein function prediction. The accuracy is defined as the percentage of proteins predicted with correct function terms. A protein is considered to be correctly annotated if there is a match between the predicted and known function terms.

Generally, the MR algorithm assigns a protein with the function terms that occur most frequently amongst its direct interaction partners. Assessing the MR algorithm on the same test dataset that we used in CSIDOP, MR made functional predictions with an accuracy of 59.50% (Table 4.8). As for the domain based method, considering the fact that a number of protein domains are annotated in Pfam [160] with specific functions, it is possible to make protein function predictions according to the functional terms of its constituent domains. Using the same set of proteins, only 61.98% were assigned with correct functions using the simple domain based scheme (Table 4.8). Lastly, for the orthology based method, we attempted to assign functions to proteins according to their annotated orthologs in other species. The orthologs were retrieved using Inparanoid [179]. The orthology based method achieved a prediction accuracy of 83.86%, and among the novel predictions, it only covered 56.35% of our novel discoveries. Therefore, this demonstrates that our CSIDOP method can provide extra power in protein function prediction compared to the orthology detection.

Most existing methods have been evaluated on the *S. cerevisae* proteome using a much smaller number of functional categories. Schwikowski et al. [103], Hishigaki et al [104], and Brun et al [110] used 42, 41, and 44 "cellular role" categories in the Yeast Protein Database (YPD) [180], and the accuracies achieved were 72%, 64%, and 67%, respectively. Vazquez et al. [108] evaluated their method using two different level of functional classification in MIPS [115]. In the coarse-grained level containing only 20 functional categories, the accuracy was about 83%. In the finest level containing 424 functional categories, the accuracy decreased to 65%. Noticeably, the CSIDOP prediction is made over 2,972 GO functional categories. This is significantly larger than those employed in other methods. Accordingly, the assigned functions are specific rather than generic. In principle, the more coarse-grained the classification, the easier the prediction is. Applying the same definition of success, our CSIDOP method is able to make correct predictions an astounding 95.42% of the time using the full 2,972 GO molecular function categories. However, in the GO function tree, the closer a node is to the root, the lower the level in the GO tree which means that the corresponding function is more abstract and the farther it is from the root, the higher the level in the GO tree, and thus, the more detailed. An important advantage of the CSIDOP method is that it can be tailored to different levels in the GO database based upon demand. For example, suppose that the GO level is set to five, then all predicted terms at GO tree levels higher than or equal to five will be generalized to the corresponding function at level five. In other words, the more specific functional terms that reside at higher levels of the tree are replaced with their ancestor terms which are located at level five. Higher prediction accuracy is expected as we lower the GO depth. Consistent with the expectation, the prediction accuracy in the test dataset reached 98.85% when the depth parameter is set to 2, which still contains 129 GO functional categories (Table 4.9). Table 4.9 shows the prediction accuracies

as a function of GO level for this test dataset and indicates the robustness and reliability of the CSIDOP method. This depth parameter allows users to assign function terms for a protein at different resolutions according to their individual needs.

Table 4.9 Evaluation of the CSIDOP algorithm at different prediction resolution

Depth in the GO graph	# of unique GO function categories	# of correctly predicted proteins	# of proteins w/ predicted terms different from their GO terms	Prediction accuracy
2	129	432	5	98.85%
3	473	427	10	97.71%
4	961	422	15	96.56%
5	1996	419	18	95.88%
6	2598	418	19	95.65%
7	2816	417	20	95.42%
8	2938	417	20	95.42%
9	2957	417	20	95.42%
10	2972	417	20	95.42%

Accuracy is assessed over a number of values for the depth parameter (i.e. generalizing annotated terms when parameter decreases). A protein is considered to be correctly annotated if the known function occurred among the predicted terms.

4.4.2 CSIDOP Contribution to Current GO Annotation

A protein often exhibits multiple molecular functions, so its annotation in GO may not be complete. CSIDOP may provide additional functional terms to existing proteins. For example, the Alpha-2-macroglobulin precursor (Swiss-Prot: P01023) was predicted by CSIDOP to be involved in protease inhibitor activity (GO:0030414) which is not among the current list of functions annotated in GO. Consistent with this prediction, alpha-2-macroglobulin is found to be a major human plasma protease inhibitor capable of inhibiting most endopeptidases tested [181]. Another example is the PRS7 (Swiss-Prot: P35998) gene in human. This gene is currently annotated in GO to participate in protein binding (GO:0005515) with no other listed terms. Our CSIDOP method predicted that it is also involved in ATP binding (GO:0005524), hydrolase activity (GO:0016787), nucleotide binding (GO:0000166), and nucleoside_triphosphatase activity (GO:0017111), all of which can be verified in InterPro

[182]. Other assigned terms for PRS7 by CSIDOP included endopeptidase activity (GO:0004175) and ATPase activity (GO:0016887), which were observed in the orthologous protein of PRS7. An orthologous protein in *D. melanogaster*, RPT1 (Fly-Base: FBgn0028687), is annotated with endopeptidase activity inferred from direct assay [183]. Another orthologous protein in *S. cerevisiae*, YKL145W, is also annotated with the function terms endopeptidase activity and ATPase activity.

Moreover, for the 20 proteins (in previous section 4.4.1) with predicted functions that do not match with their 'true' terms, the differences between the predicted terms and the 'true' terms may be due to the incompleteness in current GO annotations. To gain insight into the 20 proteins that were "incorrectly" annotated by CSIDOP, we analyzed the relationship between the predicted terms and their true GO terms. Figure 4.8 shows a histogram of distances between the predicted terms and the 'true' GO terms, which is defined as the number of edges between these two terms in the GO graph. As illustrated in Figure 4.8, 15 out of the 20 proteins were predicted with function distances of one or two. A distance of one means that the two terms have a direct parent-child relationship; for instance, protein binding (GO:0005515) is a known function of Furin precursor protein (Swiss-Prot: P09958), and our method predicted it to be involved in protein domain specific binding (GO:0019904) which is a direct child term of protein binding in GO. If we consider such case to also be a successful prediction, then the accuracy improves from 95.42% to 97.71%. A distance of two indicates that the two terms share a parent. For example, suppressor of cytokine signaling 1 (Swiss-Prot: O15524) was identified in GO to be associated with insulin-like growth factor receptor binding (GO:0005159), whereas we assigned the function term, sevenless binding (GO:0005118). The two terms share a parent term, receptor binding (GO:0005102). In this

case, if the more general terms were used, a correct functional annotation would have been achieved.

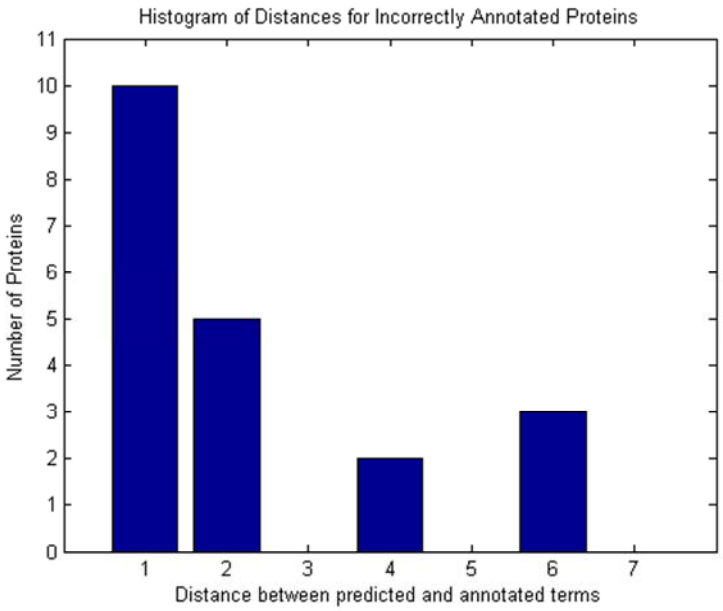

Figure 4.8 Histogram of distances between the wrongly predicted terms and the 'true' terms.

Moreover, we analyzed correlations between the predicted function terms and the 'true' terms. In GO, gene products can be associated with more than one term. Therefore, the correlation between two GO terms is defined based on the number of gene products in common. The larger the correlation value is, the closer the two GO terms are. In order to assess the significance of the correlation scores between predicted and 'true' terms, 10,000 GO term pairs were randomly selected, and a correlation score was computed for each pair. E-value is described as the probability of random GO term pairs achieving at least a certain correlation score. For instance, an E-value of 0.0008 implies that only eight out of the 10,000 random GO term pairs have scores equal to or higher than a particular correlation score. As a result, we observed that among the 20 "incorrectly" annotated proteins, many predicted terms are closely correlated to the true GO terms with significant E-values. Table 4.10 shows the

number of proteins versus different E-values. Examples of proteins in which extremely high correlation exists between the predicted and 'true' terms (E-value ≤ 0.0008) are illustrated in Table 4.11.

Table 4.10 Correlation analysis of proteins with known terms that differ from the predicted ones

E-value	Correlation score (≥ score)	# of proteins
0.0116	1	19
0.0028	10	16
0.0021	20	14
0.0014	50	12
0.0008	100	10
0.0006	200	8
0.0005	300	6
0.0003	500	5
0.0001	3000	3
0.0000	10000	1

Correlation score between two GO terms is defined as the number of gene products in common. E-value is defined as the probability of random GO term pairs achieving at least a certain correlation score. The third column shows number of the wrongly predicted proteins reaching different correlation scores between predicted and 'true' terms.

Table 4.11 Examples of proteins with high correlation scores between predicted and 'true' terms

Protein	'True' GO term	Predicted GO term	Correlation
Partitioning defective 6 homolog alpha (Swiss-Prot: Q9NPB6)	Rho GTPase binding (GO:0017048)	Actin binding (GO:0003779)	186
SH3-containing GRB2-like protein 2 (Swiss-Prot: Q99962)	Transferase activity (GO:0016740)	Calcium ion binding (GO:0005509)	425
Hepatocyte growth factor precursor (Swiss-Prot: P14210)	Serine-type endopeptidase activity (GO:0004252)	Peptidase activity (GO:0008233)	6430
Erythrocyte membrane protein band 4.2 (Swiss-Prot: P16452)	ATP binding (GO:0005524)	Transferase activity (GO:0016740)	33762

Examples of proteins with predicted terms different from their 'true' terms but sharing high correlation scores (i.e. E-value ≤ 0.0008).

4.4.3 Novel Protein Function Assignment

Most importantly, CSIDOP has made novel functional annotations for 181 *H. sapiens* proteins that are not currently described in the GO database. Some of these novel annotations are supported with evidence provided by QuickGO, a web browser of gene ontology data maintained by the European Bioinformatics Institute. For instance, the gene FHL1, four and a half LIM domains protein (Swiss-Prot: Q13642), is identified by CSIDOP to participate in metal ion binding (GO:0046872) and zinc ion binding (GO:0008270). The metal ion binding annotation can be found in QuickGO which was inferred from UniProt keywords. The zinc ion binding term was found by both the UniProt keywords and in InterPro [182], which is a database of protein families, domains and functional sites in which identifiable features found in known proteins can be applied to unknown protein sequences.

Many novel functional annotations are supported by evidences found in their orthologous protein annotations. Orthologous proteins are generally believed to have similar functions, and the orthologs can be obtained from Inparanoid [179]. For example, the *H. sapiens* gene POLA2, DNA polymerase subunit alpha B (Swiss-Prot: Q14181), was predicted by CSIDOP to exhibit alpha DNA polymerase activity (GO:0003889). Orthologs of POLA2 found by Inparanoid include: POL12 (ORF: YBL035C; SGD:S000000131) in *S. cerevisiae*, POLA2 (RGD:621817) in *R. norvegicus*, and CG5923 (FlyBase: FBgn0005696) in *D. melanogaster*. All three orthologs are associated with the alpha DNA polymerase activity (GO:0003889).

Furthermore, the CSIDOP method detected three molecular function terms for the human protein SLY, SH3 protein expressed in lymphocytes homolog (Swiss-Prot: O75995), while no information was found anywhere else. The three functions identified are DNA binding (GO:0003677), chromatin binding (GO:0003682), and zinc ion binding (GO:0008270). The SLY protein contains a COR1 chromatin-binding domain, and it was suggested by Ellis et al.

[184] that SLY may be targeted to the gonosomes in spermatids and may regulate gonosomal chromatin conformation and expression. Another protein, CCNB3 (Swiss-Prot: Q8WWL7) in the human genome, is predicted by the CSIDOP method to be involved in cyclin-dependent protein kinase regulator activity (GO:0016538) and protein binding (GO:0005515). An orthologous protein found in *D. melanogaster*, CG5814 (FlyBase: FBgn0015625), shared both functional annotations which were inferred from sequence or structural similarity and physical interaction [185] respectively. In the literature, CCNB3 was described as sharing properties with both A- and B-type cyclins. Cyclins play a key role in controlling progression through the cell cycle. They act as regulatory subunits of p34cdc2/CD28 and related cyclin-dependent protein kinases (cdks) [186]. Tschop et al. found CCNB3 to interact with the cyclin-dependent kinase CDK2 which implies that it indeed participates in protein binding and cyclin-dependent protein kinase regulator activity [187]. Some of the 181 novel functional annotations found with supporting evidences can be found in supplementary Table S1 of our published paper [188]. A complete list of the novel predictions for proteins in *H. sapiens* can also be found in supplementary material (Text S3) of our paper [188].

4.4.4 Robustness of CSIDOP in Function Prediction

The CSIDOP method is shown above to produce highly accurate function predictions for proteins in *H. sapiens*. To demonstrate its robustness, we further analyzed the method for its performance on *D. melanogaster*. For this study, we integrated protein interaction data from *S. cerevisiae*, *C. elegans*, and *H. sapiens* to form the training dataset to determine functional annotations of proteins in *D. melanogaster*, the test dataset. None of the protein pairs in *D. melanogaster* were involved in training our model. In other words, the interacting domain patterns are extracted purely from interaction pairs in *S. cerevisiae*, *C. elegans*, and *H.*

sapiens species only. Function annotations are effectively assigned to 447 *D. melanogaster* proteins. Among the 447 proteins, CSIDOP accurately assigned function annotations to 419 proteins (i.e. 93.73% in accuracy).

In addition, we are able to discover novel annotations for some proteins. For example, the *D. melanogaster* protein CG15912 (Swiss-Prot: Q9W4J7) is detected by CSIDOP to exhibit ATPase activity, coupled to transmembrane movement of ions, phosphorylative mechanism (GO:0015662). Its orthologs, Haloacid dehalogenase-like hydrolase domain containing 3 (Swiss-Prot: Q9BSH5) in *H. sapiens* and (Swiss-Prot: Q9CYW4) in *M. musculus*, were both found to be associated with phosphoglycolate phosphatase activity (GO:0008967) and hydrolase activity (GO:0016787) which is an ancestor term of our predicted term (GO:0015662). Moreover, for the protein CG18445 (Swiss-Prot: Q9V5F2), a multispan transmembrane protein related to fly Porcupine, our algorithm identified it to carry out the O-acyltransferase activity (GO:0008374). Through literature search, we discovered that biological experiments conducted by Kraut et al. [189] confirmed the findings for CG18445.

Since the CSIDOP method only keeps the most significant interacting domain patterns from the closely related protein interaction pairs across species, PPI pairs in the test dataset without patterns in the lookup table will result in no prediction which subsequently leads to lower prediction coverage. To enlarge the coverage, we further refine our interacting domain pattern based algorithm by devising a two-step prediction method: the first step will predict functions for a large number of proteins with lower confidence, and the second step uses CSIDOP for more accurate predictions. In the first step, for each protein pair in the test dataset, we construct a list of all interacting domain patterns. Then for each of these plausible domain patterns, we try to collect a list of protein interaction pairs in the training dataset that contain the pattern. Numerous interaction pairs with shared patterns may exist in the training

dataset, and certain functions annotated to those pairs may be more likely to be associated with the target protein pair than other functions. Thus in order to assess the probability of each functional assignment, we calculate the conditional probability of a protein interaction pair having function pair F_1-F_2 given interacting domain pattern D_1-D_2 (Eq. 4.7), where '-' denotes interaction. In other words, F_1 and F_2 represent function assignments to proteins in the query interaction pair with modular domains D_1 and D_2, respectively.

$$P(F_1 - F_2 \mid D_1 - D_2) = \frac{P(F_1 - F_2, D_1 - D_2)}{P(D_1 - D_2)} \tag{4.7}$$

P(F_1-F_2, D_1-D_2) is calculated by counting the number of interaction pairs in the reference dataset that contain the interacting domain pattern D_1-D_2 and have the corresponding functional annotation of F_1-F_2, and P(D_1-D_2) is computed by counting the number of pairs that contain the interacting domain pattern D_1-D_2. For a query protein interaction pair, the posterior probabilities of all possible function pairs are calculated, and finally, the top ranking function pairs are assigned. In this step, we are able to assign functions to 1,546 proteins, but with lower accuracy of 90%.

Since this prediction step is based on probability of a gene *g* having term *t*, terms with probabilities above a certain threshold can be treated as a positive prediction, and terms below the specified threshold can be treated as a negative prediction; thus, sensitivity and specificity measures can be calculated. Applying the same idea in Nariai et al. [102], sensitivity is defined as TP/(TP+FN), which corresponds to recall, and specificity is defined as TN/(TN+FP), which corresponds to precision. A set of observed positive *g*-*t* associations is obtained from the GO. The observed negative association set is defined as the association not found in the positive set and term *t* is neither an ancestor nor a descendant of the known function terms in GO hierarchy for gene *g*. Intuitively, true positives (TP) in this case refer to the overlaps between our positive predictions and the observed positive set, and true

negatives (TN) are the overlaps between our negative predictions and the observed negative set. False positives are the *g-t* associations in our positive prediction list which are observed to be in the negative set by GO. Lastly, false negatives are the *g-t* associations in our negative prediction list that should be in the positive list. For varying posterior probability cutoffs, sensitivity and 1-specificity is plotted in a ROC curve (Fig. 4.9). As shown, the lowest sensitivity of 57% is achieved with specificity equal to 96%. When the specificity is lowered to 78%, the sensitivity increases dramatically to 93%.

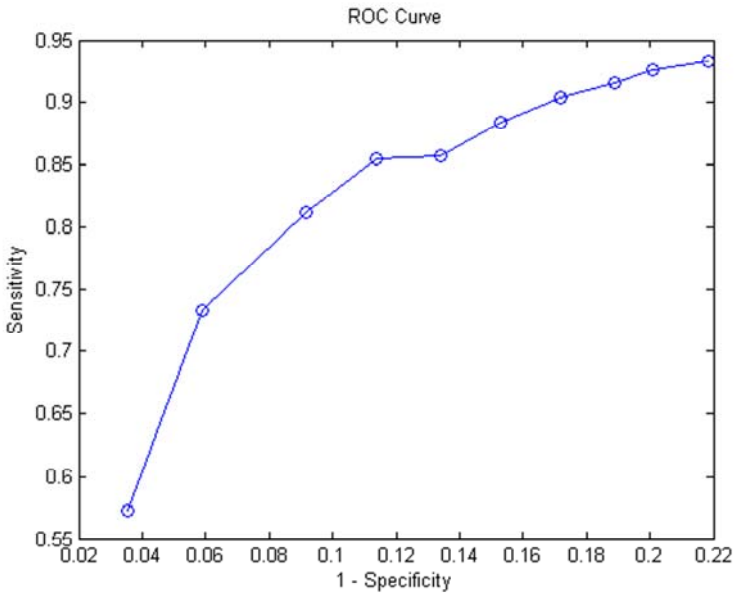

Figure 4.9 ROC curve in function prediction. Sensitivity = TP/(TP+FN), Specificity = TN/(TN+FP). Function terms with probability above certain threshold are considered to be positive predictions and terms below the specified threshold are treated as negative predictions. The observed positive set of *g-t* association is obtained from GO. The negative association set is defined as follows: if the association is not found in the positive set and term *t* is neither ancestor nor descendant of the known function terms in GO hierarchy for gene *g*. Therefore, true positives (TP) in this case refer to the overlaps between our positive predictions and observed positive set. True negatives (TN) are the overlaps between our negative predictions and the observed negative set. False positives describe *g-t* associations exist in our positive prediction list, but should be in the negative set. False negatives are *g-t* associations in our negative prediction list, but should be in the positive list.

Chapter 5. PINFUN Online System

In Chapters 3 & 4, we proposed novel domain-domain interaction based methods for protein-protein interaction prediction and protein function annotation, and experimental results demonstrated their robustness and reliability. In this chapter, we integrate all domain-domain interactions determined by the proposed frameworks (RDFF & CSIDOP) with those identified by other methods (i.e. computational and PDB) into an online system called **PINFUN** for **P**rotein **IN**teraction and **FUN**ction predictions.

5.1 System Overview

PINFUN is a web-based systematic tool to infer protein-protein interactions GO function annotation for given query proteins on the basis of underlying protein domains and their interactions. An overview of the entire PINFUN system is presented in Figure 5.1.

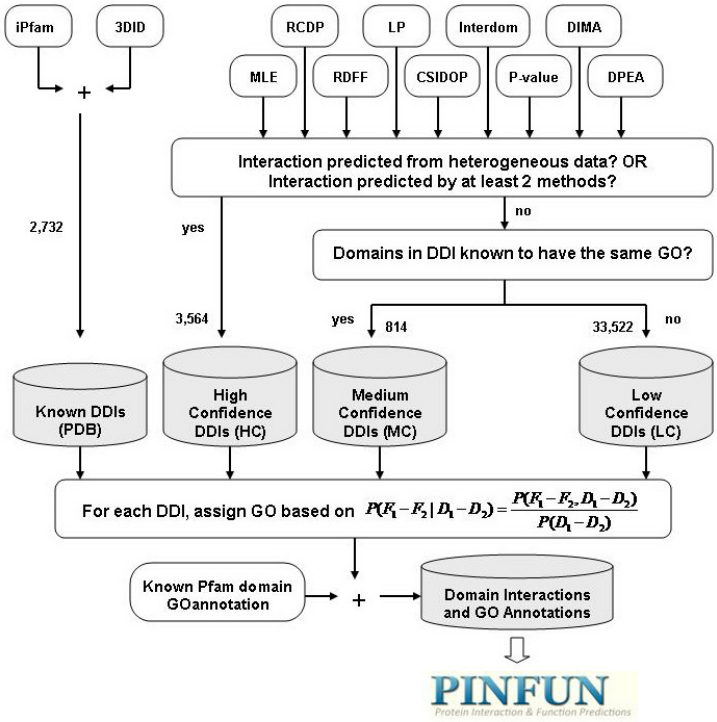

Figure 5.1 PINFUN system overview.

To ensure accuracy and coverage, we not only utilized domain-domain interactions identified by our aforementioned RDFF and CSIDOP approaches, but we also took advantage of the available known domain interactions observed in PDB complexes from iPfam [161] and 3DID [176] and putative DDIs discovered by seven other computational approaches: MLE [136], RCDP [129], LP [150], InterDom [137], DIMA [128], P-value [145], and DPEA [147] (Figure 5.1). Complete known and putative domain-domain interaction data for various methods is obtained from DOMINE [178] except for the data from our CSIDOP method. Details regarding each source are discussed in Table 5.1. Overall, we gathered 40,632 interaction pairs among 8,238 domains including both Pfam-A and Pfam-B domains.

The collected domain interaction pairs are then assigned to four confidence bins: PDB, HC, MC, and LC (Figure 5.1). Domain interaction pairs inferred from PDB entries (i.e. DDIs from iPfam and 3DID) are regarded as known interactions and assigned to the 'PDB' bin. The putative interactions by computational approaches are classified into three other categories: 'HC', 'MC', and 'LC', which refer to High-, Medium-, or Low-confidence bins, respectively. The high-confidence 'HC' bin consists of those DDIs inferred from heterogeneous data sources of information or by at least two sufficiently different computational methods. The medium-confidence 'MC' bin includes domain pairs predicted by just one approach but both domains are a part of the same GO biological process. Lastly, the low-confidence 'LC' bin encompasses domain pairs predicted simply by one computational approach.

Table 5.1 PINFUN domain-domain interaction sources

Source	# of DDIs	Source Detail
iPfam	4,030	iPfam is a database of physical interactions between domains that have representative structures in PDB.
3DID	3,034	3DID is a database of domain-domain interactions in proteins for which high-resolution 3D structures are known. Data from August 2005 is used here.
MLE	2,391	MLE refers to the method by Lee et al. [136] that integrates gene ontology, domain fusion information, and protein interactions from multiple organisms through maximum likelihood estimation (MLE) .
RCDP	960	Given two interacting proteins, Relative Co-evolution of Domain Pairs (RCDP) by Jothi et al. [129] computes the degree of sequence co-evolution among all pairs of domains, and predicts the pair with the highest degree of co-evolution to interact.
LP	2,588	Linear Programming (LP) approach [150] by Guimaraes et al. seeks the minimal set of domain pairs necessary to explain a given protein interaction data. Thresholds LP-score ≥ 0.5 and $0.0 \leq$ p-score ≤ 0.1 are used. After discarding Pfam-B domains, 2,588 DDIs remained.
Interdom	2,768	InterDom [162] is a database of putative domain interactions from multiple sources. Here, only those DDIs inferred by domain fusion analysis from version 1.1 February 15, 2003 release are used.
DPEA	1,812	Domain Pair Exclusion Analysis (DPEA) by Riley et al. [147] infers DDIs by assessing the contribution of each potential interacting domain pairs to the likelihood of a set of observed protein interactions across multiple organisms. Here, only pairs with Pfam-A domains and log odds score ≥ 3.0 are used, resulting in 1,812 DDIs.
P-value	596	Refers to the statistical approach by Nye et al. [145] which tests the null hypothesis that presence of a domain pair in a protein pair do not affect whether two proteins interact or not. Since their DDIs are predicted between SCOP domain families [190], SGD [191] was used to map the SCOP domains back to Pfam domains.
DIMA	8,012	DIMA by Pagel et al. [192] infers DDIs based on phylogenetic profiling.
RDFF	2,475	Random Decision Forest Framework (RDFF) approach [163]. Interactions with Pfam-B domains are discarded.
CSIDOP	20,837	Cross-Species Interacting Domain Pattern (CSIDOP) method [188]. Both Pfam-A and Pfam-B domains are used.

After collecting the DDI data, we attempt to assign GO function annotations to each domain-domain interaction pair in our database using a probabilistic approach discussed in Section 4.4.4. Basically, for each domain interaction pair in the collection, we compile a set of protein-protein interaction (PPI) pairs containing the domain interaction pattern across

multiple organisms. Since every protein in a pair is annotated with GO function terms and numerous pairs of PPIs may share the same domain interaction pattern, certain functions annotated to those protein pairs may be more likely to be associated with the domain interaction pattern than other functions. Thus, in order to assess the probability of each functional assignment, we calculate the conditional probability of a protein interaction pair having function pair F_1-F_2 given the two proteins interact through domain pattern D_1-D_2 where '-' denotes interaction (Eq. 4.7 in Chapter 4 Page 81). For each domain interaction pair, posterior probabilities of all possible function pairs among the associated PPIs are calculated, and in the end, function annotations with the highest posterior probabilities are assigned to the corresponding domains.

It is important to note that the above mentioned approach to protein GO function prediction is based on domain interaction patterns. Thus, GO annotations of a single domain can be derived from the predicted annotations of all DDIs containing the domain. For example, assume that domain interaction pair D_1-D_2 is predicted with function pair F_1-F_2 and D_1-D_3 is predicted with F_3-F_4. Then the individual domain D_1 would have both annotations F_1 and F_3. Besides the putative domain GO annotations, we also collected known domain annotations from Pfam version 22.0. Finally, all collected DDIs and domain GO annotations, both inferred and known, are utilized in the PINFUN system (Figure 5.1).

5.2 Database Design

The previous section discussed how we prepared and processed various types of data that are necessary to infer protein-protein interactions and functions in PINFUN. Here, we describe the structure of our database system: how to store the prepared data. The central database system of PINFUN is implemented using MySQL. Since PINFUN deduces protein-protein

interactions and protein functions on the basis of Pfam domains, protein domain assignments are quintessential, and we downloaded the Pfam database from release 22.0 [170]. The Pfam database is a large repository of protein domain families where each family is represented by multiple sequence alignments and hidden Markov models (HMMs).

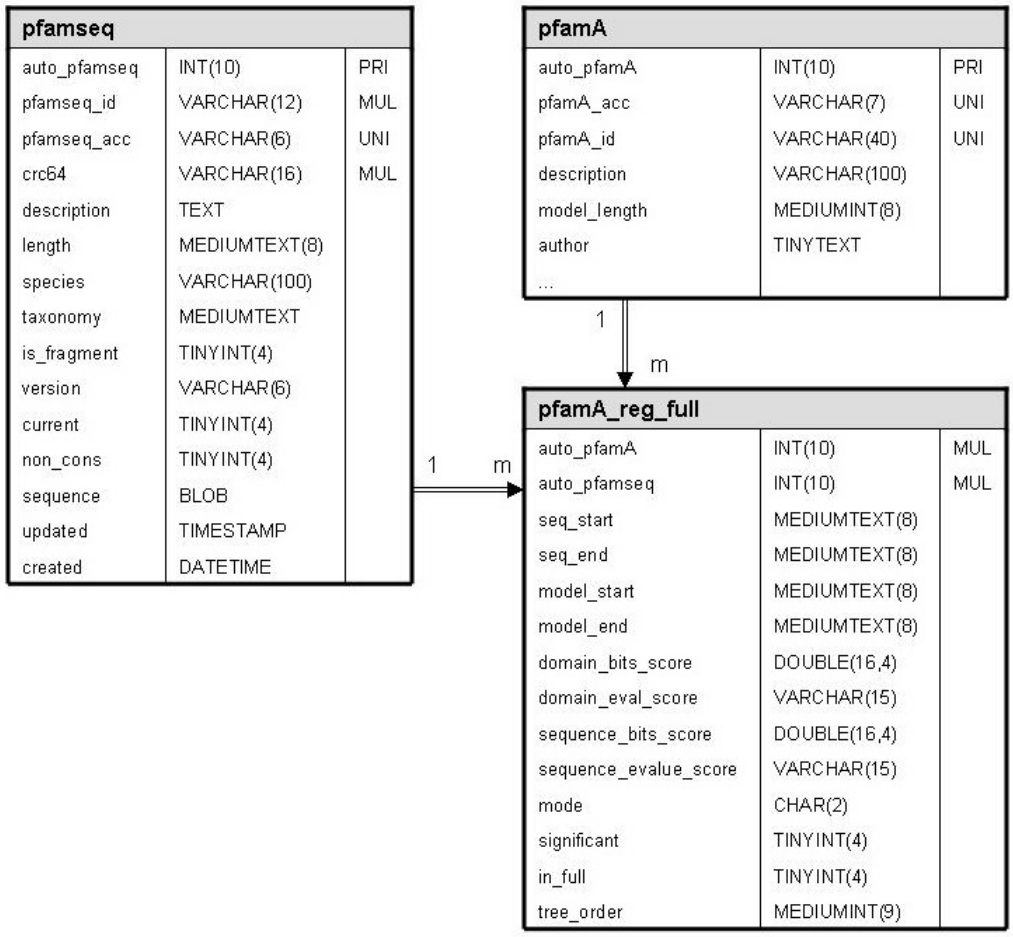

Figure 5.2 Central tables of Pfam database: pfamseq & pfamA. Each table contains three columns. Column 1 contains the table fields and column 2 contains the corresponding field types. Column 3 has information on the relational keys: PRI for primary key; UNI for unique key; and MUL for multiple key. Complete description of pfamA can be found in Appendix B.

The Pfam database contains three central tables: '**pfamseq**', '**pfamA**', and '**pfamB**'. The '**pfamseq**' table is an underlying sequence database built from the UniProtKB [193]. The

'**pfamA**' table contains information on the Pfam-A domain families, and '**pfamB**' table contains Pfam-B families. There are two supporting tables, '**pfamA_reg_full**' and '**pfamB_reg**', which contain all of the sequence regions that match the HMM for each Pfam domain family. Details regarding each table and their relationships are depicted in Figure 5.2 & 5.3. A complete description of the table '**pfamA**' can be found in Appendix B. With these Pfam tables, one can retrieve protein domain information easily. For example, obtaining Pfam-A domains of a UniProtKB protein can be accomplished with simple MySQL queries in the following text boxes.

Retrieve all domains for a UniProtKB protein with id = 'VAV_HUMAN'

```
SELECT      pfamA_acc
FROM        pfamseq, pfamA, pfamA_reg_full
WHERE       pfamseq_id = 'VAV_HUMAN'
AND         pfamseq.auto_pfamseq = pfamA_reg_full.auto_pfamseq
AND         pfamA_reg_full.auto_pfamA = pfamA.auto_pfamA
AND         in_full = '1';
```

Retrieve all domains for a UniProtKB protein with accession = 'P15455'

```
SELECT      pfamA_acc
FROM        pfamseq, pfamA, pfamA_reg_full
WHERE       pfamseq_acc = 'P15455'
AND         pfamseq.auto_pfamseq = pfamA_reg_full.auto_pfamseq
AND         pfamA_reg_full.auto_pfamA = pfamA.auto_pfamA
AND         in_full = '1'";
```

In addition to the Pfam central tables in the core database of PINFUN, we also created several tables to store other necessary information such as domain-domain interactions and domain annotations (Table 5.2). Details about the tables are shown in Figure 5.4. For a given domain, its interacting partner can be easily retrieved using the following SQL command.

Table 5.2 PINFUN tables

Table	Purpose
DDINet_Interaction	Stores information regarding domain-domain interactions, domain interaction confidence level, GO function assignments, and associated function assignment score computed using Eq. 4.7.
Domain_Annot	It contains single domain annotations compiled from the GO functions assigned to each domain interaction pairs (details discussed at the end of Section 5.1).
GO_Terms	Provides a full description map for each GO accession ids. Useful in display of results.

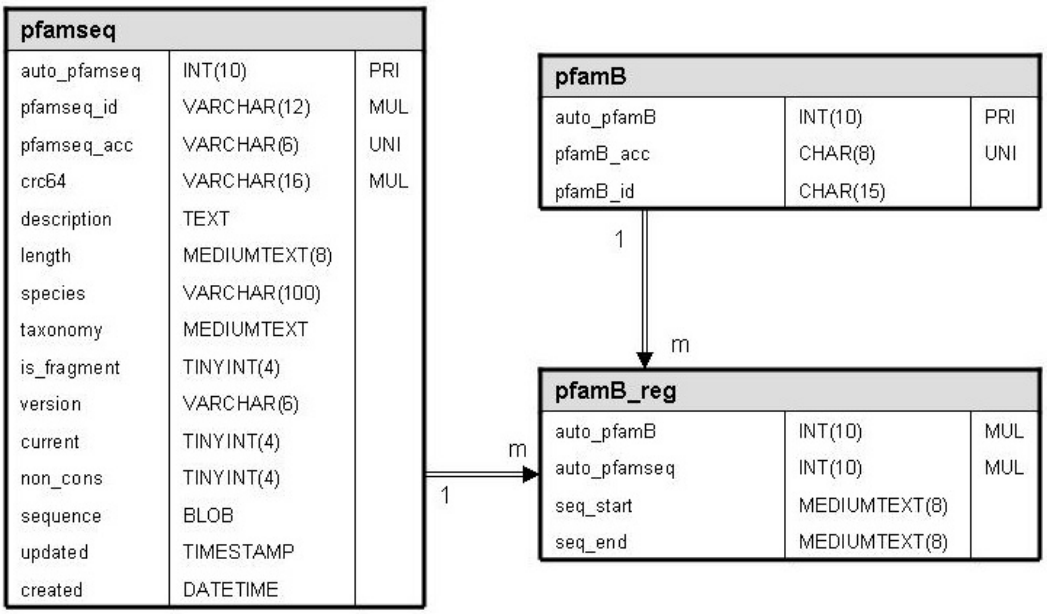

Figure 5.3 Central tables of Pfam database: pfamseq & pfamB. Each table contains three columns. Column 1 contains the table fields and column 2 contains the corresponding field types. Column 3 has information on the relational keys: PRI for primary key; UNI for unique key; and MUL for multiple key.

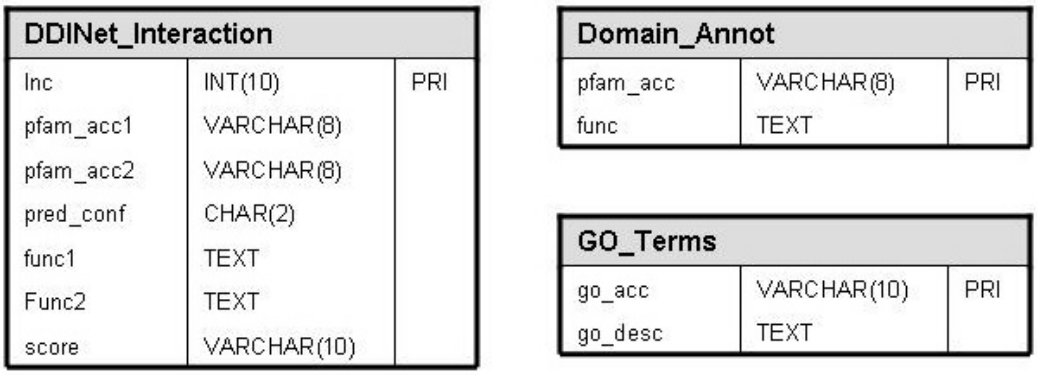

Figure 5.4 Remaining tables in the PINFUN database – contribute to protein interaction and function predictions. Each table has three columns. Column 1 contains the table fields and column 2 contains the corresponding field types. Column 3 has information on the relational keys: PRI for primary key; UNI for unique key; and MUL for multiple key.

5.3 Web Interface Design

The underlying PINFUN system is connected to the outside world via a web interface constructed using HTML and PHP scripts. It is currently available at http://www.ittc.ku.edu/~meiliu/PINFUN/pinfun.html. Figure 5.5 depicts the entire design of the web interface functionalities implemented in PINFUN. Primarily, PINFUN allows for two specific functionalities: protein-protein interaction prediction and protein function annotation. Details regarding how PINFUN achieves these two tasks are presented in the following sections.

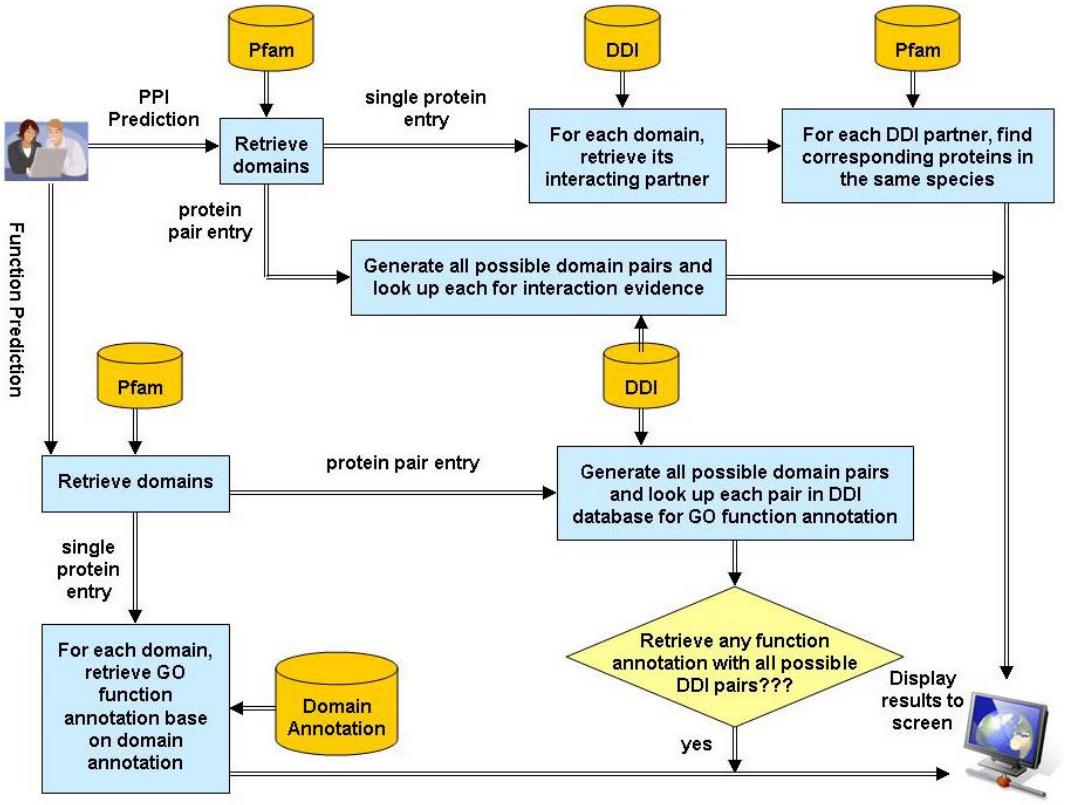

Figure 5.5 Main processes of PINFUN

5.3.1 PINFUN Protein-Protein Interaction Prediction

The first objective of PINFUN is to infer protein-protein interactions based on domain-domain interactions. For a given protein, a user may request to identify its possible interacting partners. In this case, the user simply needs to enter the query protein either in UniProtKB id or accession number and select the species to which the protein belongs (Figure 5.6). PINFUN presently only supports four organisms: *S. cerevisiae* (yeast), *C. elegans* (worm), *D. melanogaster* (fruit fly), and *H. sapiens* (human). After the user submits their query, PINFUN starts the protein-protein interaction prediction process by retrieving all Pfam-A domains belonging to the query protein. Then, it searches through the

'DDINet_interaction' table for all domain-domain interactions that each of the protein domains participates in. The task can be accomplished using the following SQL query:

Retrieve all domain-domain interactions for Pfam domain 'PF00069'

```
SELECT      *
FROM        DDINet_Interaction
WHERE       pfam_acc1 = 'PF00069'
OR          pfam_acc2 = 'PF00069';
```

From the above database search, a set of domain interaction partners of the constituent domains in query protein domains is compiled as a consequence. Intuitively, when domains of two proteins interact, the two proteins are sure to interact. Based on this concept, PINFUN identifies interaction partners of the query protein by seeking proteins that contain at least one of the domain interaction partners. For each domain interaction partner of the query protein, corresponding proteins can be retrieved with the following SQL statement.

Retrieve all proteins in yeast with Pfam domain 'PF00069'

```
SELECT      pfamseq_acc
FROM        pfamseq, pfamA, pfamA_reg_full
WHERE       pfam_acc = 'PF00069'
AND         pfamA.auto_pfamA = pfamA_reg_full.auto_pfamA
AND         pfamA_reg_full.auto_pfamseq = pfamseq.auto_pfamseq
AND         in_full = '1'
AND         pfamseq_id REGEXP 'YEAST';
```

As a result of the protein-protein interaction prediction from one query protein, PINFUN outputs all identified DDIs ordered by confidence level (Figure 5.7). The user may click on any one of the DDIs to acquire a list of possible interaction partners of the query protein (Figure 5.8).

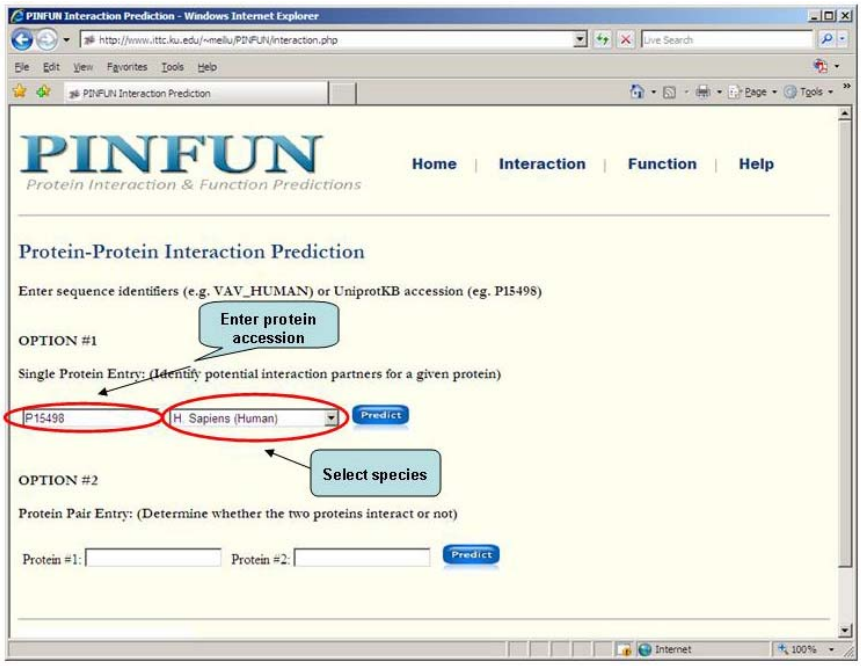

Figure 5.6 PPI prediction option #1 query – determines possible interaction partners of a query protein.

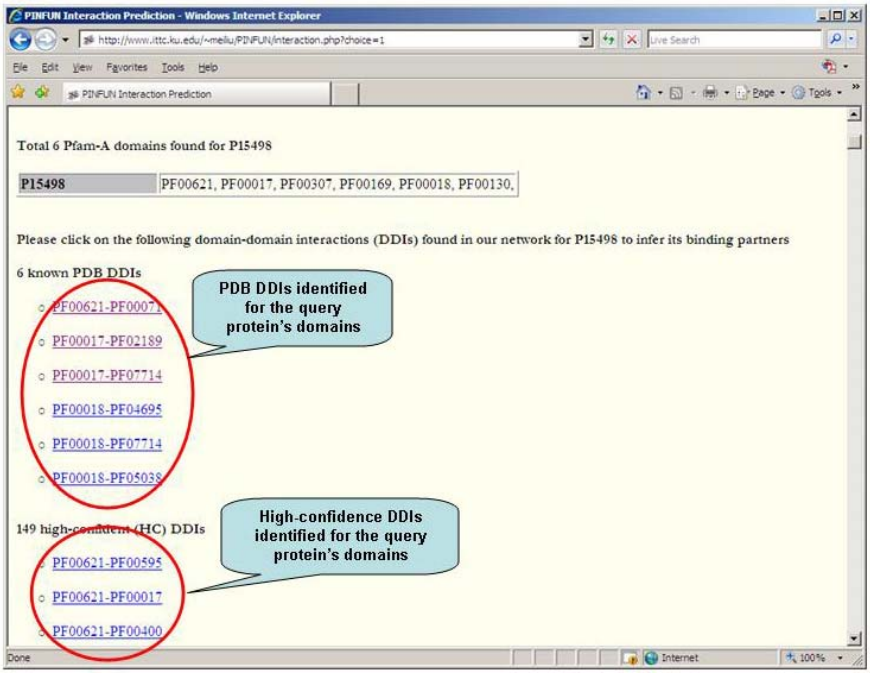

Figure 5.7 PPI prediction option #1 results – DDIs identified for the query protein's domains

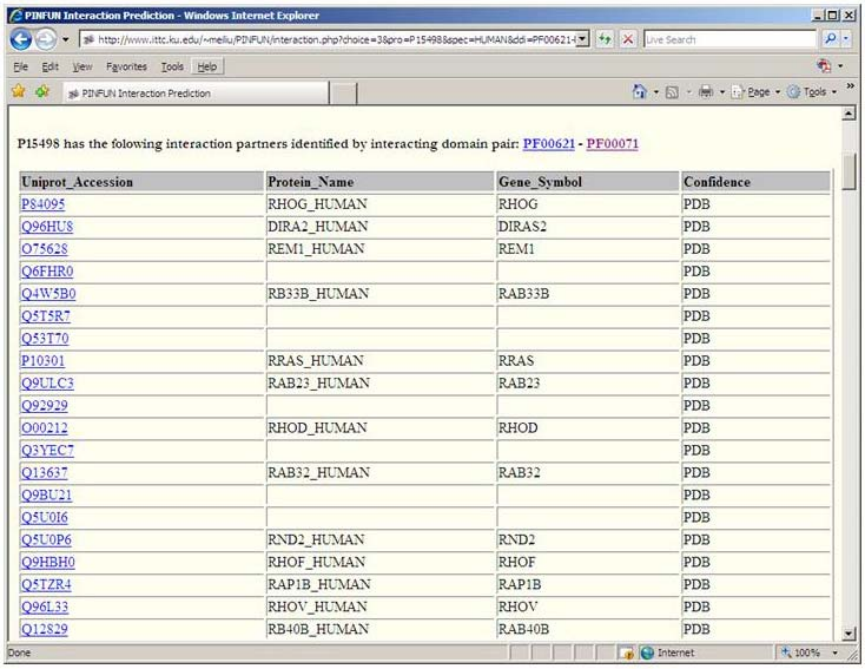

Figure 5.8 PINFUN PPI prediction option #1 results – partners identified for a specific DDI

A user may simply want to determine whether two proteins interact or not which is referred here as PPI prediction option #2 of PINFUN (Figure 5.9). In this particular case, PINFUN proceeds by generating all possible pairs of domains between the two query proteins and looking up each domain pair in our collected DDI database to find evidences of interaction. If two query proteins are found to be interacting, PINFUN displays the domain pairs that mediate such interaction (Figure 5.10).

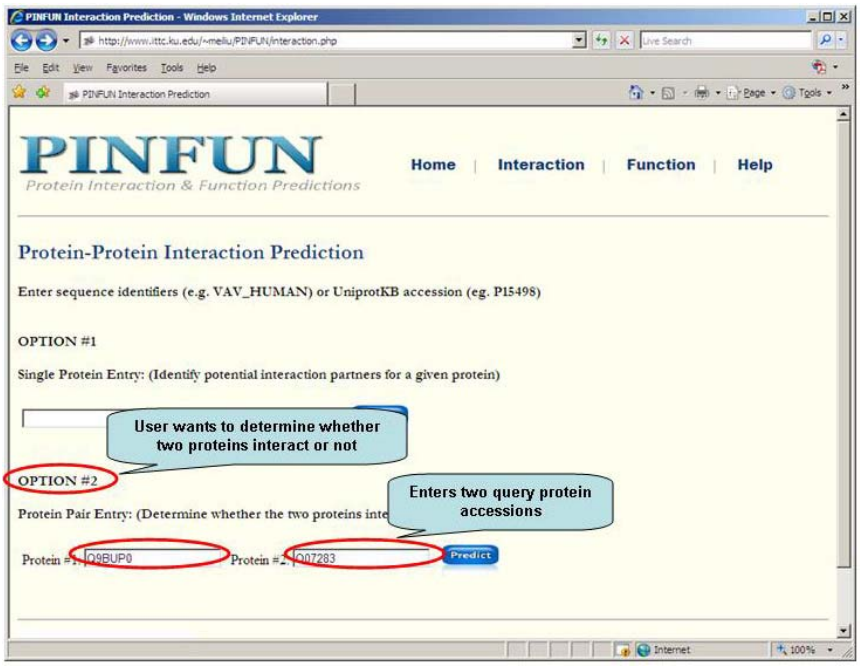

Figure 5.9 PINFUN PPI prediction option #2 query – determines whether two query proteins interact or not.

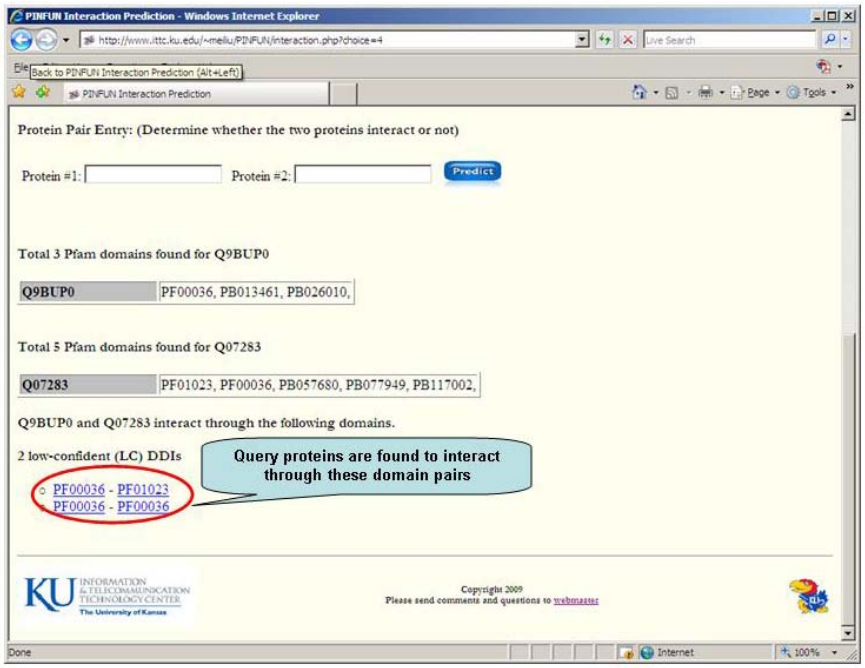

Figure 5.10 PINFUN PPI prediction option #2 results – displays the DDIs that mediate the interaction between the query proteins.

5.3.2 PINFUN Protein Function Prediction

The second goal of PINFUN is to infer GO function annotations for proteins on the basis of their constituent domains and interaction patterns. For protein function prediction, PINFUN also supports two query options. The user can either infer GO functions of a given protein based on the annotations of their constituent domains (Figure 5.11) or infer functions of a protein based on its specific domain interaction patterns with other proteins (Figure 5.13).

The query option #1 is usually pursued when interaction information is unknown and the user wants to simply determine functions based on constituent domain annotations of the query protein (Figure 5.12). Since we already have derived and known domain annotations stored in database, the task can be easily accomplished with the following SQL query:

Retrieve annotations of a Pfam domain 'PF00069'

```
SELECT      func
FROM        Domain_Annot
WHERE       pfam_acc = 'PF00069'
```

The query option #2 is normally recommended when interactions among proteins are already known so that function predictions can be made based on domain interaction patterns between the two query proteins. When a pair of proteins is submitted, PINFUN generates all possible pairs of domains between the two proteins and check each domain pair to see whether they are truly interacting or not according to our DDI database. If at least one domain pair is found in our domain interaction database, associated GO annotations would be assigned to the query protein pair. PINFUN outputs inferred annotation results to a table where columns 1 and 3 are the predicted GO annotations for each of the query proteins (Figure 5.14). Column 5 displays the DDIs from which the GO annotations are derived and

their corresponding confidence level and posterior probability of the domain pair having the predicted annotation pair (i.e. score) are shown in columns 6 and 7, respectively.

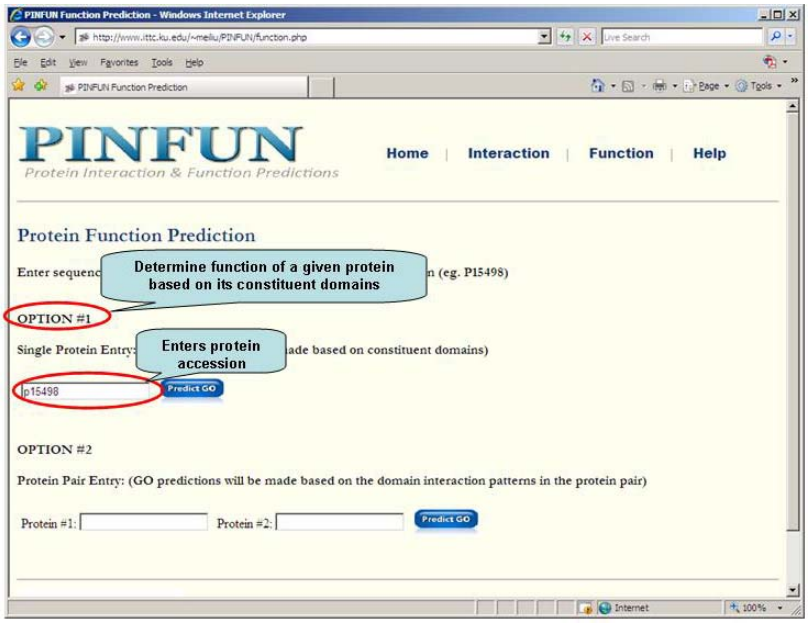

Figure 5.11 PINFUN Protein function prediction option #1 query – infer protein GO functions for a single protein.

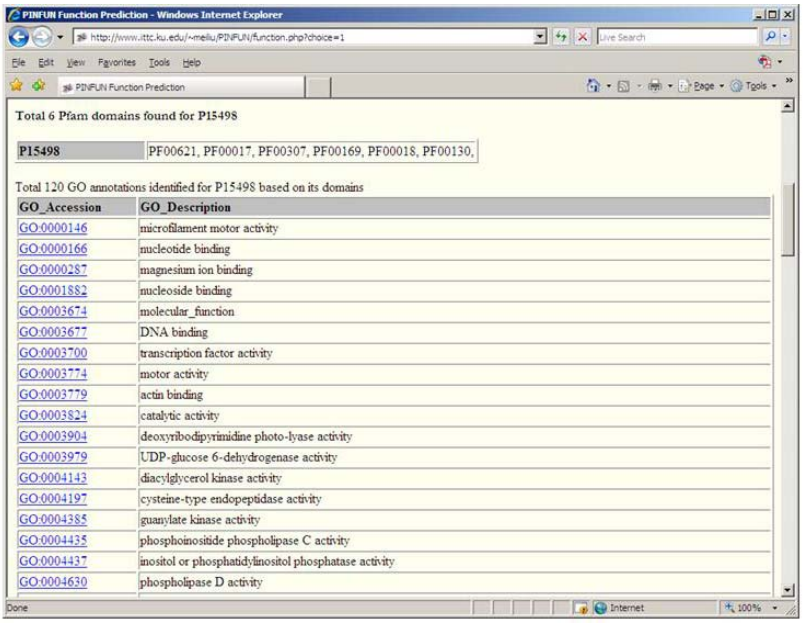

Figure 5.12 PINFUN Protein function prediction option #1 results – GO annotations assigned based on constituent domains of a single query protein.

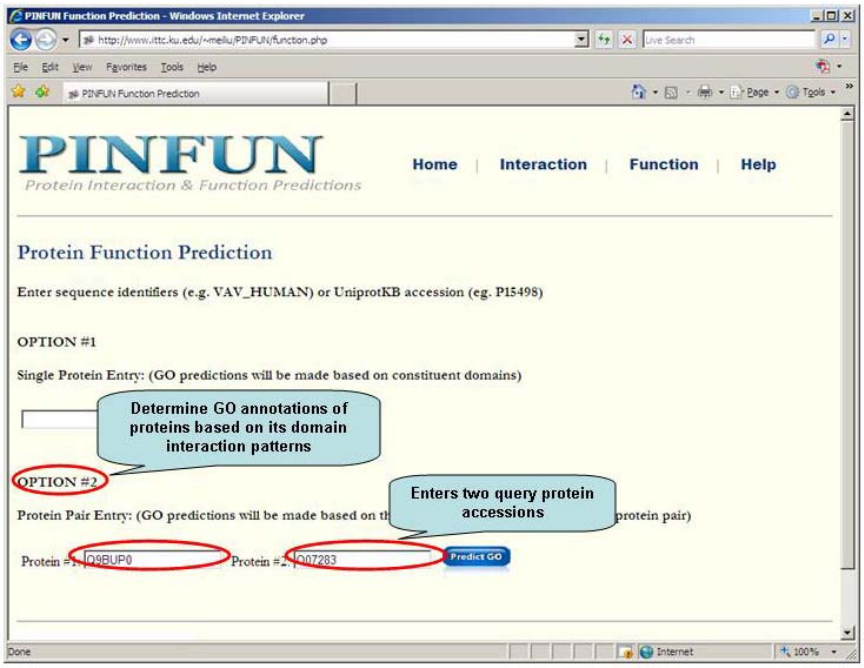

Figure 5.13 PINFUN Protein function prediction option #2 query – infer protein GO functions for a pair of proteins.

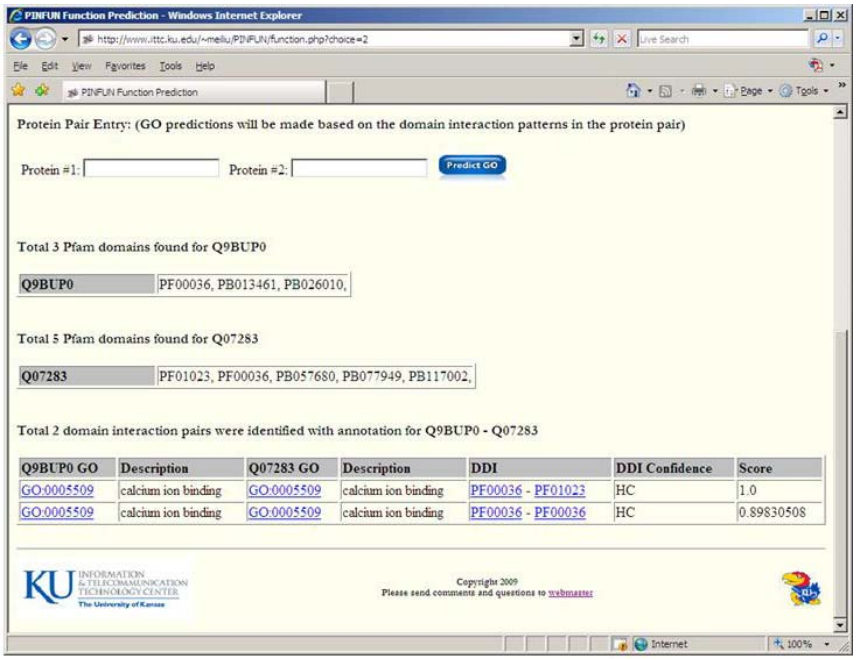

Figure 5.14 PINFUN Protein function prediction option #2 results – GO annotations assigned based on specific domain interaction patterns between the query proteins.

Chapter 6. Conclusion

Proteins play a central role in nearly all cell functions such as composing cellular structure, promoting chemical reactions, carrying messages from one cell to another and acting as antibodies. However, proteins rarely act in isolation. The multiplicity of functions that proteins execute in most cellular processes and biochemical events is attributed to their interactions with other proteins. Thus, to better understand protein functions and the underlying cellular processes, it is essential to understand protein-protein interactions at a genome scale. Insights into protein functions will subsequently help us to gain further understanding of human diseases and may directly contribute to future developments of drug and therapeutic treatments.

6.1 Summary of Research

In this dissertation, I proposed protein domain-based computational approaches to solve the two challenging problems in bioinformatics: protein-protein interaction and protein function predictions. It is believed that proteins interact with each other through specific intermolecular interactions that are localized to specific structural domains within each protein. Determination of the interaction sites is critical as mutations at the sites can disrupt normal interactions or create new undesirable interactions between proteins which may lead to many human diseases. Hence, understanding interactions at the domain level is not only a critical step towards thorough understanding of protein interaction networks and their evolution, but it is also one step closer to acquiring insights into the functions of proteins and causes of human diseases.

First of all, I introduced a new domain-based random decision forest framework (RDFF) for the prediction of protein-protein interactions. The method is particularly useful because

biologically relevant domain–domain interactions can be inferred from the domains involved in existing protein interactions. It allows discovery of interactions between modular domains (i.e. two or more domains forming a unit to participate in interactions). The learned domain interaction patterns can be utilized to reliably determine protein-protein interactions. In terms of RDFF performance in protein interaction prediction, we compared it to the MLE method by Deng et al. [140], and experimental results on *Saccharomyces cerevisiae* dataset demonstrate that our RDFF approach can predict protein–protein interactions with higher specificity (64.38%) and sensitivity (79.78%) than the MLE method.

Secondly, for the protein function annotation problem, I proposed a novel approach called CSIDOP that extracts conserved interacting domain patterns from cross-species protein interaction data. As a result of the CSIDOP method, a weighted web of domain interactions is constructed and statistically analyzed with metrics that combine the static network topology and weights of the underlying interactions. Similar to other biological systems, the domain-domain interaction network built here exhibits scale-free degree distribution. It has been hypothesized that the prevalence of the scale-free property may be the reason that biological systems are more robust toward random perturbations but vulnerable to targeted attacks that may subsequently result in catastrophic system failures [172]. For metrics that consider the interplays between links and their weights, we observe that the unweighted measures and their weighted counterparts generally follow the same trends. The same phenomena were observed by Wuchty [133].

Furthermore, the CSIDOP method was assessed, both biologically and statistically, on the *Homo sapiens* genome for function annotation based on domain patterns extracted from interacting protein pairs in *S. cerevisiae*, *C. elegans*, *D. melanogaster* and *H. sapiens*. Functional assignments were made from a pool of 2,972 unique functional categories. This

number is considerably larger than the number of categories utilized in attempts by other researchers. Using the *H. sapiens* genome, the CSIDOP method accurately assigned functions to 95.42% of the proteins when 2,972 function terms were used, which is highly reliable and is of practical use. The accuracy increased to 98.85% when the number of function terms was decreased to 129. In contrast, with the same testing dataset, the Majority Rule algorithm, the simple domain based method, and orthology based method achieved only 59.50%, 61.98%, and 83.86% in accuracy, respectively. The CSIDOP method can not only provide additional functions to the incomplete GO annotations, but also assign functions for 181 human proteins that currently do not have GO functional terms. Supporting evidences for several of these newly annotated proteins can be found from other data sources or biological experiments, which confirms the utility of this approach.

As more genomes are sequenced, there will be a growing need for better analysis of protein-protein interactions and protein functional annotations of these genomes. In this dissertation, I have shown that *in silico* methods are capable of making reliable large-scale protein-protein interaction and protein function discoveries based on common domain interaction patterns.

6.2 Future Work

Despite the fact that existing protein interaction prediction methods have generated promising results, we are still far away from obtaining complete interactomes especially for those less studied organisms where little information is known. In this case, a cross-organism computational model for PPI prediction would be attractive and crucial so that we can infer interactions for proteins in target organisms using known features from the well studied model organisms. Moreover, I only explored domain-domain interactions to explain protein-

protein interactions in this dissertation. In fact, integration of multiple data sources in protein interaction prediction has recently attracted a lot of attention. The main idea is that different data sources may cover different part of the interactome; therefore, integrating various data sources may increase both prediction accuracy and coverage. There is still a great need in developing more efficient algorithms for data integration to predict protein interactions. For future studies, I would like to elaborate my protein interaction research by investigating different algorithms in cross-species heterogeneous data integration. My goal is to assess the existing PPIs and roles that various data features play in protein interactions. Moreover, protein interaction prediction is a typical imbalanced data classification problem. There exists much greater number of protein pairs that do not interact than the ones that do interact. Apparently, failing to identify one of the few true interacting protein pairs is much worse than inaccurately classifying one of the many non-interacting pairs as interacting. Thus, it is another important topic to address in my future research.

For protein function prediction, the CSIDOP method is shown to be reliable; however, it has its shortcomings. It is limited in predicting functions of proteins with *a priori* knowledge of their interactions. Nevertheless, CSIDOP should continue to improve as protein-protein interaction data are increased both in quality and quantity, and it will readily scale to a genome-wide application. Another drawback of CSIDOP is that it cannot make predictions if domain patterns of a protein pair are not found among the list of derived domain interaction patterns. As part of my future research, I would like to address the function prediction coverage problem by investigating similar match search for domain interaction patterns rather than exact match. For instance, multiple protein domains may be similar in terms of sequence, structure, or the function they perform and thereby should be grouped together. Then when we perform domain pattern search later, we are not looking for exact pattern matches

anymore, but for the group it most likely belongs to. This is a relatively new research field which would require considerable investigations.

Finally, there are several improvements can be made to the PINFUN system in the future. The current version has the following limitations:

1. Since the Pfam database is used to retrieve domain information for query proteins, if the query protein is not currently annotated in Pfam, PINFUN would not be able to make any predictions. This can be resolved in the future by adding sequence alignment capabilities for domain discovery.
2. Predictions are currently restricted to the four species: *S. cerevisiae*, *C. elegans*, *D. melanogaster*, and *H. sapiens*. This definitely can be expanded in the future.
3. For protein interaction prediction, only interactions between Pfam-A domains are considered to ensure prediction quality.
4. For protein function predictions, the derived domain annotations are restricted to GO 'molecular function' term. Only the downloaded known annotations of domains from Pfam include GO 'biological process' and 'cellular component' terms. The problem can be solved by retraining the CSIDOP model with 'biological process' and 'cellular component' terms.

References

1. Noble, D., *The music of life : biology beyond the genome*. 2006, Oxford ; New York: Oxford University Press. xiii, 153 p.
2. Carlson, J.M. and Doyle, J., *Complexity and robustness.* Proc Natl Acad Sci U S A, 2002. **99 Suppl 1**: p. 2538-45.
3. Auffray, C., Imbeaud, S., Roux-Rouquie, M., and Hood, L., *From functional genomics to systems biology: concepts and practices.* C R Biol, 2003. **326**(10-11): p. 879-92.
4. Aggarwal, K. and Lee, K.H., *Functional genomics and proteomics as a foundation for systems biology.* Brief Funct Genomic Proteomic, 2003. **2**(3): p. 175-84.
5. Kitano, H., *Computational systems biology.* Nature, 2002. **420**(6912): p. 206-10.
6. Rousseau, F. and Schymkowitz, J., *A systems biology perspective on protein structural dynamics and signal transduction.* Curr Opin Struct Biol, 2005. **15**(1): p. 23-30.
7. Aloy, P. and Russell, R.B., *Structural systems biology: modelling protein interactions.* Nat Rev Mol Cell Biol, 2006. **7**(3): p. 188-97.
8. *Finishing the euchromatic sequence of the human genome.* Nature, 2004. **431**(7011): p. 931-45.
9. Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., et al., *A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.* Nature, 2000. **403**(6770): p. 623-7.
10. Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., et al., *Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid*

interactions in all possible combinations between the yeast proteins. Proc Natl Acad Sci U S A, 2000. **97**(3): p. 1143-7.

11. Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., et al., *Functional organization of the yeast proteome by systematic analysis of protein complexes.* Nature, 2002. **415**(6868): p. 141-7.
12. Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., et al., *Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry.* Nature, 2002. **415**(6868): p. 180-3.
13. Grigoriev, A., *On the number of protein-protein interactions in the yeast proteome.* Nucleic Acids Res, 2003. **31**(14): p. 4157-61.
14. Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A., et al., *BioGRID: a general repository for interaction datasets.* Nucleic Acids Res, 2006. **34**(Database issue): p. D535-9.
15. Peri, S., Navarro, J.D., Amanchy, R., Kristiansen, T.Z., Jonnalagadda, C.K., et al., *Development of human protein reference database as an initial platform for approaching systems biology in humans.* Genome Res, 2003. **13**(10): p. 2363-71.
16. Crosby, M.A., Goodman, J.L., Strelets, V.B., Zhang, P., and Gelbart, W.M., *FlyBase: genomes by the dozen.* Nucleic Acids Res, 2007. **35**(Database issue): p. D486-91.
17. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
18. Kini, R.M. and Evans, H.J., *Prediction of potential protein-protein interaction sites from amino acid sequence. Identification of a fibrin polymerization site.* FEBS Lett, 1996. **385**(1-2): p. 81-6.

19. Jones, S. and Thornton, J.M., *Prediction of protein-protein interaction sites using patch analysis.* J Mol Biol, 1997. **272**(1): p. 133-43.

20. Jones, S. and Thornton, J.M., *Analysis of protein-protein interaction sites using surface patches.* J Mol Biol, 1997. **272**(1): p. 121-32.

21. Jansen, R., Greenbaum, D., and Gerstein, M., *Relating whole-genome expression data with protein-protein interactions.* Genome Res, 2002. **12**(1): p. 37-46.

22. Deane, C.M., Salwinski, L., Xenarios, I., and Eisenberg, D., *Protein interactions: two methods for assessment of the reliability of high throughput observations.* Mol Cell Proteomics, 2002. **1**(5): p. 349-56.

23. Kim, S.K., Lund, J., Kiraly, M., Duke, K., Jiang, M., et al., *A gene expression map for Caenorhabditis elegans.* Science, 2001. **293**(5537): p. 2087-92.

24. Troyanskaya, O.G., Garber, M.E., Brown, P.O., Botstein, D., and Altman, R.B., *Nonparametric methods for identifying differentially expressed genes in microarray data.* Bioinformatics, 2002. **18**(11): p. 1454-61.

25. Bhardwaj, N. and Lu, H., *Correlation between gene expression profiles and protein-protein interactions within and across genomes.* Bioinformatics, 2005. **21**(11): p. 2730-8.

26. Tornow, S. and Mewes, H.W., *Functional modules by relating protein interaction networks and gene expression.* Nucleic Acids Res, 2003. **31**(21): p. 6283-9.

27. Teichmann, S.A. and Babu, M.M., *Conservation of gene co-regulation in prokaryotes and eukaryotes.* Trends Biotechnol, 2002. **20**(10): p. 407-10; discussion 410.

28. Grigoriev, A., *A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae.* Nucleic Acids Res, 2001. **29**(17): p. 3513-9.

29. Mrowka, R., Patzak, A., and Herzel, H., *Is there a bias in proteome research?* Genome Res, 2001. **11**(12): p. 1971-3.

30. Stuart, J.M., Segal, E., Koller, D., and Kim, S.K., *A gene-coexpression network for global discovery of conserved genetic modules.* Science, 2003. **302**(5643): p. 249-55.

31. Fraser, H.B., Hirsh, A.E., Wall, D.P., and Eisen, M.B., *Coevolution of gene expression among interacting proteins.* Proc Natl Acad Sci U S A, 2004. **101**(24): p. 9033-8.

32. Bowers, P.M., Pellegrini, M., Thompson, M.J., Fierro, J., Yeates, T.O., et al., *Prolinks: a database of protein functional linkages derived from coevolution.* Genome Biol, 2004. **5**(5): p. R35.

33. Ermolaeva, M.D., White, O., and Salzberg, S.L., *Prediction of operons in microbial genomes.* Nucleic Acids Res, 2001. **29**(5): p. 1216-21.

34. Moreno-Hagelsieb, G. and Collado-Vides, J., *A powerful non-homology method for the prediction of operons in prokaryotes.* Bioinformatics, 2002. **18 Suppl 1**: p. S329-36.

35. Strong, M., Mallick, P., Pellegrini, M., Thompson, M.J., and Eisenberg, D., *Inference of protein function and protein linkages in Mycobacterium tuberculosis based on prokaryotic genome organization: a combined computational approach.* Genome Biol, 2003. **4**(9): p. R59.

36. Salgado, H., Moreno-Hagelsieb, G., Smith, T.F., and Collado-Vides, J., *Operons in Escherichia coli: genomic analyses and predictions.* Proc Natl Acad Sci U S A, 2000. **97**(12): p. 6652-7.

37. Dandekar, T., Snel, B., Huynen, M., and Bork, P., *Conservation of gene order: a fingerprint of proteins that physically interact.* Trends Biochem Sci, 1998. **23**(9): p. 324-8.

38. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., and Maltsev, N., *The use of gene clusters to infer functional coupling.* Proc Natl Acad Sci U S A, 1999. **96**(6): p. 2896-901.

39. Galperin, M.Y. and Koonin, E.V., *Who's your neighbor? New computational approaches for functional genomics.* Nat Biotechnol, 2000. **18**(6): p. 609-13.

40. Rogozin, I.B., Makarova, K.S., Murvai, J., Czabarka, E., Wolf, Y.I., et al., *Connected gene neighborhoods in prokaryotic genomes.* Nucleic Acids Res, 2002. **30**(10): p. 2212-23.

41. Huynen, M., Snel, B., Lathe, W., 3rd, and Bork, P., *Predicting protein function by genomic context: quantitative evaluation and qualitative inferences.* Genome Res, 2000. **10**(8): p. 1204-10.

42. Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O., et al., *Detecting protein function and protein-protein interactions from genome sequences.* Science, 1999. **285**(5428): p. 751-3.

43. Enright, A.J., Iliopoulos, I., Kyrpides, N.C., and Ouzounis, C.A., *Protein interaction maps for complete genomes based on gene fusion events.* Nature, 1999. **402**(6757): p. 86-90.

44. Marcotte, C.J. and Marcotte, E.M., *Predicting functional linkages from gene fusions with confidence.* Appl Bioinformatics, 2002. **1**(2): p. 93-100.

45. Yanai, I., Derti, A., and DeLisi, C., *Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes.* Proc Natl Acad Sci U S A, 2001. **98**(14): p. 7940-5.

46. Chia, J.M. and Kolatkar, P.R., *Implications for domain fusion protein-protein interactions based on structural information.* BMC Bioinformatics, 2004. **5**: p. 161.

47. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., and Yeates, T.O., *Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.* Proc Natl Acad Sci U S A, 1999. **96**(8): p. 4285-8.

48. Eisenberg, D., Marcotte, E.M., Xenarios, I., and Yeates, T.O., *Protein function in the post-genomic era.* Nature, 2000. **405**(6788): p. 823-6.

49. Date, S.V. and Marcotte, E.M., *Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages.* Nat Biotechnol, 2003. **21**(9): p. 1055-62.

50. Sun, J., Xu, J., Liu, Z., Liu, Q., Zhao, A., et al., *Refined phylogenetic profiles method for predicting protein-protein interactions.* Bioinformatics, 2005. **21**(16): p. 3409-15.

51. Snitkin, E.S., Gustafson, A.M., Mellor, J., Wu, J., and DeLisi, C., *Comparative assessment of performance and genome dependence among phylogenetic profiling methods.* BMC Bioinformatics, 2006. **7**: p. 420.

52. Bowers, P.M., O'Connor, B.D., Cokus, S.J., Sprinzak, E., Yeates, T.O., et al., *Utilizing logical relationships in genomic data to decipher cellular processes.* FEBS J, 2005. **272**(20): p. 5110-8.

53. Bowers, P.M., Cokus, S.J., Eisenberg, D., and Yeates, T.O., *Use of logic relationships to decipher protein network organization.* Science, 2004. **306**(5705): p. 2246-9.

54. Barker, D. and Pagel, M., *Predicting functional gene links from phylogenetic-statistical analyses of whole genomes.* PLoS Comput Biol, 2005. **1**(1): p. e3.

55. Pazos, F., Helmer-Citterich, M., Ausiello, G., and Valencia, A., *Correlated mutations contain information about protein-protein interaction.* J Mol Biol, 1997. **271**(4): p. 511-23.

56. Pazos, F. and Valencia, A., *In silico two-hybrid system for the selection of physically interacting protein pairs.* Proteins, 2002. **47**(2): p. 219-27.

57. Goh, C.S., Bogan, A.A., Joachimiak, M., Walther, D., and Cohen, F.E., *Co-evolution of proteins with their interaction partners.* J Mol Biol, 2000. **299**(2): p. 283-93.

58. Walhout, A.J., Sordella, R., Lu, X., Hartley, J.L., Temple, G.F., et al., *Protein interaction mapping in C. elegans using proteins involved in vulval development.* Science, 2000. **287**(5450): p. 116-22.

59. Pazos, F. and Valencia, A., *Similarity of phylogenetic trees as indicator of protein-protein interaction.* Protein Eng, 2001. **14**(9): p. 609-14.

60. Ramani, A.K. and Marcotte, E.M., *Exploiting the co-evolution of interacting proteins to discover interaction specificity.* J Mol Biol, 2003. **327**(1): p. 273-84.

61. Gertz, J., Elfond, G., Shustrova, A., Weisinger, M., Pellegrini, M., et al., *Inferring protein interactions from phylogenetic distance matrices.* Bioinformatics, 2003. **19**(16): p. 2039-45.

62. Jothi, R., Kann, M.G., and Przytycka, T.M., *Predicting protein-protein interaction by searching evolutionary tree automorphism space.* Bioinformatics, 2005. **21 Suppl 1**: p. i241-50.

63. Pazos, F., Ranea, J.A., Juan, D., and Sternberg, M.J., *Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome.* J Mol Biol, 2005. **352**(4): p. 1002-15.

64. Sato, T., Yamanishi, Y., Kanehisa, M., and Toh, H., *The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships.* Bioinformatics, 2005. **21**(17): p. 3482-9.

65. Goh, C.S. and Cohen, F.E., *Co-evolutionary analysis reveals insights into protein-protein interactions.* J Mol Biol, 2002. **324**(1): p. 177-92.

66. Lu, L., Lu, H., and Skolnick, J., *MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading.* Proteins, 2002. **49**(3): p. 350-64.

67. Lu, L., Arakaki, A.K., Lu, H., and Skolnick, J., *Multimeric threading-based prediction of protein-protein interactions on a genomic scale: application to the Saccharomyces cerevisiae proteome.* Genome Res, 2003. **13**(6A): p. 1146-54.

68. Aloy, P. and Russell, R.B., *InterPreTS: protein interaction prediction through tertiary structure.* Bioinformatics, 2003. **19**(1): p. 161-2.

69. Aytuna, A.S., Gursoy, A., and Keskin, O., *Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces.* Bioinformatics, 2005. **21**(12): p. 2850-5.

70. Matthews, L.R., Vaglio, P., Reboul, J., Ge, H., Davis, B.P., et al., *Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs".* Genome Res, 2001. **11**(12): p. 2120-6.

71. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.

72. Bu, D., Zhao, Y., Cai, L., Xue, H., Zhu, X., et al., *Topological structure analysis of the protein-protein interaction network in budding yeast.* Nucleic Acids Res, 2003. **31**(9): p. 2443-50.

73. Albert, I. and Albert, R., *Conserved network motifs allow protein-protein interaction prediction.* Bioinformatics, 2004. **20**(18): p. 3346-52.

74. Sharan, R., Suthram, S., Kelley, R.M., Kuhn, T., McCuine, S., et al., *Conserved patterns of protein interaction in multiple species.* Proc Natl Acad Sci U S A, 2005. **102**(6): p. 1974-9.

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

76. Bock, J.R. and Gough, D.A., *Predicting protein--protein interactions from primary structure.* Bioinformatics, 2001. **17**(5): p. 455-60.

77. Martin, S., Roe, D., and Faulon, J.L., *Predicting protein-protein interactions using signature products.* Bioinformatics, 2005. **21**(2): p. 218-26.

78. von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., et al., *Comparative assessment of large-scale data sets of protein-protein interactions.* Nature, 2002. **417**(6887): p. 399-403.

79. Tong, A.H., Drees, B., Nardelli, G., Bader, G.D., Brannetti, B., et al., *A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules.* Science, 2002. **295**(5553): p. 321-4.

80. Oyama, T., Kitano, K., Satou, K., and Ito, T., *Extraction of knowledge on protein-protein interaction by association rule discovery.* Bioinformatics, 2002. **18**(5): p. 705-14.

81. Jansen, R., Lan, N., Qian, J., and Gerstein, M., *Integration of genomic datasets to predict protein complexes in yeast.* J Struct Funct Genomics, 2002. **2**(2): p. 71-81.

82. Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., et al., *A Bayesian networks approach for predicting protein-protein interactions from genomic data.* Science, 2003. **302**(5644): p. 449-53.

83. Lee, I., Date, S.V., Adai, A.T., and Marcotte, E.M., *A probabilistic functional network of yeast genes.* Science, 2004. **306**(5701): p. 1555-8.

84. Huttenhower, C., Hibbs, M., Myers, C., and Troyanskaya, O.G., *A scalable method for integration and functional analysis of multiple microarray datasets.* Bioinformatics, 2006. **22**(23): p. 2890-7.

85. Myers, C.L. and Troyanskaya, O.G., *Context-sensitive data integration and prediction of biological networks.* Bioinformatics, 2007. **23**(17): p. 2322-30.

86. Singh, R., Xu, J., and Berger, B., *Struct2net: integrating structure into protein-protein interaction prediction.* Pac Symp Biocomput, 2006: p. 403-14.

87. Gilchrist, M.A., Salter, L.A., and Wagner, A., *A statistical framework for combining and interpreting proteomic datasets.* Bioinformatics, 2004. **20**(5): p. 689-700.

88. Yamanishi, Y., Vert, J.P., and Kanehisa, M., *Protein network inference from multiple genomic data: a supervised approach.* Bioinformatics, 2004. **20 Suppl 1**: p. i363-70.

89. Ben-Hur, A. and Noble, W.S., *Kernel methods for predicting protein-protein interactions.* Bioinformatics, 2005. **21 Suppl 1**: p. i38-46.

90. Zhang, L.V., Wong, S.L., King, O.D., and Roth, F.P., *Predicting co-complexed protein pairs using genomic and proteomic data integration.* BMC Bioinformatics, 2004. **5**: p. 38.

91. Wong, S.L., Zhang, L.V., Tong, A.H., Li, Z., Goldberg, D.S., et al., *Combining biological networks to predict genetic interactions.* Proc Natl Acad Sci U S A, 2004. **101**(44): p. 15682-7.

92. Bader, J.S., Chaudhuri, A., Rothberg, J.M., and Chant, J., *Gaining confidence in high-throughput protein interaction networks.* Nat Biotechnol, 2004. **22**(1): p. 78-85.

93. Jaimovich, A., Elidan, G., Margalit, H., and Friedman, N., *Towards an integrated protein-protein interaction network: a relational Markov network approach.* J Comput Biol, 2006. **13**(2): p. 145-64.

94. Qi, Y., Klein-Seetharaman, J., and Bar-Joseph, Z., *Random forest similarity for protein-protein interaction prediction from multiple sources.* Pac Symp Biocomput, 2005: p. 531-42.

95. Rhodes, D.R., Tomlins, S.A., Varambally, S., Mahavisno, V., Barrette, T., et al., *Probabilistic model of the human protein-protein interaction network.* Nat Biotechnol, 2005. **23**(8): p. 951-9.

96. Zhong, W. and Sternberg, P.W., *Genome-wide prediction of C. elegans genetic interactions.* Science, 2006. **311**(5766): p. 1481-4.

97. Pearson, W.R. and Lipman, D.J., *Improved tools for biological sequence comparison.* Proc Natl Acad Sci U S A, 1988. **85**(8): p. 2444-8.

98. Brown, M.P., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., et al., *Knowledge-based analysis of microarray gene expression data by using support vector machines.* Proc Natl Acad Sci U S A, 2000. **97**(1): p. 262-7.

99. Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O., and Eisenberg, D., *A combined algorithm for genome-wide prediction of protein function.* Nature, 1999. **402**(6757): p. 83-6.

100. Troyanskaya, O.G., Dolinski, K., Owen, A.B., Altman, R.B., and Botstein, D., *A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae).* Proc Natl Acad Sci U S A, 2003. **100**(14): p. 8348-53.

101. Chen, Y. and Xu, D., *Global protein function annotation through mining genome-scale data in yeast Saccharomyces cerevisiae.* Nucleic Acids Res, 2004. **32**(21): p. 6414-24.

102. Nariai, N., Kolaczyk, E.D., and Kasif, S., *Probabilistic protein function prediction from heterogeneous genome-wide data.* PLoS ONE, 2007. **2**(3): p. e337.

103. Schwikowski, B., Uetz, P., and Fields, S., *A network of protein-protein interactions in yeast.* Nat Biotechnol, 2000. **18**(12): p. 1257-61.

104. Hishigaki, H., Nakai, K., Ono, T., Tanigami, A., and Takagi, T., *Assessment of prediction accuracy of protein function from protein--protein interaction data.* Yeast, 2001. **18**(6): p. 523-31.

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

106. Letovsky, S. and Kasif, S., *Predicting protein function from protein/protein interaction data: a probabilistic approach.* Bioinformatics, 2003. **19 Suppl 1**: p. i197-204.

107. Deng, M., Tu, Z., Sun, F., and Chen, T., *Mapping Gene Ontology to proteins based on protein-protein interaction data.* Bioinformatics, 2004. **20**(6): p. 895-902.

108. Vazquez, A., Flammini, A., Maritan, A., and Vespignani, A., *Global protein function prediction from protein-protein interaction networks.* Nat Biotechnol, 2003. **21**(6): p. 697-700.

109. McDermott, J., Bumgarner, R., and Samudrala, R., *Functional annotation from predicted protein interaction networks.* Bioinformatics, 2005. **21**(15): p. 3217-26.

110. Brun, C., Chevenet, F., Martin, D., Wojcik, J., Guenoche, A., et al., *Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network.* Genome Biol, 2003. **5**(1): p. R6.

111. Bader, G.D. and Hogue, C.W., *An automated method for finding molecular complexes in large protein interaction networks.* BMC Bioinformatics, 2003. **4**: p. 2.

112. Spirin, V. and Mirny, L.A., *Protein complexes and functional modules in molecular networks.* Proc Natl Acad Sci U S A, 2003. **100**(21): p. 12123-8.

113. Pereira-Leal, J.B., Enright, A.J., and Ouzounis, C.A., *Detection of functional modules from protein interaction networks.* Proteins, 2004. **54**(1): p. 49-57.

114. Kann, M.G., *Protein interactions and disease: computational approaches to uncover the etiology of diseases.* Brief Bioinform, 2007. **8**(5): p. 333-46.

115. Mewes, H.W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., et al., *MIPS: a database for genomes and protein sequences.* Nucleic Acids Res, 2002. **30**(1): p. 31-4.

116. Chothia, C., Gough, J., Vogel, C., and Teichmann, S.A., *Evolution of the protein repertoire.* Science, 2003. **300**(5626): p. 1701-3.

117. Chothia, C., *Proteins. One thousand families for the molecular biologist.* Nature, 1992. **357**(6379): p. 543-4.

118. Wolf, Y.I., Grishin, N.V., and Koonin, E.V., *Estimating the number of protein folds and families from complete genome data.* J Mol Biol, 2000. **299**(4): p. 897-905.

119. Apic, G., Gough, J., and Teichmann, S.A., *An insight into domain combinations.* Bioinformatics, 2001. **17 Suppl 1**: p. S83-9.

120. Koonin, E.V., Wolf, Y.I., and Karev, G.P., *The structure of the protein universe and genome evolution.* Nature, 2002. **420**(6912): p. 218-23.

121. Vogel, C., Bashton, M., Kerrison, N.D., Chothia, C., and Teichmann, S.A., *Structure, function and evolution of multidomain proteins.* Curr Opin Struct Biol, 2004. **14**(2): p. 208-16.

122. Ye, Y. and Godzik, A., *Comparative analysis of protein domain organization.* Genome Res, 2004. **14**(3): p. 343-53.

123. Koonin, E.V., Aravind, L., and Kondrashov, A.S., *The impact of comparative genomics on our understanding of evolution.* Cell, 2000. **101**(6): p. 573-6.

124. Teichmann, S.A., Park, J., and Chothia, C., *Structural assignments to the Mycoplasma genitalium proteins show extensive gene duplications and domain rearrangements.* Proc Natl Acad Sci U S A, 1998. **95**(25): p. 14658-63.

125. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.

126. Reiss, D.J. and Schwikowski, B., *Predicting protein-peptide interactions via a network-based motif sampler.* Bioinformatics, 2004. **20 Suppl 1**: p. i274-82.

127. Lehrach, W.P., Husmeier, D., and Williams, C.K., *A regularized discriminative model for the prediction of protein-peptide interactions.* Bioinformatics, 2006. **22**(5): p. 532-40.

128. Pagel, P., Wong, P., and Frishman, D., *A domain interaction map based on phylogenetic profiling.* J Mol Biol, 2004. **344**(5): p. 1331-46.

129. Jothi, R., Cherukuri, P.F., Tasneem, A., and Przytycka, T.M., *Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-domain interactions mediating protein-protein interactions.* J Mol Biol, 2006. **362**(4): p. 861-75.

130. Kann, M.G., Jothi, R., Cherukuri, P.F., and Przytycka, T.M., *Predicting protein domain interactions from coevolution of conserved regions.* Proteins, 2007. **67**(4): p. 811-20.

131. Wuchty, S., *Interaction and domain networks of yeast.* Proteomics, 2002. **2**(12): p. 1715-23.

132. Wuchty, S. and Almaas, E., *Evolutionary cores of domain co-occurrence networks.* BMC Evol Biol, 2005. **5**(1): p. 24.

133. Wuchty, S., *Topology and weights in a protein domain interaction network--a novel way to predict protein interactions.* BMC Genomics, 2006. **7**: p. 122.

134. Pasek, S., Bergeron, A., Risler, J.L., Louis, A., Ollivier, E., et al., *Identification of genomic features using microsyntenies of domains: domain teams.* Genome Res, 2005. **15**(6): p. 867-74.

135. Moon, H.S., Bhak, J., Lee, K.H., and Lee, D., *Architecture of basic building blocks in protein and domain structural interaction networks.* Bioinformatics, 2005. **21**(8): p. 1479-86.

136. Lee, H., Deng, M., Sun, F., and Chen, T., *An integrated approach to the prediction of domain-domain interactions.* BMC Bioinformatics, 2006. **7**: p. 269.

137. Ng, S.K., Zhang, Z., and Tan, S.H., *Integrative approach for computationally inferring protein domain interactions.* Bioinformatics, 2003. **19**(8): p. 923-9.

138. Sprinzak, E. and Margalit, H., *Correlated sequence-signatures as markers of protein-protein interaction.* J Mol Biol, 2001. **311**(4): p. 681-92.

139. Kim, W.K., Park, J., and Suh, J.K., *Large scale statistical prediction of protein-protein interaction by potentially interacting domain (PID) pair.* Genome Inform, 2002. **13**: p. 42-50.

140. Deng, M., Mehta, S., Sun, F., and Chen, T., *Inferring domain-domain interactions from protein-protein interactions.* Genome Res, 2002. **12**(10): p. 1540-8.

141. Liu, Y., Liu, N., and Zhao, H., *Inferring protein-protein interactions through high-throughput interaction data from diverse organisms.* Bioinformatics, 2005. **21**(15): p. 3279-85.

142. Gomez, S.M., Lo, S.H., and Rzhetsky, A., *Probabilistic prediction of unknown metabolic and signal-transduction networks.* Genetics, 2001. **159**(3): p. 1291-8.

143. Gomez, S.M. and Rzhetsky, A., *Towards the prediction of complete protein--protein interaction networks.* Pac Symp Biocomput, 2002: p. 413-24.

144. Greenbaum, D., Colangelo, C., Williams, K., and Gerstein, M., *Comparing protein abundance and mRNA expression levels on a genomic scale.* Genome Biol, 2003. **4**(9): p. 117.

145. Nye, T.M., Berzuini, C., Gilks, W.R., Babu, M.M., and Teichmann, S.A., *Statistical analysis of domains in interacting protein pairs.* Bioinformatics, 2005. **21**(7): p. 993-1001.

146. Wojcik, J. and Schachter, V., *Protein-protein interaction map inference using interacting domain profile pairs.* Bioinformatics, 2001. **17 Suppl 1**: p. S296-305.

147. Riley, R., Lee, C., Sabatti, C., and Eisenberg, D., *Inferring protein domain interactions from databases of interacting proteins.* Genome Biol, 2005. **6**(10): p. R89.

148. Wang, H., Segal, E., Ben-Hur, A., Li, Q.R., Vidal, M., et al., *InSite: a computational method for identifying protein-protein interaction binding sites on a proteome-wide scale.* Genome Biol, 2007. **8**(9): p. R192.

149. Huang, C., Morcos, F., Kanaan, S.P., Wuchty, S., Chen, D.Z., et al., *Predicting protein-protein interactions from protein domains using a set cover approach.* IEEE/ACM Trans Comput Biol Bioinform, 2007. **4**(1): p. 78-87.

150. Guimaraes, K.S., Jothi, R., Zotenko, E., and Przytycka, T.M., *Predicting domain-domain interactions using a parsimony approach.* Genome Biol, 2006. **7**(11): p. R104.

151. Guimaraes, K.S. and Przytycka, T.M., *Interrogating domain-domain interactions with parsimony based approaches.* BMC Bioinformatics, 2008. **9**: p. 171.

152. Breiman, L., *Random forests.* Machine Learning, 2001. **45**(1): p. 5-32.

153. Ho, T.K., Hull, J.J., and Srihari, S.N., *Decision Combination in Multiple Classifier Systems.* IEEE Trans. Pattern Anal. and Mach. Intel., 1994. **16**(1): p. 66-75.

154. Ho, T.K., *Random Decision Forests.* Proc. Third Int'l Conf. Document Analysis and Recognition, 1995: p. 278-282.

155. Ho, T.K., *The random subspace method for constructing decision forests.* IEEE Trans. Pattern Anal. and Mach. Intel., 1998. **20**(8): p. 832-844.

156. Quinlan, J.R., *Discovering rules from large collections of examples: a case study*. Expert Systems in the Micro Electronic Age, ed. D. Michie. 1979: Edinburgh University Press. p. 168-201.

157. Quinlan, J.R., *Learning efficient classification procedures and their application to chess end games*. Machine Learning: An Artificial Intelligence Approach, ed. R.S. Michalski, Carbonell, J.G., and Mitchell, T.M. 1983: Morgan Kaufmann, Los Altos. p. 463-482.

158. Xenarios, I., Fernandez, E., Salwinski, L., Duan, X.J., Thompson, M.J., et al., *DIP: The Database of Interacting Proteins: 2001 update.* Nucleic Acids Res, 2001. **29**(1): p. 239-41.

159. Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., et al., *The Database of Interacting Proteins: 2004 update.* Nucleic Acids Res, 2004. **32**(Database issue): p. D449-51.

160. Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., et al., *The Pfam protein families database.* Nucleic Acids Res, 2004. **32**(Database issue): p. D138-41.

161. Finn, R.D., Marshall, M., and Bateman, A., *iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions.* Bioinformatics, 2005. **21**(3): p. 410-2.

162. Ng, S.K., Zhang, Z., Tan, S.H., and Lin, K., *InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes.* Nucleic Acids Res, 2003. **31**(1): p. 251-4.

163. Chen, X.W. and Liu, M., *Prediction of protein-protein interactions using random decision forest framework.* Bioinformatics, 2005. **21**(24): p. 4394-400.

164. Bailis, J.M., Bernard, P., Antonelli, R., Allshire, R.C., and Forsburg, S.L., *Hsk1-Dfp1 is required for heterochromatin-mediated cohesion at centromeres.* Nat Cell Biol, 2003. **5**(12): p. 1111-6.

165. Lee, S.E., Frenz, L.M., Wells, N.J., Johnson, A.L., and Johnston, L.H., *Order of function of the budding-yeast mitotic exit-network proteins Tem1, Cdc15, Mob1, Dbf2, and Cdc5.* Curr Biol, 2001. **11**(10): p. 784-8.

166. Pawson, T. and Nash, P., *Assembly of cell regulatory systems through protein interaction domains.* Science, 2003. **300**(5618): p. 445-52.

167. Pereira-Leal, J.B. and Teichmann, S.A., *Novel specificities emerge by stepwise duplication of functional modules.* Genome Res, 2005. **15**(4): p. 552-9.

168. Li, S., Armstrong, C.M., Bertin, N., Ge, H., Milstein, S., et al., *A map of the interactome network of the metazoan C. elegans.* Science, 2004. **303**(5657): p. 540-3.

169. Lee, M.O., Kim, E.O., Kwon, H.J., Kim, Y.M., Kang, H.J., et al., *Radicicol represses the transcriptional function of the estrogen receptor by suppressing the stabilization of the receptor by heat shock protein 90.* Mol Cell Endocrinol, 2002. **188**(1-2): p. 47-54.

170. Finn, R.D., Tate, J., Mistry, J., Coggill, P.C., Sammut, S.J., et al., *The Pfam protein families database.* Nucleic Acids Res, 2008. **36**(Database issue): p. D281-8.

171. Barabasi, A.L. and Albert, R., *Emergence of scaling in random networks.* Science, 1999. **286**(5439): p. 509-12.

172. Albert, R., Jeong, H., and Barabasi, A.L., *Error and attack tolerance of complex networks.* Nature, 2000. **406**(6794): p. 378-82.

173. Barrat, A., Barthelemy, M., Pastor-Satorras, R., and Vespignani, A., *The architecture of complex weighted networks.* Proc Natl Acad Sci U S A, 2004. **101**(11): p. 3747-52.

174. Albert, R. and Barabasi, A.-L., *Statistical mechanics of complex networks.* Rev. Mod. Phys., 2002. **74**(1): p. 47-97.

175. Newman, M.E., *Assortative mixing in networks.* Phys Rev Lett, 2002. **89**(20): p. 208701.

176. Stein, A., Russell, R.B., and Aloy, P., *3did: interacting protein domains of known three-dimensional structure.* Nucleic Acids Res, 2005. **33**(Database issue): p. D413-7.

177. Itzhaki, Z., Akiva, E., Altuvia, Y., and Margalit, H., *Evolutionary conservation of domain-domain interactions.* Genome Biol, 2006. **7**(12): p. R125.

178. Raghavachari, B., Tasneem, A., Przytycka, T.M., and Jothi, R., *DOMINE: a database of protein domain interactions.* Nucleic Acids Res, 2008. **36**(Database issue): p. D656-61.

179. O'Brien, K.P., Remm, M., and Sonnhammer, E.L., *Inparanoid: a comprehensive database of eukaryotic orthologs.* Nucleic Acids Res, 2005. **33**(Database issue): p. D476-80.

180. Costanzo, M.C., Hogan, J.D., Cusick, M.E., Davis, B.P., Fancher, A.M., et al., *The yeast proteome database (YPD) and Caenorhabditis elegans proteome database (WormPD): comprehensive resources for the organization and comparison of model organism protein information.* Nucleic Acids Res, 2000. **28**(1): p. 73-6.

181. Poller, W., Barth, J., and Voss, B., *Detection of an alteration of the alpha 2-macroglobulin gene in a patient with chronic lung disease and serum alpha 2-macroglobulin deficiency.* Hum Genet, 1989. **83**(1): p. 93-6.

182. Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., et al., *The InterPro database, an integrated documentation resource for protein families, domains and functional sites.* Nucleic Acids Res, 2001. **29**(1): p. 37-40.

183. Holzl, H., Kapelari, B., Kellermann, J., Seemuller, E., Sumegi, M., et al., *The regulatory complex of Drosophila melanogaster 26S proteasomes. Subunit composition and localization of a deubiquitylating enzyme.* J Cell Biol, 2000. **150**(1): p. 119-30.

184. Ellis, P.J., Clemente, E.J., Ball, P., Toure, A., Ferguson, L., et al., *Deletions on mouse Yq lead to upregulation of multiple X- and Y-linked transcripts in spermatids.* Hum Mol Genet, 2005. **14**(18): p. 2705-15.

185. Jacobs, H.W., Knoblich, J.A., and Lehner, C.F., *Drosophila Cyclin B3 is required for female fertility and is dispensable for mitosis like Cyclin B.* Genes Dev, 1998. **12**(23): p. 3741-51.

186. Gallant, P. and Nigg, E.A., *Identification of a novel vertebrate cyclin: cyclin B3 shares properties with both A- and B-type cyclins.* Embo J, 1994. **13**(3): p. 595-605.

187. Tschop, K., Muller, G.A., Grosche, J., and Engeland, K., *Human cyclin B3. mRNA expression during the cell cycle and identification of three novel nonclassical nuclear localization signals.* Febs J, 2006. **273**(8): p. 1681-95.

188. Chen, X.W., Liu, M., and Ward, R., *Protein Function Assignment through Mining Cross-Species Protein-Protein Interactions.* PLoS ONE, 2008. **3**(2): p. e1562.

189. Kraut, R., Menon, K., and Zinn, K., *A gain-of-function screen for genes controlling motor axon guidance and synaptogenesis in Drosophila.* Curr Biol, 2001. **11**(6): p. 417-30.

190. Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C., *SCOP: a structural classification of proteins database for the investigation of sequences and structures.* J Mol Biol, 1995. **247**(4): p. 536-40.

191. Cherry, J.M., Ball, C., Weng, S., Juvik, G., Schmidt, R., et al., *Genetic and physical maps of Saccharomyces cerevisiae.* Nature, 1997. **387**(6632 Suppl): p. 67-73.

192. Pagel, P., Oesterheld, M., Tovstukhina, O., Strack, N., Stumpflen, V., et al., *DIMA 2.0--predicted and known domain interactions.* Nucleic Acids Res, 2008. **36**(Database issue): p. D651-5.

193. *The universal protein resource (UniProt).* Nucleic Acids Res, 2008. **36**(Database issue): p. D190-5.

Appendix A: Distance analysis

Table shows a GO annotation distance analysis of total 329 (i.e. top 10% scoring) domain-domain interactions predicted by CSIDOP. The GO annotations for each domain are obtained from Pfam. The first column has the identified domain interaction pairs where two domains are separated by a semicolon ';'. The second column contains GO annotations for the domain pairs that give rise to the closest distance. It has the following format:

dom1_annot1 ; dom2_annot1 | dom1_annot2 ; dom2_annot2 | …

'NA' means no annotation found for the domains in Pfam. The last column contains the closest distance between annotation terms of the two domains.

Domain Pair	Annotation Pair	Distance
PF00628;PF00105	GO:0008270;GO:0008270	0
PF00628;PF00104	GO:0005515;GO:0003700	4
PF00089;PF00102	GO:0006508;GO:0006470	4
PF00627;PF00017	NA	
PF00010;PF00008	NA	
PF00010;PF00047	NA	
PF00010;PF00069	GO:0030528;GO:0004672	6
PF01833;PF02178	NA	
PF00271;PF00036	GO:0003676;GO:0005509	4
PF02891;PF01017	GO:0008270;GO:0004871|GO:0008270;GO:0003700	7
PF00010;PF00110	GO:0030528;GO:0004871	3
PF00089;PF00373	GO:0004252;GO:0005856|GO:0006508;GO:0005856	11
PF00225;PF03145	GO:0005875;GO:0005634	5
PF01477;PF00038	NA	
PF00010;PF00320	GO:0005634;GO:0005634	0
PF00069;PF03165	GO:0006468;GO:0006355	8
PF00569;PF00134	NA	
PF01044;PF00076	GO:0005198;GO:0003676	3
PF00105;PF00850	NA	
PF00170;PF01749	GO:0005634;GO:0005634	0
PF00096;PF06001	NA	
PF00536;PF08513	NA	
PF00439;PF00856	NA	
PF00554;PF03719	GO:0045449;GO:0006412	3
PF00443;PF00415	NA	
PF00010;PF00665	GO:0030528;GO:0003677|GO:0045449;GO:0015074	4
PF00041;PF07714	NA	
PF00688;PF07714	GO:0040007;GO:0006468	7

PF00069;PF03508	NA	
PF00046;PF01833	NA	
PF04408;PF00520	GO:0004386;GO:0016020	6
PF07145;PF00134	NA	
PF01392;PF00702	NA	
PF00009;PF02984	GO:0005525;GO:0005634	12
PF00178;PF00071	NA	
PF00028;PF00439	NA	
PF00130;PF00130	GO:0007242;GO:0007242	0
PF00433;PF00089	GO:0006468;GO:0006508	4
PF00433;PF00096	GO:0005524;GO:0003676	6
PF00130;PF00089	GO:0007242;GO:0006508	7
PF00028;PF00569	GO:0005509;GO:0008270	3
PF08513;PF05485	NA	
PF00412;PF00047	NA	
PF04433;PF00041	NA	
PF00069;PF03957	NA	
PF00433;PF00282	GO:0006468;GO:0019752	6
PF02198;PF02178	NA	
PF02944;PF01749	GO:0003677;GO:0008565	6
PF00178;PF00443	GO:0003700;GO:0004221	7
PF00439;PF00041	NA	
PF04408;PF00036	GO:0004386;GO:0005509	6
PF02171;PF02170	NA	
PF00069;PF00029	NA	
PF00089;PF07716	GO:0004252;GO:0003700|GO:0006508;GO:0006355	7
PF00069;PF00038	NA	
PF00089;PF07732	GO:0004252;GO:0016491	5
PF00069;PF00060	GO:0004672;GO:0016020	9
PF00069;PF00079	GO:0004672;GO:0004867	10
PF00069;PF00089	GO:0006468;GO:0006508	4
PF00071;PF01085	NA	
PF00071;PF01079	NA	
PF00379;PF02363	NA	
PF00006;PF00076	GO:0016469;GO:0003676	7
PF03166;PF00076	GO:0005622;GO:0003676	6
PF00010;PF07727	NA	
PF00324;PF00128	GO:0016020;GO:0003824|GO:0006810;GO:0005975|GO:0006810;GO:0003824	5
PF00069;PF00229	GO:0006468;GO:0006955	8
PF00178;PF00757	GO:0005634;GO:0016020	5
PF00069;PF00282	GO:0006468;GO:0019752	6
PF00855;PF00105	NA	
PF00855;PF00104	NA	
PF01138;PF00270	GO:0003723;GO:0003676	1
PF00093;PF00514	NA	
PF00102;PF02985	NA	

PF02259;PF03874	NA	
PF00017;PF00612	NA	
PF00514;PF05485	NA	
PF00069;PF00569	GO:0005524;GO:0008270	9
PF00651;PF04062	GO:0005515;GO:0005856	8
PF00856;PF00097	GO:0005634;GO:0005515	8
PF00069;PF00656	GO:0006468;GO:0006508	4
PF00856;PF00104	GO:0005634;GO:0005634	0
PF00856;PF00105	GO:0005634;GO:0005634	0
PF00630;PF00104	NA	
PF00435;PF02984	NA	
PF00688;PF01030	GO:0040007;GO:0016020	5
PF02892;PF00514	NA	
PF02170;PF02170	NA	
PF00096;PF02172	GO:0003676;GO:0005515	2
PF00096;PF02135	GO:0008270;GO:0008270	0
PF02037;PF02864	GO:0003676;GO:0003700	2
PF02037;PF02865	GO:0003676;GO:0003700	2
PF02874;PF00076	GO:0016469;GO:0003676	7
PF01049;PF00569	GO:0005509;GO:0008270	3
PF00178;PF07714	GO:0006355;GO:0006468|GO:0043565;GO:0005524|GO:0003700;GO:0004713|GO:0003700;GO:0005524	8
PF05965;PF00439	NA	
PF01602;PF00861	GO:0030117;GO:0005840	2
PF05033;PF00850	NA	
PF03931;PF04858	GO:0006511;GO:0016481	9
PF03957;PF07714	NA	
PF02260;PF02269	NA	
PF01049;PF00439	NA	
PF00850;PF02826	NA	
PF07714;PF00387	GO:0006468;GO:0006629	6
PF07714;PF00388	GO:0006468;GO:0007165	7
PF00170;PF07714	GO:0046983;GO:0005524	7
PF03920;PF00389	NA	
PF08447;PF00010	NA	
PF00751;PF02874	GO:0005634;GO:0016469	6
PF07714;PF00616	GO:0006468;GO:0051056	8
PF07714;PF00620	GO:0006468;GO:0007165	7
PF07714;PF00621	GO:0004713;GO:0005622|GO:0005524;GO:0005622|GO:0006468;GO:0005622|GO:0006468;GO:0035023	10
PF02196;PF07714	GO:0007165;GO:0006468	7
PF00385;PF00145	GO:0003682;GO:0003677	3
PF03931;PF01133	NA	
PF00357;PF00013	NA	
PF07714;PF00757	GO:0005524;GO:0005524|GO:0006468;GO:0006468	0
PF00305;PF00038	NA	
PF01466;PF04858	GO:0006511;GO:0016481	9

PF02210;PF00569	NA	
PF02210;PF00514	NA	
PF03920;PF00010	NA	
PF00194;PF07565	GO:0006730;GO:0006820	7
PF02210;PF00439	NA	
PF04990;PF00076	GO:0003677;GO:0003676	1
PF00110;PF01079	GO:0007275;GO:0007275	0
PF00110;PF01085	GO:0007275;GO:0007275	0
PF03096;PF07647	NA	
PF02826;PF00439	NA	
PF07717;PF00520	NA	
PF00271;PF02877	GO:0004386;GO:0003950	5
PF01466;PF01133	NA	
PF00494;PF00004	GO:0016740;GO:0005524	8
PF00023;PF00560	NA	
PF07529;PF07714	NA	
PF00170;PF00514	NA	
PF00027;PF01133	NA	
PF08441;PF08441	NA	
PF07717;PF00036	NA	
PF00751;PF05986	GO:0005634;GO:0031012|GO:0003700;GO:0031012	5
PF00271;PF02319	GO:0003676;GO:0003700	2
PF02891;PF02864	GO:0008270;GO:0004871|GO:0008270;GO:0003700	7
PF02891;PF02865	GO:0008270;GO:0004871|GO:0008270;GO:0003700	7
PF00320;PF00008	NA	
PF00178;PF01030	GO:0005634;GO:0016020	5
PF00178;PF01166	GO:0006355;GO:0006355|GO:0003700;GO:0003700	0
PF03957;PF00018	NA	
PF03957;PF00017	NA	
PF00071;PF00443	NA	
PF00076;PF04983	GO:0003676;GO:0003677	1
PF00046;PF00569	GO:0043565;GO:0008270|GO:0003700;GO:0008270	7
PF00046;PF00554	GO:0005634;GO:0005634|GO:0003700;GO:0003700	0
PF00554;PF02178	NA	
PF00400;PF00105	NA	
PF00400;PF00104	NA	
PF07653;PF00612	NA	
PF00170;PF00018	NA	
PF00170;PF00017	GO:0046983;GO:0005515	1
PF00046;PF00439	NA	
PF00178;PF01388	GO:0043565;GO:0003677|GO:0003700;GO:0003677	1
PF00400;PF00010	NA	
PF01812;PF03096	NA	
PF00046;PF00379	GO:0003700;GO:0042302	4
PF01749;PF05485	GO:0008565;GO:0003676	5
PF00071;PF00130	NA	
PF03143;PF00134	NA	
PF02196;PF01079	GO:0007165;GO:0007154	1

PF02196;PF01085	GO:0007165;GO:0007267	2
PF02172;PF00271	GO:0005515;GO:0003676	2
PF00069;PF01079	GO:0006468;GO:0006508	4
PF00069;PF01085	GO:0006468;GO:0007267	7
PF00090;PF00514	NA	
PF02944;PF00514	NA	
PF00536;PF00400	NA	
PF00046;PF00134	NA	
PF00046;PF00110	GO:0003700;GO:0004871	4
PF00098;PF00105	GO:0008270;GO:0008270	0
PF00098;PF00104	GO:0003676;GO:0003700	2
PF07529;PF00628	NA	
PF00046;PF00047	NA	
PF00917;PF00010	NA	
PF00271;PF05406	NA	
PF00389;PF00439	NA	
PF00249;PF07714	NA	
PF04433;PF02319	NA	
PF00130;PF02201	GO:0007242;GO:0005634	10
PF01049;PF06001	NA	
PF07529;PF00041	NA	
PF01839;PF00013	NA	
PF02198;PF00071	NA	
PF02319;PF03719	GO:0005667;GO:0005840|GO:0006355;GO:0006412|GO:0003700;GO:0003735	4
PF03153;PF00046	GO:0003702;GO:0003700	2
PF02735;PF00125	GO:0003677;GO:0003677	0
PF07714;PF01030	GO:0004713;GO:0016020|GO:0005524;GO:0016020|GO:0006468;GO:0016020	10
PF02259;PF02269	NA	
PF01812;PF00022	GO:0005524;GO:0005515	6
PF00097;PF00010	GO:0005515;GO:0030528	3
PF00071;PF03166	NA	
PF00071;PF03165	NA	
PF02197;PF01133	NA	
PF00041;PF02751	NA	
PF00019;PF01030	GO:0008083;GO:0016020	8
PF00373;PF00505	GO:0005856;GO:0003677	9
PF03731;PF00125	NA	
PF00494;PF06068	GO:0016740;GO:0003678	3
PF02210;PF06001	NA	
PF01833;PF00439	NA	
PF00702;PF00046	GO:0003824;GO:0003700	3
PF01833;PF00333	NA	
PF02198;PF00757	GO:0005634;GO:0016020	5
PF02319;PF00333	GO:0003700;GO:0003723	3
PF00041;PF02268	NA	

PF00999;PF00999	GO:0016021;GO:0016021|GO:0006885;GO:0006885|GO:0015299;GO:0015299	0
PF02260;PF03874	NA	
PF00035;PF00036	GO:0003725;GO:0005509	6
PF03144;PF02984	GO:0005525;GO:0005634	12
PF02136;PF00397	GO:0006810;GO:0005515|GO:0005622;GO:0005515	6
PF03920;PF02826	NA	
PF05185;PF07714	GO:0008168;GO:0004713	6
PF00023;PF01462	NA	
PF00023;PF01463	NA	
PF00493;PF00004	GO:0005524;GO:0005524	0
PF00514;PF00778	NA	
PF07533;PF00041	NA	
PF00096;PF04062	GO:0005622;GO:0005856	4
PF01479;PF00076	GO:0003723;GO:0003676	1
PF01833;PF03719	NA	
PF00023;PF01582	NA	
PF00035;PF00520	GO:0005622;GO:0016020	2
PF00028;PF02135	GO:0005509;GO:0008270	3
PF02198;PF07714	GO:0043565;GO:0005524	8
PF00028;PF02172	GO:0005509;GO:0005515	4
PF02518;PF00676	GO:0005524;GO:0008152	9
PF00357;PF01839	NA	
PF04992;PF00076	GO:0003677;GO:0003676	1
PF01105;PF00324	GO:0006810;GO:0006810	0
PF01157;PF02671	GO:0005840;GO:0005634|GO:0005622;GO:0005634|GO:0006412;GO:0006355	4
PF01049;PF02172	GO:0005509;GO:0005515	4
PF01049;PF02135	GO:0005509;GO:0008270	3
PF00022;PF01849	NA	
PF02922;PF00393	GO:0005975;GO:0006098	6
PF00306;PF00076	GO:0016469;GO:0003676	7
PF00178;PF02178	NA	
PF00019;PF07714	GO:0008083;GO:0005524	8
PF07533;PF00628	NA	
PF00514;PF00110	NA	
PF00433;PF02201	GO:0005524;GO:0005634|GO:0004674;GO:0005634|GO:0006468;GO:0005634	12
PF00357;PF08441	NA	
PF00569;PF02984	GO:0008270;GO:0005634	11
PF00069;PF02078	GO:0006468;GO:0007269	7
PF00458;PF00458	GO:0005524;GO:0005524|GO:0006418;GO:0006418|GO:0004812;GO:0004812	0
PF02234;PF00096	GO:0005634;GO:0005622	4
PF02210;PF02172	NA	
PF02210;PF02135	NA	
PF00069;PF02201	GO:0004672;GO:0005634	11

PF02135;PF00271	GO:0003712;GO:0004386|GO:0003712;GO:0003676	4
PF03143;PF02984	GO:0005525;GO:0005634	12
PF00090;PF01749	NA	
PF07145;PF02984	NA	
PF00009;PF00134	NA	
PF00096;PF00769	GO:0005622;GO:0005737	2
PF00019;PF00757	GO:0008083;GO:0016020|GO:0008083;GO:0005524	8
PF02922;PF03446	GO:0005975;GO:0006098	6
PF00656;PF00319	GO:0006508;GO:0006355	7
PF00096;PF00569	GO:0008270;GO:0008270	0
PF07533;PF07714	NA	
PF00069;PF02750	GO:0006468;GO:0007269	7
PF00096;PF00373	GO:0005622;GO:0005856	4
PF00130;PF01085	GO:0007242;GO:0007267	3
PF00130;PF01079	GO:0007242;GO:0007154	2
PF03730;PF00125	GO:0003677;GO:0003677	0
PF02198;PF01030	GO:0005634;GO:0016020	5
PF00751;PF07679	NA	
PF00071;PF02319	NA	
PF02198;PF01166	GO:0043565;GO:0003700	2
PF03167;PF00104	NA	
PF03167;PF00105	NA	
PF06001;PF00271	NA	
PF00046;PF02363	NA	
PF00071;PF02178	NA	
PF00005;PF00096	GO:0005524;GO:0003676	6
PF02198;PF01388	GO:0043565;GO:0003677	1
PF07714;PF02268	GO:0004713;GO:0003702|GO:0005524;GO:0003702	8
PF00028;PF06001	NA	
PF00627;PF07714	NA	
PF00093;PF01749	NA	
PF00769;PF00505	GO:0008092;GO:0003677	4
PF02037;PF01017	GO:0003676;GO:0003700	2
PF00104;PF00076	GO:0003700;GO:0003676	2
PF01161;PF00565	NA	
PF01161;PF00567	NA	
PF00128;PF00393	GO:0003824;GO:0004616	4
PF00520;PF00520	GO:0016020;GO:0016020|GO:0005216;GO:0005216|GO:0006811;GO:0006811	0
PF00595;PF00702	GO:0005515;GO:0003824	3
PF00435;PF00134	NA	
PF05351;PF00071	NA	
PF00400;PF05485	NA	
PF00564;PF02201	NA	
PF07714;PF02751	GO:0004713;GO:0003702|GO:0005524;GO:0003702	8
PF00850;PF00389	NA	
PF00595;PF00388	GO:0005515;GO:0007165	7

PF00595;PF00387	GO:0005515;GO:0007165|GO:0005515;GO:0006629	7
PF00688;PF00069	GO:0040007;GO:0006468	7
PF02892;PF01749	GO:0003677;GO:0008565	6
PF00397;PF00769	GO:0005515;GO:0008092	1
PF00595;PF00122	GO:0005515;GO:0016020|GO:0005515;GO:0016820|GO:0005515;GO:0005524	6
PF00751;PF00306	GO:0005634;GO:0016469	6
PF03144;PF00134	NA	
PF00595;PF00005	GO:0005515;GO:0005524	6
PF00104;PF00850	NA	
PF00789;PF02363	NA	
PF00271;PF00645	GO:0003676;GO:0003677	1
PF00271;PF00644	GO:0004386;GO:0003950	5
PF08441;PF00013	NA	
PF00751;PF00095	GO:0003700;GO:0030414	5
PF00751;PF00090	NA	
PF00751;PF00047	NA	
PF00130;PF07714	GO:0007242;GO:0006468	8
PF00751;PF00014	GO:0003700;GO:0004867	7
PF00751;PF00006	GO:0005634;GO:0016469	6
PF00271;PF00520	GO:0004386;GO:0016020|GO:0003676;GO:0016020	6
PF00397;PF00373	GO:0005515;GO:0005856	8
PF04433;PF07714	NA	
PF00128;PF03446	GO:0003824;GO:0004616	4
PF01193;PF00023	NA	
PF00554;PF00439	NA	
PF00688;PF00757	GO:0040007;GO:0016020	5
PF00439;PF07714	NA	
PF00554;PF00333	GO:0045449;GO:0006412|GO:0003700;GO:0003723	3

Appendix B: Complete description of Pfam-A table

pfamA		
auto_pfamA	INT(10)	PRI
pfamA_acc	VARCHAR(7)	UNI
pfamA_id	VARCHAR(40)	UNI
description	VARCHAR(100)	
model_length	MEDIUMINT(8)	
author	TINYTEXT	
seed_source	TINYTEXT	
alignment_method	TINYTEXT	
type	TINYTEXT	
ls_sequence_GA	DOUBLE(16,4)	
ls_domain_GA	DOUBLE(16,4)	
fs_sequence_GA	DOUBLE(16,4)	
fs_domain_GA	DOUBLE(16,4)	
ls_sequence_TC	DOUBLE(16,4)	
ls_domain_TC	DOUBLE(16,4)	
fs_sequence_TC	DOUBLE(16,4)	
fs_domain_TC	DOUBLE(16,4)	
ls_sequence_NC	DOUBLE(16,4)	
ls_domain_NC	DOUBLE(16,4)	
fs_sequence_NC	DOUBLE(16,4)	
fs_domain_NC	DOUBLE(16,4)	
ls_mu	DOUBLE(16,4)	
ls_kappa	DOUBLE(16,4)	
fs_mu	DOUBLE(16,4)	
fs_kappa	DOUBLE(16.4)	
comment	LONGTEXT	
previous_id	TINYTEXT	
hmmbuild_ls	TINYTEXT	
hmmcalibrate_ls	TINYTEXT	
hmmbuild_fs	TINYTEXT	
hmmcalibrate_fs	TINYTEXT	
num_seed	INT(10)	
num_full	INT(10)	
updated	TIMESTAMP	
created	DATETIME	
Version	SMALLINT(5)	