Towards the integration of structural and systems biology: structure-based studies of protein-protein interactions on a genome-wide scale

Qiangfeng Cliff Zhang

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2011

ABSTRACT

Towards the integration of structural and systems biology: structure-based studies of protein-protein interactions on a genome-wide scale

Qiangfeng Cliff Zhang

Knowledge of protein-protein interactions (PPIs) is essential to understanding regulatory processes in a cell. High-throughput experimental methods have made significant contributions to PPI determination, but they are known to have many false positives and fail to identify a signification portion of *bona fide* interactions. The same is true for the many computational tools that have been developed. Significantly, although protein structures provide atomic details of PPIs, they have had relatively little impact in large-scale PPI predictions and there has been only limited overlap between structural and systems biology. Here in this thesis, I present our progress in combining structural biology and systems biology in the context of studies analyzing, coarse-grained modeling and prediction of protein-protein interactions.

I first report a comprehensive analysis of the degree to which the location of a protein interface is conserved in sets of proteins that share different levels of similarities. Our results show that while, in general, the interface conservation is most significant among close neighbors, it is still significant even for remote

structural neighbors. Based on this finding, we designed PredUs, a method to predict protein interface simply by "mapping" the interface information from its structural neighbors (*i.e.*, "templates") to the target structure. We developed the PredUs web server to predict protein interfaces using this "template-based" method and a support vector machine (SVM) to further improve predictions. The PredUs webserver outperforms other state-of-the-art methods that are typically based on amino acid properties in terms of both prediction precision and recall. Meanwhile, PredUs runs very fast and can be used to study protein interfaces in a high throughput fashion. Maybe more importantly, it is not sensitive to local conformational changes and small errors in structures and thus can be applied to predict interface of protein homology models, when experimental structures are not available.

I then describe a novel structural modeling method that uses geometric relationships between protein structures, including both PDB structures and homology models, to accurately predict PPIs on a genome-wide scale. We applied the method with considerable success to both the yeast and the human genomes. We found that the accuracy and the coverage of our structure-based prediction compare favorably with the methods derived from sequence and functional clues, *e.g.* sequence similarity, co-expression, phylogenetic similarity, *etc*. Results further improve when using a naive Bayesian classifier to combine structural information with non-structural clues (PREPPI), yielding predictions of

comparable quality to high-throughput experiments. Our data further suggests that PREPPI predictions are substantially complementary to those by experimental methods thus providing a way to dissect interactions that would be hard to identify on a purely high-throughput experimental basis.

We have for the first time designed a "template-based" method that predicts protein interface with high precision and recall. We have also for the first time used 3D structure as part of the repertoire of experimental and computational information and find a way to accurately infer PPIs on a large scale. The success of PredUs and PREPPI can be attributed to the exploitation of both the information contained in imperfect models and the remote structure-function relationships between proteins that have been usually considered to be unrelated. Our results constitute a significant paradigm shift in both structural and systems biology and suggest that they can be integrated to an extent that has not been possible in the past.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

FIGURES

TABLES

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, Prof. Barry Honig, for his caring, supporting and mentorship. He showed me what a professional scientist is by being curious and even a little skeptical on all "good" results. He also showed me what a caring advisor is by his patience and encouragement when I was frustrated by some "bad" results. Throughout my Ph.D. studies, Barry gave me keen insights to the whole project, yet at the same time offered me the freedom to explore on my own. He taught me from asking questions, to solving questions, and to answering questions. I am indebted to him more than he knows.

I am also deeply grateful to Profs. Andrea Califano, Dennis Vitkup, Larry Shapiro, Harmen Bussemarker, Richard Mann and Arthur Palmer, for their supporting and participation on my academic development. In particular, Andrea provided many invaluable advices on our project. I am always amazed by how many and how smart ideas he has.

I would like to give special thanks to Donald Petrey. Whenever I have problems, questions, interesting findings, ranging from biology to mathematics, to computer science, and sometimes to movies, Donald always helps. Our work and this dissertation as well, were not possible without him.

I thank all of the Honig lab members, who made scientific research an enjoyable task. In particular, Katie Rosa offered me a lot of help from working to living in New York City. Jiang Zhu, Remo Rohs, Brian Chen gave me valuable comments on my research and advices on my career. I have also benefited a lot from talking with Yinghao Wu, Raquel Norel, Jeremie Vendome, Markus Fischer, Peng Liu, and Lei Deng.

Finally, I would like to thank my parents, my wife, my whole family and friends. To do a Ph.D. is not easy; to do a second one is never easier. I would not have gone this far without your continuous love and support.

DEDICATION

I dedicate the thesis to my wife, Minghui, for her love and support.

CHAPTER 1. INTRODUCTION

1.1 From genomics to functional genomics: technology drives science

Modern biological research is always driven by technology development. The past decades have witnessed how high-throughput genome-wide experimentation, most notably the next generation sequencing studies, have remarkably advanced our understanding of biological systems in many different aspects and at many different levels. Whole genome sequencing projects like the HGP (Human Genome Project, (Lander, Linton et al. 2001; Venter, Adams et al. 2001)) have generated a plethora of DNA sequences for thousands of organisms. Genome annotation projects such as ENCODE (ENCyclopedia Of DNA Elements, (Birney, Stamatoyannopoulos et al. 2007; Myers, Stamatoyannopoulos et al. 2011)) and modENCODE (Model Organism ENCODE, (Gerstein, Lu et al. 2010; Roy, Ernst et al. 2010; Elsner and Mak 2011; Muers 2011)) have been carried out that aim to find all functional elements in genomes using RNA-seq (RNA sequencing, (Mortazavi, Williams et al. 2008; Wang, Gerstein et al. 2009; Haas and Zody 2010)), CHIP-seq (Chromatin Immunoprecipitation sequencing, (Mardis 2007; Kharchenko, Tolstorukov et al. 2008; Park 2009)), and MeDIP-seq (Methylated DNA immunoprecipitation sequencing, (Down, Rakyan et al. 2008)) techniques. And functional genomics, by its broadest definition, promises to

provide a complete picture of how these genetic elements function together to make a living organism. Ultimately, the ability to decipher the relationship between an organism's genome and its phenotype and to manipulate genetic circuits that dictate different cellular and organismal activity and behavior will have important implications for the understanding of genetic diseases and their treatment.

Many systematic or genome-wide studies have been conducted to detect the functions of individual genetic elements and their interactions. Loss-of-function studies, which systematically "knock out" genes one by one using mutagenesis (Brown and Balling 2001; Vidan and Snyder 2001; Bochner 2003) or RNAi (RNA interference, (Hannon 2002; Bartel 2009)) techniques can provide clues to the functions of the lost gene based on resulting phenotypes. More complicatedly, gene function can be investigated in the context of genetic interactions, which represent the degree to which the presence of a mutation in one gene modulates the phenotype of a mutation in a second gene. Systematic and quantitative approaches for measuring genetic interactions, such as SGA (Synthetic Genetic Arrays, (Tong, Evangelista et al. 2001; Tong, Lesage et al. 2004), dSLAM (diploid Synthetic Lethality Analysis by Microarray (Ooi, Shoemaker et al. 2003)), and E-MAP (Epistatic MiniArray Profile, (Collins,

Miller et al. 2007; Roguev, Bandyopadhyay et al. 2008)), are effective tools to study genetic interactions.

Traditionally, we think of phenotypes as observable characteristics or traits such as morphology, development, behavior, or biochemical or physiological properties. However, for many genes/organisms, an obvious phenotype is hard to define, or especially, hard to quantitatively characterize. Nevertheless, a gene is almost always transcribed into RNA molecules. And the abundance of the transcripts of a gene is usually tightly regulated by the interplay of mutations or polymorphisms in its DNA sequence and regulatory RNAs and/or proteins in the same cellular environment. Consequently, the expression levels of a gene could be used to quantitatively define a phenotype. The invention of microarray techniques that can probe the expression landscape of the entire genome and accomplish many genetic tests in parallel has dramatically changed our way to study gene functions (Brown and Botstein 1999; Heller 2002). Procedures to measure and analyze the expression of tens of thousands of genes simultaneously and under hundreds of different environmental conditions have been streamlined and could be conveniently carried out in thousands of laboratories all over the world.

Many genes need to be translated into proteins to carry out their functions. As the workhorses of a cell factory, proteins take part in essentially every

structure and activity of life, by interacting with other proteins, DNA, RNA and small molecule ligands. Much effort has therefore been devoted to experimental determination of protein-protein interactions (PPIs) using both small scale pull-down experiments or high-throughput approaches like yeast-two-hybrid screenings, affinity purifications, and protein-fragment complementation assays (see reviews in (Salwinski and Eisenberg 2003; Shoemaker and Panchenko 2007)), and protein–DNA interactions by DNA EMSA (Electrophoretic Mobility Shift Assay, (Hellman and Fried 2007)), ChIP (Chromatin Immunoprecipitation, (O'Neill and Turner 1996)) and its high-throughput variants ChIP-chip (Zhang, Guo et al. 2008) and ChIP-seq (Mardis 2007; Kharchenko, Tolstorukov et al. 2008; Park 2009).

Three dimensional structures, obtained mainly using X-ray crystallography (Woolfson 1997) and NMR (Nuclear Magnetic Resonance) spectroscopy (Cavanagh 2007), are essential to a full understanding of protein functions. Since year 2000, structural genomics initiatives have been carried out that aim to solve 3-dimensional structures for a set of representative proteins and to draw a full image of the whole structural space with the aid of high-throughput structure determination pipelines (Baker and Sali 2001; Vitkup, Melamud et al. 2001; Gerstein, Edwards et al. 2003; Chandonia and Brenner 2006; Terwilliger, Stuart et al. 2009). Together, the steady progress of traditional structural biology

and structural genomics efforts has generated many tens of thousands of structures deposited in the Protein Data Bank (PDB, (Berman, Westbrook et al. 2000)) database, covering the majority of known protein families.

A daunting quantity of data has been produced by these experimental techniques. This flood of information poses an array of challenges but also opportunities for biological scientists. The needs to store, organize, and analyze it demand new computational and informatics tools. More importantly, data by itself alone is not knowledge. Thus, to mine the data for biologically meaningful patterns that are comprehensible to humans is among the most challenging missions of functional genomics. Computational techniques based on sequence analysis, graph theory, machine learning, and statistical inference are crucial to this endeavor.

1.2 From bioinformatics to systems biology: the whole is more than the sum of its parts

Bioinformatics is the discipline that applies computational and informatics techniques to biological research. Conventionally, the major topics of bioinformatics include sequence alignment and assembly, gene and motif finding, protein structure modeling and docking, drug design, protein function prediction, gene expression analysis, disease gene finding, association mapping, and phylogenic tree reconstruction *etc.* (Jones and Pevzner 2004; Pevsner 2009).

Three decades of development of bioinformatics have generated a battery of databases, webservers and software that play key roles in almost all sub-disciplines of biology. For example, it has been a routine for a scientist to search on the NCBI (National Center for Biotechnology Information) genome database for genes of similar sequences using the Basic Local Alignment Search Tool (BLAST, (Altschul, Madden et al. 1997)), when one is interested in an DNA sequence of unknown function. It is also very common for a researcher to generate testable hypotheses using a number of structure-based function annotation servers (Laskowski, Watson et al. 2005; Pal and Eisenberg 2005; Fischer, Zhang et al. 2011), if the structure of a protein is known or a reliable homology model can be built.

For decades, biologists have been highly successful in studying biological systems through a reductionist approach, deconstructing systems into individual components and focusing on specific aspects of the systems. However, with the unprecedented growth of biological data and the development of analytic tools, the breadth and the depth of information and means available now have for the first time afforded us the ability to address biology at an integrative systems level. We have now reconstructed large physical and functional interaction networks for many cellular systems through biochemical, biophysical and genetic approaches. These networks, revealed by the connectivity of individual genes and proteins,

can help to identify repeating motifs of biological significance and modules of specific functions (Barabasi and Oltvai 2004). For example, Tang and colleagues (Ma, Trusina et al. 2009) found that among tens of thousands of all possible three-node enzyme network topologies, only two major core motifs, a negative feedback loop with a buffering node and an incoherent feed-forward loop with a proportioner node, could perform biochemical adaptation, the ability to reset after responding to a stimulus. This phenomenon of adaption is a so-called *emergent property* of the whole system, *i.e.*, it cannot be achieved and analyzed on the level of individual genes.

The so-called "systems biology" may mean different things to different people. Some people think of systems biology as large-scale research, i.e., research at the "omics"-scale. Others may focus on quantitative modeling of relatively small systems. In spite of large-scale research or quantitative modeling, the idea of "integrative study" plays an essential role in systems biology. Indeed, Sauer and colleagues wrote that (Sauer, Heinemann et al. 2007):

"*…the pluralism of causes and effects in biological networks is better addressed by observing, through quantitative measures, multiple components simultaneously and by rigorous data integration with mathematical models.*"

Integration means gathering all relevant information on whole systems, which could be the same type of information on different individual components

or different types of information on the same system. For example, for complex diseases such as cancer, diagnosis could be difficult because a disease usually involves many different genes. Systems approaches are gaining increasingly important roles in identifying disease genes from the perspective of the whole networks of protein-protein interactions and protein-DNA interactions (Adler, Lin et al. 2006; Franke, van Bakel et al. 2006; Oti, Snel et al. 2006; Bergholdt, Storling et al. 2007; Ergun, Lawrence et al. 2007; Lage, Karlberg et al. 2007; Amit, Garber et al. 2009). Moreover, the integration of more types of information, like gene expression and genome variation profiles, may help to improve diagnostics or prognostics, or even elaborate the disease mechanism, such as how a handful of master regulators or driver genes could cause a complex disease (Calvano, Xiao et al. 2005; Tian, Greenberg et al. 2005; Anastassiou 2007; Mani, Lefebvre et al. 2008; Nibbe, Markowitz et al. 2009; Wang, Saito et al. 2009; Akavia, Litvin et al. 2010; Carro, Lim et al. 2010).

Systems biology is more than reductionism but not its antithesis. For example, system-wide pathways or interaction maps are mainly built by integration of knowledge about individual genes and proteins. Also, systems biology is not only data-driven. Rather, it is an integrative framework to make discoveries, as well to build predictive models and testable hypotheses using system-wide data and perturbation techniques.

1.3 Structural biology meets systems biology

To date, structural information has not been widely exploited in systems biology, mainly because of the limited number of protein structures available, especially of complexes which are particularly relevant to systems biology. However, combining techniques from computational structural biology and systems biology has the potential to address the shortcomings of each. On one hand, structural biology provides atomic level descriptions of protein function but studies tend to focus on only a few proteins at a time. On the other hand systems biology can generate functional hypotheses for large numbers of proteins simultaneously but with questionable reliability, a problem which could be addressed by using structural information to confirm, negate or suggest alternate hypotheses.

Protein-protein interactions represent a key connection point between structural and systems biology (Aloy and Russell 2006; Kiel, Beltrao et al. 2008). In past years, there has been much interest in the generation of comprehensive networks of interacting proteins, i.e., “interactomes”, of different organisms, using large-scale, high-throughput experimental approaches (Uetz, Giot et al. 2000; Ito, Chiba et al. 2001; Rain, Selig et al. 2001; Gavin, Bosche et al. 2002; Ho, Gruhler et al. 2002; Giot, Bader et al. 2003; Li, Armstrong et al. 2004; Butland, Peregrin-Alvarez et al. 2005; Rual, Venkatesan et al. 2005; Stelzl, Worm et al.

2005; Gavin, Aloy et al. 2006; Krogan, Cagney et al. 2006; Ewing, Chu et al. 2007; Tarassov, Messier et al. 2008; Yu, Braun et al. 2008; Dreze, Carvunis et al. 2011) and manual curation of small-scale experiments reported in the literature (Reguly, Breitkreutz et al. 2006; Cusick, Yu et al. 2009). In parallel, approaches that use indirect evidence such as sequence homology (Matthews, Vaglio et al. 2001; Yu, Luscombe et al. 2004), gene co-expression (Qian, Dolled-Filhart et al. 2001; Jansen, Greenbaum et al. 2002; Soong, Wrzeszczynski et al. 2008), function similarity (Wu, Zhu et al. 2006), gene fusion (Enright, Iliopoulos et al. 1999; Marcotte, Pellegrini et al. 1999), genomic context (Dandekar, Snel et al. 1998; Huynen, Snel et al. 2000), and phylogenetic profile/tree similarity (Huynen and Bork 1998; Pellegrini, Marcotte et al. 1999; Pazos and Valencia 2001; Goh and Cohen 2002) have also been developed to computationally infer PPIs on a large scale (see reviews (Valencia and Pazos 2002; Salwinski and Eisenberg 2003; Shoemaker and Panchenko 2007; Skrabanek, Saini et al. 2008)).

Despite significant progress, however, comparative studies (Aloy and Russell 2002; Bader and Hogue 2002; von Mering, Krause et al. 2002; Sprinzak, Sattath et al. 2003; Braun, Tasan et al. 2009; Cusick, Yu et al. 2009; Salwinski, Licata et al. 2009) suggest that there is still a long way to go in developing a complete and error-free understanding of even the widely studied yeast and human interactomes. For example, high-throughput experimental approaches

produce many false positives while failing to identify the majority of true interactions. Indeed, while more than 75,000 PPIs for yeast can be extracted from existing databases, there is only limited overlap between PPI maps assembled by distinct groups (von Mering, Krause et al. 2002). It also has been suggested that the false negative may be quite high, for example in the 80% range for Y2H experiments (Yu, Braun et al. 2008).

Can structural information be applied to the problem? At present, the PDB (Berman, Westbrook et al. 2000) structure database contains more than 70,000 structures, which have been classified into different 4,198 SCOP families (ver 1.75 as of June 2009; about 10,000 families in total are expected), or covered 5,084 Pfam families (ver 24.0 as of Dec 2009; 11,912 in total). In addition, a big portion of the PDB structures are protein complexes of more than one protein chains. Currently, the PDB database contains about 37,000 protein complexes, representing >5,200 different pairs of Pfam families, based on the 3did database (Stein, Panjkovich et al. 2009). The large number of structures and complexes and the significant coverage on structural space suggest that approaches based on comparative complex modeling could be useful. Such approaches use experimentally determined protein complexes in the PDB and PQS (Protein Quaternary Structure, (Henrick and Thornton 1998)) databases as “templates” to

model potential interactions (Lu, Lu et al. 2002; Aloy and Russell 2003; Davis, Braberg et al. 2006).

An important question, however, is how well these templates represent "interaction space" (Aloy and Russell 2004). Systematic studies have highlighted the variability of the binding modes for proteins of the same pair of families (Aloy, Ceulemans et al. 2003; Jefferson, Walsh et al. 2006; Kim, Henschel et al. 2006; Shoemaker, Panchenko et al. 2006), complicating the development of reliability measures for predictions, which may require accurate modeling; structure-based interaction prediction methods have thus tended to rely heavily on closely-related proteins, limiting the number of templates that may be used to model a particular interaction, and consequently the number predictions that can be made.

Moreover, structures of protein complexes are indispensable towards a full understanding to these interactions. Many studies have been carried out to study the physico-chemical properties that govern PPIs, which have been used in computational protein docking, protein interface prediction, protein complex assembly and modeling (Chothia and Janin 1975; Jones and Thornton 1996; Jones and Thornton 1997; Lo Conte, Chothia et al. 1999; Nooren and Thornton 2003). In particular, our group has been studying the structural and energetic basis of PPIs for some time (Sheinerman and Honig 2002; Sheinerman, Al-Lazikani et al. 2003) most recently in the context of specificity determinants in cadherins (Patel,

Chen et al. 2003; Chen, Posy et al. 2005; Patel, Ciatto et al. 2006; Shapiro and Honig 2007). However, this level of detail in structural modeling is not yet achievable on a genome-wide scale.

1.4 Specific aims of this thesis

Our lab has worked on the research area of homology modeling (Petrey and Honig 2005; Forrest, Tang et al. 2006; Xiang, Steinbach et al. 2007; Soto, Fasnacht et al. 2008; Zhu, Fan et al. 2008), the structural and energetic basis of PPIs (Sheinerman and Honig 2002; Sheinerman, Al-Lazikani et al. 2003), and the relationships of protein structures and functions (Petrey and Honig 2009; Petrey, Markus et al. 2009). In this thesis, I will describe my work related to the development of new methods to predict the function of a given protein based on its three-dimensional structure and the application of these methods to the study of networks of interacting proteins.

1.4.1 High-throughput prediction of protein-protein interfaces from structural neighbors

The ability to predict protein-protein interfaces from monomer structures is important for understanding their functions and further help to the prediction of PPIs. Early efforts in this area are represented by the work of Thornton and coworkers to predict surface patches overlapping with interfaces (Jones and

Thornton 1997). Since then, many papers have been published (Armon, Graur et al. 2001; Zhou and Shan 2001; Neuvirth, Raz et al. 2004; Bordner and Abagyan 2005; Bradford and Westhead 2005; de Vries, van Dijk et al. 2006; Liang, Zhang et al. 2006; Ofran and Rost 2007; Porollo and Meller 2007), which use different sets of residue characteristics and different machine learning algorithms to predict protein interfaces.

We studied protein interface conservation in sets of proteins that share varying degrees of sequence and structural similarities (Zhang, Petrey et al. 2010). Our results confirm the most significant conservation among close neighbors, but also find surprisingly high level of conservation even for remote structural neighbors. We used this finding to develop PredUs, a method to predict protein interface simply by "mapping" the interface information from its structural neighbors to the target structure. Our method outperforms other state-of-the-art methods that are typically based on amino acid properties.

We also developed the PredUs web server to predict protein interfaces using this template-based method (Zhang, Deng et al. 2011). In the webserver, we use a support vector machine (SVM) to further improve interface predictions. The server allows users to visualize their predictions and interactively apply different ranking operators and different functional and structural filters to tailor the prediction to a particular hypothesis.

PredUs runs very fast and can be used to investigate protein interface in a high-throughput fashion. Moreover, it is not sensitive to local conformational changes and small errors in structures and thus could be applied to homology models. The success of PredUs suggests the possibility of using structural information as a basis for predicting PPIs on a genome-wide scale.

1.4.2 Structure-based prediction of protein-protein interactions on a genome-wide scale

Despite recent progress in exploiting the idea of using structural modeling to predict PPIs and to model protein complexes (Lu, Lu et al. 2002; Aloy, Bottcher et al. 2004; Davis, Braberg et al. 2006; Fukuhara, Go et al. 2007; Gunther, May et al. 2007), the number of interactions that could be identified remains small and the overlap of these predictions and the experimental interactions is very low.

Advances in the understanding of the nature of protein sequence/ structure/function space offered an opportunity to integrate structural and systems biology methods in the context of PPI prediction. I have focused on exploiting protein homology models and remote structural relationships to increase the coverage of structural modeling methods for use in the prediction of PPIs. We developed a novel computational method based on geometric relationships between protein structures, which can be used to accurately predict PPI on a

genome-wide scale. Indeed, the comparative study shows that the coverage and the accuracy compare favorably with methods derived from sequence and functional clues. Moreover structural information provides orthogonal clues to these non-structure-based methods. We thus use a Bayesian evidence learning model to combine structural information with other non-structural clues. The resulting method, called PREPPI, yields surprisingly high quality predictions that are comparable to high-throughput experiments.

The effectiveness of three-dimensional structural information can be attributed to the use of homology models combined with the exploitation of both close and remote geometric relationships between proteins. Our results suggest that Structural Biology and molecular systems biology can be integrated to an extent that has not been possible in the past.

1.5 Thesis outline

I will introduce the relevant biology and computer science background knowledge in Chapter 2, mainly focusing on experimental and computational methods that identify protein interface and PPIs. In Chapter 3, I present our study on protein interface conservation, and the idea of using it for the prediction of interface of a given protein. In Chapter 4, I describe the PredUs protein interface prediction webserver which based on interface conservation and a SVM. In Chapter 5, I present our work on PPI prediction using structural modeling and the

combine of structural information with other functional clues into a PPI prediction framework. Finally, in Chapter 6, I summarize my thesis and propose several potential future directions suggested by my work.

CHAPTER 2. GENERAL BACKGROUND

2.1 Protein structure space

The function of a protein is closely dependent on its three dimensional structure. Traditionally, there have primarily been two techniques that allow the determination of a protein structure at a resolution of the level of distinguishing individual atoms, *i.e.*, X-ray crystallography (Woolfson 1997) and NMR technique (Cavanagh 2007). But recently, Cryo-electron microscopy (Cryo-EM) has become another important means of determining protein structures with high resolution (Liu, Jin et al. 2010; Zhang, Jin et al. 2010). With further improvement, it is expected to be a tool with increasing significance in the future, especially for solving structures of large protein complexes.

Since the determination of the first protein structure, myoglobin, more than 50 years of effort has accumulated about 74,000 protein structures in the present PDB, among which around 90% are determined by X-ray crystallography (data of Aug. 2011). Comparative structural analyses show that many of these structures share similarities to some extent. How to detect and use these relationships is an intriguing aspect of the study of protein sequence/structure/ function space. One possible answer is to classify proteins into different levels of similarities based on both sequence and structural relatedness. For example, the

SCOP (Structural Classification of Proteins, (Andreeva, Howorth et al. 2004)) database classifies all protein structures into different classes, folds, superfamilies and families. Usually, structures in the same SCOP class only have some extent of similarities in the general structural architecture; those in the same fold share similar arrangement of regular secondary structures; and in the same superfamily, sufficient structural and functional similarity; in the same family, some extent of sequence homology. Currently, the SCOP database (ver 1.75 as of June 2009) contains 1195 folds, 1962 superfamilies and 3902 families. The number of folds has even surpassed its original speculation of one thousand structural folds in total (Chothia 1992), although in the last decade, the increase in SCOP categories has been slowing down (Levitt 2007). It appears that the structures in the current PDB database have covered a big portion of protein structure "space".

By organizing protein structures in an easily comprehensible hierarchical manner, efforts like SCOP and CATH (Protein Structure Classification, (Pearl, Bennett et al. 2003)) have helped researchers to easily locate their structures of interest in the sequence/structure/function space and identify proteins in their neighborhood. However, to classify proteins into families, superfamilies, folds and classes may obscure the relationships between different categories (Xie and Bourne 2008; Petrey and Honig 2009; Petrey, Markus et al. 2009; Skolnick, Arakaki et al. 2009). There have been many examples that protein structures in

different folds turn out to share significant geometric similarities detected by structural alignment tools (Shindyalov and Bourne 1998; Petrey and Honig 2003; Zhang and Skolnick 2005; Holm, Kaariainen et al. 2006). These geometric similarities often implicate important functional relationships. Increasingly, studies have suggested that, rather than a sum of different folds, protein structural space should be regarded as continuous.

Protein structure is directly determined by its primary amino acid sequence, and similar sequences usually result in similar structures. This observations is the basis of homology modeling, a technique to develop three-dimensional models for a target protein sequence based on the structures of homologous proteins (for reviews see (Marti-Renom, Stuart et al. 2000; Petrey and Honig 2005; Ginalski 2006; Zhang 2008)). Databases have been generated to store homology models of a much bigger number of proteins than those in the PDB database (Pieper, Eswar et al. 2006; Kiefer, Arnold et al. 2009; Lee, Li et al. 2010). For example, our analysis shows that about one tenth of yeast proteins or one fifth of the human proteins have associated PDB structures. However, for both of them, about two thirds of them have reliable homology models in the ModBase (Pieper, Eswar et al. 2006) and the SkyBase (Lee, Li et al. 2010) databases. The increase of structural coverage on proteins of other less studied organisms from PDB structures to homology models is even much more

significant. For example, the current PFAM (protein family, (Finn, Mistry et al. 2010)) database contains a little more than 12,000 sequence families among which about 5000 have PDB structures for at least one family member (ver 25.0 as of Mar 2011). Since proteins in the same PFAM family are expected to have similar structures, this implies that millions of protein sequences could potentially been covered by homology models.

2.2 Prediction of protein interfaces

Identification of protein-protein interfaces is necessary for understanding how proteins interact with other molecules. Experimental methods of determining protein interfaces include *in situ* hybridization and mutation studies, both of which are labor intensive and time consuming, highlighting the need for computational approaches.

Structural analyses of protein complexes revealed general principles that govern protein-protein interactions (Chothia and Janin 1975; Jones and Thornton 1996; Jones and Thornton 1997; Lo Conte, Chothia et al. 1999; Nooren and Thornton 2003). It has been shown that protein interface share common properties that can distinguish them from the rest of protein surface. For example, protein interfaces are usually enriched of hydrophobic (and aromatic) residues and arginine, especially for obligate complexes (Lo Conte, Chothia et al. 1999; Glaser, Steinberg et al. 2001; Zhou and Shan 2001; Crowley and Golovin 2005). Interface

residues also appear to have higher side-chain energies (*i.e.* less stable) than the other surface residues (Cole and Warwicker 2002; Liang, Zhang et al. 2006). They also tend to be more conserved (Lichtarge, Bourne et al. 1996; Hu, Ma et al. 2000; Valdar and Thornton 2001; Zhou and Shan 2001; Pupko, Bell et al. 2002), especially for those structural and functional important sites. Most interfaces are spatially continuous patches of a number of residues, which are often among the most planar and most accessible patches (Jones and Thornton 1997). And interestingly, they have a preference for β-sheets and relatively long non-structured chains, but not for α-helices (Neuvirth, Raz et al. 2004).

However, no single property is sufficient for complete and accurate prediction whether a surface residue is on interface or not. The characteristics distinguishing interface residues generally are weak and even, sometimes, controversial. For example, it has been shown that hydrophobicity at the interfaces of transient complexes is not as distinguishable from the remainder of the surface as hydrophobicity at the interfaces of the obligate complexes (Jones and Thornton 1996; Lo Conte, Chothia et al. 1999). It was also argued that interface is rarely significantly more conserved than other surface patches (Bradford and Westhead 2003; Caffrey, Somaroo et al. 2004), and transient interfaces evolve faster than obligate ones (Mintseris and Weng 2005).

A combination of different residue properties considered over surface patches of multiple residues is thus usually necessary for protein interface prediction. Many methods classify residues as interfacial or non-interfacial using different machine learning algorithms such as linear regression (de Vries, van Dijk et al. 2006; Liang, Zhang et al. 2006; Murakami and Jones 2006; Kufareva, Budagyan et al. 2007), neural network (Zhou and Shan 2001; Fariselli, Pazos et al. 2002; Chen and Zhou 2005; Porollo and Meller 2007), support vector machines (Koike and Takagi 2004; Bordner and Abagyan 2005; Bradford and Westhead 2005), Bayesian networks (Neuvirth, Raz et al. 2004; Bradford, Needham et al. 2006), and random forest (Sikic, Tomic et al. 2009). These methods generally take a set of residue properties as input and train classifiers on a set of protein complexes.

The computational prediction of protein interfaces has been a very hot topic in bioinformatics research. The reported performances of these different methods, however, are not directly comparable, because of the different benchmark datasets, different performance evaluation methods, and different definitions of protein interface used in their evaluations. A number of comparative studies have compared different prediction methods on the same benchmarks (Zhou and Qin 2007; de Vries and Bonvin 2008). It is shown that many of these methods have been very successful, especially for the prediction of some specific

types of interfaces, *e.g*. interfaces between enzymes and inhibitors. Yet challenges remain. For example, these methods generally do not perform very well for protein interfaces with large conformation changes during complex formation, and large interfaces that formed between large proteins or multiple binders.

These methods rely on physical-chemical features of individual residues, and can be sensitive to their spatial positions. In addition, some interface residues may have very distinguishing characteristics while others may not. In this thesis, we report a protein interface prediction method that is mechanistically different from the above-mentioned methods. Our method, called PredUs, is a "template-based" prediction method (by contrast, we may call the methods mentioned here "*ab initio*" methods), in which an interface for a given query protein is inferred based on some similarity to another protein or set of proteins with known interfaces. PredUs may overcome some difficulties of those "*ab initio*" methods. For example, it is capable of identifying interface residues of less distinguishing properties, as can be seen from the much higher prediction recalls. It also seems to be insensitive to conformational changes that occur upon binding, as can be seen from the small difference between the performances of PredUs on the bound and unbound CAPRI targets. Please see Chapter 3 and Chapter 4 for detailed discussion.

2.3 Experimental determination of protein-protein interactions

A multitude of methods have been developed for the determination of direct physical interactions between proteins. As a community effort that aims to define exchange standards for molecular interaction data, HUPO's (Human Proteome Organization) PSI-MI (Proteomics Standards Initiative – Molecular Interactions: (Hermjakob, Montecchi-Palazzi et al. 2004)) lists tens of different methods, which can be broadly classified into biochemical and biophysical methods (Kerrien, Orchard et al. 2007). Each method has its own strengths and weaknesses in identifying protein-protein interactions (PPIs). One method may identify some interactions but fail on others. Some methods may detect direct interactions between two proteins while others may only identify a group of proteins that form a complex. And some methods may be accurate but can only be carried out in small scale; others can be easily scaled up but are not as reliable. A number of reviews have been written that discuss these methods (Phizicky and Fields 1995; Aloy and Russell 2002; Deane, Salwinski et al. 2002; Fields 2005; Piehler 2005; Berggard, Linse et al. 2007; Gingras, Gstaiger et al. 2007; Shoemaker and Panchenko 2007). Here I only give a brief introduction to the yeast two-hybrid (Y2H) and the affinity purification followed by mass spectroscopy (AP-MS) methods, which are among the most important methods for high-throughput PPI screening.

The yeast two-hybrid (Y2H) method was originally developed by Fields and colleagues (Fields and Song 1989). It implements a system in which a functional transcription factor (TF) is split into two separate fragments, the DNA binding domain (DBD) and activating domain (AD), which are independently fused with a "bait" protein X and a "prey" protein Y in study. Upon the binding of proteins X and Y, the AD is brought in close proximity to its DBD counterpart and restores the TF's function to activate the transcription of a reporter gene.

There are many advantages of the Y2H systems. First of all, it is a eukaryotic in vivo technique, and is easily to be carried out. In addition, it can detect weak and transient interactions, and can be scaled up to apply on a genomic scale. However, there are also several disadvantages with the Y2H method. The main criticism is the possibility of a high number of false negative and false positive identifications. The reasons lie in the protocols of the systems. For examples, the testing is usually carried out in a heterologous environment such as yeast; the fusion proteins must be targeted to the nucleus; and the fusion itself may affect the structural conformation of the protein. However, both the false negative rate and the false positive rate are difficult to estimate due to the fact that our knowledge on PPIs is incomplete and noisy.

In contrast to Y2H methods which detect direct physical interactions between a pair of proteins, affinity purification (AP) methods identify prey

proteins that form stable complexes with a selected bait protein by virtue of an affinity tag. The complexes are isolated from cell lysate through one or more AP steps and the components are then identified usually by a subsequent mass spectroscopy (MS) step.

AP-MS methods capture PPIs in near physiological conditions. They can determine the quantitative composition of protein complexes, and can also be easily applied in large scale studies. However, AP-MS methods, by definition, identify protein complexes (*i.e.* usually involving more than two proteins). The data obtained from AP-MS experiments needs further processing to infer direct physical interactions, often using the "spoke" or the "matrix" model with some heuristic algorithms (Bader and Hogue 2002). They usually work very well in identification of stable complexes, but cannot detect transient interactions. It is also possible that the addition of an affinity tag brings errors into their results.

Despite the problems and disadvantages, the invention of the Y2H and the AP-MS methods have revolutionized the way PPIs are detected, and have been the primary experimental techniques in genome-wide investigation of PPIs for many organisms (Uetz, Giot et al. 2000; Ito, Chiba et al. 2001; Rain, Selig et al. 2001; Gavin, Bosche et al. 2002; Ho, Gruhler et al. 2002; Giot, Bader et al. 2003; Li, Armstrong et al. 2004; Butland, Peregrin-Alvarez et al. 2005; Rual, Venkatesan et al. 2005; Stelzl, Worm et al. 2005; Gavin, Aloy et al. 2006; Krogan,

Cagney et al. 2006; Ewing, Chu et al. 2007; Yu, Braun et al. 2008; Dreze, Carvunis et al. 2011). These high-throughput screenings aimed to generate large PPI interaction set in an unbiased fashion. However, comparative studies showed that their overlaps are surprisingly low, even if restricted to the same set of proteins (Bader and Hogue 2002; von Mering, Krause et al. 2002).

2.4 Protein-protein interaction curation and databases

In parallel to high throughput screenings, substantial efforts have been devoted to characterize protein-protein interactions (PPIs) with small-scale experiments. Since 1990s, some databases that originally focus on genomics of individual organisms, for example, the Yeast Proteome Database (YPD, (Garrels 1996)), have started to include PPI information generated by these experiments from literature. As more and more interaction data accumulated, databases mainly dedicated to PPIs were created to systematically collect the information, for example the Munich Information Center for Protein Sequence (MIPS) protein interaction database (Mewes, Albermann et al. 1997), the Biomolecular Interaction Network Database (BIND, (Bader, Betel et al. 2003)), the Database of Interacting Proteins (DIP, (Salwinski, Miller et al. 2004)), the Protein Interaction Database (IntAct, (Kerrien, Alam-Faruque et al. 2007)), the Molecular Interaction Database (MINT, (Chatr-aryamontri, Ceol et al. 2007)), the Human Protein Reference Database (HPRD, (Keshava Prasad, Goel et al. 2009)) and the

Biological General Repository for Interaction Datasets (BioGRID, (Stark, Breitkreutz et al. 2006)).

As we mentioned, there have been community efforts in creating a common framework for standardizing PPI data representation and curation policies. The PSI-MI provided controlled vocabulary and data structure to reduce the ambiguities in data collection. The International Molecular Exchange Consortium (IMEx) organizes the collaboration between major public interaction data providers including all above-mentioned ones. These efforts have been vital to the curation quality and the easy-exchange of PPI data across different databases, and have made it possible to aggregate PPI data from different sources into large-scale systematic networks.

Table 2-1 gives some statistics of the major PPI databases that are available to public as of Aug. 2011. Only direct "physical interactions" are considered here although some databases also contain information of "genetic interactions". Redundancy has been removed, i.e., evidences of the same interactions have been merged. We use data from the curator's website when available. Among these databases, DIP, IntAct and MINT are active members and BioGRID is an observer of the IMEx initiative. Some databases contain interactions of multiple organisms, among which IntAct is the most comprehensive one; and the others (MIPS and HPRD) only focus on interactions

of one organism (yeast and human respectively). PPIs in these databases include both high throughput screenings and small scale experiments, but not computational predictions (which will be discussed in the following sections). The majority of them account for proteins of yeast and human. Please see (Tsai, Rohl et al. 2006; Lehne and Schlitt 2009) for reviews.

Table 2-1. PPI databases (Aug., 2011).

Database	Proteins	Interactions	Publications	Organisms	URL
MIPS	4,162	9,119	668	1	http://www.mips.com/
DIP	23,201	71,276	4,607	372	http://dip.doe-mbi.ucla.edu
IntAct	57,741	268,981	13,802	341	http://www.ebi.ac.uk/intact
MINT	33,439	92,170	4,108	389	http://mint.bio.uniroma2.it/mint
HPRD	30,047	39,194	20,074	1	http://www.hprd.org/
BioGRID	32,142	143,964	20,960	25	http://www.thebiogrid.org/

Similar to high-throughput studies, discrepancies have been noticed in different curation efforts (Reguly, Breitkreutz et al. 2006; Cusick, Yu et al. 2009; Lehne and Schlitt 2009; Turinsky, Razick et al. 2010). These discrepancies are mainly because different databases usually focus on different sets of publications. However, Wodak and colleagues showed that, even for the same set of publications, two databases only fully agree on 42% of the interactions and 62% of the proteins on average (Turinsky, Razick et al. 2010). The main reason for this is the use of different gene/protein identifiers in different databases, which

sometimes cannot be mapped in a perfect one-to-one match. Another reason is the different confidence sets or thresholds used to decide on interactions in different databases. Without any doubt that these interaction databases are crucial to systems biology studies, users should keep in mind that they contain some level of false interactions and they are largely incomplete for interactomes of most organisms.

2.5 Computational prediction of protein-protein interactions

As it is easy for experiments to produce many false positives and difficult to identify all true interactions, computational predictions are used both to validate experimentally identified interactions and to infer new interactions from indirect clues.

2.5.1 Prediction using non-structural clues

Information like sequence homology (Matthews, Vaglio et al. 2001; Yu, Luscombe et al. 2004), domain-domain interaction profile (Sprinzak and Margalit 2001; Ng, Zhang et al. 2003), genomic context (Dandekar, Snel et al. 1998; Huynen, Snel et al. 2000), gene fusion (Enright, Iliopoulos et al. 1999; Marcotte, Pellegrini et al. 1999; Marcotte, Pellegrini et al. 1999), phylogenetic profile/tree similarity (Huynen and Bork 1998; Pellegrini, Marcotte et al. 1999; Pazos and Valencia 2001; Goh and Cohen 2002), gene co-expression (Qian, Dolled-Filhart

et al. 2001; Jansen, Greenbaum et al. 2002; Soong, Wrzeszczynski et al. 2008), and function similarity (Wu, Zhu et al. 2006) has been effectively exploited to predict protein-protein interactions (PPIs) in large scale. There have been several reviews that discuss the principles used in these methods and compare their advantages and disadvantages (Valencia and Pazos 2002; Salwinski and Eisenberg 2003; Szilagyi, Grimm et al. 2005; Musso, Zhang et al. 2007; Shoemaker and Panchenko 2007).

Briefly, these methods are based on the following assumptions:

- **Sequence homology** – PPIs can be transferred from some organism to another through the homology relationship between proteins. Interactions of homologous proteins in different organisms are sometimes called "interologs" and thus this method is also referred as "interolog" method. It has benefited from the dramatic increase of genomic data due to recent advances in DNA sequencing.
- **Domain-domain interaction profile** – there are certain domains whose most common function is to mediate PPIs. Hence if two proteins each contain one of these domains, the chance that they will interact is higher. Information about domain-domain interactions can be obtained directly from structure complexes or inferred from PPI data. It can be regarded as a

development of sequence homology method in that the presence of domains is frequently determined by sequence similarity.

- **Genomic context** – genes that are near each other on the chromosome tend to interact. This method is based on the concept of a transcription operon. It is usually useful for the prediction of PPIs in prokaryotes.
- **Gene fusion** – proteins that interact in one organism may be fused into a single protein in another organism, thus protein pairs that are fused in other organisms tend to interact.
- **Phylogenetic profile/tree similarity** – interacting proteins tend to co-evolve. The co-occurrence of proteins in the same set of organisms thus is an indicative of PPI (phylogenetic profile similarity method). Taking a step further, the similarity between the phylogenetic trees of a pair of proteins also suggests a higher likelihood for them to interact (phylogenetic tree similarity or mirror tree method).
- **Co-expression** – interacting proteins tend to have a correlated expression pattern in different conditions, especially for permanent protein complexes.
- **Function similarity** – proteins coordinate to perform functions thus similarity in function (e.g. GO annotation) is an indicative of PPI.

Usually, every indirect evidence by itself is only a very weak PPI predictor; and predictions could be improved by integrating different evidences,

using a variety of machine learning methods such as logistic regression (Bader, Chaudhuri et al. 2004), decision tree (Zhang, Wong et al. 2004), random forest (Lin, Wu et al. 2004), naïve Bayes classifier (Jansen, Yu et al. 2003; Lefebvre, Lim et al. 2007), and support vector machines (Ben-Hur and Noble 2005).

There have been online databases or servers that store or perform PPI predictions using the above-mentioned indirect interaction clues and machine learning methods, notably the STRING (Search Tool for the Retrieval of Interacting Genes/Proteins, (von Mering, Huynen et al. 2003)), PIPs (protein-protein interactions predictions, (McDowall, Scott et al. 2009)), and PPISearch (Chen, Lin et al. 2009). STRING contains both experimentally solved PPIs from a variety of interaction databases and predicted PPIs using a naïve Bayes classifier to integrate interaction clues of mainly sequence homology, genomic context, gene fusion, phylogenetic profile similarity, and gene co-expression. The version 9.0 of STRING has more than 57 million predictions covering more than 1,100 organisms (Szklarczyk, Franceschini et al. 2010). The majority of these are predicted from phylogenetic profiles and thus are indicatives more of protein functional associations than of direct physical interactions. PIPs also used a naïve Bayes classifier to combine interaction clues including sequence homology, domain-domain interaction profile, gene co-expression and other information like protein co-localization, protein post-translational modification and interaction

network property. However, the server only focuses on human proteins and provides predictions of only ~80,000 interactions at the lowest cutoff. Different to STRING and PIPs, PPISearch is an online PPI prediction server which first performs sequence homology search and then filters the interologs for conserved domain–domain pairs and function pairs.

According to our analysis, none of these prediction servers are satisfactory. The overlaps of the prediction results with known interaction reference datasets are small (data unpublished). This is consistent with some other observations, for example, Recent Dialog for Reverse Engineering and Assessment of Methods (DREAM) challenges have highlighted that the inference of PPIs is significantly less accurate and sensitive than the inference of other, for example transcriptional interactions (Stolovitzky, Prill et al. 2009).

2.5.2 Prediction using structural clues

Despite that structural information provides atomic details of PPIs, it has had relatively little impact in constructing protein-protein interactomes, primarily because there is a dramatic difference between the number of proteins with known sequence and those with an experimentally known structure. The discrepancy suggests that if structure is to be useful on a large scale, it is essential that modeling be exploited.

The traditional method is to use a procedure called “docking” that attempts to evaluate the interacting complex mainly on the basis of shape or electrostatic complementarity between monomer structures (Smith and Sternberg 2002; Wodak and Mendez 2004; Gray 2006). The success of this methodology requires the availability of high resolution structures of both monomers, a fast way to generate a set of docking configurations which includes at least one nearly correct one and an accurate scoring function that reliably distinguishes nearly correct configurations from the others. However, despite some recent progress that takes advantage of known interfaces (Sacquin-Mora, Carbone et al. 2008) or identifies interaction partners from a distribution of docking scores of non-binders (Wass, Fuentes et al. 2011), the potential of using docking to predict PPIs on a genome-wide scale remains in question.

Experimental structures of protein complexes can also possibly be used to predict interactions between sequence and structural homologs of the proteins involved by comparative modeling, since their binding modes tend to be similar as well (Aloy and Russell 2002; Lu, Lu et al. 2002; Aloy, Ceulemans et al. 2003; Davis, Braberg et al. 2006; Gunther, May et al. 2007; Singh, Park et al. 2010). In essence, such approaches align a pair of target proteins with their sequence or structural neighbors in a template complex, and evaluate the model with a set of empirical scoring functions. The success of this methodology, however, depends

on a number of factors: the availability of high quality protein complex structures that contain the close sequence and structural neighbors of the target proteins; correct alignments of the target proteins on the template chains and scoring functions that capture the characteristics of the interaction; and atomic details of the constructed model.

As a consequence, the number of interactions that could be predicted by these methods and also the overlap of their predictions with the known PPI datasets are small, although the prediction accuracy based on structures is usually higher than those based on non-structural clues. There have been a number of prediction servers using structural modeling in PPI prediction, such as InterPreTS (Aloy and Russell 2003), PRISM (Ogmen, Keskin et al. 2005), 3D-partner (Chen, Lo et al. 2007), Struct2Net (Singh, Park et al. 2010), HOMCOS (Fukuhara and Kawabata 2008), and Protinfo PPC (Kittichotirat, Guerquin et al. 2009). However, structural information has only been used alone, and not contained in either of the above integrative servers, STRING and PIPs (STRING contains information of protein structure complex but it is used as evidence of experimental interactions but not as a basis for prediction).

In Chapter 5, I present our approach to PPI prediction using structural information with two major improvements. First, predictions are not limited to pairs of proteins for which another pair with high sequence and/or structure

similarities exists in the PDB; instead, we seek local geometric relationships between groups of secondary structure elements identified by local structural alignment. Second, candidate interacting proteins were evaluated using empirical scores measuring features only weakly dependent on atomistic details. Our benchmarks show that the method has greatly increased the coverage on the whole interaction space and known interactions as well. In fact, the prediction coverage is now comparable to non-structural evidence and yet the prediction accuracy remains much higher.

2.6 Machine learning and its applications in computational biology

Machine learning is a branch of artificial intelligence that is concerned with the design and development of computer systems that automatically improve their performance based on empirical data or past experience (Mitchell 1997; Bishop 2006). Usually, these data are examples with attributes that illustrate relations between observed variables. Given a set of examples (referred as training set in machine learning terminology), a machine learning algorithm learns to capture characteristics of their unknown underlying probability distribution (this learning process is called training), which then could be applied to unseen examples and make predictions. In practice, the underlying probability distribution is usually too large to be covered by the set of observed examples or

is too complex to manually implement. For example, no simple algorithm can identify Mr. Bill Gates from photographs containing his picture, although it is a relatively easy task for most of us human beings. In this case, a machine learning algorithm must learn characteristics of Gates' face from a set of his pictures so as to be able to recognize him in new photographs.

There has been a long history of applying machine learning algorithms in computational biology. Early work includes the use of the perceptron, a type of artificial neural network, in the search of translation start sites in *E. coli* (Stormo, Schneider et al. 1982). In the intervening years, with the development of many different computational learning techniques and theories, machine learning has become an important tool in genomics, proteomics, and systems biology (for reviews please see (Larranaga, Calvo et al. 2006; Tarca, Carey et al. 2007)). For example, different machine learning techniques have been used to find protein-coding and RNA genes from DNA sequences including gene boundaries, intron-exon structures, and functional elements in non-coding regions as well (Mathe, Sagot et al. 2002; Bockhorst, Craven et al. 2003; Won, Prugel-Bennett et al. 2004; Das and Dai 2007). They also have been used to predict protein and RNA secondary structures (Rost and Sander 1993; Fogel, Porto et al. 2002) and protein functions (Troyanskaya, Dolinski et al. 2003; Lee, Date et al. 2004), and to analyze microarray profiles (Butte 2002; Allison, Cui et al. 2006). Recently, they

have also been exploited in systems biology including the reverse engineering of protein interaction networks, regulation networks, and signaling networks (Muggleton 2005; Kaski, Rousu et al. 2007). Essentially, machine learning techniques have been applied to almost all fields in computational biology.

Although it can vary in different applications, a machine learning system usually consists of a learning element that receives and processes the input, a knowledge base that may contain some knowledge in the beginning and is able to update with new knowledge, a performance element that uses the knowledge base to perform some tasks and to produce the corresponding output, an idealized system that produces correct solutions for a set of training examples, and a feedback element that compares the outputs of the learning element and the idealized system and updates the knowledge base so as to produce the correct output (this process is called training).

The knowledge base plays a key role in the whole process, and its representation affects the algorithms of learning. A multitude of different knowledge representation schema, including linear algebra, decision trees, artificial neural networks (ANN), logic programs, hidden Markov models (HMM), support vector machines (SVM), Bayesian networks, have been exploited in many different machine learning applications. Here we only briefly introduce SVMs

and Bayesian network classifiers (including Naïve Bayesian classifiers) in the context of protein interface and protein-protein interaction (PPIs) predictions.

The problem of prediction protein interfaces and PPIs can be formulated as a classification problem, a type of the problem of supervised machine learning: given a training set of labeled instances of the form $<a_1, a_2, \ldots, a_n, c>$ (here a_i is a property of a residue in the case of interface prediction or a property of a protein-pair in the case of interaction prediction; and c is whether the residue is on interface or the protein-pair is an interaction), construct a classifier f that is capable of predicting the value of c, given an instance of $< a_1, a_2, \ldots, a_n >$.

2.6.1 SVM

Originally invented by Vapnik and Cortes, the SVM algorithm is typically used to classify data (Cortes and Vapnik 1995). Suppose the given training data are a set of points in a p-dimensional space, and we want to separate these data points with a $(p-1)$-dimensional hyperplane (which is a line when $p=2$, Figure 2-1) so that those points in the same class are on the same side of the hyperplane. If such a hyperplane exists, we can use it to separate new data points of unknown classes in the future (Figure 2-1A). In the SVM method, we choose the hyperplane that represents the largest separation, or margin, between the two classes. The margin is defined as the shortest distance between a hyperplane and

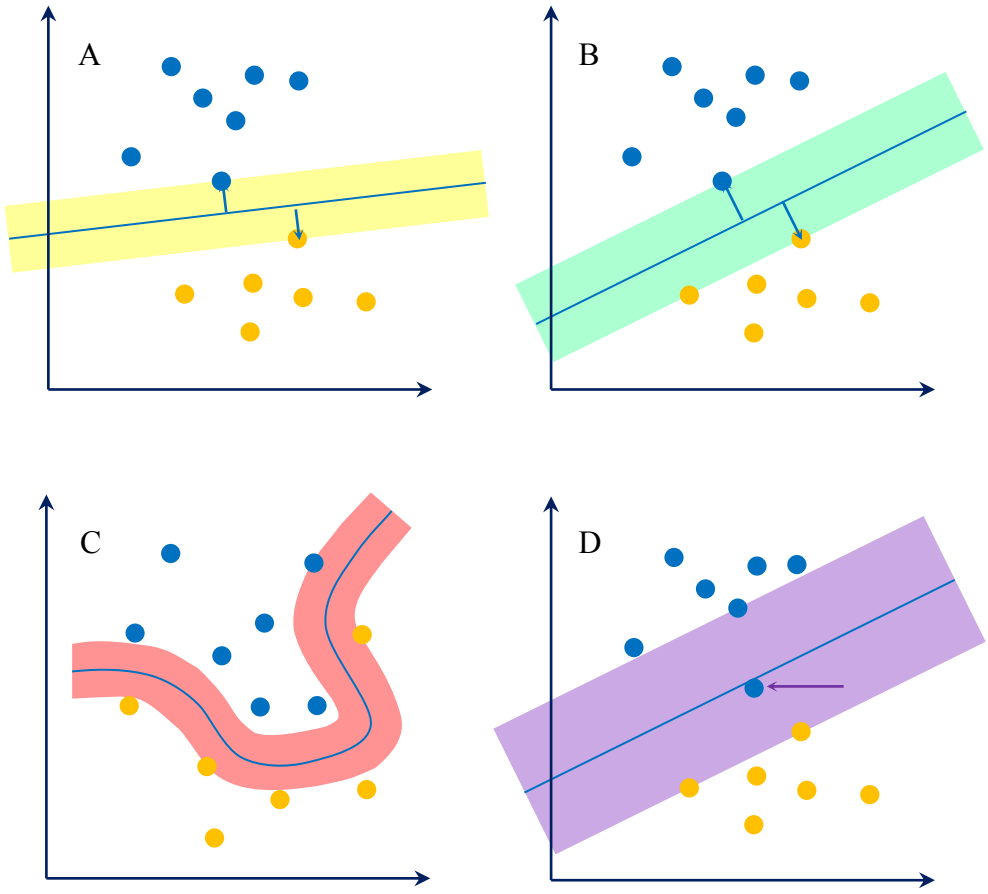


Figure 2-1. SVM as classifiers. Data points belonging to two different classes are shown in blue and orange circles. Classifying hyperplanes are shown as blue lines within shadowed boxes. Margins are distances from classifying hyperplanes to the nearest data points of each side, shown as blue arrows. Support vectors are data points on the edge of the shadowed boxes. Here SVM classifiers are shown with (A) small margin; (B) maximum margin; (C) kernel function that transfer the original space to a high dimensional space where the separating line is a non-linear curve in the original space (Note that the original space but not high dimensional space is shown here); (D) soft margin where the data point with purple arrow will receive a penalty for being misclassified.

the training data points on each side (Figure 2-1A and B). The data points that lie on the margin are the support vectors, from which the name SVM comes.

It is often possible that the given data points are not linearly separable using a ($p-1$)-dimensional hyperplane. More recent approaches to SVMs map the original vector space into a much higher-dimensional space using "kernel functions" where the data points may be linearly separable (Figure 2-1C). In addition, a SVM model often includes a penalty function that allows some data points to be misclassified (Figure 2-1D). The construction of a SVM model thus involves the training of the parameters associated with the penalty function and the kernel function.

Thanks to the use of kernels, SVMs are especially suitable for biological data since they can easily handle high-dimensional, noisy, or non-vector biological data. They have been widely applied in computational biology for gene sequence and protein structure/function classification, protein functional site identification, PPI prediction, and microarray classification. Please see (Ben-Hur, Ong et al. 2008) for a review. In particular, SVM methods have been used to predict protein interface (Koike and Takagi 2004; Yan, Honavar et al. 2004; Bordner and Abagyan 2005; Bradford and Westhead 2005; Res, Mihalek et al. 2005; Chung, Wang et al. 2006; Wang, Chen et al. 2006; Wang, Wong et al.

2006). In our study, we also used SVM to improve the prediction of protein interface based on conservation, please see Chapter 4 for details.

2.6.2 Bayesian network classifiers

A Bayesian network or belief network is a type of probabilistic graphical model that denotes a set of random variables and their conditional dependencies via a directed acyclic graph (Figure 2-2), where nodes represent random variables and edges represent conditional dependencies; each variable is conditionally independent of its non-descendants given the values of their parent variables nodes (Neapolitan 2004). Given a training set $<a_1, a_2, \ldots, a_n, c>$, the problem of learning a Bayesian network is to learn the structure and parameters, *i.e.*, the conditional dependencies and probabilities, of the graph that "best describes" the training data.

An advantage of Bayesian networks is its great interpretability due to explicitly specifying direct dependencies and distributions of different variables. However, learning unrestricted Bayesian networks can be a difficult task, and may results in poor classifiers especially in case of many attributes. The alternative approaches is to design the network by experts (however, this is also difficult when the number of attributes is big) or to use restricted networks. The naïve Bayesian classifier is the simplest Bayesian network classifier where the only dependency is between the class variable *C* and all attributes (Figure 2-2A).

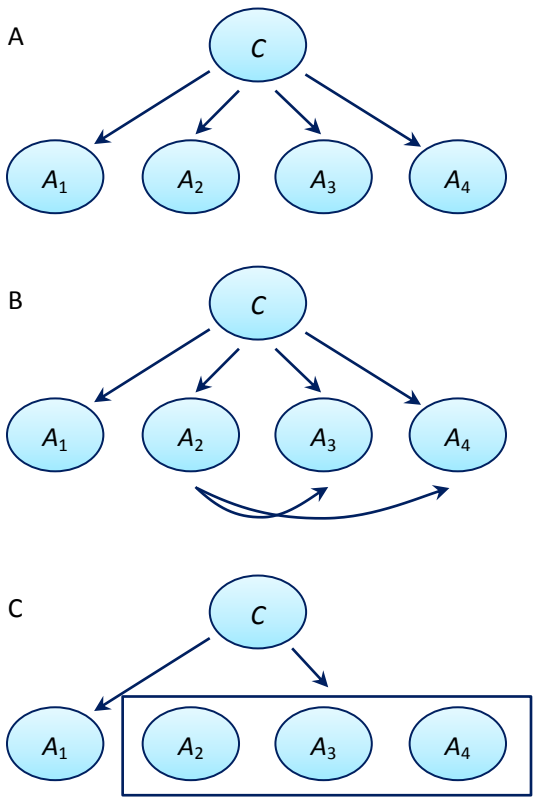


Figure 2-2. Bayesian network classifiers: (A) naïve Bayesian classifier; (B) tree augmented naïve Bayesian classifier; (C) naïve Bayesian classifier with a fully connected component (A_2, A_3, and A_4).

Despite the apparently over-simplified assumption and the simple design, naïve Bayes classifiers have worked surprisingly well in many complex real-world situations. Of course, it is also very common that the correlations between different attributes are too strong to be neglected. And thus there also have been many improvements on naïve Bayesian classifiers by adding correlations among attributes (e.g., tree augmented naïve Bayesian classifier, Figure 2-2B), or to

select a subset of independent attributes (e.g., selective Bayesian classifier). Sometimes, a subset of attributes are correlated to each other but to no other attribute, then they can be separated from others and form a fully connected component, where a joint distribution containing these variables can be used to describe their correlations and dependencies with the label *C* (Figure 2-2C). In our study of PPI structural modeling, we have used such a Bayesian classifier (see Chapter 5).

Bayesian networks including naïve Bayes classifiers are becoming increasingly important in biological research, for example genome analysis (Sandberg, Winberg et al. 2001), protein interface prediction (Bradford, Needham et al. 2006), genetic data analysis (Beaumont and Rannala 2004), cellular network inference (Friedman 2004), and protein signaling pathway modeling (Sachs, Perez et al. 2005). In particular, the naïve Bayes classifier is widely used as an integrative method for protein function and especially PPI prediction due to its simplicity in algorithm implementation, its efficiency, its scalability to easily incorporate more types of information, and its interpretability for contribution of each component (Jansen, Yu et al. 2003; Troyanskaya, Dolinski et al. 2003; Lee, Date et al. 2004; Lefebvre, Rajbhandari et al. 2010). In our study of PPI prediction, we also used a naïve Bayesian classifier to combine different types of PPI evidences. Please see Chapter 5 for details.

CHAPTER 3. PROTEIN INTERFACE CONSERVATION ACROSS STRUCTURAL SPACE

The following chapter is a paper published in the *Proceedings of the National Academy of Sciences of the USA* (Volume 107, Issue 24, 15 June 2010, pp. 10896-10901).

3.1 Abstract

With the advent of systems biology, the prediction of whether two proteins form a complex has become a problem of increased importance. A variety of experimental techniques have been applied to the problem but three-dimensional structural information has not been widely exploited. Here we explore the range of applicability of such information by analyzing the extent to which the location of binding sites on protein surfaces is conserved among structural neighbors. We find, as expected, that interface conservation is most significant among proteins that have a clear evolutionary relationship but that there is a significant level of conservation even among remote structural neighbors. This finding is consistent with recent evidence that information available from structural neighbors, independent of classification, should be exploited in the search for functional insights. The value of such structural information is highlighted through the development of a new protein interface prediction method, PredUs, that identifies

what residues on protein surfaces are likely to participate in complexes with other proteins. The performance of PredUs, as measured through comparisons with other methods, suggests that relationships across protein structure space can be successfully exploited in the prediction of protein-protein interactions.

3.2 Introduction

The knowledge of whether two proteins form a complex is a problem of central importance in the description of cellular networks and in a large number of other biological applications. Much effort has been devoted recently to high-throughput experimental determination and literature curation of protein-protein interactions (see references (Shoemaker and Panchenko 2007; Shoemaker and Panchenko 2007) for a review) and the results have been deposited into numerous databases (Stark, Breitkreutz et al. 2006; Kerrien, Alam-Faruque et al. 2007). In addition, a variety of computational approaches have been developed to predict protein interaction partners (Salwinski and Eisenberg 2003; Fields 2005; Shoemaker and Panchenko 2007; Skrabanek, Saini et al. 2008). Three-dimensional structural information has not been widely used in large scale studies, in part because the number of complexes for which such information is available is far smaller than the number of interactions that can be inferred by other techniques.

A number of groups have shown that the use of homologous relationships can expand the range of structural information by providing plausible models for a protein complex that can then be evaluated with other methods (Lu, Lu et al. 2002; Aloy, Bottcher et al. 2004; Davis, Braberg et al. 2006). However, the extent to which a known 3D structure of a complex can be used reliably as a template for a model of two related proteins is unclear, especially if the relevant sequence and/or structural relationship is remote. Model reliability should, in general, increase if the proteins involved are closely related but the use of close homologs necessarily limits the number of possible interactions that can be detected. We have recently shown (Petrey, Fischer et al. 2009) that the use of remote structural relationships can detect functional relationships between proteins that are obscured by classification schemes. One of the aims of the current paper is to evaluate whether structural relationships that can go beyond classification can be exploited in the structure-based prediction of protein-protein interactions. Our longer range goal is to expand the range of applicability of structural information to the point that it can be used on a scale comparable to that of other, non-structure-based methods.

Most current methods that build models of complexes by homology rely in part on criteria for model reliability that have been established by comparative studies of different complexes (Bashton and Chothia 2002; Aloy, Ceulemans et al.

2003; Kim and Ison 2005; Korkin, Davis et al. 2005; Littler and Hubbard 2005; Han, Kerrison et al. 2006; Kim, Henschel et al. 2006; Shoemaker, Panchenko et al. 2006). A nagging reality of such studies is that there is no unambiguous way of determining whether two complexes are similar. Figure 3-1 illustrates some of the underlying the issues. In the figure, a representative protein complex, A, is compared to three others (see the caption for general details on how this comparison is carried out). Although each of the complexes B, C, and D has some relationship with complex A, this will not necessarily be identified by every measure of similarity. For example, measures that rely on translations/rotations of individual subunits (Aloy, Ceulemans et al. 2003; Han, Kerrison et al. 2006; Jefferson, Walsh et al. 2006) would characterize A and B as similar complexes but not A and C since a 90 degree rotation would be required to superpose C2 on A2. Criteria that depend on the relative location of the centers of mass (Littler and Hubbard 2005) would characterize A and C as similar but not A and D.

Other similarity measures rely on the equivalence of interfacial residues once the proteins in two complexes have been rotated into a common coordinate frame. Using a residue equivalency measure, A and B are clearly similar while A and C might also be considered similar since some of the residues on both sides of the interface are aligned. There is a relationship between complexes A and D since some interfacial residues in one of the monomers are well-aligned. This

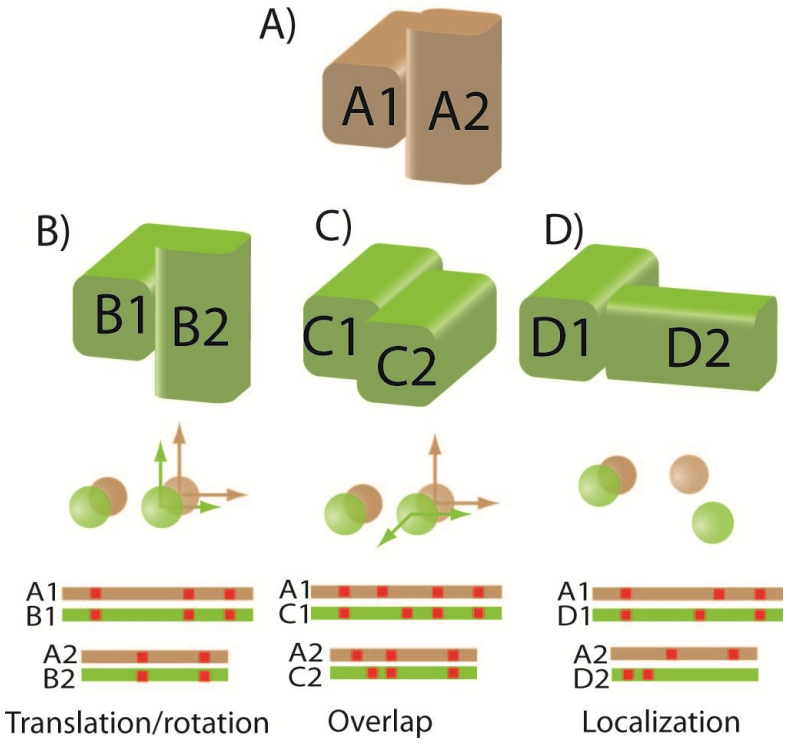


Figure 3-1. Types of geometric conservation and their measures. Protein complex A is compared here to three other complexes B, C, and D. Typically one subunit is superposed on a structurally similar subunit in the complex to which it is being compared (i.e. A1 would be superposed on B1) and the transformation that relates the first subunits is applied to the second so that all proteins are in the same coordinate system. Measures of conservation generally involve calculating: the transformation (translation/rotation) required to optimally superimpose the second subunits on each other (brown/green arrows); distances and angles between the centers of mass of the second subunit (brown/green spheres); and the alignment (independent of residue identity) of interfacial residues in a primary sequence alignment of the two subunits (red squares). Although there is some similarity between A and each of the other three complexes, recognizing it will depend on which measure is used (see text).

feature is a property of only one subunit of the complex and would only be recognized by a criterion such as the "localization index" introduced by Sali and coworkers (Korkin, Davis et al. 2005). Throughout the text we refer to this phenomenon as "interface conservation" and take it to mean that two proteins interact with their partners at geometrically similar locations (independent of the identity of the residues involved).

In order to correlate structural relationships between complexes with standard measures of sequence and structural similarity, complexes have been classified based on the properties of the individual subunits. Using a measure of geometric conservation that depend on translations/rotations, Aloy *et al*. (Aloy, Ceulemans et al. 2003) found that below 30% pairwise sequence identity, little geometric conservation is expected. Other studies using different measures of interface similarity and protein classification have been reported (Han, Kerrison et al. 2006; Jefferson, Walsh et al. 2006; Kim, Henschel et al. 2006; Shoemaker, Panchenko et al. 2006) but general rules have been difficult to establish. Nevertheless, it seems clear from the reported results that little interface conservation is to be expected in the absence of an obvious evolutionary relationship between the proteins that form the two complexes. However, the type of relationship that exists between complexes A and D in Figure 3-1 (conservation of the interface locations in just one of the subunits) has not been extensively

studied. In this case, the underlying question is whether two proteins that share a geometric relationship, e.g. A1 and D1, use a common region of their surface to form an interface independent of the identity or orientation of the second member of the complex. Significant localization of interfaces has been found at the family (Korkin, Davis et al. 2005) and superfamily (Littler and Hubbard) level however there has not, to our knowledge, been a systematic study of the extent to which protein structural similarity can be used as a basis for predicting the interfacial residues.

A number of studies have suggested that this may be possible. Nussinov and co-workers (Tsai, Lin et al. 1996; Keskin and Nussinov 2005) identified similarities in the relative positions of small sets of secondary structural elements within the interfaces of structurally dissimilar interacting proteins suggesting a relationship between patterns of secondary structure and interface formation. Russell *et. al.* (Russell, Sasieni et al. 1998) showed that groups of proteins classified as belonging to different superfamilies or folds interact with their ligands in structurally equivalent locations. Remote similarities such as these have been exploited in a wide range of applications including the prediction of protein-ligand interactions (Brylinski and Skolnick 2008), protein-protein interactions (Lu, Lu et al. 2002), and function annotation (Friedberg and Godzik 2005; Petrey, Fischer et al. 2009).

In this study, we report a comprehensive analysis of the degree to which the location of protein-protein interaction sites is conserved in sets of proteins that share varying degrees of similarity. We start by identifying structural neighbors of the query protein independent of classification and then, using the statistical approach developed by Russell *et al*. (Russell, Sasieni et al. 1998), quantify interface conservation both among close homologs and among remote structural neighbors. Our results show that while, in general, the conservation of interface locations is greatest among close neighbors, significant information is also provided by remote structural neighbors that have no obvious evolutionary relationship to the query. Based on these findings we develop PredUs (http://wiki.c2b2.columbia.edu/honiglab_public/index.php/Software:PredUs), a method for predicting a protein binding region on the surface of a query protein based entirely on information derived from structural neighbors. PredUs compares favorably with methods that, given a three-dimensional structure, predict interfacial regions based on specific features (e.g. sequence conservation, amino-acid properties) of clusters of surface residues. Our findings have important implications, both regarding the nature of protein sequence/structure/function space and for the possibility of using structural information as a basis for predicting protein-protein interactions on a genome-wide scale.

3.3 Results

3.3.1 Interface conservation

We used the procedure described in Methods to quantify interface conservation. Briefly, structural neighbors are identified for a given query protein, and the locations of interfacial residues of the neighbors that are part of a complex are "mapped" to residues in the query protein to generate a "contact map" associated with each structural neighbor. Interface conservation can be visualized by summing individual contact maps and generating a contact frequency heat map. Figure 3-2 shows the surface of the T-cell receptor protein CD8 (PDB code 1akj, chain D) with each residue colored according to the frequency with which interactions are mapped to it when structural neighbors are taken from the same SCOP (Structural Classification Of Proteins) family, superfamily and fold.

Using the approach of Russell *et al.* (Russell, Sasieni et al. 1998), a Z-score that reflects overlap in the set of contact maps (*i.e.*, whether or not there is a set of residues in the query that preferentially has interactions mapped to it) is then calculated. Figure 3-3 shows the distribution of Z-scores for the proteins in our test set (188 protein chains curated from a docking benchmark dataset (Hwang, Pierce et al. 2008), see Methods). To ensure reasonable statistics, at least 6 structural neighbors are needed to calculate Z-scores (83 structures had at least 6 structural neighbors in the same family, 106 in the same superfamily, and 130 in

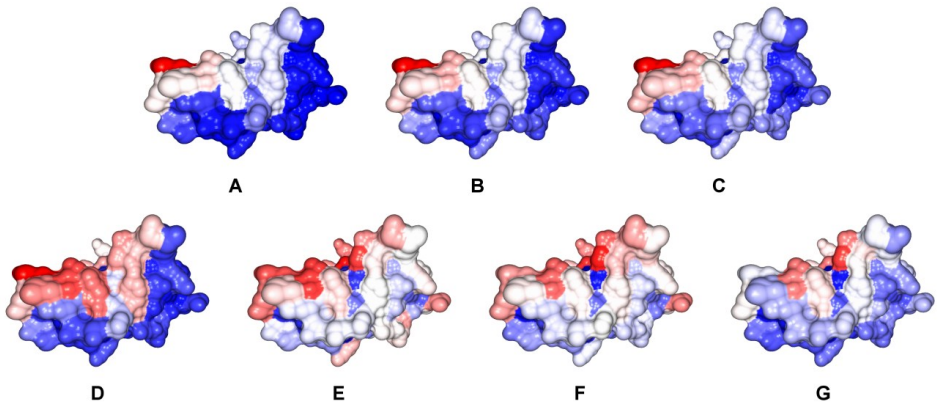


Figure 3-2. The surface of T-cell receptor protein CD8 (PDB code 1akj, chain D) colored according to the frequency with which interactions made by its structural neighbors are "mapped" to individual residues on its surface (red/white/blue = high/intermediate/low frequency). Each surface is colored based on a different set of structural neighbors: (A) SCOP family b.1.1.1; (B) superfamily b.1.1; (C) fold b.1; (D) PSD<0.6 (found by Ska); (E) PSD<0.6 in different families; (F) PSD<0.6 in different superfamilies; (G) PSD<0.6 in different folds. The red high contacting frequency regions show conserved protein interface.

the same fold). As can be seen from the figure, most of the proteins in the test set have Z-scores larger than 3 which is our cutoff for statistical significance (78 out of 83, 95 out of 106 and 118 out of 130, for the same family, superfamily and fold respectively).

As expected, less conservation is observed when more remote structural neighbors are considered, with average Z-scores decreasing as neighbors are

taken from the same family, superfamily, or fold (average Z-score 34, 25, 22, respectively). However, there are many individual cases where the opposite is true and the Z-scores are still significant, suggesting that while there is certainly increased variability in the location of interfaces in the more remote neighbors, significant interface conservation remains. Details about each query protein in our test set including individual Z-scores, the number of structural neighbors, and the highest residue contacting frequencies are given in SI Table S3-1 at http://honiglab.c2b2.columbia.edu/PredUs/html/pnas_si.html.

We also identified structural neighbors using the structure alignment program Ska (Yang and Honig 2000; Petrey and Honig 2003) independent of classification into family, superfamily or fold groups. The average Z-score for the 176 query proteins that had more than 5 structural neighbors is 28, and 166 have Z-score larger than 3 (see Figure 3-3 and SI Table S3-1). The set of structural neighbors identified by Ska was generally significantly larger than the number of proteins classified as belonging to a given grouping in SCOP and contained significant structural diversity. For example, Ska found 978 structure neighbors contained in at least one complex for the structure 1akj.D. These proteins came from 87 different SCOP families, 71 superfamilies and 57 folds. Despite the structural diversity, the difference in average Z-scores for structural neighbors identified independent of classification and for those classified as belonging to the

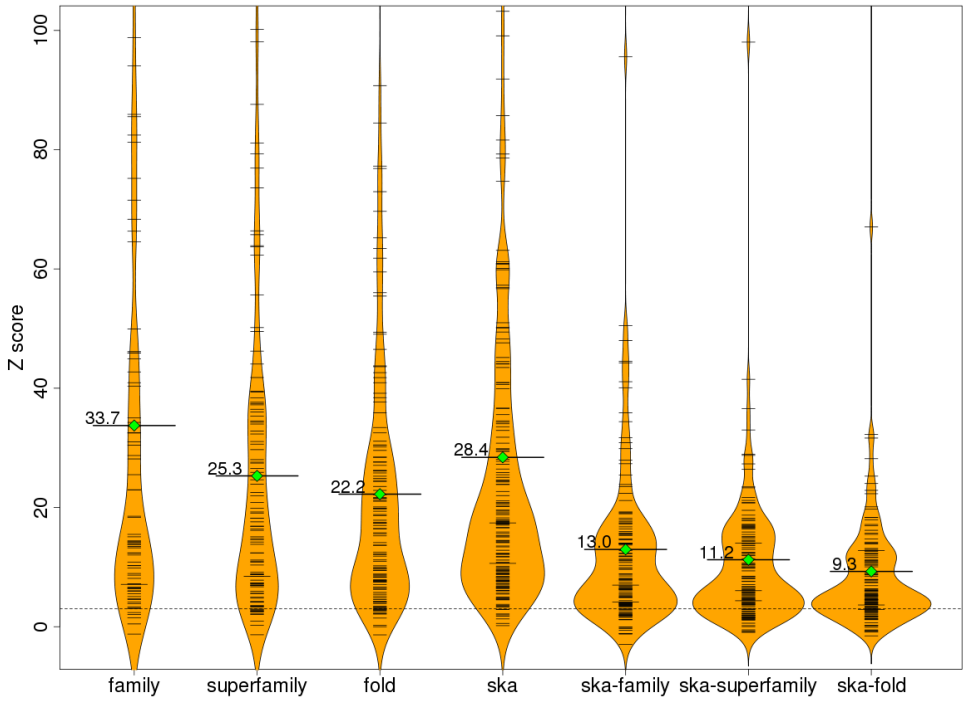


Figure 3-3. Distributions of Z-scores reflecting interface conservation. Each column in the graph shows a Z-score distribution when interface conservation for proteins in our docking benchmark set is calculated based on a different set of structural neighbors. The black bars and the width of each plot reflects the density of Z-scores near the corresponding value on the y-axis. Solid lines with green diamonds show the mean value of each distribution. The dashed line corresponds to a Z-score of 3 which we take as the cutoff of statistical significance. The individual plots have been scaled so that their areas are proportional to the number of proteins for which a valid Z-score could be calculated.

same family, superfamily or fold was small. Since Z-scores reflect overlap in the contact maps calculated for each structural neighbor, these results suggest that

there are a significant number of structures classified differently whose protein-protein interactions sites overlap those of even the close sequence neighbors of the query.

It is possible, of course, that the results obtained independent of classification are due to the presence of family and superfamily members in the set of structural neighbors we identify for each query protein. In order to determine the contribution of neighbors outside of a particular grouping, we carried out a further analysis in which proteins belonging to a particular SCOP classification were excluded (structures with no SCOP annotation were also excluded). Although the Z-scores were not as high as for families, superfamilies and folds, they were still statistically significant (i.e. Z-score >3) with mean values of 13/11/9 (over 138/135/129 structures) when family, superfamily and fold were respectively excluded (see Figure 3-3 and SI Table S3-1 for details).

As described above, this can be visualized using a heat map. For example, for the T-cell receptor CD8 (1akj.D), we identified 254 structural neighbors in 86 families different from that of 1akj.D, 143 structures in 70 different superfamilies, and 90 structures in 56 different folds. Although all these structures come from different families, superfamilies and folds, there is still a well-defined set of residues which preferentially has interactions mapped to it and overlaps with that obtained by considering only more closely related structures (Figure 3-2).

3.3.2 Interface prediction

Based on the above results, we developed a method, PredUs, to predict interfacial residues based entirely on structural neighbors (only the top 50 Ska hits are used, see Methods). Our approach was tested on the docking benchmark described in Materials and Methods and also on the set of structures used in the CAPRI exercise (Janin and Wodak 2007). Results were compared to the top three programs (cons-PPISP (Chen and Zhou 2005), PINUP (Liang, Zhang et al. 2006), and ProMate (Neuvirth, Raz et al. 2004)) reported in a recent comparative study of interface prediction methods (Zhou and Qin 2007), which also performed best in a small-scale evaluation we carried out. We also compared a random prediction in which surface residues are classified as interfacial with a probability of 0.25, which is roughly the portion of interface residues in our test set and is consistent with other studies (Chen and Zhou 2005).

Results are summarized in Table 3-1 (see SI Tables S3-2 and S3-3 at http://honiglab.c2b2.columbia.edu/PredUs/html/pnas_si.html). PredUs results are clearly of comparable quality for both data sets and offer the best combination of precision and recall among all methods tested. This conclusion is based on inspection of Table 3-1 but it is also consistent with the Matthew's Correlation Coefficient (MCC, SI Table S3-4 at http://honiglab.c2b2.columbia.edu/PredUs/html/pnas_si.html). The precision of PredUs is similar to that of other methods

Table 3-1. Precision and recall averages of different interface prediction methods on the docking benchmark dataset and CAPRI bound/unbound targets. Here DKBM stands for the dataset of docking benchmark, N_p and N_c stand for the numbers of total and correctly predicted interfacial residues.

dataset	prediction methods	cases	N_p	N_c	precision average	recall average
DKBM	**PredUs**	**185**	**7,862**	**3,429**	**43.6%**	**45.7%**
	Promate	90	689	322	46.7%	4.3%
	cons-PPISP	188	4,936	2,310	46.8%	30.8%
	PINUP	188	4,227	1,798	42.5%	24.0%
	Random prediction	188	6,827	1,638	24.0%	21.9%
CAPRI bound	**PredUs**	**56**	**2,221**	**921**	**41.5%**	**42.2%**
	cons-PPISP	56	1,497	630	42.1%	28.9%
	PINUP	56	1,204	424	35.2%	19.4%
	Random prediction	56	2,155	492	22.8%	22.6%
CAPRI unbound	**PredUs**	**55**	**2,393**	**952**	**39.8%**	**44.6%**
	cons-PPISP	56	1,542	618	40.1%	29.0%
	PINUP	56	1,320	466	35.3%	21.8%
	Random prediction	56	2,167	544	25.1%	25.5%

but its recall is significantly higher. In order to evaluate the results obtained based on classification, we used PredUs to make predictions but restricted structural

neighbors to members of the same family, superfamily and fold. Results are summarized in Table 3-2. As expected, the highest precision is obtained when only members of the same family are used, and precision decreases as more distant neighbors (superfamily, fold, and the top 50 Ska hits) are included. The trend of the recall value is in the opposite direction. The significant increase in recall when Ska50 is used reflects the additional information available by going beyond SCOP fold. On average, within the Ska50 set there are only 8.6/10.5/11.9 neighbors from the same family/superfamily/fold, while 18.1/16.1/14.7 from different ones (unannotated proteins are excluded).

Table 3-2. Precision and recall averages of PredUs when using structure neighbors from the same and different SCOP groupings on the docking benchmark dataset. Here N_p and N_c stand for the numbers of total and correctly predicted interfacial residues.

prediction methods	cases	N_p	N_c	precision average	recall average
family	141	4,990	2,536	50.8%	33.8%
superfamily	147	5, 907	2,710	45.9%	36.2%
fold	153	6,948	2,904	41.8%	38.7%
Ska50-family	162	8,338	2,541	30.5%	33.9%
Ska50-superfamily	161	8,331	2,370	28.4%	31.6%
Ska50-fold	159	8,603	2,497	29.0%	33.3%

In order to gain insight as to the contributions of increasingly remote structural neighbors to the results, we used PredUs to make predictions where neighbors identified by SCOP were progressively removed from the data set (unannotated proteins also removed). Predictions made in this way are indentified in Table 3-2 as Ska50-family, superfamily and fold, respectively. As is evident from Table 3-2, not considering close family members significantly decreases prediction accuracy but the results are very similar when members of the same fold and superfamily are also removed. Even when only considering members of a different fold the results are better than random. It is clear from Tables 3-1 and 3-2 that the combined use of close and distant neighbors offers the best combination of precision and recall. Most importantly, only by combining in-fold and cross-fold information is it possible to increase recall to above 40%.

Overall, PredUs performed very well for 125 out of 188 docking benchmark proteins. In particular, whenever a successful prediction was achieved using PredUs (both precision and recall better than random) the average precision and recall significantly outperformed other methods (see Table 3-3). There were also some cases where interface information could be extracted from the structural neighbors but where PredUs still made predictions with low precision and recall (26 of the docking benchmark chains). However, the performance in these cases was not due to poor interface conservation in the set of structural

Table 3-3. Precision and recall averages of PredUs good predictions, bad predictions and the others on the docking benchmark dataset.

	prediction methods	precision average	recall average
Good predictions (125 cases)	**Pred-us**	**60.2%**	**57.2%**
	cons-PPISP	54.6%	36.5%
	PINUP	51.9%	29.0%
	ProMate	47.4%	12.1%
Bad predictions (26 cases)	**Pred-us**	**7.3%**	**8.5%**
	cons-PPISP	27.7%	24.6%
	PINUP	29.7%	24.5%
	ProMate	15.2%	5.1%
Others (37 cases)	**Pred-us**	**24.4%**	**39.7%**
	cons-PPISP	36.1%	30.5%
	PINUP	34.4%	24.2%
	ProMate	35.4%	13.5%

neighbors (since the Z-scores were still significant for those cases), but seems to be due to the fact that the particular interface to be predicted for these cases was rarely seen in the set of structural neighbors. This issue is addressed below.

3.4 Discussion

The central result of this study is that there are localized regions on protein surfaces that are conserved among structural neighbors that participate in protein-protein interactions. These regions are properties of a set of neighbors even

though the individual proteins will, in general, form complexes with different proteins using different interface geometries. Thus it is not the geometry of the complex that is conserved but rather the location of surface residues that participate in complexes. The neighbors may belong to the same family or superfamily, and thus bear a clear evolutionary relationship, or belong to the same fold or to different folds, in which case an evolutionary relationship may be present, but its existence is hard to prove. Our findings are consistent with previous work which identified cross-fold functional relationships that are properties of protein fragments and not of the entire structure (Russell, Sasieni et al. 1998; Friedberg and Godzik 2005; Keskin and Nussinov 2005; Petrey, Fischer et al. 2009).

Our results do not imply that a set of structural neighbors will always interact with their partners at a single structurally equivalent patch. Since all interfaces from all structural neighbors are mapped to the query protein in the construction of the contact frequency map, this set of positions may be localized and contiguous or may consist of multiple disjoint patches. Thus, even if there are multiple, distinct protein-protein interactions observed in a set of structurally similar proteins, a high Z-score will be obtained as long as there are enough proteins in the set under consideration that interact with their partners at some set of structurally equivalent locations.

The results in Tables 3-1 and 3-2 highlight the advantages of basing an interface prediction method entirely on information about complexes formed by structural neighbors of a protein. While it is expected that PredUs yields good precision if it is based only on neighbors in the same family or superfamily, that precision is so high when all neighbors are considered seems quite remarkable, and reflects the conservation we describe above. Moreover, using remote structural neighbors produces a significant improvement in recall at the cost of only a moderate decrease in precision. This suggests, that current structural databases are surprisingly complete, in the sense that it generally possible to find representatives of the possible binding modes of a given protein within the 36,888 complexes in the PQS (Protein Quaternary Structure) database (Henrick and Thornton 1998). This depends, however, on the large set of structural neighbors generated using our loose definitions of similarity as well as on the definition of conservation that we use.

Structural information also appears to be a principal source of the improvement in recall of PredUs relative to methods that rely primarily on differences in characteristics (e.g. hydrophobicity, sequence conservation, interface propensity, accessibility, side-chain entropy (Neuvirth, Raz et al. 2004; Chen and Zhou 2005; Liang, Zhang et al. 2006)) between interfacial and non-interfacial residues. Because it may be generally expected that not all of the

residues in a given interface will be distinct in terms of such characteristics, this factor may have a deleterious effect on recall. In our approach, all the interfacial residues from structural neighbors are mapped to the query protein regardless of their characteristics and this difficulty is thus avoided. Since the two approaches are quite distinct and use largely complementary information, it may be of value to combine them in some way in future work.

There are potential drawbacks to the heavy reliance on structural neighbors implicit in our method, but they do not appear to be significant based on an analysis of our test sets. For example, only a small percentage of the proteins did not have enough structural neighbors to enable a prediction (3 in the docking benchmark and 1 in the CAPRI set). Some proteins may have multiple binding sites, and our method depends on identifying those locations which are most frequently associated with protein-protein interactions. An important question, then, is whether or not other approaches will perform better when predicting interfaces that are distinct from the most frequently observed ones. To determine this, we calculated the average precision and recall for the 26 cases where PredUs made bad predictions (both precision and recall are less than random). They were quite low (<10%, see Table 3-3) suggesting that the interfaces to be predicted in these cases are indeed distinct from that most frequently observed. While the other methods used in this study performed better

for these cases, only cons-PPISP made predictions that on average were even slightly better than random, suggesting that these interfaces are not only geometrically distinct, but also distinct in terms of the residue characteristics typically used to describe protein-protein interaction sites. Hence, there seems to be little cost to using the most frequently observed interface, at least compared to other approaches. Moreover, for the 125 cases where a successful prediction was made, using structure resulted in a significant increase in performance (Table 3-3).

Our results have implications for how structural information may be used to analyze and characterize protein-protein interactions, especially on a large-scale. Although there may be increased variability in the geometric binding properties of pairs of proteins with increasingly remote relationships, structural similarity can be effectively used to identify the sites of protein-protein interaction. As long as structural information is available for a given pair of proteins, the accuracy of our predictions suggests that the set of "template complexes" available in the current structural databases can be used to generate coarse-grained models of protein-protein interactions. Most importantly, we see that using remote structural neighbors produces a significant improvement in recall, which suggests that remote structural relationships have the potential to yield a much large number of hypotheses for protein-protein interactions than has been previously possible (Lu, Lu et al. 2002; Aloy, Bottcher et al. 2004; Davis,

Braberg et al. 2006). Together these results suggest that the use of remote structural similarity can potentially significantly increase the number of functional relationships that can be detected, modeled and evaluated.

3.5 Materials and Methods

Protein dataset and interface definition. We used a set of proteins originally created to evaluate protein docking methods by Hwang *et. al.* (Hwang, Pierce et al. 2008). This dataset was designed to have significant diversity in both overall protein shape and binding mode and has been used by other groups to evaluate protein interface prediction methods (Liang, Zhang et al. 2006; Zhou and Qin 2007). The benchmark contains 124 pairs of interacting structures, and 309 protein chains. We created a non-redundant set at 40% sequence identity using the program cd-hit (Li and Godzik 2006) and also removed chains shorter than 50 amino acids. This left 188 individual protein chains as our test dataset, coming from 137 SCOP families, 124 superfamilies, and 105 folds. The interface in each case is determined based on its interactions with all other members of its associated complex in PQS. A residue was defined to be on the surface if its solvent accessible surface area (calculated using the isolated chain) was $\geq 10\text{Å}^2$, and it was defined to be in the interface if the distance between any of its heavy atoms and any heavy atoms from a partner chain was ≤ 5 Å (Zhou and Qin 2007). In total, the 188 chains contained 39,780 residues and 7,496 in an interface. We

also tested our interface prediction method on targets T01~T27 from the Critical Assessment of Prediction of Interactions (CAPRI, (Janin and Wodak 2007)). These 56 bound/unbound chains contain 12,124/12,181 residues with 2,180/2,134 in the interface.

Structural neighbors. Structural neighbors were defined in two ways. Structural neighbors belonging to the same family, superfamily or fold were taken from the SCOP 1.73 database (Andreeva, Howorth et al. 2008). We also used the program Ska (Yang and Honig 2000; Petrey and Honig 2003), to identify neighbors independent of classification. Neighbors were defined based on a protein structural distance (PSD) (Yang and Honig 2000) from the query of less than 0.6. In the procedures described below, only structural neighbors that are involved in any PQS complex (36,888 as of Aug. 2009) are used and if a structural neighbor has multiple binding partners, all are considered. The complete PQS database was used to identify structural neighbors, but to avoid overcounting of highly similar complexes, we applied the following procedure: PQS chains were clustered using cd-hit at a 40% sequence identity cutoff. Given structural neighbors N_1 and N_2 of a protein and their interacting partners P_1 and P_2, if N_1 belongs to the same cluster as N_2, and P_1 belongs to the same cluster as P_2 only one structural neighbor/partner would be considered.

Z-score to evaluate interface conservation. To evaluate the degree of interface conservation, we used a variant of the statistical test introduced by Russell *et al.* (Russell, Sasieni et al. 1998) in an analysis of interactions between proteins and small molecules. For each query protein, Q, and each structural neighbor N, the interactions N makes with its partner, P, are "mapped" to the surface residues of Q to create the *contact map* for this particular structural neighbor. This procedure is repeated for all structure neighbors of Q and the contact maps are then summed to form the *contact frequency map* (see Figure 3-4 for detail).

We then ask whether or not there is a statistically significant set of residues on the surface of the query protein that preferentially has interaction sites mapped to it. Following Russell *et al.* (Russell, Sasieni et al. 1998) the statistical significance is determined by counting the number of times any pair of contact maps overlap at a residue. This can be calculated as

$$T = \sum_{i=0}^{|S|} \frac{i(i-1)O_i}{2}$$

where $|S|$ is the number of structural neighbors, O_i is the number of surface residues in the query which interact with i structural neighbors. It was shown in (Russell, Sasieni et al. 1998) that this number is statistically equivalent to:

$$X = \sum_{i=0}^{|S|} (i-a)^2 O_i$$

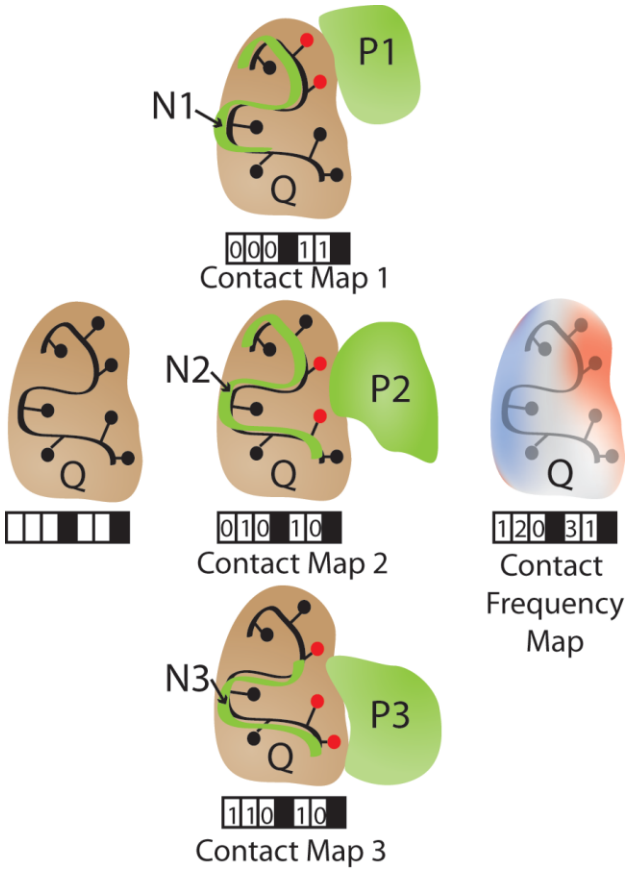


Figure 3-4. Calculating the contact map and contact frequency map. In the above example, a given query protein (Q, brown) with 7 residues has 5 residues on the surface. Structural neighbors (Ni, green lines) involved in protein complexes are superimposed on Q and the same transformation is applied to their interacting partners (Pi, green surfaces). Whenever a heavy atom from a residue of Pi is <5 angstroms of an atom of a surface residue of Q after applying the transformation, that residue is marked (red circles), generating a "contact map" for each structural neighbor (black boxes represent non-surface residues that are not included). The "contact frequency map" is generated by summing the individual contact maps.

X represents bias in the distribution of the O_is. To measure the statistical significance of X for a given query protein we calculate an approximate pivotal independent of the number of structural neighbors and the number of contacted residues:

$$Z = \frac{\sum_{i=0}^{|S|} w_i (O_i - E_i)}{(\sum_{i=0}^{|S|} w_i^2 E_i - \sum_{i=0}^{|S|} \sum_{j=0}^{|S|} w_i w_j E_i E_j / N)^{1/2}}$$

where $w_i = (i-a)^2$, and E_i is the expected value of O_i under the assumption that the contact maps are randomly distributed over the surface of the query protein (calculated as described below). This score then essentially indicates the chance of observing the value X and can be used to evaluate degrees of interface conservation (please refer to (Russell, Sasieni et al. 1998) for detail). The larger the Z-score, the more significant the conservation will be.

We estimated the values of E_i for each query protein by simulation. For each contact map generated for a structural neighbor of the query, we constructed a corresponding random surface patch that has the same number of contacting atoms using the subroutine MAKE_REGION of the program MODELLER (Sali and Blundell 1993). This is repeated 100 times and E_i is taken to be the average of the O_i's generated in each run. Ideally, the simulation should be done that each contact map and its random maps have the same number of residues. We

compared the Z-scores from simulation of the same number of atoms and the same number of residues and found little difference. Because the generation of random maps with the same number contacting residues will take much more time, we generate random maps of the same number of contacting atoms in our simulation.

Using conservation to predict interfaces. We exploited the observed conservation to develop an interface prediction method. Given a query structure, we first identified its structure neighbors using Ska, and kept only the 50 most similar neighbors that were also contained in complexes (for benchmarking purposes, complexes that contain the query protein were excluded). We calculated the contact frequency map as described above and turn the contact frequencies into residue-based *interfacial scores* using a logistic function:

$$\varsigma = \frac{1}{1 + e^{\frac{-f + \max(f)/2}{\max(f)/10}}}$$

Here f is the contacting frequency of a residue, and max(f) is its maximum value for the whole structure. We chose an interfacial score cutoff of 0.05 since this results in 20-25% of residues being predicted as interfacial (roughly the portion of interface residues in our datasets). Prediction accuracy is assessed in terms of recall=N_c/N_i and precision=N_c/N_p where N_c= the number of correctly predicted interface residues, N_i= the number of real interface residues, and N_p= the total

number of predicted interfacial residues. When comparing our approach to other methods, we used the web services Promate (http://bioinfo.weizmann.ac.il/promate/many.html), and obtained the cons-PPISP and PINUP from the developers and ran them locally.

Acknowledgements

This work is supported by NIH grants GM030518, GM074958 and CA121852. We sincerely thank Yaoqi Zhou for the protein binding site prediction program PINUP, and Huan-Xiang Zhou for cons-PPISP. We thank Brian Chen for the protein surface collision detection program SurfaceExtractor, and Yulei Zhang for helpful discussion on statistics.

CHAPTER 4. PredUS: A WEB SERVER FOR PREDICTING PROTEIN INTERFACES USING STRUCTURAL NEIGHBORS

The following chapter is a paper published in the *Nucleic Acids Research* (Volume 39, Web Server Issue, 23 May 2011, pp. W283-W287).

4.1 Abstract

We describe PredUs, an interactive web server for the prediction of protein-protein interfaces. Potential interfacial residues for a query protein are identified by "mapping" contacts from known interfaces of the query protein's structural neighbors to surface residues of the query. We calculate a score for each residue to be interfacial with a support vector machine. Results can be visualized in a molecular viewer and a number of interactive features allow users to tailor a prediction to a particular hypothesis. The PredUs server is available at: http://wiki.c2b2.columbia.edu/honiglab_public/index.php/Software:PredUs.

4.2 Introduction

Prediction of the potential locations at which proteins interact with other proteins is essential to understanding their function and has been successfully exploited in many applications, including identification of an approximate binding

mode in protein-protein docking, as a guide in site-directed mutagenesis and in the identification of pharmacological targets. Approaches to interface prediction typically depend on the recognition of differences in the properties of amino acids (e.g., residue hydrophobicity and sequence conservation) in surface patches that interact with other molecules, as compared to other surface residues (Tsai, Lin et al. 1996; Jones and Thornton 1997; Lo Conte, Chothia et al. 1999; Zhou and Qin 2007; de Vries and Bonvin 2008; Tuncbag, Kar et al. 2009).

"Template-based" prediction, in which an interface for a given query protein is inferred based on some similarity to another protein or set of proteins with known interfaces has been less extensively used. This is especially true of remote similarities which may be due to the lack of data about conservation of the location of binding sites in remote neighbors. Recently, we reported a comprehensive analysis of the degree to which the location of a protein interface is conserved in sets of proteins that share varying degrees of similarities (Zhang, Petrey et al. 2010). Our results showed that while, in general, interface conservation is most significant among close neighbors, it is still significant even for remote structural neighbors. Based on this observation, we implemented a template-based protein interface prediction method and tested it on a docking benchmark and a set of CAPRI targets. Our method offered the best combination of prediction precision and recall among all methods tested, including PINUP

(Liang, Zhang et al. 2006), cons-PPISP (Chen and Zhou 2005), and ProMate (Neuvirth, Raz et al. 2004), which were suggested to be the top three standalone protein interface prediction programs in a recent comparative study of six interface prediction methods (Zhou and Qin 2007).

Here we describe PredUs, an interactive web server using this template-based protein interface prediction method. Given a query protein structure as input, we "map" interaction sites of structural neighbors involved in a complex to residues on the surface of the query. Based on the mapped contacting frequencies, we calculate a score for residues to be interfacial. In the version of our method implemented on the server we use a support vector machine (SVM) to calculate the score, which shows superior performance compared to the original score based on logistic regression (Zhang, Petrey et al. 2010) on the same benchmarks.

4.3 PredUs Algorithms

Given a protein structure, we first find its structural neighbors using the structural alignment program Ska (Petrey and Honig 2003). We use a PSD (protein structure distance, a measure of structural similarity (Yang and Honig 2000)) cutoff of 0.6 which allows detection of both close and remote relationships. Structures that are involved in a PQS (Protein Quaternary Structures, (Henrick and Thornton 1998)) or PDB (Protein Data Bank, (Berman, Westbrook et al. 2000)) complex are kept and ranked by structural alignment score, (Kolodny,

Koehl et al. 2005), which reflects a combination of structural similarity and alignment length.

An interface from a structural neighbor is "mapped" to the query by placing any interacting partners of the structural neighbor in the coordinate system of the query, using the transformation that relates the structural neighbor to the query. If a heavy atom of a query residue is within 5.0 angstroms of an interacting partner after the transformation, we increment a counter associated with this residue with the sequence identity between the query and the structural neighbor. This is repeated for each structural neighbor ordered according to its structural alignment score. To avoid over counting of highly similar interfaces, we cluster PQS/PDB chains using cd-hit (Li and Godzik 2006) at 40% sequence identity cutoff. If two structural neighbors belong to a single cluster and their interacting partners also belong to a single cluster, only the structural neighbor with the higher structural alignment score will be considered. We sum the weighted contact frequencies at each residue of the query after interfaces of all structural neighbors have been mapped (see reference (Zhang, Petrey et al. 2010) for details).

In the current version of the PredUs server, we use a support vector machine (SVM) to predict whether or not a surface residue is in an interface. The SVM is implemented with the package libsvm 3.0 using radial basis function as

the kernel. For each surface residue, we define a patch that includes the residue and its 14 spatially nearest surface residues. The contacting frequencies (freq) and solvent accessible surface areas (ASA) of the residues in the surface patch and the maximum contacting frequency of residues of the entire protein constitute a feature profile of length 31, *i.e.* [$freq_{max}$, $freq_0$, $freq_1$, …, $freq_{14}$, ASA_0, ASA_1, …, ASA_{14}]. These profiles are used as the input to the SVM and are mapped to vectors of a high-dimensional space using the kernel function. The SVM attempts to construct a hyperplane in that space that separates the vectors associated with interfacial residues from those that are non-interfacial. The interfacial score reflects the distance above (positive score) or below (negative score) this hyperplane. The higher the score the more likely a given residue is to be in an interface. By default, PredUs predicts all residues with positive score to be interfacial, but this cutoff is adjustable by the user.

4.4 PredUs Features

Input to the PredUs web server can be a protein structure file in PDB format, or a PDB code. PredUs will check the validity of the input structure, and once confirmed, submit it for prediction. Users can submit multiple structures, and provide a job title or email address to facilitate retrieval of results.

As a unique feature, PredUs allows users to specify the structure of the binding partner. Once users provide another structure file or PDB code as

"Partner Structure", PredUs will predict the interface specifically used in the binding of the provided partner by only mapping the interfaces between structural neighbors of the query protein and structural neighbors of the partner.

A typical prediction takes a few minutes and almost all complete in no more than 30 minutes. The output consists of a list of residues and their associated

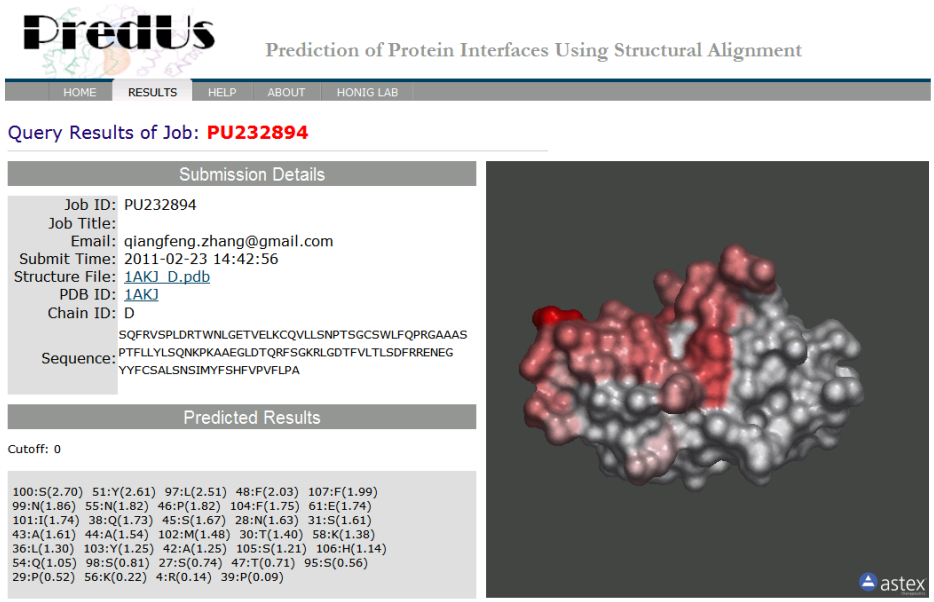


Figure 4-1. PredUs prediction output. The left of the figure shows the submission details and prediction results. All residues with interfacial score higher than 0 are shown with scores in parentheses following residue number (in the PDB structure file) and residue name. On the right is the submitted structure with its molecular surface rendered in colors according to residue interfacial score. Residues of score higher than 0 are shown from light red to red as the score increases.

score to be in an interface for each submitted structure which can be downloaded in text format. Individual predictions can be visualized in the molecular viewer AstexViewer (Hartshorn 2002) by following the "View Structure" link. Surface residues are rendered in different colors according to their predicted interfacial score (Figure 4-1).

Another unique feature of PredUs is that users can tailor a prediction to a particular hypothesis following the "Interactive prediction" link. Figure 4-2 shows structure-based sequence alignments between the query protein (on the top) and its structural neighbors on which the prediction is based. Below the alignment are tools that allow users to filter structural neighbors based on functional information including GO terms (Ashburner, Ball et al. 2000), or SCOP (Lo Conte, Ailey et al. 2000), PFAM (Finn, Mistry et al. 2010), and InterPro (Apweiler, Attwood et al. 2001) categories. It is well known that proteins can interact with different partners at distinct regions of their surfaces and these different interfaces can be associated with different functions (Keskin, Gursoy et al. 2008). By default, however, PredUs will map all interfaces of structural neighbors of a query protein without regard to sequence or functional relationships. Hence default predictions are indications of all possible places where the query may interact with other proteins and may initially be overly broad. Restricting the set of structural neighbors via filters to include only close

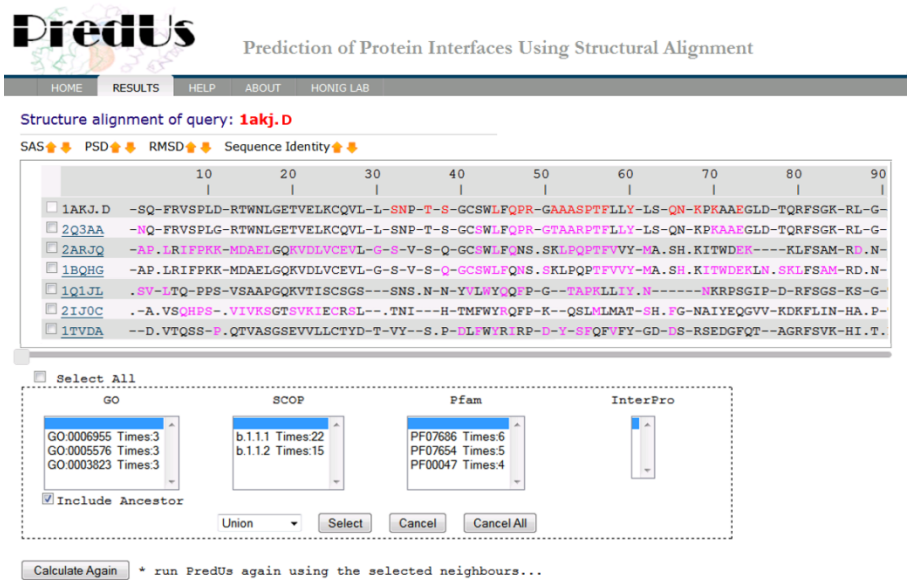


Figure 4-2. PredUs interactive prediction. The figure shows the structure-based sequence alignments of a query protein and its structural neighbors. Predicted interfacial residues in the query sequence are colored in red and the actual interfacial residues in the structural neighbors are indicated in purple. Functional terms populated in the set of structural neighbors are shown below the alignments. These can be used as functional filters to generate function-specific predictions by clicking the "Calculate Again" button. Gaps are shown as dashes. For brevity, insertions of more than one residue with respect to the query are shown as dots.

sequence neighbors, for example, or remote homologs that are associated with a specific function should in many cases produce a more accurate prediction.

On this page, users can also reorder the set of structural neighbors using different ranking operators shown above the alignments. Structural neighbors can

be ranked based on four scores: structural alignment score, the default; PSD; RMSD (root mean square deviation, based on aligned residues); and SID (sequence identity). With the different operators, users can compare predicted interfacial residues to real interfacial residues in structural neighbors ranked by different similarity measurements.

The query protein can be further analyzed in our protein function annotation server MarkUs (Petrey, Fischer et al. 2009) provided by the link "MarkUs Annotation". Interfaces predicted by PredUs can be examined in MarkUs and comparatively studied with other functional properties like ligand binding sites, enzymatic active sites and other residue and surface features, across a wide range of sequence and structural similarities.

4.5 PredUs Benchmarks

We used protein docking benchmark dataset of 188 chains in training and testing PredUs. As an independent test, we also used a set of CAPRI targets that contains 56 chains in both bound and unbound forms. Please see reference (Zhang, Petrey et al. 2010) for a detailed description of the datasets.

To assess the predictions, we calculated a variety of quantities:

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP})$$

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN})$$

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FP} + \text{TN} + \text{FN})$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Here TP, FP, TN, FN are true positive, false positive, true negative, false negative predictions; MCC is Matthews's correlation coefficient. We also drew the receiver operating characteristic (ROC) curve and calculated the area under the curve (AUC).

We used 10-fold cross validation to test PredUs on the protein docking benchmark dataset. We tested the prediction performance of the SVM in terms of AUC value using different surface patch sizes ranging from 3 to 25 and found that the best performance was achieved with a 15-residue patch. No structural and functional filters were applied in benchmarking. All quantities except AUC were calculated using an interfacial score cutoff of 0 (in principle, a score higher than 0 means the residue is more likely to be in an interface). These are also default settings in the PredUs server.

As shown in Table 4-1, PredUs can achieve a high prediction precision and recall at the same time and achieves superior performance compared to our original study (Zhang, Petrey et al. 2010) as a result of the use of the SVM

Table 4-1. PredUs prediction performance averages on the docking benchmark dataset (DKBM3) and CAPRI bound/unbound targets. Quantities in each column are defined in the description of the PredUs benchmarks in the main text.

dataset	precision	recall	accuracy	AUC	MCC	F1
10 fold Cross-validation						
DKBM3	50.3%	57.5%	72.6%	0.739	0.345	0.530
Independent test						
CAPRI bound	43.0%	53.0%	72.1%	0.713	0.290	0.474
CAPRI unbound	43.3%	53.6%	73.2%	0.729	0.304	0.479

classifier. In the current version of PredUs, we achieve a precision and recall of 50% and 58%, compared to 44% and 46% using the original scoring scheme. Here and in the following test of CAPRI targets, we only compare with the original algorithm, which had been shown to offer the best combination of precision and recall among other methods we tested, including PINUP, cons-PPISP, and ProMate (Zhang, Petrey et al. 2010).

The SVM classifier trained on the whole docking benchmark set was applied to the CAPRI test sets. The results are summarized in Table 4-1 and the performance was again improved (prediction precision and recall are 43%/43% and 53%/54% vs. 42%/40% and 42%/45% in the original prediction for bound/unbound targets respectively, (Zhang, Petrey et al. 2010)).

4.6 Discussion

PredUs predicts protein interfaces by mapping binding sites from structural neighbors. In contrast to methods based on residue properties, such as hydrophobicity and conservation, an advantage of this type of direct mapping is that it allows the identification of interfacial residues that are less distinctive in terms of such properties. This can be seen from the much higher recalls of the PredUs server than other protein interface prediction methods (Table 4-1 and reference (Zhang, Petrey et al. 2010)). This type of mapping also seems to be insensitive to conformational changes that may occur upon binding, as can be seen from the small difference between the performances of PredUs on the bound and unbound CAPRI targets (Table 4-1).

The choice of structural neighbors is an important issue affecting the performance of template-based approaches and it might be expected that restricting the set of structural neighbors to closely related sequence homologs may produce more biologically relevant results. We have shown previously (Zhang, Petrey et al. 2010) that while such a limitation improves predictive accuracy it decreases the recall at the same time. As seen in Table 4-2, a general trend is that the number of cases for which we can make predictions and also the prediction recall improves as more remote neighbors are include with little sacrifice in precision. Consequently, the prediction strategy implemented in

Table 4-2. PredUs prediction performance averages when using structure neighbors from the same and different SCOP groupings on the docking benchmark dataset. Quantities in each column are defined in the description of the PredUs benchmarks in the main text.

prediction methods	cases	precision average	recall average
PredUs(server)	**185**	**50.3%**	**57.5%**
PredUs(original)	185	43.6%	45.7%
family	141	50.8%	33.8%
superfamily	147	45.9%	36.2%
fold	153	41.8%	38.7%

PredUs is to use the widest range of structural neighbors by default, since this appears to provide the best indication of the possible binding sites on a given protein. To limit the set of structural neighbors to those that a user thinks might be more biologically relevant, they can then apply the different evolutionary, structural and functional filters, or specify a binding partner, as well as directly compare actual interfacial residues in the structural neighbors to the predictions.

A limitation of PredUs is that, for every query protein, structural neighbors in a complex are required to make predictions. By exploiting remote structural homology, however, this limitation is small with only about 5% the proteins in our benchmark having no structural neighbors with binding partners,

and this percentage should continue to decrease as more protein-protein complexes are characterized structurally.

PredUs has been set up for half a year and has been tested extensively. In an application of genome-wide modeling of protein-protein interactions, we have used it to predict interfaces for all proteins with structural information in the yeast and human proteomes.

Funding

This work was supported by National Institutes of Health [grant numbers GM030518, GM095315, and CA121852]. National Natural Science Foundation of China [60873040]; and Shuguang Scholar Program of Shanghai Education Development Foundation. Funding of open access charge was supported by Howard Hughes Medical Institute.

CHAPTER 5. STRUCTURE-BASED PREDICTION OF PROTEIN-PROTEIN INTERACTION ON A GENOME-WIDE SCALE

5.1 Introduction

The genome-wide identification of pairs of interacting proteins is an important step in the elucidation of cell regulatory mechanisms (Bonetta 2010; Vidal, Cusick et al. 2011). Much of our current knowledge derives from high-throughput techniques such as Yeast Two Hybrid and Affinity Purification (Shoemaker and Panchenko 2007), as well from manual curation of experiments on individual systems (Reguly, Breitkreutz et al. 2006). A variety of computational approaches based, for example, on sequence homology, gene co-expression, and phylogenetic profiles have also been developed for the genome-wide inference of PPIs (Salwinski and Eisenberg 2003; Shoemaker and Panchenko 2007). Yet, comparative studies suggest that the development of accurate and complete repertoires of protein-protein interactions (interactomes) is still in its early stages (Deane, Salwinski et al. 2002; von Mering, Krause et al. 2002; Braun, Tasan et al. 2009).

To date, structural information has had relatively little impact in constructing protein-protein interactomes, primarily because there is a dramatic

difference between the number of proteins with known sequence and those with an experimentally known structure. For example, the PDB (Protein Data Bank) provides structures for ~600 of the total complement of ~6,500 yeast proteins (~10%), while structural coverage of protein-protein complexes is even more sparse with only about 300 structures available out of the approximately 75,000 PPIs (<0.5%) recorded in databases. Fortunately, however, ~3,600 additional yeast proteins have homology models in either the ModBase (Pieper, Eswar et al. 2006) or SkyBase (Mirkovic, Li et al. 2007) databases. Moreover, as of early 2010, there were about 37,000 protein-protein complexes taken from multiple organisms in the PDB and PQS (Henrick and Thornton 1998) (Protein Quaternary Structure) databases, that might be used to model PPIs. Clearly, if structure is to be useful on a large scale, it is essential that modeling of individual proteins and of complexes be exploited.

A number of studies have used structurally characterized complexes as "templates" to construct models of complexes that might be formed between proteins that have obvious sequence and/or structural relationships to the proteins in the template (Aloy and Russell 2002; Lu, Lu et al. 2002; Davis, Braberg et al. 2006). But this requirement inevitably limits the number of interactions that may be inferred. The alternative strategy adopted here is not to limit ourselves to proteins that have been classified as sequence or structurally related (for example

proteins in the same SCOP family, superfamily, or fold) but rather, to search more broadly for templates identified from geometric relationships between groups of secondary structure elements as revealed by structural alignment, independently of how they are classified. It has been demonstrated that even distantly related proteins often use regions of their surface with similar arrangements of secondary structure elements to bind to other proteins (Petrey, Fischer et al. 2009; Gao and Skolnick 2010; Zhang, Petrey et al. 2010), suggesting the possibility of significantly expanding the number of putative PPIs that can be identified.

Here we show that three-dimensional structural information can be used to predict PPIs with an accuracy and coverage that are superior to predictions based on non-structural evidence. Moreover, combining structural information with other functional clues yields predictions of comparable quality to high-throughput experiments. The surprising effectiveness of three-dimensional structural information can be attributed to the use of homology models combined with the exploitation of both close and remote geometric relationships between proteins. Our results suggest that structural biology and molecular systems biology can be integrated at an extent that has not been possible in the past.

5.2 Methods

Our approach to the prediction of PPIs is embodied in an algorithm we have named PREPPI (Predicting Protein-Protein Interactions) that combines

structural and non-structural interaction clues using Bayesian statistics (see Figure 5-1 and Supplementary Materials and Methods for details). The structural component of PREPPI involves a number of steps. Briefly, given a pair of query proteins (QA and QB), we first use sequence alignment to identify structural representatives for each (MA and MB) and then use structural alignment to find the set of both close and remote structural neighbors (NA_i and NB_j) of MA and MB (an average of ~1500 neighbors are found for each structure). Whenever two (e.g. NA_1 and NB_3) of the over 2 million pairs of neighbors of MA and MB form a complex reported in the PDB, this defines a template for modeling the interaction of QA and QB. Models of the complex are created by superimposing the representative structures on their corresponding structural neighbors in the template (i.e., MA on NA_1 and MB on NB_3). Using this procedure, we built structural models for about 2.4 million potential binary interactions involving about 3,900 proteins of yeast and about 36 million interactions involving about 13,000 proteins of human (for a given interaction, there are on average 200 models for yeast or 300 models for human).

The approach we take to scoring these models is central to our entire strategy. Although our procedure produces a three dimensional model for every putative complex, we never actually evaluate the model itself with standard scoring functions (for example as used in docking (Wass, Fuentes et al. 2011)),

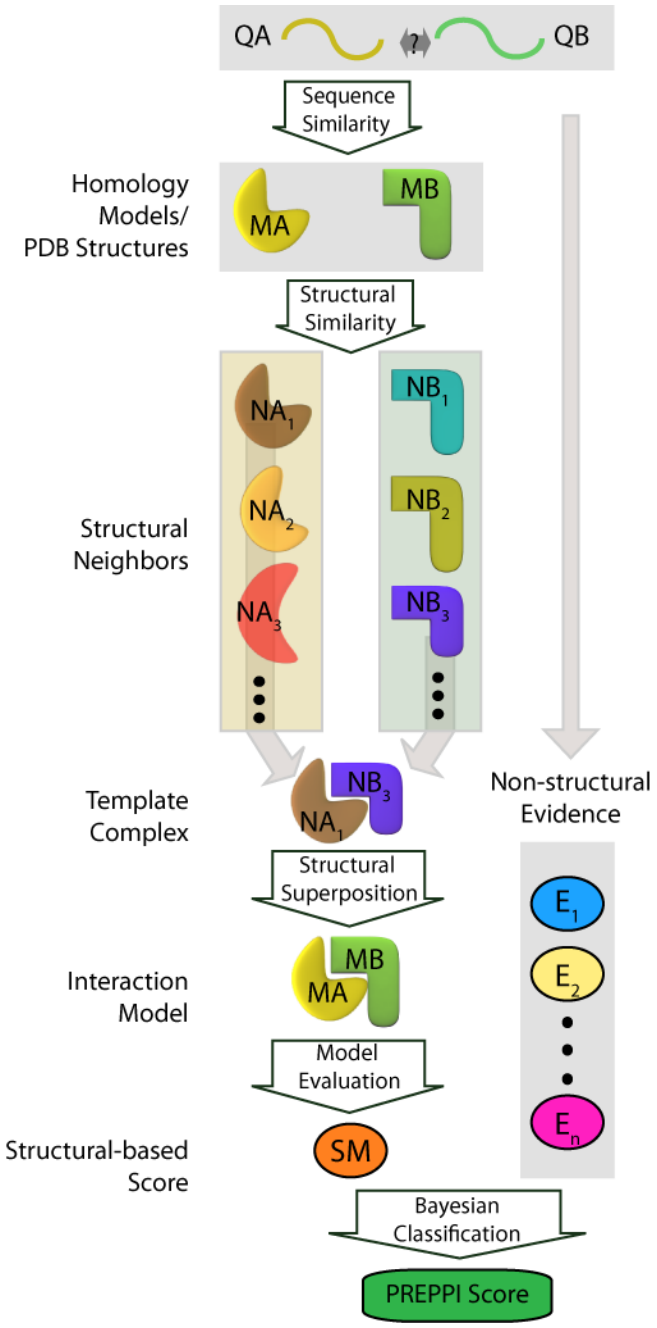


Figure 5-1. Predicting protein-protein interactions using PREPPI. Given a pair of query proteins that potentially interact (QA, QB), representative structures for the individual subunits (MA, MB) are taken from the PDB, where available, or from homology model

databases. For each subunit we find both close and remote structural neighbors. A "template" for the interaction exists whenever a PDB or PQS structure contains a pair of interacting chains (e.g. NA1-NB3) that are structural neighbors of MA and MB, respectively. A model is constructed by superposing the individual subunits, MA and MB, on their corresponding structural neighbors, NA1 and NB3. We assign five empirical structure-based scores to each interaction model (Figure S5-1) and then calculate an informative likelihood for each model to represent a true interaction by combining these scores using a Bayesian Network (Figure S5-2) trained on the HC and the N interaction reference sets. We finally combine the structure-derived score (SM) with non-structural evidence associated with the query proteins (e.g., co-expression, functional similarity) using a naïve Bayesian classifier.

since the binding mode of the two interacting proteins may not be accurately reproduced. Rather, we use a set of five empirical scores (described in Figure S5-1) that measure properties that are only weakly dependent on atomic detail. The first score, a), depends on the structural similarity between models of the two query proteins (i.e. MA and MB) and those in the template complex (i.e. NA_1 and NB_3). The next two scores determine whether the interface in the template complex actually exists in the model. They are calculated as b) the number and c) the fraction of interacting residue pairs in the template (e.g. NA_1-NB_3) that align to some pair of residues in the model (MA-MB). The final two scores reflect whether the residues that appear in the model interface have properties consistent with those that mediate known PPIs (e.g., residue type, evolutionary conservation,

or statistical propensity to be in protein-protein interfaces). This information is obtained from three publically available servers that predict interfacial residues based on the sequence and structure of the individual subunits of the model (Chen and Zhou 2005; Liang, Zhang et al. 2006; Zhang, Petrey et al. 2010). The scores are calculated as d), same as b) with the additional requirement that both residues in an interacting pair of the template align to predicted interfacial residues in MA and MB; and e) the number of template interfacial residues that align to predicted interfacial residues in MA and MB.

These scores are combined using a Bayesian network (Figure S5-2) to assign a likelihood ratio (LR, see Supplementary Materials and Methods) that each candidate protein-protein complex represents a true interaction. The network is trained on positive and negative "gold standard" reference datasets. Similar to two recent studies of the yeast and human B-cell interactomes (Yu, Braun et al. 2008; Lefebvre, Rajbhandari et al. 2010), we combine interaction data from multiple databases (73,787 PPIs for yeast and 58,772 for human, Table S5-1) to ensure the broadest coverage of true interactions in the positive reference set. We divide these sets into high-confidence (HC) and low-confidence (LC) subsets; the HC sets contain 11,851 yeast interactions and 7,409 human interactions which have more than one publication supporting their existence and the other

interactions with only one supporting publication compose the LC set. All interactions *not* in the HC+LC set form the negative (N) reference set.

5.3 Results

Figure 5-2A shows an example how an HC set interaction of serine/threonine-protein kinase D1 (PKD1) and protein kinase C epsilon type (PKCε) is recovered using homology models and remote structural relationships. Homology models of PKD1 and PKCε are superimposed on template structures taken from a crystal structure of an E2 enzyme/ubiquitin complex to produce a model of the PKD1/PKCε complex. That is, two proteins in the ubiquitin pathway (not kinases) are being used here to predict a PPI between two kinases. Note that PKD1 and PKCε are not sequence homologs of the two corresponding ubiquitin pathway proteins and are classified as belonging to different folds. The two kinases do however share some local structural similarity with their respective templates as is evident from the figure. The model interface covers the template interface quite well and contains many residue pairs independently predicted to be interfacial. As a result, the interaction model has significant PREPPI scores and indeed has an LR of 130.

To quantitatively assess the performance of structural modeling (SM), we compared it with a number of different clues previously used in the literature to infer PPIs (Jansen, Yu et al. 2003; von Mering, Jensen et al. 2005; Lefebvre,

Rajbhandari et al. 2010): a) essentiality of the proteins in the interacting pair; b) co-expression level; c) Gene Ontology (GO) functional similarity; d) MIPS functional similarity; and e) phylogenetic profile similarity. We developed our own phylogenetic profile algorithm and used the same algorithms or data for other clues as Gerstein and coworkers (Jansen, Yu et al. 2003) (see details in Supplementary Materials and Meth and Table S5-2).

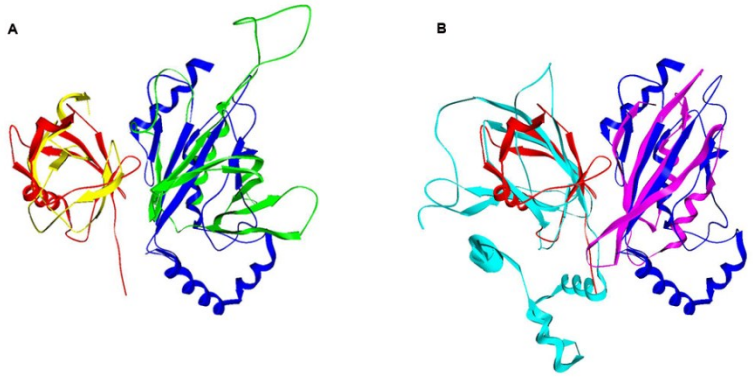


Figure 5-2. Models for the PPI formed between (A) PKD1 and PKCε, and (B) EF-1δ and pVHL using homology models and remote structural relationships. The same E2-ubiquitin template complex (PDB code: 2fuh A and B chain, shown in blue and red respectively) was used in both cases. The structures of PKD1 and EF-1δ (shown in green and purple) are homology models from ModBase; the structure of PKCε (yellow) is a homology model from SkyBase; the structure of pVHL (cyan) is from PDB (1lm8 V chain). In each case, the relevant homology models are structurally superimposed on one of the two templates in the E2-ubiqutin complex.

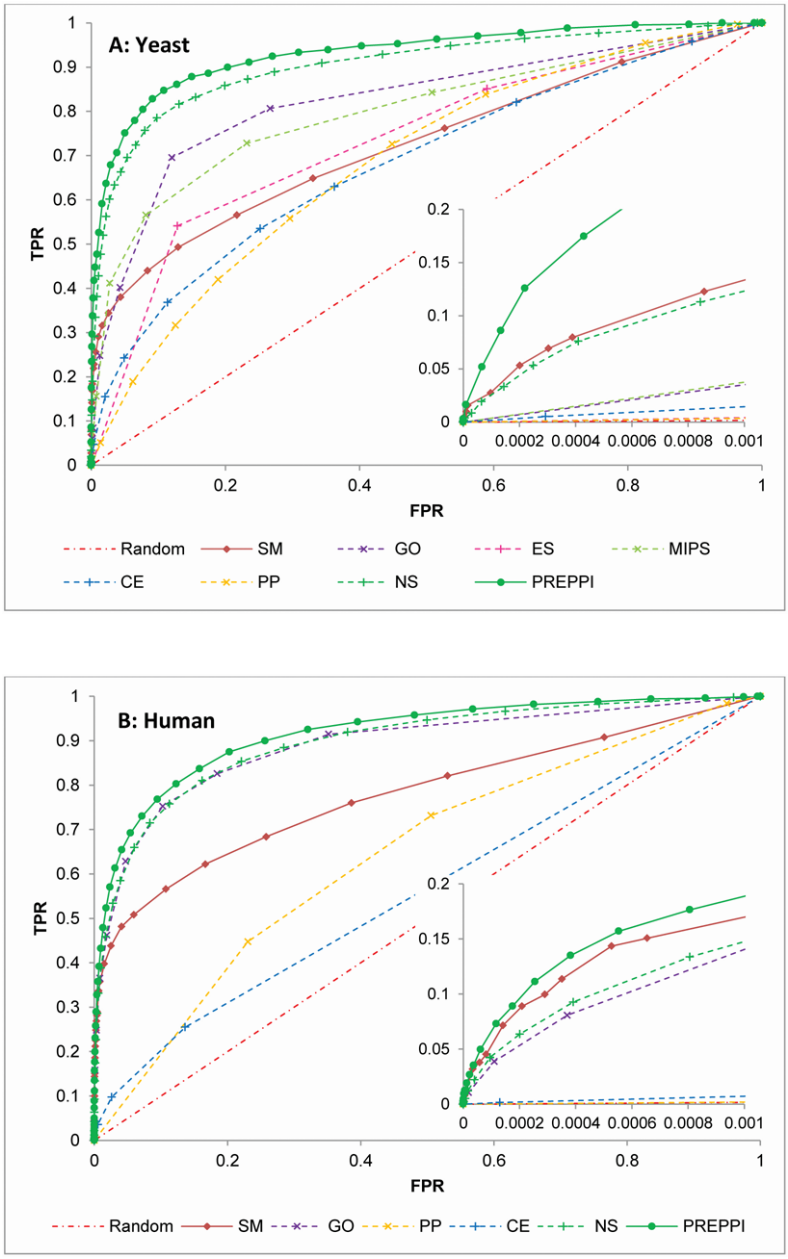


Figure 5-3. Receiver operating characteristic (ROC) curves for PPI prediction based on different clues and their combinations for yeast (A) and human (B). The curves here are calculated for subsets of protein pairs for which the indicated clues are available. ROC curves

for the full yeast and human proteomes and for subset of proteins for which structures and/or models are available are shown in Figure S5-3. The clues used are structural modeling (SM), GO and MIPS term similarities, protein essentiality (ES) relationship, co-expression (CE) and phylogenetics profile (PP) similarity. NS refers to Bayesian classifiers derived from non-structure-based clues (GO, MIPS, ES, CE and PP for yeast; and GO, CE and PP human). PREPPI combines these clues with structural modeling. The inset in each figure magnifies the curves in the low false positive rate (FPR) region.

Figure 5-3 presents ROC (receiver operating characteristic) plots of true positive rate (TPR) vs. false positive rate (FPR) for the yeast and human proteomes (results for yeast interaction were from 10-fold cross validation, for human interactions they were derived using the Bayesian network trained on yeast although virtually identical results were obtained with a cross validation on human data). As can be seen from the figure, SM yields comparable performance to other clues over the entire range of FPR but is considerably more effective at low FPR (see insets to Figures 5-3 and Figure S5-3). This is critical because the latter is the only range where predictions can be used effectively. Due to the very large number of possible PPIs, only very low FPR rates (e.g. FPR $\leq$ 0.1%) can produce an acceptable number of false positives. At low FPR, SM by itself outperforms even the naïve Bayesian classifiers that combine all non-structure-based clues (NS). Each curve in Figure 5-3 is based on the subset of yeast protein pairs for which data are available for the corresponding clue, but our conclusions

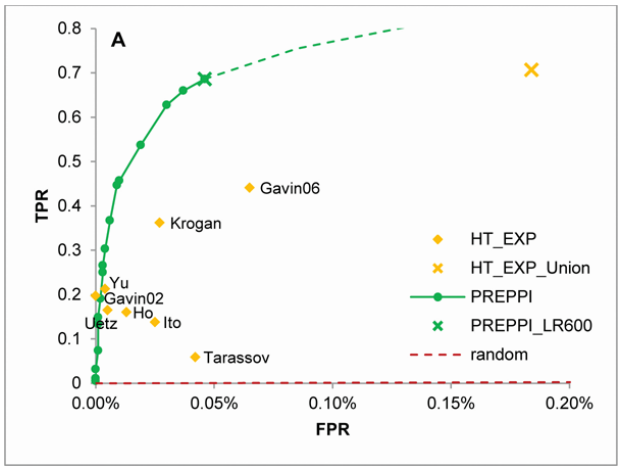
remain the same independent of the data sets used to evaluate the predictions (Figure S5-4). Moreover, the definition of the N set results in significant overestimates for the computed false positive rate as any new correctly predicted interactions will be, by definition, in the negative reference set. Indeed, looking specifically at the thousands of SM predictions of high LR (>600) in the LC and the N sets, about 70% and 50%, respectively, of them share GO biological term at, or more specific than, the 6^{th} level of the GO hierarchy, suggesting that these interactions may be real (Figure S5-5).

As mentioned above, PREPPI combines structural and non-structural clues using a naïve Bayesian network (Jansen, Yu et al. 2003; von Mering, Jensen et al. 2005; Lefebvre, Rajbhandari et al. 2010). It is evident from the figure that PREPPI's performance is superior to other methods over the entire range of false positive rates, with its performance at low FPRs, the most critical range, being due primarily to the inclusion of structural information (insets in Figure 5-3). As an independent test of PREPPI, we assessed its performance against one of the challenges in the 2009 DREAM (Dialogue for Reverse Engineering Assessments and Methods) workshop specifically aimed at protein-protein interaction predictions (Stolovitzky, Prill et al. 2009). As discussed in Table S5-3, PREPPI outperformed all other methods for cases where structural information is available.

In addition to comparisons to other computational predictions based on non-structural evidence, we have also compared the performance of PREPPI to that of high-throughput (HT) experimental techniques (Table S5-4). A detailed comparison of different HT techniques was reported by Vidal and coworkers (Yu, Braun et al. 2008). We used both their CCSB-BGS (Center for Cancer Systems Biology Binary Gold Standard, ~1,300 PPIs) and the CCSB-PRS (CCSB Positive Reference Set, a subset of CCSB-BGS of 188 highly reliable PPIs) datasets as definitions of true interactions and compiled a new negative reference set which consists of protein pairs where each protein in a pair is annotated as localized to a different cellular compartment (440,000 yeast and 1,750,000 human protein pairs, see Methods online). This was essential for comparison to experimental assays since, as constructed, our N set excludes data compiled from HT experiments, and hence the FPR for experimental assays is artificially, zero (see also related discussion in SOM of reference (Yu, Braun et al. 2008)).

Figures 5-4A shows a ROC curve calculated based on this new negative reference set and the CCSB-PRS positive reference set. This data show that, surprisingly, PREPPI outperforms all HT methods yielding higher TPRs at corresponding FPRs. With a few exceptions, the same conclusion holds for the larger CCSB-BGS and the HC reference sets (Figure S5-7).

Figure 5-4B shows a Venn diagram based on an LR cutoff of 600 (FPR ≈ 0.1%) while the HT results correspond to higher FPRs for yeast and lower for human (see Figure S5-7). Results for other LRs and additional reference sets are



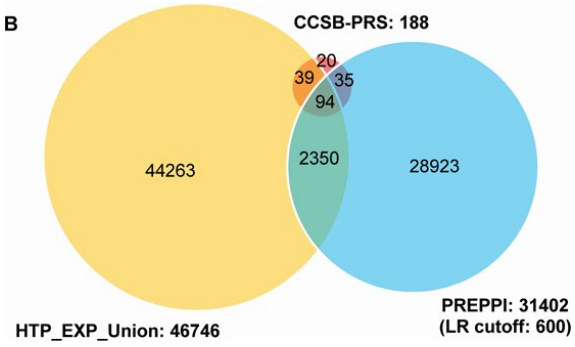


Figure 5-4. ROC curve (A) and Venn diagram (B) for PREPPI predictions and high-throughput (HT) experiments for yeast. HT experiments are labeled with the first author of the relevant publication (Table S5-3). The number of interactions in each set is given after the set label in the Venn diagram.

shown in Figure S5-8. As can be seen in the Venn diagrams in Figures 5-4B and Figures S5-8, many of the interactions inferred by PREPPI are different from those identified by HT methods. This suggests that computational prediction provides complementary clues to existing experimental assays and that methods that combine computational and high-throughput sources of evidence may prove to be highly effective. Figure 5-2B describes a prediction of an LC set interaction between the elongation factor 1-delta (EF-1δ) and the von Hippel-Lindau tumor suppressor (pVHL) based on the same template (E2-ubiquitin complex) used in Figure 5-2A. Again, there is no sequence relationship between the target and the template proteins, and they are classified into different SCOP folds but, nevertheless, the interaction model has an LR of 70. (Parenthetically, SM provides the only computational clue that makes it possible to infer the two interactions in Figure 5-2). We note that the two proteins in Figure 5-2B were found to interact in a high throughput study by mass spectroscopy (Ewing, Chu et al. 2007), a result that can now be taken with higher confidence given the PREPPI analysis.

5.4 Discussion

The accuracy and range of applicability of structural-based PPI prediction were unanticipated, but should not come as a complete surprise. Most protein complexes in the PDB have structural neighbors that share binding properties

(Zhang, Petrey et al. 2010), and protein interface space may well be close to "complete" in terms of the packing orientations of secondary structure elements (Gao and Skolnick 2010). Moreover, these elements can be identified with geometric alignment methods (Keskin, Nussinov et al. 2008; Zhang, Petrey et al. 2010), a fact that has been exploited in the approach introduced here. Although the information required to predict whether two proteins interact thus often seems present in the PDB, the question has been how to mine it.

Two key elements are responsible for our success. First, the wide exploitation of homology models significantly expands upon the number proteins for which interactions can be modeled. About 1,500 PDB structures but more than 7,000 models are found as representatives of at least one domain of ~4,000 yeast proteins and ~8,500 PDB structures and more than 31,000 models for at least one domain of ~14,000 human proteins. Had we only used experimentally determined structures in our analysis of yeast, a total of only 65,614 PPIs could potentially have been identified, of which only 498 are found in our HC data set. In contrast, the corresponding numbers when homology models are used are about 2.4 million and 3,063. For human the amplification is almost equally dramatic (a total of 2.5 million PPIs with 1,845 in the HC set using only crystal structures and 36 million PPIs and 4,032 in the HC set when homology models are used).

Second, it has been demonstrated that even distantly related proteins often use regions of their surface with similar arrangements of secondary structure elements to bind to other proteins (Petrey, Fischer et al. 2009; Gao and Skolnick 2010; Zhang, Petrey et al. 2010), and the use of such relationships here greatly amplifies the number of putative interactions that can be modeled (see examples in Figure 5-2). In fact, had we limited our definition of structural neighbors to members of the same SCOP fold or superfamily, only about 300 interactions in the yeast HC set could potentially have been identified.

The use of homology models and of remote structural relationships implies that each new structure that is determined experimentally can be used to detect large numbers of new functional relationships even if the protein in question is of only limited biological interest on its own. In this regard, our approach offers a rationale for structural genomics initiatives, which produced a large increase in the coverage of sequence families that did not have structural representatives (Levitt 2009). Moreover, since models can be built for every interaction inferred by our approach, it is now possible to predict the location of the interface on a protein surface for large numbers of protein-protein complexes and, consequently, to derive experimentally testable hypotheses as to the presence of a true physical interaction. For example, the interaction model of PKD1 and PKCε in Figure 5-2A not only predicts an interaction between the two kinases, it

also predicts that the interaction is mediated by the PKD1-PH domain and the PKCε-C1 domain, which is consistent with the observation that the PKD1-PH domain is required for the formation of a complex of PKD1 and PKCη, another member of the novel PKC family (Waldron, Iglesias et al. 1999). In conclusion, our study suggests the ability to add a structural "face" for a large number of PPIs and that structural biology can now begin playing an important role in molecular systems biology.

Acknowledgements

This work is supported by NIH grants GM030518 and CA121852. L.D. thanks CSC (China Scholarship Council) scholarship 2010626059. We thank Ursula Pieper from Andrej Sali's laboratory for help with ModBase, Hunjoong Lee for help with SkyBase and Celine Lefebvre for helpful discussions on human PPIs.

Supplementary Materials and Methods

Proteins and domains. We obtained the yeast proteome from UniProt (Apweiler, Bairoch et al. 2004), and parsed its 6,521 proteins into 7,792 domains using the SMART online server (Letunic, Doerks et al. 2009). Similarly, for human, we identified 20,318 unique proteome members, producing 49,851 individual domains.

Structures. Structural representatives of the entire protein or different individual domains were either taken directly from the PDB (Berman, Westbrook et al. 2000), where available, or from the ModBase (Pieper, Eswar et al. 2006) and SkyBase (Mirkovic, Li et al. 2007) homology model databases. PDB structures were identified by sequence homology, using a single iteration of PSI-BLAST (Altschul, Madden et al. 1997) and an E-value cutoff 0.0001; further, we required that matching structures in the PDB have >90% sequence identity and cover >80% of the query target (the entire protein or any domain). Homology models were selected based on two criteria: a) an E-value less than 1e-6, or b) an E-value less than 1 and either a structure-based pG score ≥ 0.3, for SkyBase models (Sanchez and Sali 1998), or a ModPipe protein quality score MPQS ≥ 0.5, for ModBase models. When multiple structures were available for a target/domain we choose only one representative by: a) first, the PDB structure with the best resolution, if available; b) otherwise, the ModBase model with the highest MPQS score; or c) lastly, the SkyBase model with the highest pG score. Based on these criteria, we could identify 1,361 PDB structures and 7,222 homology models for 4,193 different yeast proteins. Among these, 627 proteins could be matched to a PDB structure and 3,662 to a homology model, with some proteins having both. For human, 14,132 proteins were matched to 8,582 PDB structures and 30,912 models. Specifically, 4,286 proteins were matched to a PDB structure and 11,266 were matched to a homology model, with some proteins matched both.

Structural neighbors. We used a structural alignment tool Ska (Petrey and Honig 2003) to identify structural neighbors for these structural representatives. Ska is a local alignment tool, which allows alignments to be considered significant even if only three secondary structural elements are well aligned. At a PSD (Yang and Honig 2000) (protein structure distance) cutoff of 0.6, we identified 1,448 neighbors (both close and remote) per structure for 7,875 structures of 3,911 yeast proteins and 1,553 neighbors per structure for 36,743 structures of 13,545 human proteins.

Template complexes. As of early 2010, there were about 37,000 protein-protein complexes involving multiple organisms in the PDB and PQS (Henrick and Thornton 1998) databases. We used 28,408 and 29,012 complexes as templates during our modeling of yeast and human interactions, respectively. PQS terminated updates after Aug. 2009, and has been replaced by the PISA (Protein interfaces, surfaces and assemblies) server (Krissinel and Henrick 2007) which will be used in future work.

Interaction modeling. Given a pair of proteins or domains, we built their interaction model by superimposing their structures with the corresponding structural neighbors in the templates (Figure 5-1). For yeast, we built 550 million models for 2.4 million potential PPIs, which cover 11.3% of the total possible interaction space of all proteins (21 million), but 25.8% (3,063) of the HC

interactions. For human, we built 12 billion models for 36 million potential PPIs, which cover 17.5% of the total possible interaction space of all proteins (206 million), but 54.4% (4,032) of the HC interactions.

Interaction reference datasets. To ensure accurate and broad coverage of true interactions, we combined interaction data from multiple databases (Mewes, Albermann et al. 1997; Salwinski, Miller et al. 2004; Stark, Breitkreutz et al. 2006; Chatr-aryamontri, Ceol et al. 2007; Kerrien, Alam-Faruque et al. 2007; Keshava Prasad, Goel et al. 2009) and selected a subset of all high-confidence (HC) interactions that have multiple publications supporting their existence (11,851 yeast and 7,409 human interactions). All protein pairs with no supporting publication form the negative (N) reference set. The HC and the N sets were used as our reference datasets in all training and validations. See Table S5-1 for details.

Interaction model scoring. We calculated five empirical structure-based scores for each interaction model (Figure S5-1). We used a Bayesian network to combine these scores, into a likelihood ratio (LR) to evaluate an interaction model based on the HC and the N reference sets described above (Figure S5-2). Broadly, given some clue that reflects whether two proteins interact, the LR is an indicator of how likely it is that a pair of proteins with that clue will represent a true interaction.

Non-structural clues. For the yeast proteome, we downloaded the raw data for four different clues; protein essentiality (ES), co-expression (CE), GO (Ashburner, Ball et al. 2000) similarity and MIPS (Mewes, Albermann et al. 1997) similarity, from the Gerstein lab (http://networks.gersteinlab.org/intint/supplementary.htm). We also implemented a measure of phylogenetic profile (PP) similarity based on that introduced in reference (Huynen, Snel et al. 2000) (see below). We calculate a likelihood ratio (LR) for each non-structure clue based on our own reference sets, i.e., the HC and the N sets. Gerstein and coworkers expanded a set of 174 protein complexes from the MIPS catalog into 8,617 binary interactions and used them as the positive reference set (Jansen, Yu et al. 2003). It should be noted, however, that this procedure made no distinction between direct physical interactions (i.e., A–B) and interactions mediated by other proteins (e.g., A–C–B). Since the focus of our study is on physical interactions, we used our HC reference set which is composed primarily of direct physical interactions. For the human proteome, we calculated three different clues following the protocol of Gerstein and colleagues for GO and CE and as described below for PP. For CE, we used the expression dataset (GDS1962), which is one of the most comprehensive microarray studies of 19,803 human genes under 180 different conditions (Sun, Hui et al. 2006), from the Gene Expression Omnibus (Barrett, Troup et al. 2011).

Phylogenetic profile (PP) similarity. Similar to Enault et. al. (Enault, Suhre et al. 2005), we calculated a continuous score between 0 and 1 to measure the occurrence of a protein and/or domain in 1,156 reference organisms of complete proteome information from UniProt. These scores form a phylogenetic profile vector (PPV), and the Pearson correlation coefficient (PCC) was used to define the similarity between two vectors. For proteins with multiple domains, each domain's PPV is calculated independently, and the highest PCC score of different domain pairs is selected as the similarity score between two proteins. Similarity scores for pairs of proteins/domains with >40% sequence identity and, of course, for homomeric protein/domain pairs were not calculated.

The Naïve Bayes Classifier. We combine the different types of clues with each other and structural modeling into a single Naïve Bayes PPI classifier (Jansen, Yu et al. 2003; von Mering, Jensen et al. 2005; Lefebvre, Rajbhandari et al. 2010):

$$\mathrm{LR}(c_1, c_2, \cdots, c_n) = \prod_{i=1}^{n} \mathrm{LR}(c_i)$$

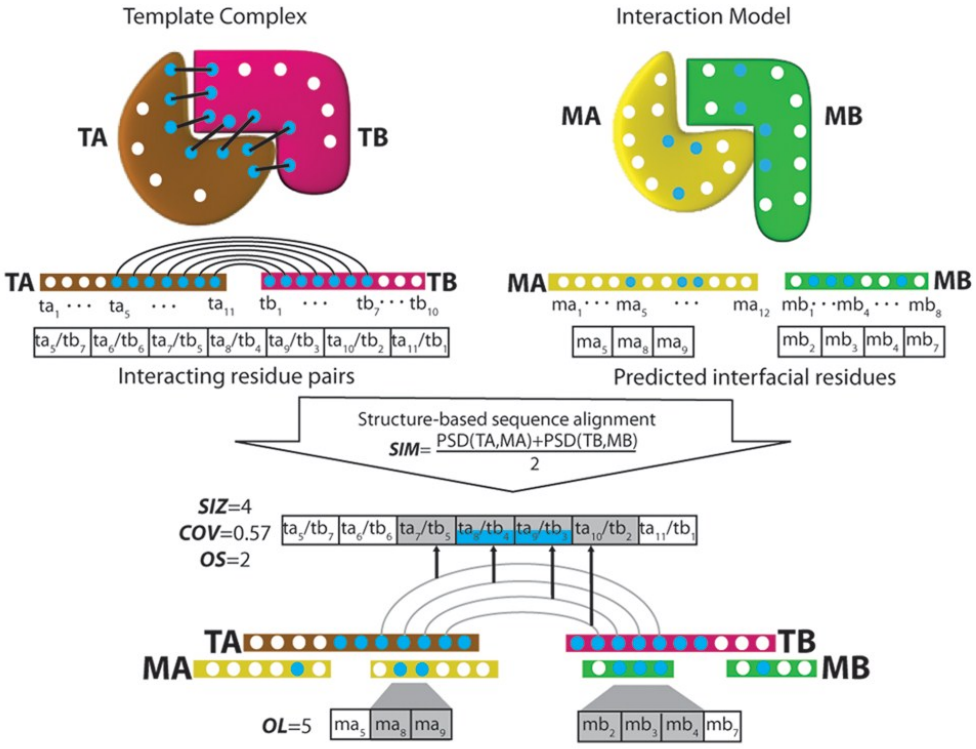
10-fold cross validation. To test the ability of a classifier to accurately and specifically predict PPIs, we carried out a 10-fold cross validation. We randomly divided the positive and negative reference sets into 10 subsets of equal size. Each time, we used 9 subsets to train the classifier, and obtained the LR for each protein pair, i.e., interaction, in the excluded subset from the trained classifier. We

repeated the procedure 10 times using different subsets as training and testing datasets and finally obtained an LR for each interaction. We counted the number of true positives (predictions in the HC set) and false positives (predictions in the N set) and calculated the prediction TPR (true positive rate) =TP/(TP+FN) and the FPR (false positive rate) =FP/(FP+TN) to plot the receiver operating characteristic (ROC) curves. Note that in all prediction performance tests, we have removed structural interaction models based on a template that corresponds to an actual crystal structure of the two target proteins.

Comparison with high-throughput (HT) experiments. We retrieved eight HT experiment datasets for yeast and three for human (Table S5-4). In our comparison, in addition to the HC sets, we also use the same reference interaction sets used in the comparative study of different HT techniques. These include ~1,300 PPIs (CCSB-BGS) and a subset of 188 highly reliable PPIs that are referenced in at least four manuscripts (CCSB-PRS). We compiled a new negative reference set, which consists of 440,000 yeast and 1,750,000 human protein pairs where each protein in a pair is annotated as localized to a different cellular compartment (Figure S5-6).

Supplementary Figures and Tables

Supplementary Figure S5-1. Interaction model evaluation scores.



The top of the figure shows a template complex (TA,TB) and an interaction model (MA,MB) obtained as described in Figure 5-1 from the main text (i.e., TA = NA_1 in Figure 5-1 and TB = NB_3). Individual residues in the different chains of the template and model are shown as dots, colored to indicate whether they are interfacial (blue) or non-interfacial (white). We also show schematic representations of the amino acid sequences below their corresponding chain in the template and model. We determine whether residues are interfacial

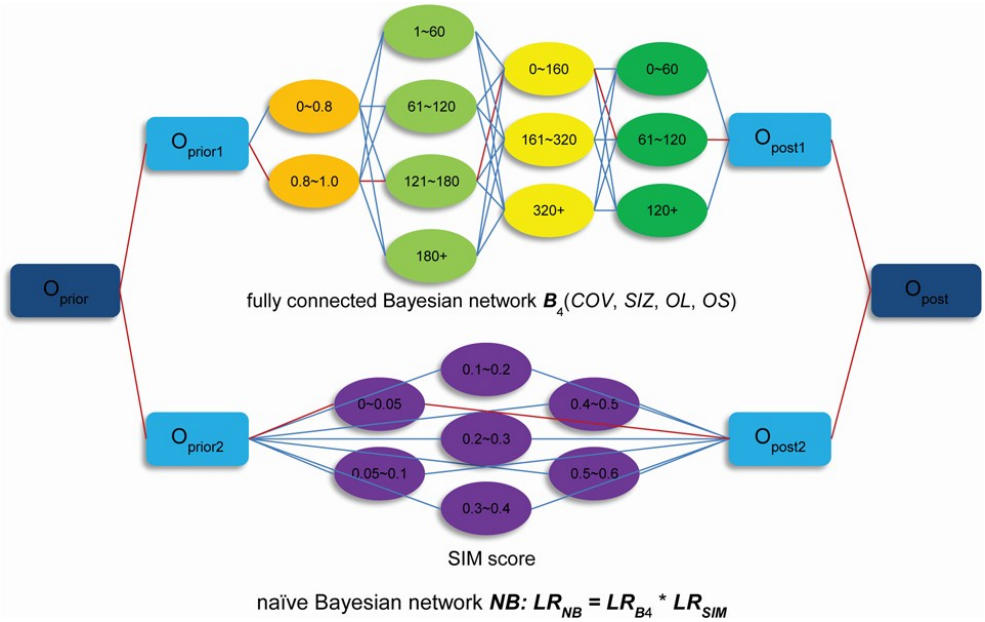
using the following criteria. For the template, this is determined directly from the associated experimentally determined structure in the PDB using a 6.05 angstrom distance cutoff between heavy atoms (Davis and Sali 2005). We also identify interacting residue pairs (ta_5/tb_7, ta_6/tb_6, etc., black lines) in the template using the same cutoff. For the model, we *predict* interfacial residues in the individual query proteins using a combination of three programs: PredUs (Zhang, Deng et al. 2011), PINUP (Liang, Zhang et al. 2006) and cons-PPISP (Chen and Zhou 2005). Note that these programs use only the structures and sequences of the individual subunits in the model (i.e., MA by itself and MB by itself) and hence are totally independent of the modeled complex. In this example, MA has 3 predicted interfacial residues (ma_2, ma_5, etc.) and MB has 4 (mb_2, mb_3, etc.). In practice, interacting residue pairs and predicted interfacial residues are pre-calculated and stored for each template complex and query protein in order to allow efficient evaluation the of the billions of models we generate. Each interaction model is associated with two structure-based sequence alignments (i.e., MA aligned to TA and MB aligned to TB). We do not evaluate the 3-dimensional model directly but rather use a set of five criteria (designated *SIM*, *SIZ*, *COV*, *OS*, *OL*), calculated from the alignments as described below.

- *SIM*: the geometric similarity between the protomers in the template and the model measured using protein structural distance (PSD, (Yang and

Honig 2000)). Since there are two geometric alignments obtained for each model (i.e., MA to TA and MB to TB), *SIM* is calculated as the average of PSD(TA,MA) and PSD(TB,MB).

- *SIZ and COV*: the number and fraction of interacting residue pairs in the template that are preserved in the model. In this example, four of the seven interacting pairs present in the template are preserved in the model (ta_7/tb_5, ta_8/tb_4, ta_9/tb_3, ta_{10}/tb_2, highlighted in grey and indicated by grey lines). Hence *SIZ*=4 and *COV*=4/7=0.57.
- *OS*: the same as *SIZ*, with the additional condition that each residue in the interacting pair aligns to a residue that is predicted to be interfacial in the model. In this example, although *SIZ*=4, only two of these interacting pairs (ta_8/tb_4 and ta_9/tb_3, highlighted in grey and blue) are present where each residue in the pair also aligns to a predicted interfacial residue in the model. Hence, *OS*=2.
- *OL:* the number of predicted interfacial residues in the model that align to template interfacial residues. In this example, MA has 2 predicted interfacial residues that align to interfacial residues in TA (ma_8 and ma_9, highlighted in grey) MB has 3 that align to interfacial residues in TB. Hence, *OL*=3+2=5.

Supplementary Figure S5-2. Bayesian network for structural modeling.



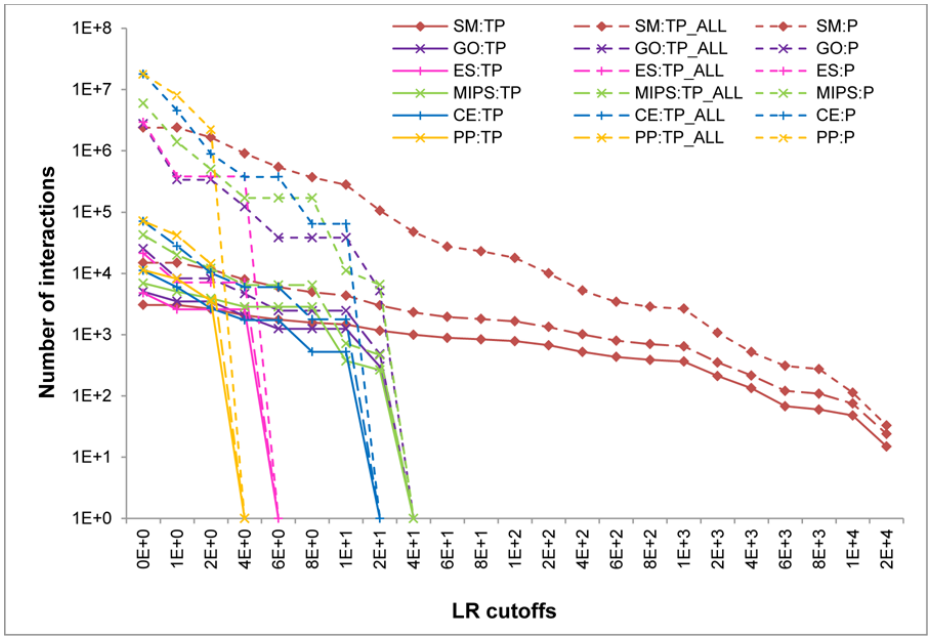
We used a Bayesian network to combine the five structure-based scores, i.e., *SIM*, *COV*, *SIZ*, *OL*, and *OS* (Figure S5-1), into a single term to evaluate an interaction model. We built a fully connected Bayesian network B_4 for *COV*, *SIZ*, *OL*, and *OS* and combined it with the *SIM* score using the naïve Bayesian approach (*NB*). (Based on a calculation of the Pearson correlation coefficients for each pair of scores using all 550 million models built for yeast, *COV*, *SIZ*, *OL*, and *OS* were correlated with each other but *SIM* was only weakly correlated with the other four.) For each score, we defined discrete bins shown conceptually in the figure (bin sizes were adjusted manually to ensure adequate coverage of each bin).

To train the network using a set of PPIs, we assigned their associated interaction models to individual bins according to the model scores. For example, an interaction model with scores *SIM*=0, *COV*=0.9, *SIZ*=150, *OL*=120, and *OS*=80 will be assigned to bin[*SIM*=0~0.05] and bin[*COV*=0.8~1.0, *SIZ*=121~180, *OL*=0~160, *OS*=61~120] shown by red lines in the figure. An interaction can have multiple models, so it is important *not* to assign different models of the same interaction to the same bin multiple times. That is, if multiple models of a single interaction have the same set of scores, only one is counted in a given bin. The likelihood ratio (LR) for any bin is then determined using Bayes theorem:

$$\mathrm{LR}(bin) = \frac{O_{post}}{O_{prior}} = \frac{P(bin \mid HC)}{P(bin \mid N)}$$

Here $P(bin \mid HC)$ (and $P(bin \mid N)$, respectively) are the probabilities that an interactions in the HC set (the N sets) is in the bin. For an interaction model, we calculate its structure-based scores and determine the LR from the associated bin. The LR represents the increase of chance that an interaction with models of particular scores to be a positive PPI, compared with a random protein pair. The maximum LR is used when an interaction has multiple models.

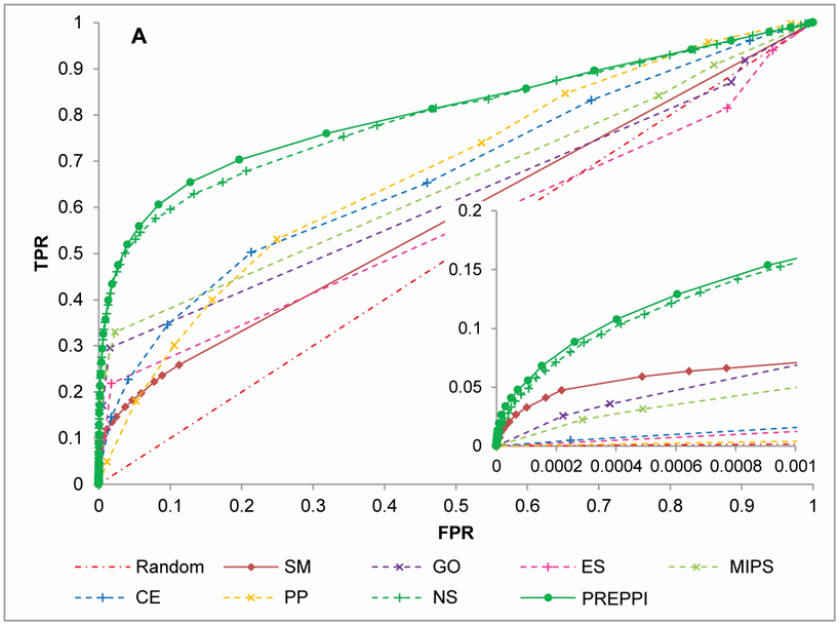
Supplementary Figure S5-3. Number of predicted interactions vs. likelihood ratio (LR) using structural modeling and non-structure based clues.

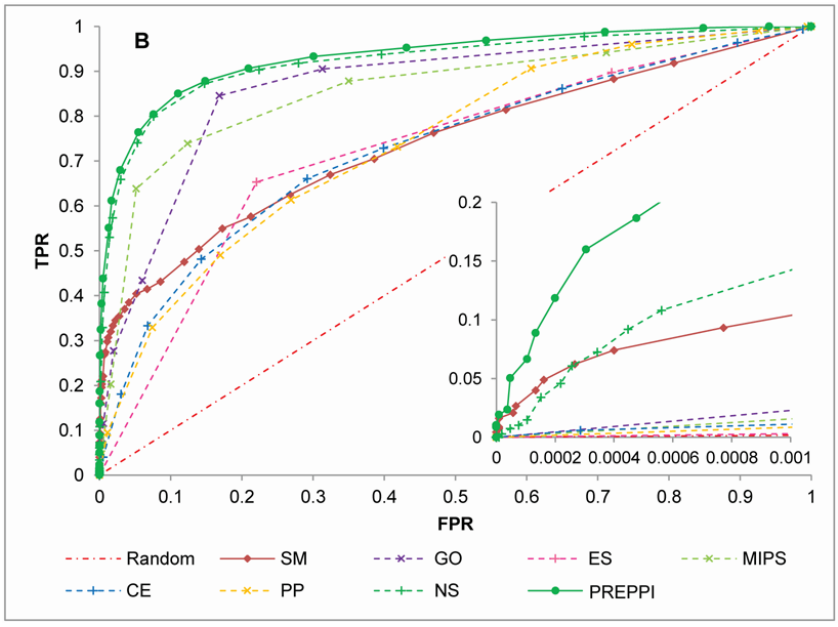


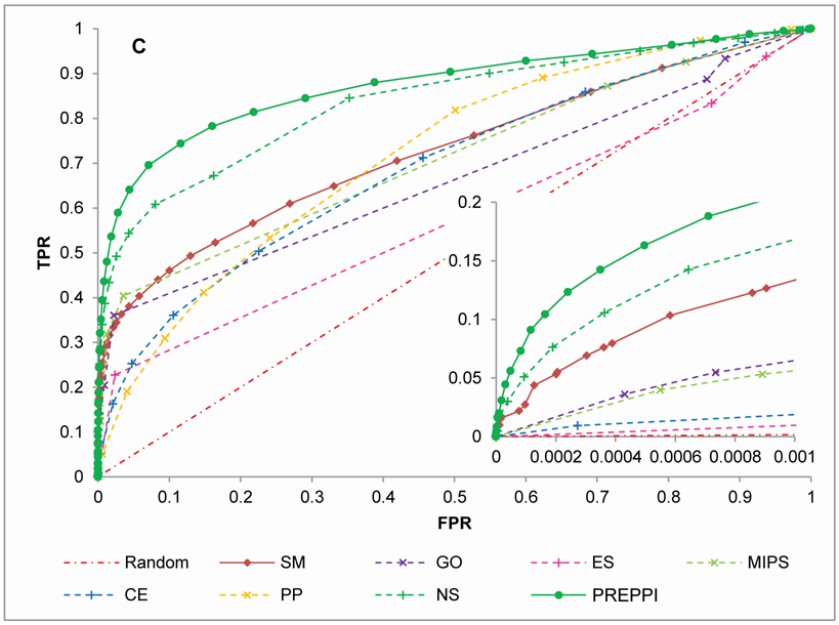
We examined different sources of information (i.e. structural modeling (SM), GO, protein essentiality (ES), MIPS, co-expression (CE), or phylogenetic profile (PP)) for their ability to predict PPIs. Any three lines of the same color and marker in the graph are associated with a particular clue and show numbers of predicted interactions with an LR above the cutoff, based on that clue. The total number of interactions predicted at a given cutoff is shown as a short-dashed line (P). The other two lines for a given clue correspond to whether the predictions are in the HC interaction set (solid line, TP), or in the union of the LC and HC interactions sets (long-dashed line, TP_ALL).

As shown in the figure, although in some cases it is possible to calculate a score for many more pairs of interactions for a given non-structural clue as compared to structural modeling, the numbers of interactions predicted with high-likelihood ratio (LR) drops much more quickly for non-structural clues. Indeed, an important property of structural information is that it is particularly effective in making predictions at high LR regime, i.e., high confidence levels.

Supplementary Figure S5-4. ROC curves for yeast PPIs predicted based on different sources of information in different interaction spaces.





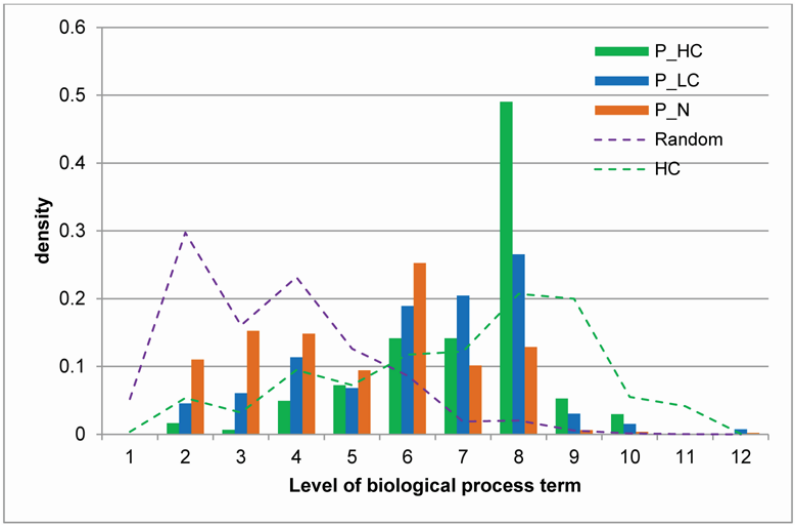


In Figure 5-2A, for yeast, we restrict each ROC curve in the plot to only those interactions for which the associated single clue or combination of clues was available. For completeness, we show here ROC curves for the different clues, but compare them using a single subset of protein pairs: (A) for the whole interaction space of 21 million protein pairs in yeast, (B) for the subset where information for all types of clues is available (116 thousand yeast protein pairs), (C) for the subset where structural information is available (2.4 million yeast interactions). The clues examined here are the same as those shown in Figure 5-2A, i.e. structural modeling (SM), GO similarity, protein essentiality (ES) relationship, MIPS similarity, co-expression (CE), phylogenetic profile (PP) similarity, or their combinations (NS for the integration of all non-structure clues,

i.e. GO, ES, MIPS, CE, and PP, and PREPPI for all structural and non-structure clues).

Figures S4A-C and Figure 5-2 consistently show that whatever data set is used, structural modeling (SM) yields comparable performance to other clues over the entire range of FPR but is considerably more effective at low FPR. In addition, the algorithm that combines structural modeling with other sources of evidence (PREPPI) shows superior performance to any method based on individual clues over the entire range of false positive rates. Obviously the performance of PREPPI at low FPRs is due primarily to structural information.

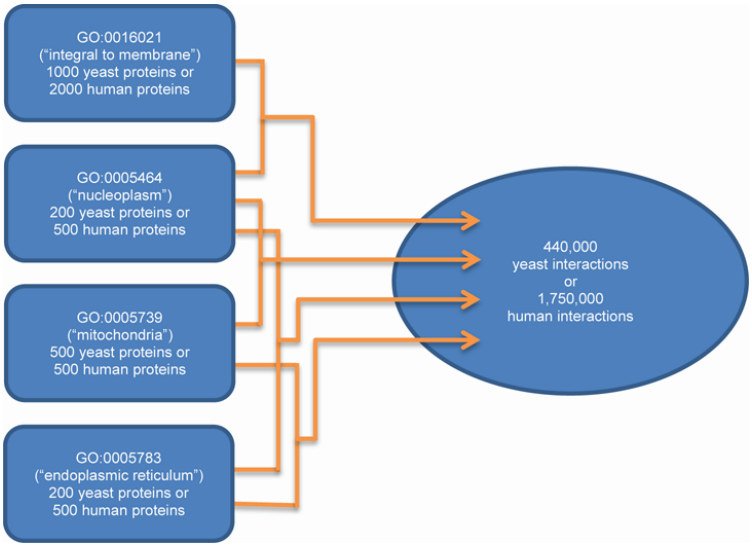
Supplementary Figure S5-5. Distributions of GO biological process (BP) similarity terms for yeast protein pairs.



We define BP similarity for two proteins as the integer representing the level of their most recent common ancestor (MRCA) in the GO hierarchy, taking the maximum if multiple MRCAs are available. We extracted GO annotation for individual yeast proteins from UniProt and calculated the similarity for different sets of pairs. The purple line shows the random distribution of similarities, i.e., for all protein pairs in yeast for which we could find GO annotations. The green line shows the distribution for protein pairs in our HC set of true interactions. The bars show the distribution of similarities for pairs of interactions predicted by structural modeling (SM) at an LR cutoff of 600 that are also in different reference sets that we use: the HC (green), LC (blue), and N (orange) sets.

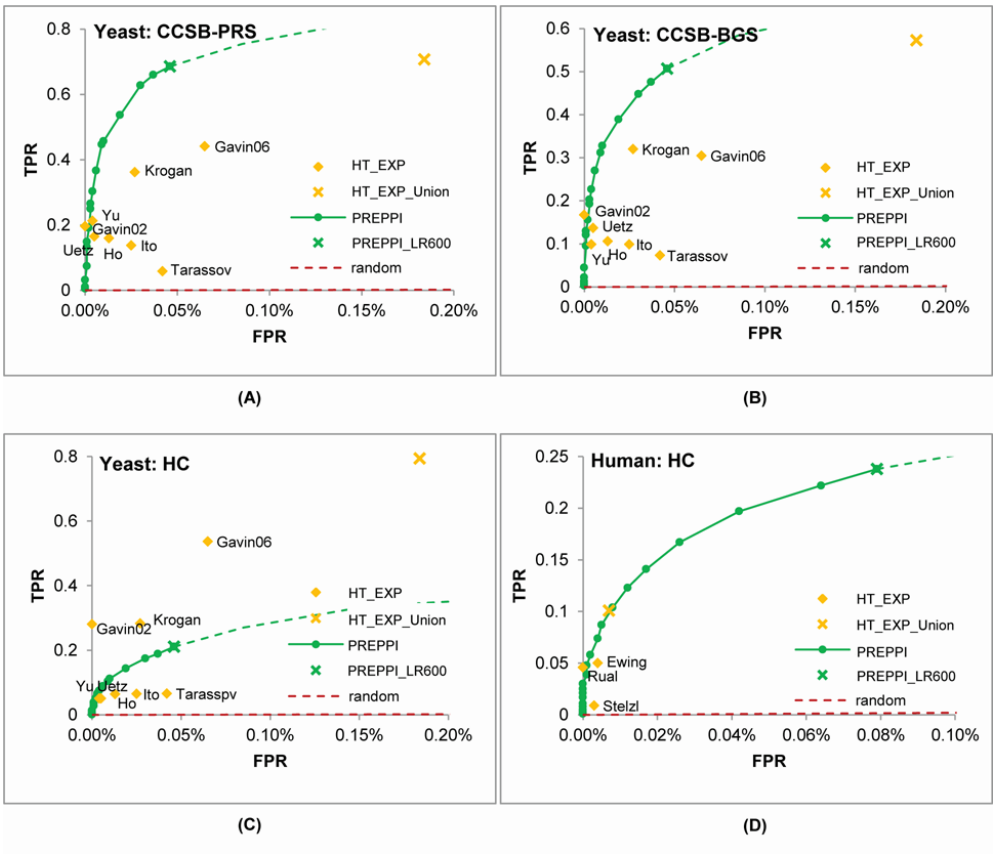
Only about 13% of random yeast interactions involve proteins that share an MRCA at least level 6 (the purple line). On the other hand, most true PPIs in the HC set (8,126 of 10,933, or 74%) share an MRCA level at least 6 (the green line). The MRCA levels for the SM predictions show similar shifts in the distribution. Specifically, at the LR cutoff 600, 434 of the predicted PPIs are in the HC data set, 363 in the LC data set and 2,640 in the N set. Of the 132 hetero-dimeric pairs in the LC set with GO annotation, 94 contain proteins that share GO biological term at, or more specific than, the 6th level of the GO hierarchy (blue bars), providing supporting evidence that these interactions are real (in addition to their presence in the LC set). Similarly, 960 of the 1,946 hetero-dimeric predictions in the N set contain proteins that share GO terms at level 6 (orange bars), suggesting that there is at least a functional relationship which may involve protein-protein interactions.

Supplementary Figure S5-6. Negative interaction reference set constructed using proteins in different cellular compartments.



We randomly chose a number of proteins based on their GO annotations and paired those from different cellular compartments to form the negative reference sets (shown as orange lines). There were several proteins annotated as belonging to two of these cellular compartments which we excluded. A very small number of interactions were also contained in the positive reference sets (e.g., HC, CCSB-PRS, and CCSB-BGS) which were removed from the new negative reference sets (i.e., the final sizes of the negative reference sets are very close to but not exactly the same number as shown in the figure).

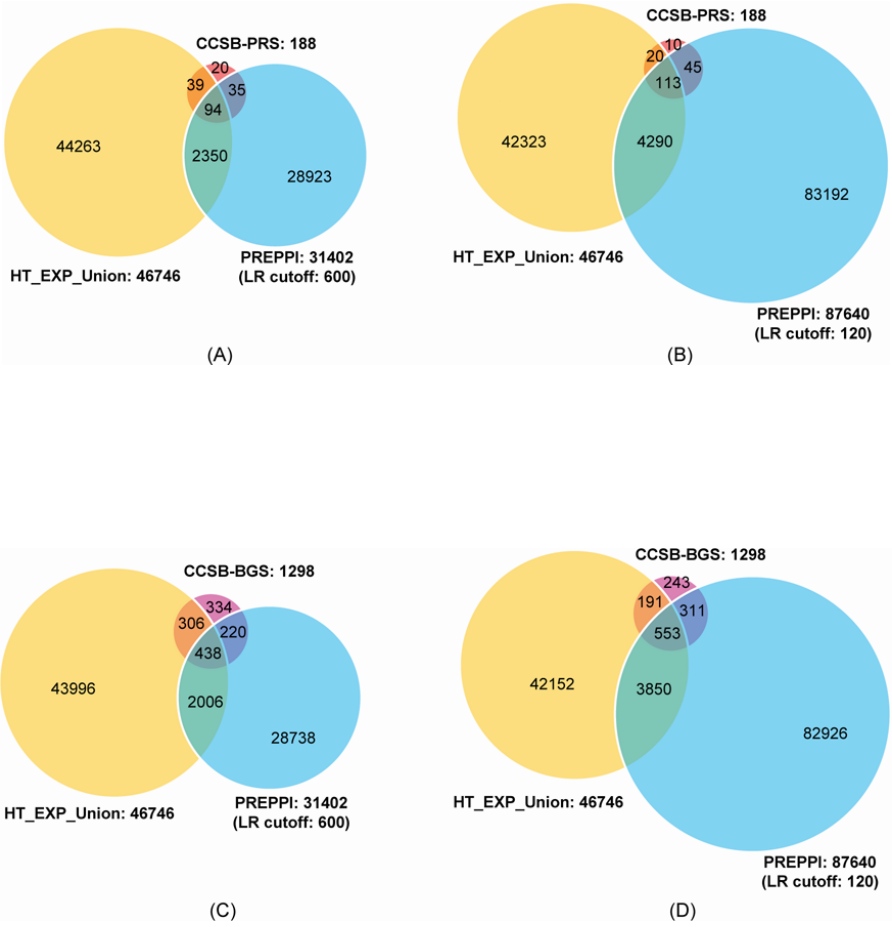
Supplementary Figure S5-7. ROC curves of PREPPI predictions and high-throughput (HT) experiments on different interaction reference datasets.

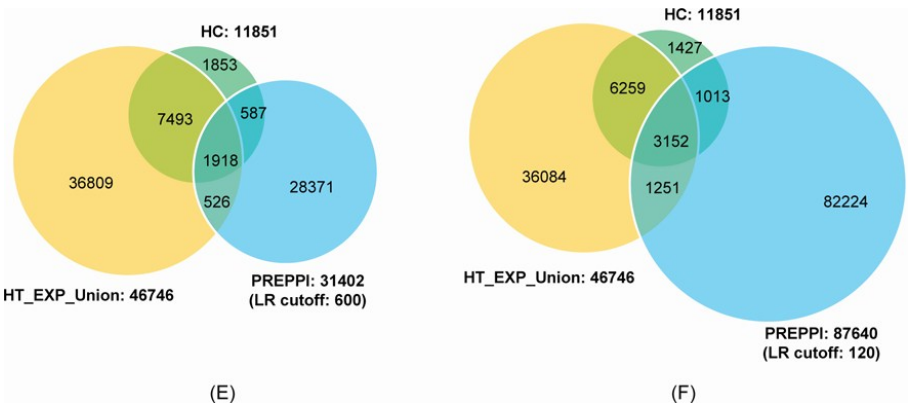


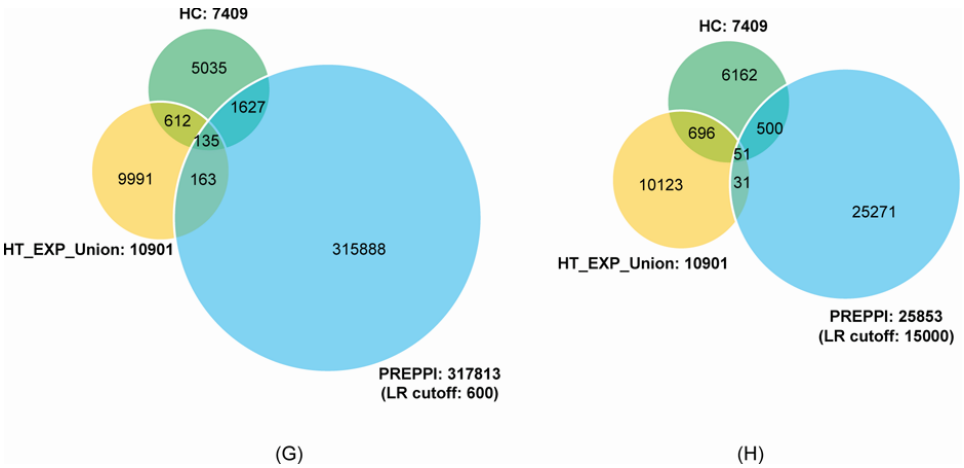
In Figure 5-3A, we show a ROC curve of PREPPI predictions and HT experiments using the CCSB-PRS reference set (reproduced here as panel A to facilitate comparisons). Here we show comparisons using additional positive reference sets: B) CCSB-BGS, and C) the yeast and D) human HC sets defined in the main text. Results from PREPPI are displayed as green curves, and the predictions at LR cutoff 600 are highlighted with green "X". HT experiments are

shown as yellow diamonds with the datasets labeled with the name of the first author of the corresponding publications (Table S5-4). The unions of HT experiments are marked with yellow "X". Our results consistently show that PREPPI predictions are comparable to most HT experimental studies.

Supplementary Figure S5-8. Venn diagrams of PREPPI predictions at different LR cutoffs, union of HT experiments, and different reference interaction datasets for yeast (A-F) and human (G-H).







In Figure 5-3B, we show a Venn diagram of PREPPI predictions at an LR cutoff of 600, unions of HT experiments, and the CCSB-PRS reference set (reproduced here as panel A for comparisons). Here we show the results of PREPPI predictions for additional positive reference sets defined in the figure along with the number of interactions they contain. The number after the label of a set shows the number of interaction in the set. The LR cutoff 600 was used in

(Jansen, Yu et al. 2003) based on the assumption that protein pairs with LR > 600 have a better than 50% chance to be a true interaction. The number of interactions of the union of HT experiments depends on individual HT experiments, which generally results in different FPRs from those obtained from PREPPI predictions at an LR cutoff of 600. For this reason we also compared PREPPI predictions at the same FPRs as unions of the HT experiments, which correspond to an LR cutoff 120 for yeast and an LR cutoff 15,000 for human.

As can be seen from the figure, PREPPI consistently predicts many interactions that are in the reference sets but not identified in any HT study. We define these interactions as the exclusive contribution of PREPPI to the reference sets (similarly, we define the exclusive contribution of the union of HT experiments to the reference sets). For most cases, the number of exclusive contributions of PREPPI is comparable to that of the union of HT experiments. The only exception is in the exclusive contributions to the yeast HC set. However, in this case the discrepancy is largely due to the fact that the yeast HC set mainly consists of interactions from HT studies (about 80% of the HC interactions are identified in at least one HT experiment). This of course biases the HC set so as to favor the evaluation of HT experiments.

Supplementary Table S5-1. Positive PPI reference sets for yeast (A) and human (B).

(A) yeast

Database	MIPS	DIP	IntAct	MINT	BioGRID	Overall
MIPS	7,539	6,955	6,379	6,349	3,910	7,539
DIP		17,511	13,305	12,731	13,149	17,511
IntAct			48,009	16,680	19,316	48,009
MINT				24,083	17,082	24,083
BioGRID					42,650	42,650
Overall						73,787

(B) human

Database	HPRD	DIP	IntAct	MINT	BioGRID	Overall
HPRD	14,977	319	4,266	3,264	7,316	14,977
DIP		1,460	430	352	706	1,460
IntAct			27,911	7,235	11,357	27,911
MINT				12,099	5,044	12,099
BioGRID					32,071	32,071
Overall						58,772

The Training and evaluation of a PPI predictor requires accurate and broad coverage gold standards for both positive and negative interactions. Yet, achieving these competing goals can pose significant challenges. Some studies have used a single, well-annotated database (Jansen, Yu et al. 2003) but bias in individual databases has been described which can complicate evaluation of the

method (Myers, Barrett et al. 2006). On the other hand, the use of all available data can also be problematic because of issues related to the accuracy of databases that incorporate interactions determined, for example, by high-throughput approaches (von Mering, Krause et al. 2002). Similar to two recent studies of the yeast and human B-cell interactomes (Yu, Braun et al. 2008; Lefebvre, Rajbhandari et al. 2010), we combine interaction data from multiple databases and select the reliable ones to ensure accurate and broad coverage of true interactions in the positive reference set. For yeast, we used the interactions databases: MIPS (Mewes, Albermann et al. 1997), DIP (Salwinski, Miller et al. 2004), BioGRID (Stark, Breitkreutz et al. 2006), intAct (Kerrien, Alam-Faruque et al. 2007) and MINT (Chatr-aryamontri, Ceol et al. 2007). We retrieved data deposited prior to Aug. 2009. For human, we used the databases: HPRD (Keshava Prasad, Goel et al. 2009), DIP, BioGRID, MINT and intAct, retrieving data deposited prior to Aug. 2010. We mapped different protein identifiers to UniProt accession numbers (AC) and used the pairs of accession numbers as the unique identifiers to all PPIs. Proteins without valid UniProt AC or not defined in the yeast and the human proteomes were removed (i.e., limited to the 6,521 proteins for yeast and the 20,318 proteins for human). The high confidence (HC) reference set for yeast contains 11,851 interactions with more than one supporting publication and the low confidence (LC) reference set contains 61,936 interactions with only one supporting publication (73,787 in total). The HC set for

human contains 7,409 unique interactions, and the LC set contains 51,363 interactions (58,772 in total). All the HC and the LC datasets are available at http://wiki.c2b2.columbia.edu/honiglab_public/index.php/Software:PREPPI. In the table below, cells on the diagonal represent the number of interactions taken from the corresponding database and the off-diagonal cells in the tables show the overlap between different data sources.

Supplementary Table S5-2. Availability of different clues for protein pairs in yeast.

Method	predictions	Coverage	HC	recall
SM	2398316	11.3%	3063	25.8%
GO	2756276	13.0%	5036	42.5%
ES	2925066	13.8%	4787	40.4%
MIPS	5962511	28.0%	6915	58.3%
CE	17967683	84.5%	11118	93.8%
PP	17848620	83.9%	11273	95.1%

Clues for GO similarity, protein essentiality (ES), MIPS similarity, and co-expression (CE) data were retrieved from (Jansen, Yu et al. 2003). We mapped the ORF names to UniProt accession numbers and only those defined in the yeast proteome were kept (i.e., limited to 6,521 yeast proteins). Coverage is the number of protein pairs for which a given clue (structural modeling (SM), GO, ES, MIPS, CE, and phylogenetic profile (PP) similarity) is available, divided by the total number of possible interactions (21 million); recall is the number of protein pairs in our HC set for which a given clue is available divided by the number of interactions in the HC set (11,851).

Supplementary Table S5-3. Predicting interactions in the DREAM exercise.

Prediction	Precision at *n*-th correct prediction			AUPR	AUROC
	1st	2nd	5th		
SM	1.00	0.67	0.71	0.49	0.74
PREPPI	0.50	0.67	0.71	0.49	0.77
Team1	1.00	1.00	1.00	0.70	0.82
Team1*	0.50	0.67	0.83	0.32	0.49
Team2	0.20	0.20	0.12	0.15	0.48
Team3	0.25	0.15	0.16	0.16	0.51
Team4	0.50	0.67	0.14	0.18	0.49
Team5	1.00	0.67	0.50	0.33	0.66

DREAM evaluates computational reverse engineering methods in Systems Biology, using double blind assessments based on experimentally assessed data, similar to CASP. In DREAM2 (Stolovitzky, Prill et al. 2009), participants were asked to predict interactions among a set of 47 proteins; 48 true interactions among these proteins had been confirmed by the DREAM organizers in at least three independent Y2H experiments by the Vidal lab. We used the DREAM2 evaluation program to benchmark all predictions. Here "precision at *n*-th correct prediction" is the precision calculated when a predictor correctly predicts the *n*-th PPI by ranking its predictions from the highest probability to the lowest. AUPR and AUROC is the area under the PR (precision-recall) curve and ROC (receiver operating characteristic) curve.

For this DREAM2 exercise, structural modeling (SM) generated models for 199 interactions between 28 proteins. Here we compare SM predictions and the prediction that integrates both structural and non-structural clues (PREPPI) with all DREAM2 participants in this subset of 199 interactions for the 28 proteins. We use the most up-to-date information in the analysis (93 true positives according to current PPI databases) and re-evaluate the performance of each team based on this gold standard. As shown in the table, SM and PREPPI both perform much better than the other methods, except for Team1. However, the performance of Team1 seems to have been due to the fact that 19 of the true positive interactions between the target proteins were known in PPI databases at the time, and these interactions were submitted by Team1 (Chua, Hugo et al.) as "predictions" with very high probability, i.e., based only on the fact that they were present in the databases as opposed to an independent computational technique. The performance of Team1 when these interactions are removed from their predictions is significantly lower (Team1*).

Supplementary Table S5-4. High-throughput (HT) experiments.

Dataset		#interactions	Type	Source database	Reference
Yeast	**Uetz**	1437	Y2H	intAct	(Uetz, Giot et al. 2000)
	Ito	4447	Y2H	intAct	(Ito, Chiba et al. 2001)
	Yu	1626	Y2H	intAct	(Yu, Braun et al. 2008)
	Ho	3614	AP/MS	intAct	(Ho, Gruhler et al. 2002)
	Gavin02	3756	AP/MS	intAct	(Gavin, Bosche et al. 2002)
	Krogan	8183	AP/MS	MINT	(Krogan, Cagney et al. 2006)
	Gavin06	21242	AP/MS	intAct	(Gavin, Aloy et al. 2006)
	Tarassov	9601	PCA	intAct	(Tarassov, Messier et al. 2008)
Human	**Rual**	2455	Y2H	intAct	(Rual, Venkatesan et al. 2005)
	Stelzl	2972	Y2H	intAct	(Stelzl, Worm et al. 2005)
	Ewing	5504	AP/MS	intAct	(Ewing, Chu et al. 2007)

We retrieved eight HT experiment datasets for yeast and three for human from the intAct (Kerrien, Alam-Faruque et al. 2007) and the MINT databases (Chatr-aryamontri, Ceol et al. 2007). Database entries without valid UniProt (Apweiler, Bairoch et al. 2004) protein accession number or not defined in the yeast and the human proteomes are removed (i.e., limited to the 6,521 proteins for yeast and the 20,318 proteins for human).

Abbreviations: Y2H, yeast two hybrid; AP/MS, affinity purification followed by mass spectroscopy; PCA, protein fragment complementation assay.

CHAPTER 6. CONCLUSION

6.1 Significance of research

Systems biology seeks a quantitative understanding for a whole biological system by integrating data from diverse sources. Thanks to biotechnology development, we are now in a phase of unparalleled data growth, especially for DNA sequences and gene expression profiles. The wealth of information comes from disparate datasets and is being analyzed and integrated through computational techniques. However, to date, structural information has remained resistant to this integration, presumably because the use of structures usually depends on accurate modeling, which is time-consuming and more importantly, only possible for a limit number of proteins.

In this thesis, I described my work that attempts to combine structural biology and systems biology by focusing on the development and the application of new methods that could use structural information in the study of protein-protein interactions (PPIs) on a genome-wide scale.

I began by introducing a comprehensive analysis that showed significant interface conservation in sets of proteins sharing varying degrees of similarities across whole structural space (Zhang, Petrey et al. 2010). We employed the conservation to design PredUs, a template-based protein-protein interface

prediction method which showed substantial improvement over existing techniques. We developed the PredUs web server to predict protein interfaces based on this method with a support vector machine (SVM) to further improve interface prediction performance (Zhang, Deng et al. 2011).

The significance of the first part of my work is the finding of functional relationships among seemingly unrelated protein structures and the development of a fast and accurate method for the prediction of protein-protein binding interfaces, which is essential to our understanding of protein functions and has been successfully exploited in many applications. To our knowledge, PredUs is the first "template-based" method that predicts protein interfaces with high precision and recall. It is not sensitive to local conformational changes and small errors in structures and thus can be applied to predict protein interface for many proteins where only homology models are available.

I then showed that 3D structure information can be used in a "high-throughput" fashion to produce comprehensive maps of PPIs. I introduced a way to use 3D structural information to predict whether two proteins interact and applied the approach to both the yeast and the human genomes. I showed that 3D structural information is superior to other sources of evidence used to computationally infer interactions, and structural information combined with other evidence using a naive Bayesian classifier (PREPPI) identifies PPIs

comparable to high-throughput experimental approaches. Our data further suggests that PREPPI predictions are substantially complementary to PPI information generated by experimental methods.

The significance of the second part of my work is the high throughput and accurate identification of protein-protein interactions, which is essential to understand regulatory processes in a cell and how their dysregulation may contribute to disease. Our success in using 3D structure to predict whether two proteins interact dramatically enhances the value of structural information and provides a computational prediction method that is competitive with the labor-intensive high-throughput experimental approaches such as yeast two-hybrid in terms of both accuracy and coverage, and providing a way to dissect interactions that would be hard to identify on a purely high-throughput experimental basis.

As mentioned, systems biology has evolved largely independently of structural biology. This thesis reports significant advances in both structural and systems biology and provides the first meaningful integration of these two disciplines. In terms of structural biology, we have achieved an enormous amplification of the information available from solved crystal structures through a novel approach that exploits imperfect homology models and that extracts functional information from geometric similarities between proteins that have generally been considered to be unrelated. From the perspective of systems

biology, we have for the first time used 3D structure as part of the repertoire of experimental and computational information and find a way to accurately infer protein interface and PPIs on a large scale.

6.2 Future directions

6.2.1 Construction of the PREPPI webserver

A more complete and accurate compendium of protein-protein interfaces and interactions would be of great interest to the biological community. This has been demonstrated by the success of PredUs, our protein-protein interface prediction server, which has been used by many hundred different users since its inception. It suggests that it would also be worthwhile to make PREPPI, our protein-protein interaction (PPI) prediction method, publicly available.

We have set up a demo version of a webserver for the PREPPI software (http://wiki.c2b2.columbia.edu/honiglab_public/index.php/Software:PREPPI). It contains all experimental interactions prior to Aug. 2011, and predicted interaction of LR higher than 100 for yeast and human. These interactions could be searched using UniProt accession number, or other commonly used names of participating genes and proteins. So far, the PREPPI webserver contains little information about the involved proteins and interactions, but we plan to include more so that PREPPI could be a “one-stop shop” for PPI studies. For example, for

experimental interactions, we will provide information about source databases and publications to facilitate further investigations; and for predicted interactions, we plan to include the associated likelihood ratio scores of each component and their integral, and reliable structural models as well, if available, for detailed studies.

So far, we have only applied the PREPPI algorithms to the yeast and the human proteomes. In the future, we can use them to predict interactions for more model organisms. Some components of PREPPI, including our structural modeling techniques, can also be used for the study of interactions between different organisms, for example, the interactions between human host and pathogen proteins, which would be useful to the development of strategies to treat and prevent infectious diseases (Davis, Barkan et al. 2007; Tastan, Qi et al. 2009). Eventually, it is expected to offer in the future a functionality of predicting the interaction likelihood for any input pair of proteins. The PREPPI webserver would be an enabling technique for studies concerning PPIs and would potentially have big impact to the whole biological community.

6.2.2 Improvement of PREPPI predictions

By combining both structural and non-structural information, PREPPI has made itself so far the most accurate PPI prediction method. It can build billions of interaction models for millions of PPIs using imperfect homology models and remote structural relationships. Nevertheless, in order to evaluate this daunting

number of models in reasonable time and limited resources, we only calculated coarse-grained model parameters in residue level. In addition, our structural modeling scores mainly focus on model interfaces and only care whether model residues are aligned with template interfaces. In other words, residue identities are not considered and little attention is paid to the rest part of interaction models. Although our results suggest that our scoring function can distinguish good and bad interaction models to a great extent, it can be improved if we can evaluate interaction models in many more different aspects.

For example, if residues or parts of the two target structures of one interaction model are not aligned to the template complex, they may overlap in 3D space. Although proteins often change their conformations to avoid these clashes, large scale of positional overlaps will forbid them from forming a complex. This suggests that if we can detect and measure conformational clashes, we can filter out many impossible interaction models. However, this detection should be done in a very fast way so that it could be applied to billions of interaction models.

Another issue is the potential over prediction of PPIs formed by the same pair of domains. If two domains D_1 and D_2 forming a complex, PREPPI often predicts that many proteins containing D_1 interact with proteins containing D_2. But in reality, bindings between different domains are often specific, *i.e.*, proteins

containing D_1 only bind to a very small set of proteins containing D_2. This binding specificity is usually mediated by mutations of a few interfacial residues. We compared the predictions by our structural modeling method and a naïve method that simply predict all proteins containing D_1 interact with proteins containing D_2. We found that structural modeling did detect some of the specific bindings, *i.e.*, giving higher scores to those true interactions. However, it is likely that it depends on whether the mutations affect protein structures, or more realistically, protein interfaces (and be captured by the three interface prediction programs). Since eventually the scoring function does not contain information of residue identities, for many cases we cannot tell the binding differences between proteins containing the same domains. Despite that predicting binding specificity between the same families of proteins is notoriously difficult studies focusing on interaction between some specific domains and their interacting peptides have shown promising results (Chen, Chang et al. 2008; Sanchez, Beltrao et al. 2008; Grigoryan, Reinke et al. 2009). We also expect that by incorporating residue evolutionary information in our structural modeling it can better distinguish specific interactions mediated by the same pair of domains.

More importantly, many interactions are formed by domains where no appropriate template complex exists even based on remote structural relationships, or are mediated by unstructured peptides. The accurate prediction of these

interactions is beyond our current structural modeling method. It is expected that by utilizing the information of domain-domain or domain-peptide interaction profiles, we can improve PREPPI's performance on these interactions.

Improvements could also be gained by integrating more types of non-structural information from independent sources. The current PREPPI only contains clues of phylogenetic profile similarity (Chapter 5, Supplementary Material and Methods). It is likely that combining the other genomic/evolutionary PPI clues such as gene fusion and genomic context (Section 2.5.1) would further improve the prediction.

6.2.3 Applications of PREPPI interactomes

The current release of PREPPI contains many predictions of new potential interactions for yeast and human; including some with structural details. These interactions and structural models could be targets of focused studies in the future to elaborate unknown functions or mechanisms of important proteins and biological processes. Although anyone can use the PREPPI webserver to search for interaction information for their own proteins of interests, there are a few types of interactions that could be particularly interesting targets for PREPPI follow-up studies.

For example, scaffold or adaptor proteins are proteins that usually mediate specific PPIs that drive the formation of protein complexes and transduce cellular signals. These proteins do not have any intrinsic enzymatic activity by themselves but instead contain domains that often bind other domains and proteins, *e.g.*, Src homology 2 (SH2) and SH3 domains. Maybe because many structure complexes of these scaffolding interactions have been crystalized, it seems that PREPPI performs especially good at predicting interactions involving these adaptor proteins. For example, the growth factor receptor-bound protein 2, known as Grb2, is an adaptor protein that is widely expressed and is essential for cell proliferation and development. It has been shown to interact with many proteins. PREPPI can recover most of the known interactions and at the same time predict many unknown ones (http://bhapp.c2b2.columbia.edu/PREPPI/cgi-bin/search.cgi?query=grb2&protein=P62993). At the LR cutoff 6,000, PREPPI predicts 107 interactions, among which 38 are validated by experiments. It would be very interesting to test whether the other predictions are true or not, and to further study their biological functions.

Many important biological processes are accomplished by macro-molecular complexes composed of a big number of proteins. In fact, cells are increasingly viewed as a collection of these modular complexes, each of which performs an independent, discrete biological function (Hartwell, Hopfield et al.

1999). Protein complexes can be inferred from PREPPI interaction networks by identifying clusters whose nodes (proteins) are densely interconnected. For example, the Califano group has identified a set of transcription factors including FOXM1 and c-MYB, which are involved in the regulation of genes that are differentially expressed in the germinal center (Lefebvre, Rajbhandari et al. 2010). Interestingly, about half of these genes encode proteins that seem to form a large supercomplex, combining the pre-replication complex with several mitotic proteins such as BUB1A/B and AURKA/B. It would be very interesting to test and to further study functions of this hypothetical complex, with the aid of information coming from available PREPPI structural interaction models.

It is very much an open question of applying the PREPPI interactions in future studies; nevertheless it is expected that a more complete image of the interactome of any organism will lead to more accurate understandings to the relationship between its genome and phenotype and also implications for network-based diagnostics and prognostics of complex disease.

BIBLIOGRAPHY

Adler, A. S., M. Lin, et al. (2006). "Genetic regulators of large-scale transcriptional signatures in cancer." Nature Genetics **38**(4): 421-430.

Akavia, U. D., O. Litvin, et al. (2010). "An integrated approach to uncover drivers of cancer." Cell **143**(6): 1005-1017.

Allison, D. B., X. Cui, et al. (2006). "Microarray data analysis: from disarray to consolidation and consensus." Nature Reviews. Genetics **7**(1): 55-65.

Aloy, P., B. Bottcher, et al. (2004). "Structure-based assembly of protein complexes in yeast." Science **303**(5666): 2026-2029.

Aloy, P., H. Ceulemans, et al. (2003). "The relationship between sequence and interaction divergence in proteins." Journal of Molecular Biology **332**(5): 989-998.

Aloy, P. and R. B. Russell (2002). "Interrogating protein interaction networks through structural biology." Proceedings of the National Academy of Sciences of the United States of America **99**(9): 5896-5901.

Aloy, P. and R. B. Russell (2002). "The third dimension for protein interactions and complexes." Trends in Biochemical Sciences **27**(12): 633-638.

Aloy, P. and R. B. Russell (2003). "InterPreTS: protein interaction prediction through tertiary structure." Bioinformatics **19**(1): 161-162.

Aloy, P. and R. B. Russell (2004). "Ten thousand interactions for the molecular biologist." Nature Biotechnology **22**(10): 1317-1321.

Aloy, P. and R. B. Russell (2006). "Structural systems biology: modelling protein interactions." Nature reviews Molecular cell biology **7**(3): 188-197.

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucl. Acids Res. **25**(17): 3389-3402.

Amit, I., M. Garber, et al. (2009). "Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses." Science **326**(5950): 257-263.

Anastassiou, D. (2007). "Computational analysis of the synergy among multiple interacting genes." Molecular Systems Biology **3**: 83.

Andreeva, A., D. Howorth, et al. (2004). "SCOP database in 2004: refinements integrate structure and sequence family data." Nucl. Acids Res. **32**(90001): D226-229.

Andreeva, A., D. Howorth, et al. (2008). "Data growth and its impact on the SCOP database: new developments." Nucleic Acids Res **36**(Database issue): D419-425.

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

Apweiler, R., A. Bairoch, et al. (2004). "UniProt: the Universal Protein knowledgebase." Nucleic Acids Research **32**(Database issue): D115-119.

Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nature Genetics **25**(1): 25-29.

Bader, G. D., D. Betel, et al. (2003). "BIND: the Biomolecular Interaction Network Database." Nucleic Acids Research **31**(1): 248-250.

Bader, G. D. and C. W. Hogue (2002). "Analyzing yeast protein-protein interaction data obtained from different sources." Nature Biotechnology **20**(10): 991-997.

Bader, J. S., A. Chaudhuri, et al. (2004). "Gaining confidence in high-throughput protein interaction networks." Nature Biotechnology **22**(1): 78-85.

Baker, D. and A. Sali (2001). "Protein structure prediction and structural genomics." Science **294**(5540): 93-96.

Barabasi, A. L. and Z. N. Oltvai (2004). "Network biology: understanding the cell's functional organization." Nature Reviews. Genetics **5**(2): 101-113.

Barrett, T., D. B. Troup, et al. (2011). "NCBI GEO: archive for functional genomics data sets--10 years on." Nucleic Acids Res **39**(Database issue): D1005-1010.

Bartel, D. P. (2009). "MicroRNAs: target recognition and regulatory functions." Cell **136**(2): 215-233.

Bashton, M. and C. Chothia (2002). "The geometry of domain combination in proteins." Journal of Molecular Biology **315**(4): 927-939.

Beaumont, M. A. and B. Rannala (2004). "The Bayesian revolution in genetics." Nature Reviews. Genetics **5**(4): 251-261.

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

Ben-Hur, A., C. S. Ong, et al. (2008). "Support vector machines and kernels for computational biology." PLoS Computational Biology **4**(10): e1000173.

Berggard, T., S. Linse, et al. (2007). "Methods for the detection and analysis of protein-protein interactions." Proteomics **7**(16): 2833-2842.

Bergholdt, R., Z. M. Storling, et al. (2007). "Integrative analysis for finding genes and networks involved in diabetes and other complex diseases." Genome Biology **8**(11): R253.

Berman, H. M., J. Westbrook, et al. (2000). "The Protein Data Bank." Nucleic Acids Research **28**(1): 235-242.

Birney, E., J. A. Stamatoyannopoulos, et al. (2007). "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project." Nature **447**(7146): 799-816.

Bishop, C. M. (2006). Pattern recognition and machine learning. New York, Springer.

Bochner, B. R. (2003). "New technologies to assess genotype-phenotype relationships." Nature Reviews. Genetics **4**(4): 309-314.

Bockhorst, J., M. Craven, et al. (2003). "A Bayesian network approach to operon prediction." Bioinformatics **19**(10): 1227-1235.

Bonetta, L. (2010). "Protein-protein interactions: Interactome under construction." Nature **468**(7325): 851-854.

Bordner, A. J. and R. Abagyan (2005). "Statistical analysis and prediction of protein-protein interfaces." Proteins **60**(3): 353-366.

Bradford, J. R., C. J. Needham, et al. (2006). "Insights into protein-protein interfaces using a Bayesian network prediction method." Journal of Molecular Biology **362**(2): 365-386.

Bradford, J. R. and D. R. Westhead (2003). "Asymmetric mutation rates at enzyme-inhibitor interfaces: Implications for the protein-protein docking problem." Protein Science **12**(9): 2099-2103.

Bradford, J. R. and D. R. Westhead (2005). "Improved prediction of protein-protein binding sites using a support vector machines approach." Bioinformatics **21**(8): 1487-1494.

Braun, P., M. Tasan, et al. (2009). "An experimentally derived confidence score for binary protein-protein interactions." Nature Methods **6**(1): 91-97.

Brown, P. O. and D. Botstein (1999). "Exploring the new world of the genome with DNA microarrays." Nature Genetics **21**(1 Suppl): 33-37.

Brown, S. D. and R. Balling (2001). "Systematic approaches to mouse mutagenesis." Current Opinion in Genetics and Development **11**(3): 268-273.

Brylinski, M. and J. Skolnick (2008). "A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation." Proceedings of the National Academy of Sciences of the United States of America **105**(1): 129-134.

Butland, G., J. M. Peregrin-Alvarez, et al. (2005). "Interaction network containing conserved and essential protein complexes in Escherichia coli." Nature **433**(7025): 531-537.

Butte, A. (2002). "The use and analysis of microarray data." Nature Reviews. Drug Discovery **1**(12): 951-960.

Caffrey, D. R., S. Somaroo, et al. (2004). "Are protein-protein interfaces more conserved in sequence than the rest of the protein surface?" Protein Science **13**(1): 190-202.

Calvano, S. E., W. Xiao, et al. (2005). "A network-based analysis of systemic inflammation in humans." Nature **437**(7061): 1032-1037.

Carro, M. S., W. K. Lim, et al. (2010). "The transcriptional network for mesenchymal transformation of brain tumours." Nature **463**(7279): 318-325.

Cavanagh, J. (2007). Protein NMR spectroscopy : principles and practice. Amsterdam ; Boston, Academic Press.

Chandonia, J. M. and S. E. Brenner (2006). "The impact of structural genomics: expectations and outcomes." Science **311**(5759): 347-351.

Chang, C. C. and C. J. Lin "LIBSVM, a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm/."

Chatr-aryamontri, A., A. Ceol, et al. (2007). "MINT: the Molecular INTeraction database." Nucleic Acids Res **35**(Database issue): D572-574.

Chen, C. C., C. Y. Lin, et al. (2009). "PPISearch: a web server for searching homologous protein-protein interactions across multiple species." Nucleic Acids Res **37**(Web Server issue): W369-375.

Chen, C. P., S. Posy, et al. (2005). "Specificity of cell-cell adhesion by classical cadherins: Critical role for low-affinity dimerization through beta-strand swapping." Proc Natl Acad Sci U S A **102**(24): 8531-8536.

Chen, H. L. and H. X. Zhou (2005). "Prediction of interface residues in protein-protein complexes by a consensus neural network method: Test against NMR data." Proteins-Structure Function and Bioinformatics **61**(1): 21-35.

Chen, J. R., B. H. Chang, et al. (2008). "Predicting PDZ domain-peptide interactions from primary sequences." Nature Biotechnology **26**(9): 1041-1045.

Chen, Y. C., Y. S. Lo, et al. (2007). "3D-partner: a web server to infer interacting partners and binding models." Nucleic Acids Res **35**(Web Server issue): W561-567.

Chothia, C. (1992). "Proteins. One thousand families for the molecular biologist." Nature **357**(6379): 543-544.

Chothia, C. and J. Janin (1975). "Principles of protein-protein recognition." Nature **256**(5520): 705-708.

Chua, H. N., W. Hugo, et al. (2009). "A probabilistic graph-theoretic approach to integrate multiple predictions for the protein-protein subnetwork prediction challenge." Annals of the New York Academy of Sciences **1158**: 224-233.

Chung, J. L., W. Wang, et al. (2006). "Exploiting sequence and structure homologs to identify protein-protein binding sites." Proteins **62**(3): 630-640.

Cole, C. and J. Warwicker (2002). "Side-chain conformational entropy at protein-protein interfaces." Protein Science **11**(12): 2860-2870.

Collins, S. R., K. M. Miller, et al. (2007). "Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map." Nature **446**(7137): 806-810.

Cortes, C. and V. Vapnik (1995). "Support-Vector Networks." Machine Learning **20**(3): 273-297.

Crowley, P. B. and A. Golovin (2005). "Cation-pi interactions in protein-protein interfaces." Proteins **59**(2): 231-239.

Cusick, M. E., H. Yu, et al. (2009). "Literature-curated protein interaction datasets." Nature Methods **6**(1): 39-46.

Dandekar, T., B. Snel, et al. (1998). "Conservation of gene order: a fingerprint of proteins that physically interact." Trends in Biochemical Sciences **23**(9): 324-328.

Das, M. K. and H. K. Dai (2007). "A survey of DNA motif finding algorithms." BMC Bioinformatics **8 Suppl 7**: S21.

Davis, F. P., D. T. Barkan, et al. (2007). "Host pathogen protein interactions predicted by comparative modeling." Protein Sci **16**(12): 2585-2596.

Davis, F. P., H. Braberg, et al. (2006). "Protein complex compositions predicted by structural similarity." Nucleic Acids Research **34**(10): 2943-2952.

Davis, F. P. and A. Sali (2005). "PIBASE: a comprehensive database of structurally defined protein interfaces." Bioinformatics **21**(9): 1901-1907.

de Vries, S. J. and A. M. Bonvin (2008). "How proteins get in touch: interface prediction in the study of biomolecular complexes." Current Protein and Peptide Science **9**(4): 394-406.

de Vries, S. J., A. D. van Dijk, et al. (2006). "WHISCY: what information does surface conservation yield? Application to data-driven docking." Proteins **63**(3): 479-489.

Deane, C. M., L. Salwinski, et al. (2002). "Protein interactions: two methods for assessment of the reliability of high throughput observations." Molecular and Cellular Proteomics **1**(5): 349-356.

Down, T. A., V. K. Rakyan, et al. (2008). "A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis." Nature Biotechnology **26**(7): 779-785.

Dreze, M., A. R. Carvunis, et al. (2011). "Evidence for network evolution in an Arabidopsis interactome map." Science **333**(6042): 601-607.

Elsner, M. and H. C. Mak (2011). "A modENCODE snapshot." Nature Biotechnology **29**(3): 238-240.

Enault, F., K. Suhre, et al. (2005). "Phydbac "Gene Function Predictor": a gene annotation tool based on genomic context analysis." BMC Bioinformatics **6**: 247.

Enright, A. J., I. Iliopoulos, et al. (1999). "Protein interaction maps for complete genomes based on gene fusion events." Nature **402**(6757): 86-90.

Ergun, A., C. A. Lawrence, et al. (2007). "A network biology approach to prostate cancer." Molecular Systems Biology **3**: 82.

Ewing, R. M., P. Chu, et al. (2007). "Large-scale mapping of human protein-protein interactions by mass spectrometry." Molecular Systems Biology **3**: 89.

Fariselli, P., F. Pazos, et al. (2002). "Prediction of protein-protein interaction sites in heterocomplexes with neural networks." European Journal of Biochemistry **269**(5): 1356-1361.

Fields, S. (2005). "High-throughput two-hybrid analysis. The promise and the peril." FEBS J **272**(21): 5391-5399.

Fields, S. and O. Song (1989). "A novel genetic system to detect protein-protein interactions." Nature **340**(6230): 245-246.

Finn, R. D., J. Mistry, et al. (2010). "The Pfam protein families database." Nucleic Acids Res **38**(Database issue): D211-222.

Fischer, M., Q. C. Zhang, et al. (2011). "MarkUs: a server to navigate sequence-structure-function space." Nucleic Acids Res **39**(Web Server issue): W357-361.

Fogel, G. B., V. W. Porto, et al. (2002). "Discovery of RNA structural elements using evolutionary computation." Nucleic Acids Res **30**(23): 5310-5317.

Forrest, L. R., C. L. Tang, et al. (2006). "On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins." Biophysical Journal **91**(2): 508-517.

Franke, L., H. van Bakel, et al. (2006). "Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes." American Journal of Human Genetics **78**(6): 1011-1025.

Friedberg, I. and A. Godzik (2005). "Connecting the protein structure universe by using sparse recurring fragments." Structure **13**(8): 1213-1224.

Friedberg, I. and A. Godzik (2005). "Fragnostic: walking through protein structure space." Nucleic Acids Res **33**(Web Server issue): W249-251.

Friedman, N. (2004). "Inferring cellular networks using probabilistic graphical models." Science **303**(5659): 799-805.

Fukuhara, N., N. Go, et al. (2007). "Prediction of interacting proteins from homology-modeled complex structures using sequence and structure scores." BIOPHYSICS **3**: 13-26.

Fukuhara, N. and T. Kawabata (2008). "HOMCOS: a server to predict interacting protein pairs and interacting sites by homology modeling of complex structures." Nucleic Acids Res **36**(Web Server issue): W185-189.

Gao, M. and J. Skolnick (2010). "Structural space of protein-protein interfaces is degenerate, close to complete, and highly connected." Proc Natl Acad Sci U S A **107**(52): 22517-22522.

Garrels, J. I. (1996). "YPD-A database for the proteins of Saccharomyces cerevisiae." Nucleic Acids Res **24**(1): 46-49.

Gavin, A. C., P. Aloy, et al. (2006). "Proteome survey reveals modularity of the yeast cell machinery." Nature **440**(7084): 631-636.

Gavin, A. C., M. Bosche, et al. (2002). "Functional organization of the yeast proteome by systematic analysis of protein complexes." Nature **415**(6868): 141-147.

Gerstein, M., A. Edwards, et al. (2003). "Structural genomics: current progress." Science **299**(5613): 1663.

Gerstein, M. B., Z. J. Lu, et al. (2010). "Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project." Science **330**(6012): 1775-1787.

Ginalski, K. (2006). "Comparative modeling for protein structure prediction." Current Opinion in Structural Biology **16**(2): 172-177.

Gingras, A. C., M. Gstaiger, et al. (2007). "Analysis of protein complexes using mass spectrometry." Nature Reviews. Molecular Cell Biology **8**(8): 645-654.

Giot, L., J. S. Bader, et al. (2003). "A protein interaction map of Drosophila melanogaster." Science **302**(5651): 1727-1736.

Glaser, F., D. M. Steinberg, et al. (2001). "Residue frequencies and pairing preferences at protein-protein interfaces." Proteins **43**(2): 89-102.

Goh, C. S. and F. E. Cohen (2002). "Co-evolutionary analysis reveals insights into protein-protein interactions." Journal of Molecular Biology **324**(1): 177-192.

Gray, J. J. (2006). "High-resolution protein-protein docking." Current Opinion in Structural Biology **16**(2): 183-193.

Grigoryan, G., A. W. Reinke, et al. (2009). "Design of protein-interaction specificity gives selective bZIP-binding peptides." Nature **458**(7240): 859-864.

Gunther, S., P. May, et al. (2007). "Docking without docking: ISEARCH--prediction of interactions using known interfaces." Proteins **69**(4): 839-844.

Haas, B. J. and M. C. Zody (2010). "Advancing RNA-Seq analysis." Nature Biotechnology **28**(5): 421-423.

Han, J. H., N. Kerrison, et al. (2006). "Divergence of interdomain geometry in two-domain proteins." Structure **14**(5): 935-945.

Hannon, G. J. (2002). "RNA interference." Nature **418**(6894): 244-251.

Hartshorn, M. J. (2002). "AstexViewer: a visualisation aid for structure-based drug design." Journal of Computer-Aided Molecular Design **16**(12): 871-881.

Hartwell, L. H., J. J. Hopfield, et al. (1999). "From molecular to modular cell biology." Nature **402**(6761 Suppl): C47-52.

Heller, M. J. (2002). "DNA microarray technology: devices, systems, and applications." Annual Review of Biomedical Engineering **4**: 129-153.

Hellman, L. M. and M. G. Fried (2007). "Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions." Nature Protocols **2**(8): 1849-1861.

Henrick, K. and J. M. Thornton (1998). "PQS: a protein quaternary structure file server." Trends in Biochemical Sciences **23**(9): 358-361.

Hermjakob, H., L. Montecchi-Palazzi, et al. (2004). "The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data." Nature Biotechnology **22**(2): 177-183.

Ho, Y., A. Gruhler, et al. (2002). "Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry." Nature **415**(6868): 180-183.

Holm, L., S. Kaariainen, et al. (2006). "Using Dali for structural comparison of proteins." Curr Protoc Bioinformatics **Chapter 5**: Unit 5 5.

Hu, Z., B. Ma, et al. (2000). "Conservation of polar residues as hot spots at protein interfaces." Proteins **39**(4): 331-342.

Huynen, M., B. Snel, et al. (2000). "Predicting protein function by genomic context: quantitative evaluation and qualitative inferences." Genome Research **10**(8): 1204-1210.

Huynen, M. A. and P. Bork (1998). "Measuring genome evolution." Proc Natl Acad Sci U S A **95**(11): 5849-5856.

Hwang, H., B. Pierce, et al. (2008). "Protein-protein docking benchmark version 3.0." Proteins-Structure Function and Bioinformatics **73**(3): 705-709.

Ito, T., T. Chiba, et al. (2001). "A comprehensive two-hybrid analysis to explore the yeast protein interactome." Proc Natl Acad Sci U S A **98**(8): 4569-4574.

Janin, J. and S. Wodak (2007). "The third CAPRI assessment meeting Toronto, Canada, April 20-21, 2007." Structure **15**(7): 755-759.

Jansen, R., D. Greenbaum, et al. (2002). "Relating whole-genome expression data with protein-protein interactions." Genome Research **12**(1): 37-46.

Jansen, R., H. Yu, et al. (2003). "A Bayesian networks approach for predicting protein-protein interactions from genomic data." Science **302**(5644): 449-453.

Jefferson, E. R., T. P. Walsh, et al. (2006). "Biological units and their effect upon the properties and prediction of protein-protein interactions." J Mol Biol **364**(5): 1118-1129.

Jones, N. C. and P. Pevzner (2004). An introduction to bioinformatics algorithms. Cambridge, MA, MIT Press.

Jones, S. and J. M. Thornton (1996). "Principles of protein-protein interactions." Proc Natl Acad Sci U S A **93**(1): 13-20.

Jones, S. and J. M. Thornton (1997). "Analysis of protein-protein interaction sites using surface patches." Journal of Molecular Biology **272**(1): 121-132.

Kaski, S., J. Rousu, et al. (2007). "Probabilistic modeling and machine learning in structural and systems biology." BMC Bioinformatics **8**.

Kerrien, S., Y. Alam-Faruque, et al. (2007). "IntAct--open source resource for molecular interaction data." Nucleic Acids Res **35**(Database issue): D561-565.

Kerrien, S., S. Orchard, et al. (2007). "Broadening the horizon--level 2.5 of the HUPO-PSI format for molecular interactions." BMC Biology **5**: 44.

Keshava Prasad, T. S., R. Goel, et al. (2009). "Human Protein Reference Database--2009 update." Nucleic Acids Res **37**(Database issue): D767-772.

Keskin, O. and R. Nussinov (2005). "Favorable scaffolds: proteins with different sequence, structure and function may associate in similar ways." Protein Engineering Design & Selection **18**(1): 11-24.

Keskin, O., R. Nussinov, et al. (2008). "PRISM: protein-protein interaction prediction by structural matching." Methods in Molecular Biology **484**: 505-521.

Keskin, Z., A. Gursoy, et al. (2008). "Principles of protein-protein interactions: What are the preferred ways for proteins to interact?" Chemical Reviews **108**(4): 1225-1244.

Kharchenko, P. V., M. Y. Tolstorukov, et al. (2008). "Design and analysis of ChIP-seq experiments for DNA-binding proteins." Nature Biotechnology **26**(12): 1351-1359.

Kiefer, F., K. Arnold, et al. (2009). "The SWISS-MODEL Repository and associated resources." Nucleic Acids Res **37**(Database issue): D387-392.

Kiel, C., P. Beltrao, et al. (2008). "Analyzing protein interaction networks using structural information." Annual Review of Biochemistry **77**: 415-441.

Kim, W. K., A. Henschel, et al. (2006). "The many faces of protein-protein interactions: A compendium of interface geometry." PLoS Computational Biology **2**(9): e124.

Kim, W. K. and J. C. Ison (2005). "Survey of the geometric association of domain-domain interfaces." Proteins **61**(4): 1075-1088.

Kittichotirat, W., M. Guerquin, et al. (2009). "Protinfo PPC: a web server for atomic level prediction of protein complexes." Nucleic Acids Res **37**(Web Server issue): W519-525.

Koike, A. and T. Takagi (2004). "Prediction of protein-protein interaction sites using support vector machines." Protein Engineering, Design and Selection **17**(2): 165-173.

Kolodny, R., P. Koehl, et al. (2005). "Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures." Journal of Molecular Biology **346**(4): 1173-1188.

Korkin, D., F. P. Davis, et al. (2005). "Localization of protein-binding sites within families of proteins." Protein Science **14**(9): 2350-2360.

Krissinel, E. and K. Henrick (2007). "Inference of macromolecular assemblies from crystalline state." Journal of Molecular Biology **372**(3): 774-797.

Krogan, N. J., G. Cagney, et al. (2006). "Global landscape of protein complexes in the yeast Saccharomyces cerevisiae." Nature **440**(7084): 637-643.

Kufareva, I., L. Budagyan, et al. (2007). "PIER: protein interface recognition for structural proteomics." Proteins **67**(2): 400-417.

Lage, K., E. O. Karlberg, et al. (2007). "A human phenome-interactome network of protein complexes implicated in genetic disorders." Nature Biotechnology **25**(3): 309-316.

Lander, E. S., L. M. Linton, et al. (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

Larranaga, P., B. Calvo, et al. (2006). "Machine learning in bioinformatics." Briefings in Bioinformatics **7**(1): 86-112.

Laskowski, R. A., J. D. Watson, et al. (2005). "ProFunc: a server for predicting protein function from 3D structure." Nucl. Acids Res. **33**(suppl_2): W89-93.

Lee, H., Z. Li, et al. (2010). "High-throughput computational structure-based characterization of protein families: START domains and implications for structural genomics." J Struct Funct Genomics **11**(1): 51-59.

Lee, I., S. V. Date, et al. (2004). "A probabilistic functional network of yeast genes." Science **306**(5701): 1555-1558.

Lefebvre, C., W. K. Lim, et al. (2007). "A context-specific network of protein-DNA and protein-protein interactions reveals new regulatory motifs in human B cells." Lecture Notes in Bioinformatics (LNCS) **4532**: 42-56.

Lefebvre, C., P. Rajbhandari, et al. (2010). "A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers." Molecular Systems Biology **6**: 377.

Lehne, B. and T. Schlitt (2009). "Protein-protein interaction databases: keeping up with growing interactomes." Human Genomics **3**(3): 291-297.

Letunic, I., T. Doerks, et al. (2009). "SMART 6: recent updates and new developments." Nucleic Acids Research **37**: D229-D232.

Levitt, M. (2007). "Growth of novel protein structural data." Proc Natl Acad Sci U S A **104**(9): 3183-3188.

Levitt, M. (2009). "Nature of the protein universe." Proc Natl Acad Sci U S A **106**(27): 11079-11084.

Li, S., C. M. Armstrong, et al. (2004). "A map of the interactome network of the metazoan C. elegans." Science **303**(5657): 540-543.

Li, W. Z. and A. Godzik (2006). "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences." Bioinformatics **22**(13): 1658-1659.

Liang, S., C. Zhang, et al. (2006). "Protein binding site prediction using an empirical scoring function." Nucleic Acids Res **34**(13): 3698-3707.

Lichtarge, O., H. R. Bourne, et al. (1996). "An evolutionary trace method defines binding surfaces common to protein families." Journal of Molecular Biology **257**(2): 342-358.

Lin, N., B. Wu, et al. (2004). "Information assessment on predicting protein-protein interactions." BMC Bioinformatics **5**: 154.

Littler, S. J. and S. J. Hubbard (2005). "Conservation of orientation and sequence in protein domain-domain interactions." Journal of Molecular Biology **345**(5): 1265-1279.

Liu, H., L. Jin, et al. (2010). "Atomic structure of human adenovirus by cryo-EM reveals interactions among protein networks." Science **329**(5995): 1038-1043.

Lo Conte, L., B. Ailey, et al. (2000). "SCOP: a Structural Classification of Proteins database." Nucl. Acids Res. **28**(1): 257-259.

Lo Conte, L., C. Chothia, et al. (1999). "The atomic structure of protein-protein recognition sites." Journal of Molecular Biology **285**(5): 2177-2198.

Lu, L., H. Lu, et al. (2002). "MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading." Proteins **49**(3): 350-364.

Ma, W., A. Trusina, et al. (2009). "Defining network topologies that can achieve biochemical adaptation." Cell **138**(4): 760-773.

Mani, K. M., C. Lefebvre, et al. (2008). "A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas." Molecular Systems Biology **4**: 169.

Marcotte, E. M., M. Pellegrini, et al. (1999). "Detecting protein function and protein-protein interactions from genome sequences." Science **285**(5428): 751-753.

Marcotte, E. M., M. Pellegrini, et al. (1999). "A combined algorithm for genome-wide prediction of protein function." Nature **402**(6757): 83-86.

Mardis, E. R. (2007). "ChIP-seq: welcome to the new frontier." Nature Methods **4**(8): 613-614.

Marti-Renom, M. A., A. C. Stuart, et al. (2000). "Comparative protein structure modeling of genes and genomes." Annual Review of Biophysics and Biomolecular Structure **29**: 291-325.

Mathe, C., M. F. Sagot, et al. (2002). "Current methods of gene prediction, their strengths and weaknesses." Nucleic Acids Res **30**(19): 4103-4117.

Matthews, L. R., P. Vaglio, et al. (2001). "Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs"." Genome Research **11**(12): 2120-2126.

McDowall, M. D., M. S. Scott, et al. (2009). "PIPs: human protein-protein interaction prediction database." Nucleic Acids Res **37**(Database issue): D651-656.

Mewes, H. W., K. Albermann, et al. (1997). "MIPS: a database for protein sequences, homology data and yeast genome information." Nucleic Acids Research **25**(1): 28-30.

Mintseris, J. and Z. P. Weng (2005). "Structure, function, and evolution of transient and obligate protein-protein interactions." Proceedings of the National Academy of Sciences of the United States of America **102**(31): 10930-10935.

Mirkovic, N., Z. Li, et al. (2007). "Strategies for high-throughput comparative modeling: applications to leverage analysis in structural genomics and protein family organization." Proteins **66**(4): 766-777.

Mitchell, T. M. (1997). Machine Learning. New York, McGraw-Hill.

Mortazavi, A., B. A. Williams, et al. (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq." Nature Methods **5**(7): 621-628.

Muers, M. (2011). "Functional genomics: the modENCODE guide to the genome." Nature Reviews. Genetics **12**(2): 80.

Muggleton, S. H. (2005). "Machine learning for systems biology." Inductive Logic Programming, Proceedings **3625**: 416-423.

Murakami, Y. and S. Jones (2006). "SHARP2: protein-protein interaction predictions using patch analysis." Bioinformatics **22**(14): 1794-1795.

Musso, G. A., Z. Zhang, et al. (2007). "Experimental and computational procedures for the assessment of protein complexes on a genome-wide scale." Chemical Reviews **107**(8): 3585-3600.

Myers, C. L., D. R. Barrett, et al. (2006). "Finding function: evaluation methods for functional genomic data." BMC Genomics **7**: 187.

Myers, R. M., J. Stamatoyannopoulos, et al. (2011). "A user's guide to the encyclopedia of DNA elements (ENCODE)." PLoS Biology **9**(4): e1001046.

Neapolitan, R. E. (2004). Learning Bayesian networks. Upper Saddle River, NJ, Pearson Prentice Hall.

Neuvirth, H., R. Raz, et al. (2004). "ProMate: a structure based prediction program to identify the location of protein-protein binding sites." J Mol Biol **338**(1): 181-199.

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

Nibbe, R. K., S. Markowitz, et al. (2009). "Discovery and scoring of protein interaction subnetworks discriminative of late stage human colon cancer." Molecular and Cellular Proteomics **8**(4): 827-845.

Nooren, I. M. and J. M. Thornton (2003). "Diversity of protein-protein interactions." EMBO Journal **22**(14): 3486-3492.

O'Neill, L. P. and B. M. Turner (1996). "Immunoprecipitation of chromatin." Methods in Enzymology **274**: 189-197.

Ogmen, U., O. Keskin, et al. (2005). "PRISM: protein interactions by structural matching." Nucleic Acids Research **33**: W331-W336.

Ooi, S. L., D. D. Shoemaker, et al. (2003). "DNA helicase gene interaction network defined using synthetic lethality analyzed by microarray." Nature Genetics **35**(3): 277-286.

Oti, M., B. Snel, et al. (2006). "Predicting disease genes using protein-protein interactions." Journal of Medical Genetics **43**(8): 691-698.

Pal, D. and D. Eisenberg (2005). "Inference of protein function from protein structure." Structure **13**(1): 121-130.

Park, P. J. (2009). "ChIP-seq: advantages and challenges of a maturing technology." Nature Reviews. Genetics **10**(10): 669-680.

Patel, S. D., C. P. Chen, et al. (2003). "Cadherin-mediated cell-cell adhesion: sticking together as a family." Curr Opin Struct Biol **13**(6): 690-698.

Patel, S. D., C. Ciatto, et al. (2006). "Type II cadherin ectodomain structures: Implications for classical cadherin specificity." Cell **124**(6): 1255-1268.

Pazos, F. and A. Valencia (2001). "Similarity of phylogenetic trees as indicator of protein-protein interaction." Protein Engineering **14**(9): 609-614.

Pearl, F. M., C. F. Bennett, et al. (2003). "The CATH database: an extended protein family resource for structural and functional genomics." Nucleic Acids Res **31**(1): 452-455.

Pellegrini, M., E. M. Marcotte, et al. (1999). "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles." Proceedings of the National Academy of Sciences of the United States of America **96**(8): 4285-4288.

Petrey, D., M. Fischer, et al. (2009). "Structural relationships among proteins with different global topologies and their implications for function annotation strategies." Proc Natl Acad Sci USA **106**(41): 17377-17382.

Petrey, D. and B. Honig (2003). GRASP2: Visualization, Surface Properties, and Electrostatics of Macromolecular Structures and Sequences. Methods in Enzymology, Academic Press. **Volume 374:** 492-509.

Petrey, D. and B. Honig (2005). "Protein structure prediction: Inroads to biology." Molecular Cell **20**(6): 811-819.

Petrey, D. and B. Honig (2009). "Is protein classification necessary? Toward alternative approaches to function annotation." Curr Opin Struct Biol **19**(3): 363-368.

Petrey, D., F. Markus, et al. (2009). "Functional relationships between apparently unrelated proteins: implications for the nature of protein structure space." Proceedings of the National Academy of Sciences of the United States of America **Submitted**.

Pevsner, J. (2009). Bioinformatics and functional genomics. Hoboken, N.J., Wiley-Blackwell.

Phizicky, E. M. and S. Fields (1995). "Protein-protein interactions: methods for detection and analysis." Microbiological Reviews **59**(1): 94-123.

Piehler, J. (2005). "New methodologies for measuring protein interactions in vivo and in vitro." Current Opinion in Structural Biology **15**(1): 4-14.

Pieper, U., N. Eswar, et al. (2006). "MODBASE: a database of annotated comparative protein structure models and associated resources." Nucleic Acids Research **34**(Database issue): D291-295.

Porollo, A. and J. Meller (2007). "Prediction-based fingerprints of protein-protein interactions." Proteins **66**(3): 630-645.

Pupko, T., R. E. Bell, et al. (2002). "Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of

evolutionary determinants within their homologues." Bioinformatics **18 Suppl 1**: S71-77.

Qian, J., M. Dolled-Filhart, et al. (2001). "Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions." Journal of Molecular Biology **314**(5): 1053-1066.

Rain, J. C., L. Selig, et al. (2001). "The protein-protein interaction map of Helicobacter pylori." Nature **409**(6817): 211-215.

Reguly, T., A. Breitkreutz, et al. (2006). "Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae." Journal of Biology **5**(4): 11.

Res, I., I. Mihalek, et al. (2005). "An evolution based classifier for prediction of protein interfaces without using protein structures." Bioinformatics **21**(10): 2496-2501.

Roguev, A., S. Bandyopadhyay, et al. (2008). "Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast." Science **322**(5900): 405-410.

Rost, B. and C. Sander (1993). "Improved prediction of protein secondary structure by use of sequence profiles and neural networks." Proc Natl Acad Sci U S A **90**(16): 7558-7562.

Roy, S., J. Ernst, et al. (2010). "Identification of functional elements and regulatory circuits by Drosophila modENCODE." Science **330**(6012): 1787-1797.

Rual, J. F., K. Venkatesan, et al. (2005). "Towards a proteome-scale map of the human protein-protein interaction network." Nature **437**(7062): 1173-1178.

Russell, R. B., P. D. Sasieni, et al. (1998). "Supersites within superfolds. Binding site similarity in the absence of homology." Journal of Molecular Biology **282**(4): 903-918.

Sachs, K., O. Perez, et al. (2005). "Causal protein-signaling networks derived from multiparameter single-cell data." Science **308**(5721): 523-529.

Sacquin-Mora, S., A. Carbone, et al. (2008). "Identification of protein interaction partners and protein-protein interaction sites." Journal of Molecular Biology **382**(5): 1276-1289.

Sali, A. and T. L. Blundell (1993). "Comparative Protein Modelling by Satisfaction of Spatial Restraints." Journal of Molecular Biology **234**(3): 779-815.

Salwinski, L. and D. Eisenberg (2003). "Computational methods of analysis of protein-protein interactions." Current Opinion in Structural Biology **13**(3): 377-382.

Salwinski, L., L. Licata, et al. (2009). "Recurated protein interaction datasets." Nature Methods **6**(12): 860-861.

Salwinski, L., C. S. Miller, et al. (2004). "The Database of Interacting Proteins: 2004 update." Nucleic Acids Res **32**(Database issue): D449-451.

Sanchez, I. E., P. Beltrao, et al. (2008). "Genome-wide prediction of SH2 domain targets using structural information and the FoldX algorithm." PLoS Computational Biology **4**(4): e1000052.

Sanchez, R. and A. Sali (1998). "Large-scale protein structure modeling of the Saccharomyces cerevisiae genome." Proc Natl Acad Sci U S A **95**(23): 13597-13602.

Sandberg, R., G. Winberg, et al. (2001). "Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier." Genome Research **11**(8): 1404-1409.

Sauer, U., M. Heinemann, et al. (2007). "Genetics. Getting closer to the whole picture." Science **316**(5824): 550-551.

Shapiro, L. and B. Honig (2007). "Cell-to-cell contact and extracellular matrix." Current Opinion in Cell Biology **19**(5): 493-494.

Sheinerman, F. B., B. Al-Lazikani, et al. (2003). "Sequence, structure and energetic determinants of phosphopeptide selectivity of SH2 domains." J Mol Biol **334**(4): 823-841.

Sheinerman, F. B. and B. Honig (2002). "On the role of electrostatic interactions in the design of protein-protein interfaces." J Mol Biol **318**(1): 161-177.

Shindyalov, I. N. and P. E. Bourne (1998). "Protein structure alignment by incremental combinatorial extension (CE) of the optimal path." Protein Engineering **11**(9): 739-747.

Shoemaker, B. A. and A. R. Panchenko (2007). "Deciphering protein-protein interactions. Part I. Experimental techniques and databases." PLoS Comput Biol **3**(3): e42.

Shoemaker, B. A. and A. R. Panchenko (2007). "Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners." PLoS Comput Biol **3**(4): e43.

Shoemaker, B. A., A. R. Panchenko, et al. (2006). "Finding biologically relevant protein domain interactions: conserved binding mode analysis." Protein Sci **15**(2): 352-361.

Sikic, M., S. Tomic, et al. (2009). "Prediction of protein-protein interaction sites in sequences and 3D structures by random forests." PLoS Computational Biology **5**(1): e1000278.

Singh, R., D. Park, et al. (2010). "Struct2Net: a web service to predict protein-protein interactions using a structure-based approach." Nucleic Acids Res **38 Suppl**: W508-515.

Skolnick, J., A. K. Arakaki, et al. (2009). "The continuity of protein structure space is an intrinsic property of proteins." Proceedings of the National Academy of Sciences of the United States of America **106**(37): 15690-15695.

Skrabanek, L., H. K. Saini, et al. (2008). "Computational prediction of protein-protein interactions." Mol Biotechnol **38**(1): 1-17.

Smith, G. R. and M. J. Sternberg (2002). "Prediction of protein-protein interactions by docking methods." Current Opinion in Structural Biology **12**(1): 28-35.

Soong, T. T., K. O. Wrzeszczynski, et al. (2008). "Physical protein-protein interactions predicted from microarrays." Bioinformatics **24**(22): 2608-2614.

Soto, C. S., M. Fasnacht, et al. (2008). "Loop modeling: Sampling, filtering, and scoring." Proteins-Structure Function and Bioinformatics **70**(3): 834-843.

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

Sprinzak, E., S. Sattath, et al. (2003). "How reliable are experimental protein-protein interaction data?" Journal of Molecular Biology **327**(5): 919-923.

Stark, C., B. J. Breitkreutz, et al. (2006). "BioGRID: a general repository for interaction datasets." Nucleic Acids Research **34**: D535-D539.

Stein, A., A. Panjkovich, et al. (2009). "3did Update: domain-domain and peptide-mediated interactions of known 3D structure." Nucleic Acids Res **37**(Database issue): D300-304.

Stelzl, U., U. Worm, et al. (2005). "A human protein-protein interaction network: a resource for annotating the proteome." Cell **122**(6): 957-968.

Stolovitzky, G., R. J. Prill, et al. (2009). "Lessons from the DREAM2 Challenges." Annals of the New York Academy of Sciences **1158**: 159-195.

Stormo, G. D., T. D. Schneider, et al. (1982). "Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli." Nucleic Acids Res **10**(9): 2997-3011.

Sun, L., A. M. Hui, et al. (2006). "Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain." Cancer Cell **9**(4): 287-300.

Szilagyi, A., V. Grimm, et al. (2005). "Prediction of physical protein-protein interactions." Physical Biology **2**(2): S1-16.

Szklarczyk, D., A. Franceschini, et al. (2010). "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored." Nucleic Acids Res.

Tarassov, K., V. Messier, et al. (2008). "An in vivo map of the yeast protein interactome." Science **320**(5882): 1465-1470.

Tarca, A. L., V. J. Carey, et al. (2007). "Machine learning and its applications to biology." PLoS Computational Biology **3**(6): e116.

Tastan, O., Y. Qi, et al. (2009). "Prediction of interactions between HIV-1 and human proteins by information integration." Pacific Symposium on Biocomputing: 516-527.

Terwilliger, T. C., D. Stuart, et al. (2009). "Lessons from structural genomics." Annu Rev Biophys **38**: 371-383.

Tian, L., S. A. Greenberg, et al. (2005). "Discovering statistically significant pathways in expression profiling studies." Proc Natl Acad Sci U S A **102**(38): 13544-13549.

Tong, A. H., M. Evangelista, et al. (2001). "Systematic genetic analysis with ordered arrays of yeast deletion mutants." Science **294**(5550): 2364-2368.

Tong, A. H., G. Lesage, et al. (2004). "Global mapping of the yeast genetic interaction network." Science **303**(5659): 808-813.

Troyanskaya, O. G., K. Dolinski, et al. (2003). "A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae)." Proc Natl Acad Sci U S A **100**(14): 8348-8353.

Tsai, C. J., S. L. Lin, et al. (1996). "Protein-protein interfaces: Architectures and interactions in protein-protein interfaces and in protein cores. Their similarities and differences." Critical Reviews in Biochemistry and Molecular Biology **31**(2): 127-152.

Tsai, J., C. Rohl, et al. (2006). "Cataloging the relationships between proteins." Molecular Biotechnology **34**(1): 69-93.

Tuncbag, N., G. Kar, et al. (2009). "A survey of available tools and web servers for analysis of protein-protein interactions and interfaces." Briefings in Bioinformatics **10**(3): 217-232.

Turinsky, A. L., S. Razick, et al. (2010). "Literature curation of protein interactions: measuring agreement across major public databases." Database (Oxford) **2010**: baq026.

Uetz, P., L. Giot, et al. (2000). "A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae." Nature **403**(6770): 623-627.

Valdar, W. S. and J. M. Thornton (2001). "Protein-protein interfaces: analysis of amino acid conservation in homodimers." Proteins **42**(1): 108-124.

Valencia, A. and F. Pazos (2002). "Computational methods for the prediction of protein interactions." Current Opinion in Structural Biology **12**(3): 368-373.

Venter, J. C., M. D. Adams, et al. (2001). "The sequence of the human genome." Science **291**(5507): 1304-1351.

Vidal, M., M. E. Cusick, et al. (2011). "Interactome networks and human disease." Cell **144**(6): 986-998.

Vidan, S. and M. Snyder (2001). "Large-scale mutagenesis: yeast genetics in the genome era." Current Opinion in Biotechnology **12**(1): 28-34.

Vitkup, D., E. Melamud, et al. (2001). "Completeness in structural genomics." Nature Structural Biology **8**(6): 559-566.

von Mering, C., M. Huynen, et al. (2003). "STRING: a database of predicted functional associations between proteins." Nucleic Acids Res **31**(1): 258-261.

von Mering, C., L. J. Jensen, et al. (2005). "STRING: known and predicted protein-protein associations, integrated and transferred across organisms." Nucleic Acids Res **33**(Database issue): D433-437.

von Mering, C., R. Krause, et al. (2002). "Comparative assessment of large-scale data sets of protein-protein interactions." Nature **417**(6887): 399-403.

Waldron, R. T., T. Iglesias, et al. (1999). "The pleckstrin homology domain of protein kinase D interacts preferentially with the eta isoform of protein kinase C." Journal of Biological Chemistry **274**(14): 9224-9230.

Wang, B., P. Chen, et al. (2006). "Predicting protein interaction sites from residue spatial sequence profile and evolution rate." FEBS Letters **580**(2): 380-384.

Wang, B., H. S. Wong, et al. (2006). "Inferring protein-protein interacting sites using residue conservation and evolutionary information." Protein and Peptide Letters **13**(10): 999-1005.

Wang, K., M. Saito, et al. (2009). "Genome-wide identification of post-translational modulators of transcription factor activity in human B cells." Nature Biotechnology **27**(9): 829-839.

Wang, Z., M. Gerstein, et al. (2009). "RNA-Seq: a revolutionary tool for transcriptomics." Nature Reviews. Genetics **10**(1): 57-63.

Wass, M. N., G. Fuentes, et al. (2011). "Towards the prediction of protein interaction partners using physical docking." Molecular Systems Biology **7**: 469.

Wodak, S. J. and R. Mendez (2004). "Prediction of protein-protein interactions: the CAPRI experiment, its evaluation and implications." Current Opinion in Structural Biology **14**(2): 242-249.

Won, K. J., A. Prugel-Bennett, et al. (2004). "Training HMM structure with genetic algorithm for biological sequence analysis." Bioinformatics **20**(18): 3613-3619.

Woolfson, M. M. (1997). An introduction to X-ray crystallography. Cambridge ; New York, NY, USA, Cambridge University Press.

Wu, X., L. Zhu, et al. (2006). "Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations." Nucleic Acids Res **34**(7): 2137-2150.

Xiang, Z. X., P. J. Steinbach, et al. (2007). "Prediction of side-chain conformations on protein surfaces." Proteins-Structure Function and Bioinformatics **66**(4): 814-823.

Xie, L. and P. E. Bourne (2008). "Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments." Proc Natl Acad Sci U S A **105**(14): 5441-5446.

Yan, C., V. Honavar, et al. (2004). "Identification of interface residues in protease-inhibitor and antigen-antibody complexes: a support vector machine approach." Neural Comput Appl **13**(2): 123-129.

Yang, A. S. and B. Honig (2000). "An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance." Journal of Molecular Biology **301**(3): 665-678.

Yu, H., P. Braun, et al. (2008). "High-quality binary protein interaction map of the yeast interactome network." Science **322**(5898): 104-110.

Yu, H., N. M. Luscombe, et al. (2004). "Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs." Genome Research **14**(6): 1107-1118.

Zhang, L. V., S. L. Wong, et al. (2004). "Predicting co-complexed protein pairs using genomic and proteomic data integration." BMC Bioinformatics **5**: 38.

Zhang, Q. C., L. Deng, et al. (2011). "PredUs: a web server for predicting protein interfaces using structural neighbors." Nucleic Acids Res **39**(Web Server issue): W283-287.

Zhang, Q. C., D. Petrey, et al. (2010). "Protein interface conservation across structure space." Proc Natl Acad Sci U S A **107**(24): 10896-10901.

Zhang, X., C. Guo, et al. (2008). "Epitope tagging of endogenous proteins for genome-wide ChIP-chip studies." Nature Methods **5**(2): 163-165.

Zhang, X., L. Jin, et al. (2010). "3.3 A cryo-EM structure of a nonenveloped virus reveals a priming mechanism for cell entry." Cell **141**(3): 472-482.

Zhang, Y. (2008). "Progress and challenges in protein structure prediction." Current Opinion in Structural Biology **18**(3): 342-348.

Zhang, Y. and J. Skolnick (2005). "TM-align: a protein structure alignment algorithm based on the TM-score." Nucleic Acids Res **33**(7): 2302-2309.

Zhou, H. X. and S. Qin (2007). "Interaction-site prediction for protein complexes: a critical assessment." Bioinformatics **23**(17): 2203-2209.

Zhou, H. X. and Y. Shan (2001). "Prediction of protein interaction sites from sequence profile and residue neighbor list." Proteins **44**(3): 336-343.

Zhu, J., H. Fan, et al. (2008). "Refining homology models by combining replica-exchange molecular dynamics and statistical potentials." Proteins-Structure Function and Bioinformatics **72**(4): 1171-1188.