# Structure-based algorithms for protein-protein interaction prediction

by

Raghavendra Hosur

Submitted to the Department of Materials Science and Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2012



Author .....
Department of Materials Science and Engineering
May 17, 2012

Certified by .....
Bonnie Berger
Professor of Applied Mathematics
Thesis Supervisor

Accepted by .....
Prof. Gerbrand Ceder
Chairman, Department Committee on Graduate Students

**Structure-based algorithms for protein-protein interaction prediction**

by

Raghavendra Hosur

Submitted to the Department of Materials Science and Engineering
on May 17, 2012, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

## Abstract

Protein-protein interactions (PPIs) play a central role in all biological processes. Akin to the complete sequencing of genomes, complete descriptions of interactomes is a fundamental step towards a deeper understanding of biological processes, and has a vast potential to impact systems biology, genomics, molecular biology and therapeutics. PPIs are critical in maintenance of cellular integrity, metabolism, transcription/translation, and cell-cell communication.

This thesis develops new methods that significantly advance our efforts at structure-based approaches to predict PPIs and boost confidence in emerging high-throughput (HTP) data. The aims of this thesis are, 1) to utilize physicochemical properties of protein interfaces to better predict the putative interacting regions and increase coverage of PPI prediction, 2) increase confidence in HTP datasets by identifying likely experimental errors, and 3) provide residue-level information that gives us insights into structure-function relationships in PPIs. Taken together, these methods will vastly expand our understanding of macromolecular networks.

In this thesis, I introduce two computational approaches for structure-based protein-protein interaction prediction: iWRAP and Coev2Net. iWRAP is an interface threading approach that utilizes biophysical properties specific to protein interfaces to improve PPI prediction. Unlike previous structure-based approaches that use single structures to make predictions, iWRAP first builds profiles that characterize the hydrophobic, electrostatic and structural properties specific to protein interfaces from multiple interface alignments. Compatibility with these profiles is used to predict the putative interface region between the two proteins. In addition to improved interface prediction, iWRAP provides better accuracy and close to 50% increase in coverage on genome-scale PPI prediction tasks. As an application, we effectively combine iWRAP with genomic data to identify novel cancer related genes involved in chromatin remodeling, nucleosome organization and ribonuclear complex assembly – processes known to be critical in cancer.

Coev2Net addresses some of the limitations of iWRAP, and provides techniques

to increase coverage and accuracy even further. Unlike earlier sequence and structure profiles, Coev2Net explicitly models long-distance correlations at protein interfaces. By formulating interface co-evolution as a high-dimensional sampling problem, we enrich sequence/structure profiles with artificial interacting homologus sequences for families which do not have known multiple interacting homologs. We build a spanning-tree based graphical model induced by the simulated sequences as our interface profile. Cross-validation results indicate that this approach is as good as previous methods at PPI prediction. We show that Coev2Net's predictions correlate with experimental observations and experimentally validate some of the high-confidence predictions. Furthermore, we demonstrate how analysis of the predicted interfaces together with human genomic variation data can help us understand the role of these mutations in disease and normal cells.

Thesis Supervisor: Bonnie Berger
Title: Professor of Applied Mathematics

# Publications

Some ideas and figures have appeared previously in the following publications:

Park, D., Singh, R., Xu, J., **Hosur, R.** and Berger, B. Struct2Net: a webservice to predict protein-protein interactions using a structure-based approach, *Nucleic Acids Research*, 38(W508-15), Jul 2010.

**Hosur, R.**, Xu, J., Bienkowska, J. and Berger, B. iWRAP: an interface threading approach with application to cancer-related protein-protein interactions, *Journal of Molecular Biology*, 405(5):1295-1310. Feb 2011. *The article and image were featured on the cover*

**Hosur, R.**, Peng, J., Arunachalam, V., Stelzl, U., Xu, J., Perrimon, N., Bienkowska, J. and Berger, B. A computational framework for boosting confidence in high-throughput protein-protein interaction datasets, *submitted*

**Hosur, R.**, Singh, R. and Berger, B. Sparse estimation for structural variability, *Algorithms for Molecular Biology*, 6:12. Jun 2011.
*Selected for fast-track publication from WABI 2010*

Bryan, A., Starner, J., **Hosur, R.**, Clark, P. and Berger, B. Structure-based prediction reveals capping motifs that inhibit $\beta$-helix aggregation, *Proceedings of the National Academy of Sciences*, 108(27):11099-11104, Jul 2011.

Daniels, N., **Hosur, R.**, Cowen, L. and Berger, B. SMURFLite: combining simplified Markov random fields with simulated evolution improves remote homology detection for beta-structural proteins into the twilight zone, *Bioinformatics*, 2012,

doi: 10.1093/bioinformatics/bts110.

# Acknowledgments

I would like to take this opportunity to express my heartfelt appreciation and gratitude to many who have supported me throughout my graduate studies at MIT. First, I am indebted to my advisor, Prof Bonnie Berger, for her support, advice and mentorship. None of this would have been possible without her constant guidance, motivation and encouragement, even during the tough times of my PhD. Her never-ending enthusiasm for research, positive attitude and belief in me have helped me evolve both academically and personally. It is something that I will always remember and cherish.

I would like to especially thank, Jadwiga Bienkowska, for her support, help and encouragement during my thesis. Her knowledge, combined with her enthusiasm in helping answer all the technical challenges that we faced during the course of this thesis helped me learn a lot about the field. I also appreciate her taking time out from her busy schedule to meet us frequently to discuss about my thesis and provide invaluable suggestions.

I would like to thank my thesis committee, Prof Collin Stultz, Prof Samuel Allen and Prof Polina Anikeeva for their feedback and suggestions on my thesis.

I want to thank the current as well as past members of the Berger lab. My discussions and brain-storming sessions with Jian over the last year have helped me gain from his vast knowledge. Discussions with Rohit, Vinu, Luke, Jason, Michael, Jerome and Charlie have also helped me think about problems from a broader perspective.

I also want to thank George, Po-Ru, Mark, Hadar and Patrick for comments and suggestions during group meetings that have helped me improve as a researcher. I want to thank Patrice Macaluso, who always welcomed me to the lab with a smile and some home-made treats.

I want to thank all my friends here at MIT – Aniruddh, Ahmed, Andrea, Vikrant, Vaibhav, Vivek, Kashi, Harshad, Pari, Chaitanya, Oshani, Sahil, Kedia, Karthik, Murali, Varun, Neeraj who have made the stay a fun and exciting experience. I will always cherish the "latt", "dudo" and "muddy" sessions with the gang. I want to especially thank Manas, Navin and Angad for their constant support and friendship.

Finally, I want to thank my parents who have made immense sacrifices so that me and my sisters can be where we are today. Their unwavering love and belief in me makes me want to be a better person every day. This thesis is as much for them as it is for me.

# Contents

# List of Figures

# List of Tables

# Abbreviations

| | |
|---|---|
| **AUC** | Area under receiver operator characteristic curve (ROC) |
| **ASA** | Accessible surface area (by solvent) |
| **FDR** | False discovery rate |
| **GO** | Gene Ontology |
| **HMM** | Hidden Markov model |
| **HTP** | High-throughput |
| **LUMIER** | Luminescence-based mammalian interactome mapping |
| **MSA** | Multiple sequence alignment |
| **mRNA** | Messenger ribonucleic acid |
| **NR** | Non-redundant |
| **PDB** | Protein data bank |
| **PPI** | Protein-protein interaction |
| **ROC** | Receiver operator characteristic curve |
| **SNP** | Single nucleotide polymorphism |
| **TAP** | Tandem affinity purification |
| **Y2H** | Yeast two-hybrid |

# Glossary

**Interactome** The whole set of protein-protein interactions in an organism.

**Alignment** A one to one mapping of characters (amino acids or nucleotides) between two protein (or DNA) sequences. The mapping respects the ordering of the characters in the individual sequences. If a character cannot be mapped, it is usually aligned to a "gap". A multiple sequence (or structure) alignment (MSA) is an alignment between multiple sequences (or structures). MSA is usually visualized as a matrix with the number of rows equal to the number of sequences that are aligned and the number of columns equal to the alignment length.

**Sequence profile** The set of distributions (or frequencies) describing the composition of the columns of a multiple sequence alignment (MSA).

**Genotype** The genetic makeup of a cell (i.e. the specific genetic sequence), usually with reference to a particular trait under consideration.

**Phenotype** The composite of an organism's observable traits and characteristics.

**Homologs** Two protein (or DNA) sequences are said to be homologous if they share an ancestor. Homology is usually determined by sequence similarity between the two proteins.

**Orthologs** Two homologous proteins are orthologous if they are present in different species and resulted from a speciation event.

**Non-redundant** If two protein sequences have less than a threshold sequence similarity (typically 30-40%), they are said to be non-redundant. Non-redundant databases

imply that each sequence in the database is different from every other sequence in the database with respect to the threshold sequence similarity.

**Complex** Complex refers to a group of proteins involved in an interaction. In structural bioinformatics and in this thesis, a complex refers to the structure of a binary protein interaction.

**Phylogenetic tree** A tree showing the evolutionary relationships (inferred) between protein sequences (or DNA sequences). It is usually based on sequence similarities between the sequences.

# Chapter 1

# Introduction

A genome of an organism encodes for tens of thousands of proteins (proteome) that make specific interactions with other proteins and bio-molecules. Systematically mapping these interactions (the interactome) is a major challenge in post-genomic biology. Elucidation of the interactome of a cell is an essential first step in understanding protein function and cellular behavior. Sustained focus on reconstructing the interactomes of various model organisms in recent years has resulted in a wealth of information. Indeed, in this new era of high-throughput (HTP) technologies, molecular biology is dominated by studies on pathways, complexes or even an entire organism.

A mechanistic understanding of how molecules interact comes only from three dimensional (3D) structures, as they provide a high resolution picture of the binding. Such an understanding allows us to design experiments that perturb systems in an intelligent way. Consequently, knowledge of these atomic details provides us with a rational way of developing therapies by repairing and/or inhibiting interactions [5]. Despite their invaluable contributions, atomic details of interactions are beyond the scope of current HTP protein-protein interaction (PPI) detection techniques. Structural-genomics initiatives and advancements in structural biology are steps in the right direction, but are lagging behind other technologies for PPI de-

tection. Moreover, the sheer number of interactions to test (e.g. 50 million possible pairs for an organism with 10000 proteins) makes it an insurmountable task for any one experimental technique alone.

The thesis aims to bridge this gap between new-era HTP systems biology and traditional computational molecular biology to give a high-resolution understanding of the interactome. By developing protein-protein interaction (PPI) prediction techniques based on atomic details of protein 3D structures, the thesis provides a deeper understanding of structure-function relationships in biological systems. Moreover, the methods developed in this thesis help overcome the limited and biased sampling of experimentally verified interactions, ultimately leading to a complete high-quality mapping of the interactome.

The rest of the chapter is structured as follows. First, in section 1.1, basics of protein structure, structure determination and computational aspects of protein structure relevant to the thesis are introduced. Then, in section 1.2, aspects of protein-protein interactions (PPI), including experimental and computational methods for PPI prediction are discussed. Finally, I explain how understanding the structure-function relationship in the context of PPIs will help design better therapies for human diseases.

## 1.1 Proteins

The central dogma of molecular biology states that genes in the DNA are transcribed to mRNA, which are then translated to a sequence of amino acids called proteins (Figure 1-1). There are 20 different types of amino acids, each differing only in their side-chains. This difference leads to differences in the physicochemical properties of the amino-acids, ultimately influencing the function of the protein. Each amino acid, also called a residue, has a bonded sequence of three atoms, a nitrogen and two

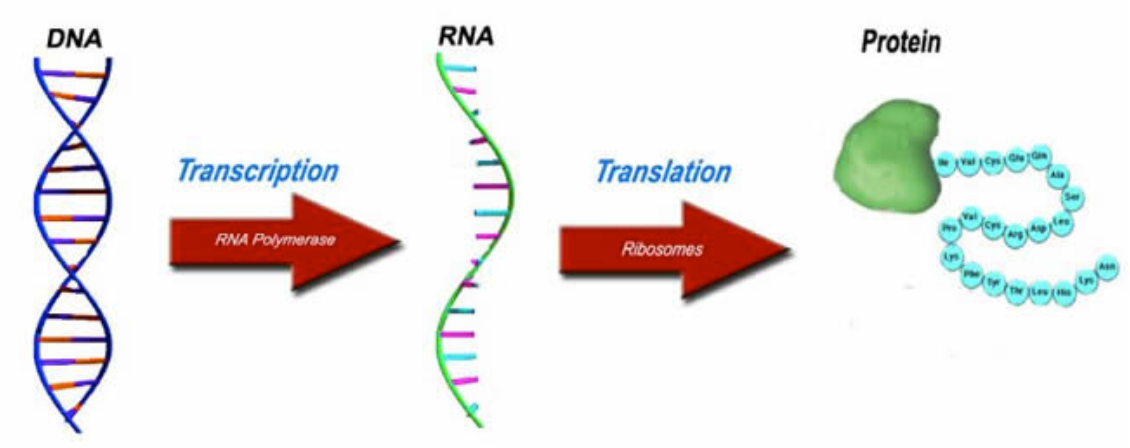
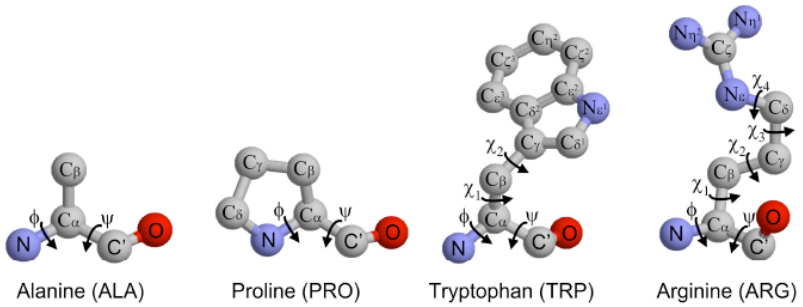


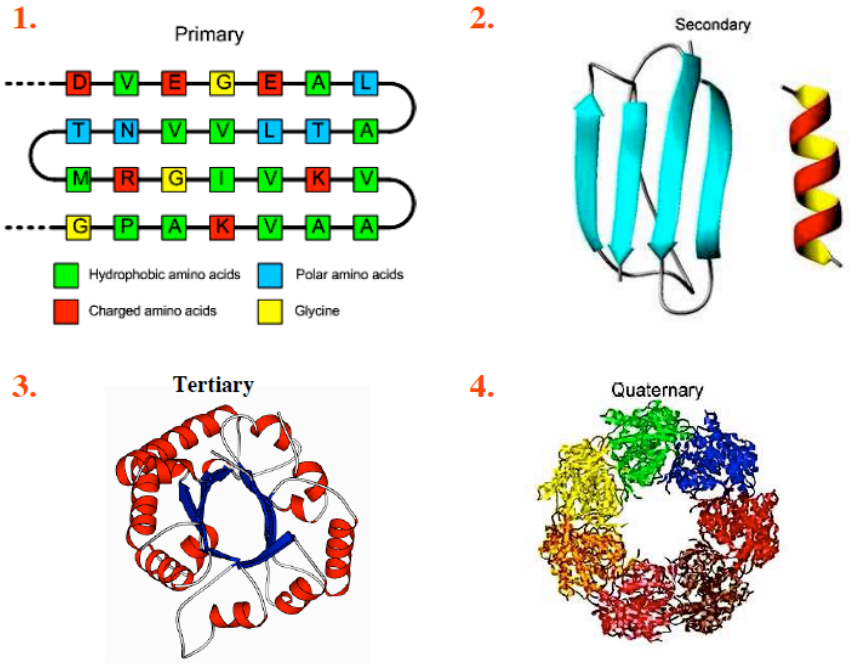

Figure 1-1: Central dogma of molecular biology. From `http://www.lhsc.on.ca`

carbons. The same triplet of atoms from each residue of a protein are concatenated together via a peptide bond to form the backbone of the protein. Each residue has a side-chain containing 0 to 10 heavy atoms branching out of the middle carbon atom (C$\alpha$). Some examples of amino acids (residues) can be seen in Figure 1-2 [21].

In its most basic form, a protein can be thought of as a linear copolymer formed by the concatenation of amino acids (primary structure). More generally, protein structure is described in a hierarchical manner: ranging from a "primary" structure to a "quaternary" structure. Under physiological conditions, the primary structure folds to a unique, compact and relatively stable 3-D structure, which determines its specific biological function. The sequence of the protein is believed to completely encode its folded structure, which arguably corresponds to the minimum free energy of the molecule. Furthermore, different regions of the sequence form one of two local regular structures - alpha helix ($\alpha$) and beta sheets ($\beta$). These locally compact structures are referred to as secondary structure. The tertiary structure is obtained by packing such structural elements into one or more compact globular units called domains. In many proteins several polypeptide chains forming different domains are brought together to form a quaternary structure (see Figure 1-2) [21].




**A.** Amino Acids




**B.** Hierarchy in protein structure

Figure 1-2: A) Some examples of amino acids. B) Protein structure is described in a hierarchical manner, ranging from a primary structure to a quaternary structure.

### 1.1.1 Protein structure determination

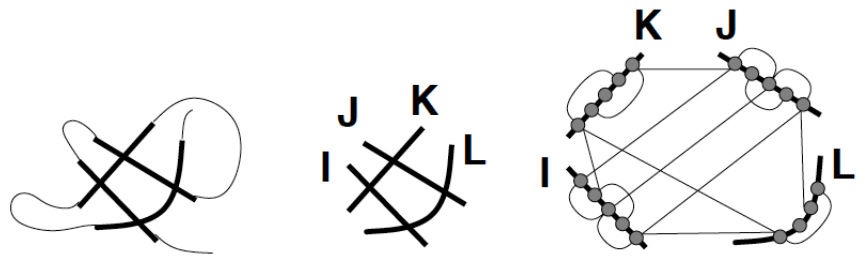
Protein function is determined by its structure. As a result, a lot of effort has been devoted to determining the 3D structure of proteins [18]. The most popular techniques for structure determination are X-ray crystallography and NMR spectroscopy [42]. Close to 85% of protein structures deposited in the Protein Data Bank (PDB) are determined by X-ray crystallography [14]. Although these techniques have given us invaluable information about protein structure and function, they are laborious and time consuming. For example, it is not uncommon to take on the order of 6 months to a year to solve a protein's structure using X-ray crystallography [21, 42, 67]. Furthermore, not all proteins are amenable to crystallography or NMR spectroscopy (e.g transmembrane proteins) [42].

Since protein structure is encoded in its sequence, it should be possible to computationally predict the 3D structure of a protein just from its sequence, using laws of physics and chemistry. To overcome limitations of structure determination techniques and to better understand protein folding, structure prediction has remained one of the most active areas in computational molecular biology [18]. There are three broad categories into which the various structure prediction methods are divided: 1) homology modeling, 2) protein threading and 3) *ab initio* folding. In homology modeling, the structure of a protein is predicted by identifying a homologous protein in the PDB. It exploits the common rule of thumb that sequences that are similar, fold in a similar way. Therefore, given the target sequence (for which the structure is to be determined), a database of solved structures is searched for similar sequences. The predicted structure is then built using large fragments of these related structures. As more and more structures are solved, homology modeling will become increasingly accurate as there is a greater chance of finding a similar sequence in the database. Depending on sequence similarity, it is sometimes possible to get structures as good as a medium resolution X-ray crystallographic structure [183]. But usually, as the se-

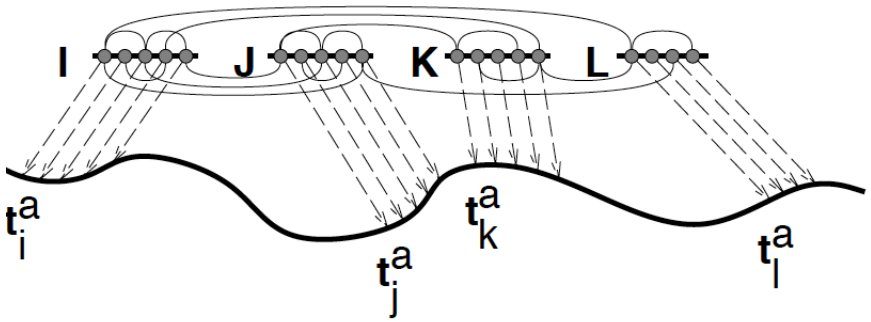
quence similarity between a target and the candidate structure goes down, it becomes an incorrect representation of the actual structure of the target sequence.

For sequences that do not have any clear homologs in the PDB, protein threading is the method of choice for structure prediction. Compared to homology modeling, which considers only the sequence similarity between the target and a candidate structure, protein threading makes use of structural information encoded in the candidate structure to improve prediction accuracy. The main components of a threading approach are a template (a simplified representation of the protein 3D structure) and a scoring function to evaluate an alignment. The goal of a threading algorithm is to find the best alignment of the target sequence to the template structure in the space of all possible alignments. Figure 1-3 gives a schematic of the components of a threading approach. First, a template is constructed from the 3D structure of the protein. Then, alignments are scored using a scoring function that evaluates the compatibility of the aligned target residue in the structural environment of the corresponding aligned template residue. In Figure 1-3b, regions of the target sequence aligned to corresponding fragments in the template are indicated by the letters $t^a$ and the alignments are represented as dashed arrows. Computing the optimal alignment (i.e. best alignment score) is formulated as a combinatorial optimization problem, and a variety of mathematical techniques are used to solve it [183]. Different threading programs use different scoring functions. All of them usually include secondary structure, solvent accessibility and pair-wise interactions in scoring an alignment. One of the best threading programs used in structural bioinformatics is RAPTOR [177]. RAPTOR formulates the alignment problem as an integer linear programming problem (ILP), and uses a branch and bound technique to efficiently solve the ILP [177]. The methods developed in the thesis use this formulation of RAPTOR for structure prediction.

Ab initio structure prediction is the most difficult method and does not use any




**A.** Template




**B.** Sequence to structure alignment

Figure 1-3: Schematic of protein threading. A) A protein 3D structure is first reduced to a simplified representation as a graph, with residues as the nodes and edges between residues that are physically close in the 3D structure. This simplified representation is known as the template. B) The target sequence (query) is then "threaded" onto the template to find the best sequence-structure alignment. This is usually formulated as an optimization problem, with both sequence and structure features in the objective function. The dashed lines represent alignment of the residues of the template to residues of the target sequence [95]

complete structure from the PDB. The main difficulty arises because conformational search space increases dramatically with respect to protein size. The optimization problem is usually non-convex and requires techniques based on Monte Carlo methods and genetic algorithms to tackle the inherent complexity [183].

## 1.2 Protein-protein interactions

Proteins interact with other proteins and molecules to perform their function. In this thesis, we are mainly concerned with understanding protein-protein interactions (PPIs) and using that knowledge to predict PPIs. Our knowledge about the rules

of association of protein molecules comes mainly from studying structures of protein complexes. We will look at methods to characterize structural features of interfaces, their chemical composition and their evolutionary histories. All these aspects of PPIs are relevant for subsequent chapters in the thesis.

### 1.2.1 Types of PPIs

Interactions between two proteins (binary PPIs) are usually divided based on the type of proteins that interact, stability of the proteins and duration of interaction. Interaction between two identical protein chains is called a homo-oligomeric complex (or homo-dimer or homomer), whereas interaction between different proteins is called a hetero-oligomeric complex (or heteromer). Based on stability, interactions are divided as obligate and non-obligate complexes [112]. Proteins forming obligate complexes do not form stable functional structures on their own, whereas proteins in non-obligate complexes can form stable structures. Many heteromers are non-obligate, while homomers are often obligate [112, 78]. In terms of duration, interactions are divided as transient or permanent. Transient interactions last for seconds or less, and typically regulate critical cellular processes by protein phosphorylation or acetylation. Permanent (stable) interactions have a typical half-life of 12 minutes to 19 hours and include some of the biggest structures in a cell such as core RNA polymerase, DNA replication complexes, etc [118]. Permanent complexes can be readily detected by common experimental techniques such as co-purification and yeast-2-hybrid (Y2H). Transient interactions are more difficult to detect, requiring some prior knowledge of the two interacting proteins and the conditions under which they interact [118, 131].

### 1.2.2 Structural features

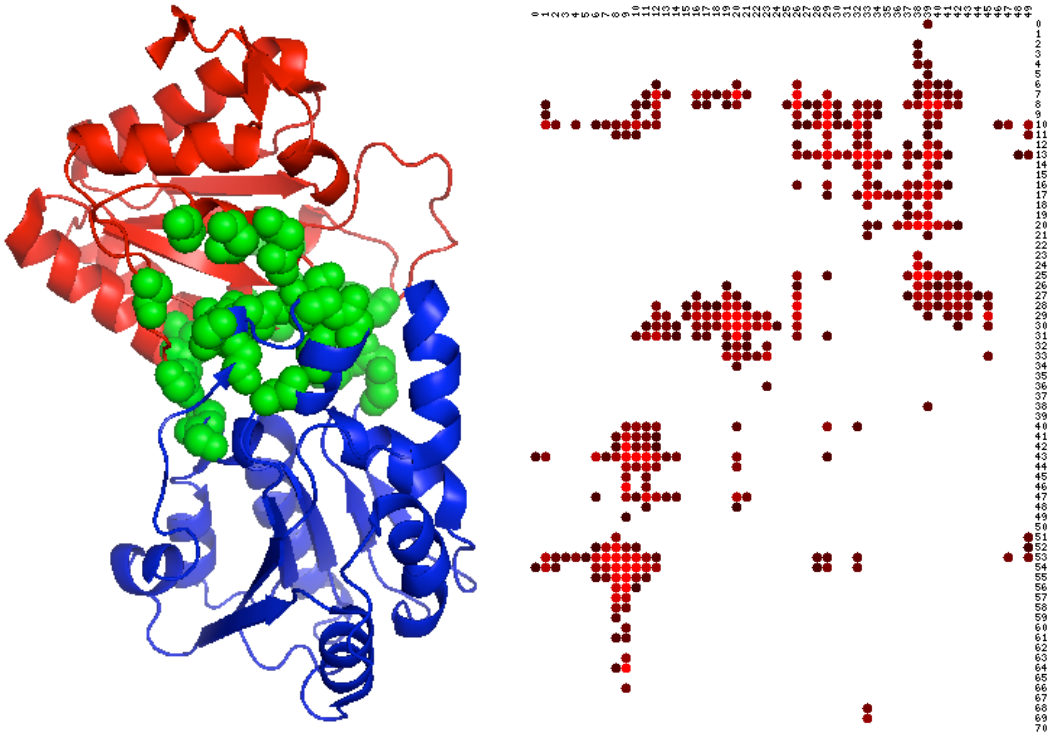
The interface between two interacting proteins in a complex can be defined in a variety of ways. The most popular and simplistic definition is that of a minimum distance –

if the distance between any two heavy atoms of two residues on either protein is less than 5Å, the two residues are said to be interacting and part of the interface. Another characterization of interface residues is in terms of accessible solvent area (ASA). In protein complexes, it has been observed that 20-45% of the interface residues have very low ASA (close to zero), and they tend to be hydrophobic [73]. Sometimes these residues are also referred to as the "interface core", with the "interface rim" consisting of residues that are less than 10Åapart in the 3D structure and having non-zero ASAs. Interfaces can also be described to reflect shape complementarity between the interacting proteins, cavities on the surface of the two proteins and atomic packing. Typically, such representations are based on embedding a 3D grid on top of the structure and measuring the volumes occupied by each atom. More generally, techniques such as voronoi diagrams allow for a precise estimation of the atomic packing [73, 89].

Another popular method of representing an interface is that of a contact map. A contact map is a 2D representation of the 3D interface, with the aim of representing only residue-level information. A contact map is a matrix of dimension $mxn$, where $m$ and $n$ are the lengths of the two proteins. An entry $ij$ in the contact map is the minimum distance between any two heavy atoms in residue $i$ in one protein and residue $j$ in the other. If the minimum distance is greater than 10, a zero is used instead. For a cleaner visualization and tractable computations, only rows/columns that have at least one non-zero entry are retained, the rest are discarded (see Figure 1-4). As can be seen, such a representation allows one to design fast search and alignment algorithms without having to deal with the more complicated 3D topology.

### 1.2.3 Physico-chemical features

The physico-chemical features of an interface depend on the relative abundance of different amino acids at the interface. Amino acids are usually described as non-



**A.** Protein-protein complex **B.** Contact Map

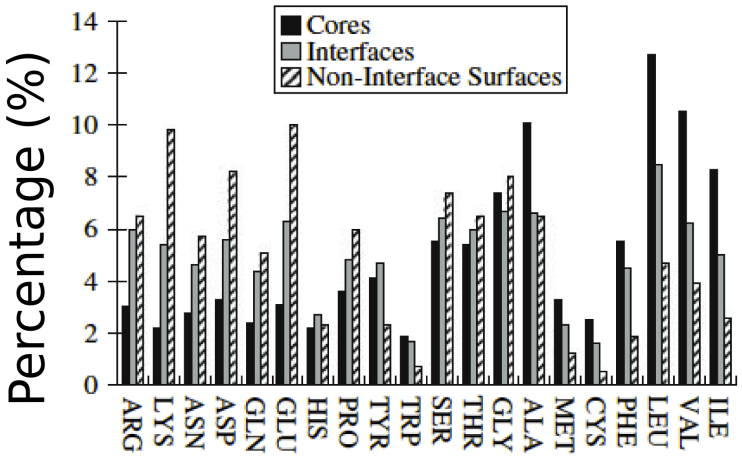
Figure 1-4: A) A binary PPI complex. Red and blue are two proteins, and the interface residues are highlighted in green. B) A contact map representation of the complex in a. The entries in the map are color-coded ranging from red (low) to black (10Å). Distances greater than 10Å are not relevant and are indicated by white.

polar/polar/charged, or in a more coarse-grained model as hydrophobic/hydrophilic. The composition of an interface is usually computed by counting interface atoms or residues, or by weighting their numbers by their buried surface area (BSA). The advantage of area-based composition is that it accounts for the amino acid size as well. Quantitatively, interface propensity is calculated as:

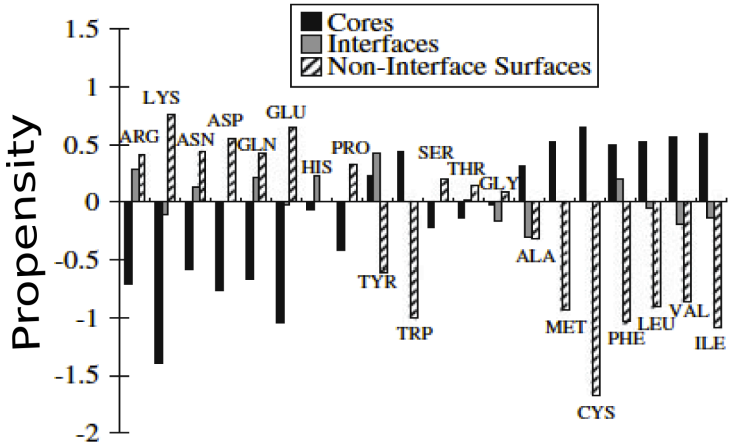
$$p_i = log(\frac{f_i}{f_i^o}) \tag{1.1}$$

where $p_i$ is the propensity of amino acid of type i at the interface, $f_i$ is the number or area fraction of type i at the interface, and $f_i^o$ is the corresponding number in a reference set (can be the whole protein, or its interior, or surface). Interpreting $p_i$ is straightforward: if $p_i > 0$, then the interface is enriched for atoms or residues of type i compared to the reference set, and $p_i < 0$ implies that the interface is depleted of type i. Figure 1-5 shows the composition and propensities calculated from a non-redundant set of protein complexes [178].

Hydrophobicity plays an important role at protein surfaces and interface. Amino acids containing groups with 'O' and 'N' in their side chains are polar and hydrophilic, the rest are non-polar and hydrophobic. Hydrophobic patches enriched for such residues are frequently found at protein interfaces, indicating that their contribution to PPIs is significant. It has been argued that capping motifs which bury otherwise-exposed hydrophobic patches have specifically evolved in certain proteins to prevent aggregation [24]. The extent of the hydrophobic contribution depends on the type of interaction and the driving force (i.e long-range electrostatics or short-range desolvation effects). For homo-dimers/obligate complexes, the monomer (protein) molecules do not exist individually and hence their interfaces tend to be always buried (from the solvent). Therefore, such interfaces can have large hydrophobic patches. Interactions in which each participating protein exists as a functional unit by itself cannot admit large hydrophobic patches on its surface as it would be energetically unfavorable [78].




**A.** Interface composition




**B.** Interface propensities

Figure 1-5: A) Residue composition at the interface in a non-redundant set of protein complexes. B) Residue propensities at the interface in the non-redundant set of protein complexes. "Core" of the protein refers to interior of the protein. Hydrophobic residues have a higher propensity at the core, whereas polar amino acids are enriched at the interfaces and surfaces. Figures taken from [178]. The residues are arranged in increasing order of their hydrophobicity (Kyte-Doolittle scale) from left to right [93].

Electrostatics also plays an important role in PPIs, as can be seen from Figure 1-5. Enrichment of polar amino acids indicates that protein interfaces tend to be more similar to protein surfaces than to protein interiors. One hypothesis for explaining this apparent anomaly is that desolvation effects are partially compensated in interfaces through the formation of networks of ion-pairs and hydrogen bonds, which are positioned so as to interact favorably with one another. Electrostatics is also known to play a significant role in the rate of protein-protein association [137]. Computationally, accounting for electrostatics requires elaborate calculations that are highly sensitive to the solvent model and local structural environment. The representation of electrostatics is not accurate enough yet, although it has resulted in a few excellent models for protein-protein complexes [20, 134]. Such calculations are usually computationally intensive and left for later stages of structure prediction, with earlier steps relying on statistical (knowledge) potentials [73].

### 1.2.4 Evolutionary features

As we have seen in the previous two sections, in order for two proteins to interact, there has to be structural as well as chemical compatibility at the interface. One can then argue that nature will try to maintain this compatibility over the course of evolution of the species. Indeed, evolutionary conservation has been observed at three levels in PPIs: 1) interface residues are conserved across orthologs in different species, 2) co-evolution of residues at the interface of the interacting proteins (correlated mutations), and 3) similarity of phylogenetic trees (evolutionary histories) for the two proteins [83, 79, 164, 72, 80]. Notice that the first two observations are at a residue-level and hence require multiple interacting proteins from different species to give any meaningful statistics. This kind of information is usually not available for all possible protein pairs, and hence these insights have generally not been used for prediction purposes till now. We will however develop techniques that overcome

this difficulty. Compared to the first two, similarity between phylogenetic trees is an indirect evidence for interaction. Correlation between phylogenetic trees could arise due to a variety of reasons not necessarily related to PPIs. It does not give us any residue-level insight into the interface compatibility of the two proteins [163, 164].

## 1.3 Experimental methods for PPI detection

### 1.3.1 Low throughput screens

Low throughput (LTP) experimental techniques include affinity chromatography, affinity precipitation, dosage lethality, biochemical assays, synthetic lethality and structure [151]. Interactions detected by these experiments are usually reliable and used as gold-standard [35, 168]. However, it is difficult to curate interactions detected by LTP experiments since one has to manually go though the publication to extract the interacting pairs. Text mining is still in its infancy, and leads to numerous false positives and negatives [111]. Moreover, the number of interactions that need to be identified to map the entire interactome is too large to be done using LTP screens alone.

### 1.3.2 High throughput screens

High throughput (HTP) screens have provided the bulk of the interactions that we know today [151, 133]. These methods are called HTP because thousands of pairwise interactions can be tested simultaneously. HTP methods include yeast-2-hybrid (Y2H) [54, 130, 162, 155, 181, 141], mass spectrometry based methods [88, 64, 31, 48, 58], protein chips [176, 185, 186] and LUMIER assays [12].

### Yeast-2-hybrid method

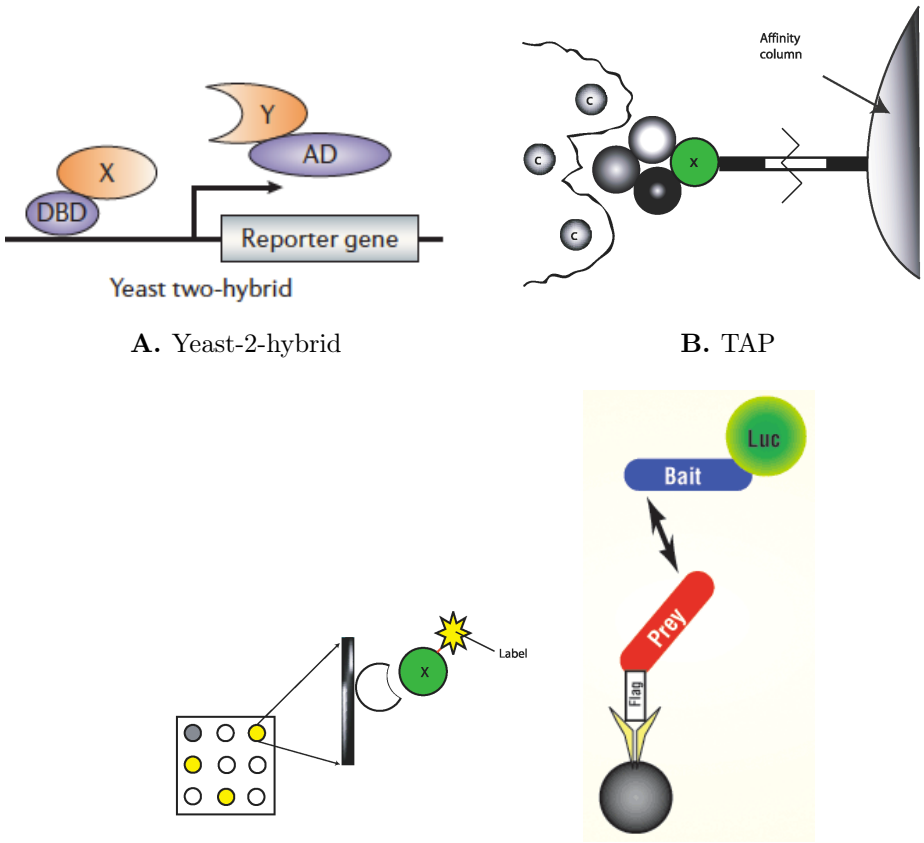
The two-hybrid system is one of the most widely used HTP screen for PPI detection. It is based on the observation that gene transcription requires the binding of two domains of a transcriptional activating protein [51]. These domains are called DNA binding domain and activator domain (Figure 1-6). The candidate proteins are fused to one of the two domains. If the two candidate proteins interact, then the DNA binding domain and activator domain are close enough to interact and result in a functional transcription complex. This activates expression of the reporter gene, leading to an observable change in phenotype (e.g. fluorescence).

### Mass spectrometry based methods

Mass spectrometry (MS) is an analytical technique to identify the chemical composition of proteins and peptides. For PPI detection, two types of MS-based HTP techniques are popular - tandem affinity purification (TAP-MS) and protein complex identification (HMS-PCI) [19, 58, 66]. In these techniques, the protein whose interacting partners are sought is called the bait and its interacting partners are called prey. Both the techniques first fuse short tags to the bait so that they can be extracted from a mixture of cellular contents (Figure 1-6). If a bait is part of a complex, then the complex is first extracted and its constituents are separated by gel electrophoresis and identified by MS.

### Protein chips

Protein chips (or microarrays) involve thousands of proteins immobilized on the surface of a microscope slide. Labelled target proteins are then added to the chip and may bind to some of the proteins on the chip. Unbound proteins are washed away and the bound ones are detected using a fluorescent dye [185, 186].




**A.** Yeast-2-hybrid **B.** TAP

**C.** Protein array **D.** Lumier

Figure 1-6: Schematics of the popular HTP techniques for PPI detection. A) Yeast-2-hybrid method involves fusing the two candidate proteins (X and Y) to a DNA binding domain (DBD) and an activator domain (AD) of a transcriptional factor. Interaction between the proteins results in a functional transcriptional complex, ultimately leading to the expression of the reporter gene [5]. B) In the TAP-MS method, the bait (X) along with its partner proteins are extracted from the cellular contents with the help of a fused tag. The constituents are then separated and identified using MS [140]. C) Protein chips involve immobilizing prey proteins by fixing them on a chip. The bait protein (X) is fused with a fluorescent tag to help visually identify the PPIs [140]. D) In a Lumier assay, the bait protein is fused with a luminescence protein, and the prey is fused with a tag to help in purification. After extraction of the bait-prey complex from cellular contents, the interaction is detected by monitoring the luminescence observed [47].

#### Lumier assay

Luminescence based mammalian interactome mapping (LUMIER) is a new technique developed to detect even transient interactions in signaling networks [12]. In a Lumier assay, a luciferase-tagged bait protein is screened against a series of flag-tagged prey proteins; an antibody against flag is used to affinity-purify the prey, and the prey-associated luminescence on exposure to an appropriate luciferin substrate is monitored to detect interaction [47]. The technique is known to be more sensitive than previous approaches, and comparatively easier to quantify dynamic shifts in PPI networks [47].

### 1.3.3 Limitations of experimental techniques

HTP screens look very promising as they identify thousands of interactions, but they suffer from high false-positive and false-negative rates [65, 16, 165, 148]. Estimates on the false discovery rates (FDR) [1] in HTP techniques are still debated as there is no gold-standard for negative data (i.e. proteins that do not interact) to evaluate against. Initial estimates computed from re-testing interactions detected by HTP experiments obtained FDRs in the range 20-40% [130, 155]. More recently, with improvements in experimental protocols, HTP studies were able to achieve FDRs between 0 to 11% [22]. Although these values seem reasonable, the more serious issue is with sensitivity of these assays. Braun et al. evaluated 5 HTP methods and obtained sensitivities of 21 to 36% [22]. Combining the methods resulted in a sensitivity of around 59% [22]. The main strategies for improving the FDR and sensitivity of HTP methods are by repeating screens, using several HTP screens or combining HTP and computational approaches for PPI prediction [181, 43]. However, conducting repeated screens or using multiple HTP screens is time-consuming and not cost-effective [135]. For example, some estimates put the time required for completing the *Drosophilia melanogaster* interactome at 1700 person-years, using the current

[1] expected fraction of false positives amongst the predicted true positives

experimental protocols [135]. Using a ranked list of interactions to test reduces this estimate considerably to 385 person-years [135]. Moreover, it has been argued that limited overlaps of interactions identified using different HTP techniques highlight the biases of those experiments rather than identify true/false positives [166, 165]. More importantly, non-physiological conditions in most experimental techniques limit our ability to translate detected PPIs into *in vivo* hypotheses.

## 1.4 Computational methods for PPI prediction

Limitations in experimental techniques combined with the sheer number of interactions to verify has provided much impetus to the development of complementary computational methods for PPI prediction. These methods usually use a variety of machine learning, statistical and graph-theory based approaches. Computational methods for PPI prediction can be roughly divided into three broad categories - indirect methods, direct methods and methods based on data-integration.

### 1.4.1 Indirect methods

Indirect methods for PPI prediction are methods that try to infer physical interaction between two proteins based on evidence for their functional association. One of the popular methods for detecting functional association is by correlation of gene expression profiles (co-expression) [96]. The idea here is that genes showing a high correlation in their expression patterns under different conditions are more likely to physically interact than random pairs. On a genomic level, functional association is usually detected by conservation of gene neighborhood, similar pattern of presence or absence across multiple genomes or gene fusion [96]. The intuition behind such approaches is that if the genes are functionally related, they will tend to be inherited as a unit since the loss of one gene would disrupt the function they are involved

in. However, such methods are always used as additional sources of evidence since functional association need not always imply direct physical interaction [5, 96].

### 1.4.2 Direct methods

Direct methods for PPI prediction usually use the primary sequence or tertiary structure in a direct way to infer PPIs. Methods that use protein sequence generally consider the physicochemical properties of the constituent residues and/or frequencies of residue combinations to quantitatively predict PPIs [17, 63, 105, 120, 180]. Most techniques map the protein sequences onto a multi-dimensional feature space, and use machine learning based classification algorithms such as support vector machines, logistic regression, neural networks to quantitatively predict PPIs. The classifiers are usually trained on a small set of high-confidence interactions and evaluated on a separate dataset (i.e. cross-validation) [17, 63, 105, 120, 180].

Another popular method for PPI prediction utilizing protein features is based on the "guilt-by-association" principle. In this method, protein pairs similar to known interacting pairs are predicted to interact. The "association" could be based on sequence similarity or other properties and annotations [96]. Such associations based on sequence similarity are called "interologs" (Figure 1-7). Predictions made using this approach quickly break down as the sequence similarity between the query and known interactors decreases.

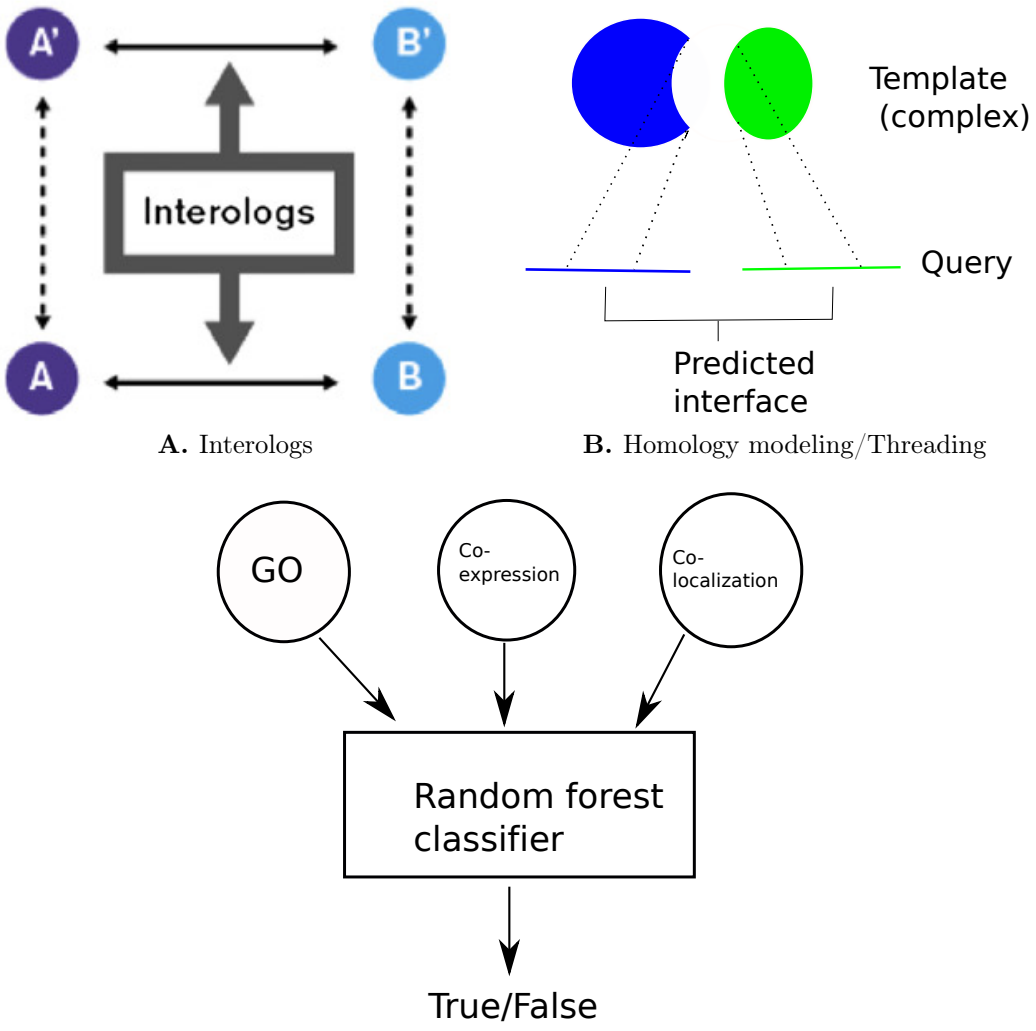
Structure-based approaches are becoming increasingly popular as the number of structures deposited in the PDB is rapidly increasing. In the past 4 years the number of complexes in SCOPPI, a database of protein interfaces, has grown by 60% [173]. For proteins whose structures are not known, homology models or threading based models are typically used to first identify the putative interface. Predictions are then made by evaluating the quality of the interface using a variety of different scoring functions [3, 56, 103, 143, 123, 7]. As shown in Figure 1-7 , such methods proceed

by first identifying a suitable template for the proteins of interest by scanning the structural database. Optimal sequence-structure alignment then gives the predicted interface, which can be evaluated using either statistical potentials or physics-based potentials for interaction suitability. For proteins that have a solved structure, the challenge is to predict the structure of the bound complex, and evaluate its energy to predict if the two proteins interact. These methods are termed as "docking" methods, with a lot of popular methods available to the community [41, 146, 174, 171].

### 1.4.3 Data integration methods

Different experimental and computational methods have their own biases, strengths and weaknesses. In general, it is natural to expect an interaction to be true if multiple observations or predictions support it. The data integration methods exploit this intuition by making a prediction based on other predictions or a number of different features. The key challenge here is to integrate different sources of information in an intelligent way by taking into account their individual accuracies. There are many different methods to do this - Fisher's, Bayesian, logistic regression, random forests, etc [96]. One example of such a method, that integrates co-expression, co-localization and functional similarity using a random forest classifier is shown in Figure 1-7 [143].

Predictions from many of these approaches have been aggregated into a number of databases/web-services offering predicted PPIs. The STRING database [76] combines experimental datasets (e.g. BioGRID [151]) with computational predictions based on co-expression, interologs, and text-mining, etc. The entries in this database correspond to functional interactions, and may not always be directly interpretable as PPIs. Another database, IntAct [85], focuses more on inferring interactions from expert curation of data from the literature. Other public services include DOMINO [29], InterDom [109], and I2D [23]. However, all of these databases suffer from a common selection bias: often, the proteins that have been selected for PPI experiments




**A.** Interologs

**B.** Homology modeling/Threading

**C.** Data Integration

Figure 1-7: A) Interaction between A and B is transferred to A' and B' using orthology assignments [96]. B) Structure-based prediction of the putative interface using homology modeling or threading. The candidate proteins are first aligned to a complex template, and the putative interface is inferred from the structure and the alignment. c) One example of a data integration method. Multiple features such as Gene Ontology (GO) annotation similarity, co-expression and co-localization for a pair of query proteins are input into a random forest classifier that makes a prediction.

are usually genes/proteins that have received some attention before and, as such, are also more likely to have functional genomic data.

## 1.5 Medical impact

Whole genome sequencing and cancer sequencing projects have given us a lot of biological insights into what genes and mutations are associated with diseases. However, this insight has very rarely led to the development of any new therapy to treat such diseases [170]. One main reason for this lack of translational breakthrough has been the difficulty in unraveling the complex genotype-to-phenotype relationships among diseases and their associated genes. Knowledge of PPIs and genome-scale interactome has enabled us to tackle this problem like never before, but there is still a long way to go. To gain a complete understanding of biology, it is not enough to know that two proteins interact; it is imperative that we know why and how they interact. This knowledge will enable us to repair disrupted interactions or inhibit aberrant interactions that are often common in many diseases. By doing so, we can design therapies that attack the source of the problem, rather than just treat the symptoms. The methods developed in this thesis not only enable researchers to know whether two proteins interact, they also give insights into why and how they interact. The advantage of this is that experiments can be carried out by mutating the predicted interface residues to further gain an understanding of interaction specificity. There is no doubt that such knowledge will become the basis for designing more efficient drugs and developing new drugs against diseases for which we don't have any therapies. As an example, Pertuzumab, a drug developed by Genentech is designed to inhibit interactions of ERBB3 with other proteins by binding to the same interface, thereby preventing cell division and tumor growth [84].

## 1.6 Organization of the thesis

The rest of the thesis is organized as follows: in chapter 2 we take a detailed look at the structure-based PPI prediction methods, which will set the stage for the methods developed in this thesis. In chapter 3, I will introduce a novel algorithm for structure-based PPI prediction that predicts interfaces and PPIs better than previous methods. In chapter 4, we will look at another PPI prediction method that utilizes evolutionary insights to overcome some of the limitations of the previous approaches.

# Chapter 2

# Struct2Net: structure-based approach to PPI prediction

## 2.1 Background

The paucity of interactome coverage (Table 2.1) and errors associated with HTP techniques has motivated significant research interest in methods for supplementing experimentally determined PPI data with interactions inferred or predicted from other sources. A wide variety of methods have been proposed including the use of "interologs", functional genomic data such as gene expression, cellular localization and GO annotation (see section 1.4, Figure 1-7).

| Organism | Number of interactions | % of proteins with at least 1 interaction |
|---|---|---|
| **Mouse** | 7794 | 15 |
| **Human** | 65846 | 57 |
| **Fly** | 24375 | 46 |
| **Yeast** | 69728 | 99 |
| **Worm** | 4692 | 15 |

Table 2.1: Number of interactions in Biogrid [151] for common eukaryotic organisms.

The use of structure-based approaches to predict interaction has been previously

proposed. Aloy and Russell suggested the use of structure-based approaches to predicting PPIs [3]. They have described InterPreTS, a web-server to predict PPIs for a given protein, using a homology modeling approach [4]. Lu, Lu and Skolnick constructed statistical potential functions to evaluate potential PPIs [102] and later described MultiProspector, a structure-based prediction algorithm [103]. More recently, Fukuhara and Kawabata have described HOMCOS [57] a web-server that performs a similar task, again by homology modeling. Tuncbag et al. have described a method that utilizes evolutionary constraints at the interfaces along with homology modeling to predict PPIs [159]. MODBase is a database of homology models for protein complexes that have high sequence similarity to known structures [119]. ADAN is a specialized database for prediction of PPIs mediated by linear motifs and utilizes position-specific matrices to assess putative interactions [49]. Other sequence-based methods utilize genetic information and multiple sequence alignments to predict specific protein-protein interactions [163, 164, 25, 138]. There have been methods to predict PPIs based on co-occurrence of sequence domains in the candidate proteins [169]. Other researchers have aimed to understand these domains from a structural perspective. Prieto and Las Rivas [121] have reviewed publicly available databases that facilitate analysis of domain-based PPIs: 3did [154], SNAPPI-DB [75], iPfam [52], PIBASE [37] and PSIBase [61]. While Struct2Net approach has some parallels with these approaches, the goal is significantly different. The domain-interaction databases are essentially repositories of known structural data, analyzed specifically from a PPI perspective. Prediction— which is the core goal— is usually beyond the scope of these approaches.

In this chapter, I describe Struct2Net (Structure-to-Network), a structure-based method for predicting protein-protein interactions. Struct2Net predicts interactions by threading each pair of protein sequences onto potential structures in the Protein Data Bank (PDB). Struct2Net provides PPI predictions that are independent of all

the non-structure-based approaches and may thus be combined with any of them. Another key advantage of Struct2Net is that, apart from the PDB data, the prediction algorithm only requires protein sequence data as input. It can thus be applied to proteins for which no functional data is available provided there is a suitable PDB structural template available. Struct2Net offers a significant advantage over other homology modeling approaches. Successful use of homology modeling requires relatively high sequence similarity between the query and template protein-pairs. In contrast, a threading-based approach widens the range of proteins for which predictions can be made. The use of threading also offers an improved performance: Fukuhara and Kawabata reported that HOMCOS achieves a recall [1] of 80% with a precision [2] of about 10%; in comparison, Struct2Net achieves a recall of 80% with a precision of 30% .

## 2.2 Methods overview

The Struct2Net method proceeds in two stages: 1) identification of the putative interface and 2) computing interaction probability from the predicted interface. The basic framework of these two stages is common (with some variations) to all the methods we describe in this thesis.

### Predicting the interface

Given any two query proteins, the interface is predicted by threading the sequences onto templates in a database (see Figure 1-7b). First, the set of complexes in the PDB is clustered based on their SCOP domains and sequence identities [108]. Then only a representative complex (chosen randomly) is retained from each cluster, to increase computational efficiency. The query sequences are then thread onto each complex in

[1] recall = True Positives/(True Positives+False Negatives)
[2] precision = True Positives/(True Positives+False Positives)

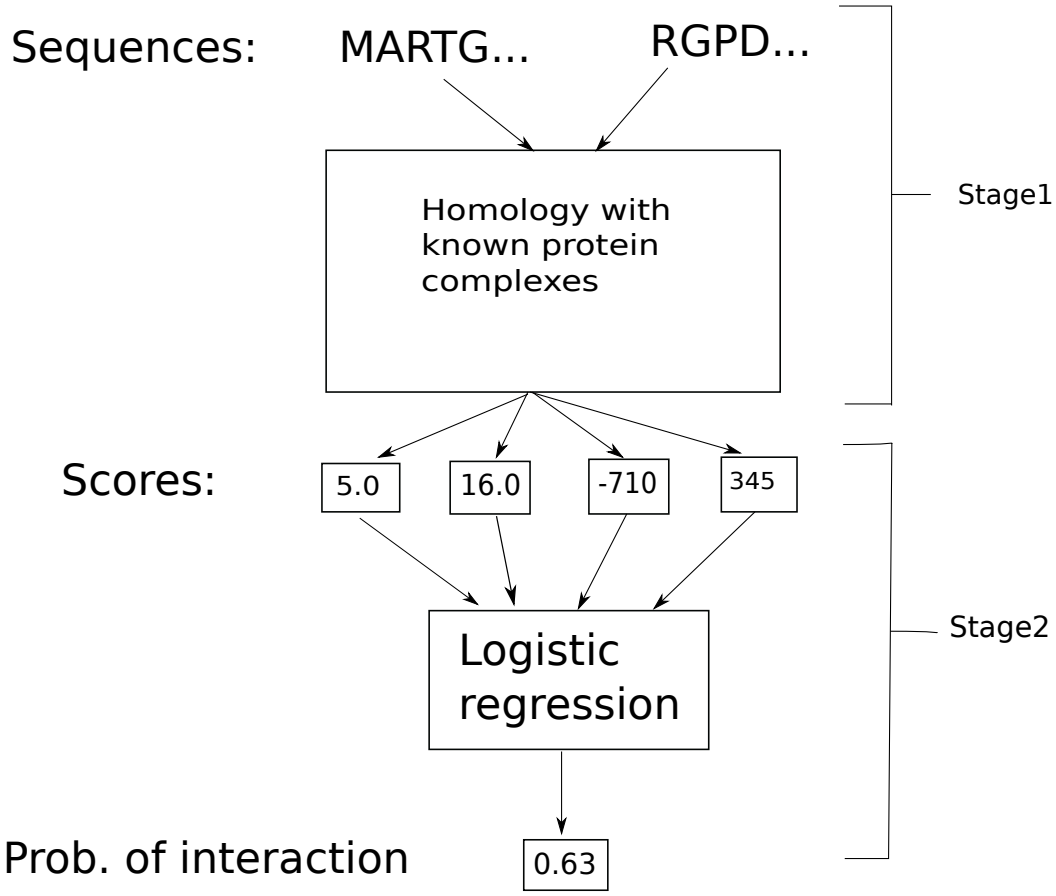



Figure 2-1: Struct2Net algorithm. The input to the algorithm are two protein sequences. The first stage consists of identifying the best complex template for the two proteins, and alignment of the proteins to the template using DBLRAP (Double RAPTOR) [177, 143]. In stage 2, a set of scores quantifying the quality of the alignment and predicted interface are extracted from the sequence-structure alignment of stage 1 and input into a classifier that predicts the probability of interaction.

the template database using RAPTOR [177], and the best template is chosen based on the alignment score. The alignments to the best template give us the predicted interface for the two sequences. Using this alignment and the interaction pattern between the complex's constituent subunits, we can also calculate the interfacial energy between our input proteins. The interfacial potential parameters are taken from Lu, Lu, and Skolnick's paper[102].

In summary, for any given sequence pair ($p$ and $q$), the threading-based interface prediction method will generate two alignment scores ($E_p$, $E_q$), their associated z-scores ($z_p$,$z_q$) , and an interfacial energy ($E_{pq}$) evaluated using the statistical potential. z-scores measure the significance of the alignment score, with the background distribution of alignment scores computed by randomizing the residues at the aligned positions. This vector of scores is then used to represent the predicted interface (Figure 2-1).

**From predicted interface to interaction probability**

Struct2Net uses a binary logistic regression to classify whether a set of interface scores (from above) corresponds to an interaction or not. In binary logistic regression, the goal is to predict a binary output variable $Y$, given a set of $r$ predictor variables $X = X_1, X_2...X_r$. For an instance $i$, suppose $y_i$ and $\mathbf{x_i} = x_{i1}, x_{2i}, ...x_{ri}$ are the random variables corresponding to $Y$ and $X$ respectively. Let $\theta_i = P(y_i = 1|\mathbf{x_i})$. In this model, the dependence of $\theta_i$ on $\mathbf{x_i}$ is expressed by the logit function:

$$logit(\theta_i) = log(\frac{\theta_i}{1-\theta_i}) = \alpha + \beta^T \mathbf{x_i} = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_r x_{ri} \qquad (2.1)$$

This can be rewritten as:

$$\frac{P(y_i = 1|\mathbf{x_i})}{P(y_i = 0|\mathbf{x_i})} = e^{\alpha+\beta^T \mathbf{x_i}} \quad or \quad P(y_i = 1|\mathbf{x_i}) = \frac{e^{\alpha+\beta^T \mathbf{x_i}}}{1 + e^{\alpha+\beta^T \mathbf{x_i}}} \tag{2.2}$$

The parameters $\beta$ are learned by maximizing the likelihood of a set of "training" examples under the model. In the context of our problem, $y_i$ is the interaction probability of two proteins $p$ and $q$. The predictor variables, $\mathbf{x_i}$, come from the first stage. For proteins $p$ and $q$, the first stage provides their interfacial energy $E_{pq}$, their respective alignment scores $E_p$ and $E_q$, as well as the associated z-scores, $z_p$ and $z_q$. In addition, the sequence lengths of the two proteins, and various functions and combinations of the existing terms are introduced as predictor variables [143].

The most informative subset of predictor variables are identified using the Akaike information criterion (AIC). The AIC score is defined as:

$$AIC = -2log - likelihood + 2\frac{k}{N} \tag{2.3}$$

where k=number of predictor variables, N=number of instances in the dataset, and the log-likelihood of the data under the model is computed using Eq 2.2. The subset of predictor variables with the lowest AIC was chosen as the final model. This model is the optimal trade-off between complexity of the model (i.e. number of independent parameters) and prediction accuracy (likelihood).

## 2.3 Evaluation

To evaluate the algorithm, we need gold-standard positive (interacting) and negative (i.e. non-interacting) datasets. Unfortunately, there are no standard procedures to construct such datasets. In order to construct our high-confidence datasets, we require

that the positive examples either come from a small set of trustworthy protocols, or from low-throughput experiments, or roughly correspond to co-clustered protein pairs in the PPI network. For negative examples, we required that the two proteins either be disconnected in the PPI network or be at least 3 hops away from each other. Using these criteria, we had a training set of 62,519 pairs and a test set of 15635 pairs (with a positive:negative ratio of 1:6 approximately, in both sets). We believe that these datasets provide good evidence of validation [142].

The datasets (both positive and negative) are separated into two groups - one for training and one for testing. The parameters for the logistic regression are learned by maximizing the log-likelihood of the training group under the model. The optimized model is then used to predict the probabilities of interaction on the test set. To evaluate the classifier and model, a probability threshold is varied, and statistics such as number of true positives predicted, number of true negatives predicted, etc are calculated for each threshold. The results are displayed using a receiver operator characteristic curve (ROC) (Figure 2-2). An ideal classifier will display a step function, with a sensitivity of 100% at zero to 100% specificity values. The area under the curve (AUC) is usually used to compare different classification algorithms tested on the same dataset; greater the AUC, better is the classifier.

## 2.4 Conclusion

Although high-throughput biochemical approaches for discovering PPIs have proven very successful, the current experimental coverage of the interactome remains inadequate and would benefit from computational tools. The Struct2Net algorithm allows the user to easily query for high-probability structure-based interactions as a potentially high-quality, high-coverage data source for large-scale integrative approaches to interactome construction. The predicted interactions also include a numeric score,

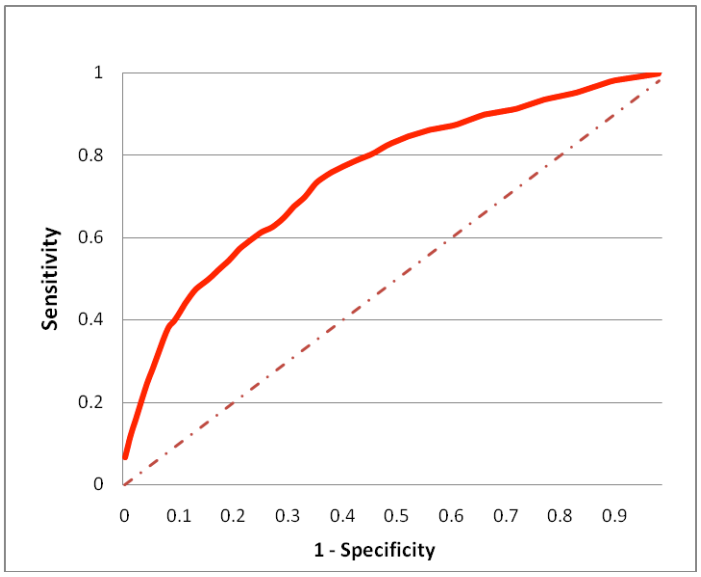



Figure 2-2: The prediction algorithm can achieve 60% sensitivity while maintaining 75% specificity as measured on the test set. Here, sensitivity = (true positives)/ (true positives + false negatives) and specificity = (true negatives)/ (true negatives + false positives). We constructed a training set and test set of positive and negative examples from yeast and fly, using criteria we have developed to identify high-confidence positive and negative examples of PPIs [142]. After training the logistic regression model on the training set, its performance was measured on the test set.

allowing users to further filter the data. Struct2Net's predictions may be used by themselves or as one of the inputs into a computational framework that combines them with other sources (e.g., low-quality experimental data or predictions from functional genomic data). For example, Jensen et al. [76], Qi et al. [125] and Srinivasan et al. [150] have described some general approaches for combining various predictors of PPI data. Struct2Net's predicted interaction scores can easily be integrated into such models.

# Chapter 3

# iWRAP: an interface threading approach for PPI prediction

## 3.1 Introduction

There has been considerable interest to harness the information provided by structure-based computational approaches as a potentially high-quality, high-coverage data source for large-scale integrative approaches to interactome construction [143, 3, 86, 5, 7]. Prieto, Las and Rivas [121] have reviewed publicly available interaction databases of known structural data that facilitate analysis of PPIs [154, 75, 52]. In the absence of a solved structure for a pair of protein query sequences, structure-based approaches typically rely on aligning the query sequences to either sequence or structure-based templates for solved structures in the Protein Data Bank (PDB) [14]. Homology modeling and threading-based approaches are the commonly used techniques for structure-based PPI prediction.

While homology modeling/threading approaches work well and have good overall accuracy when sequences are somewhat similar to their putative templates, they perform poorly in the "twilight zone" ($< 40\%$) of sequence identities. In particular,

they often give inaccurate alignments in the putative interaction regions for sequences with low similarity and therefore are unable to predict interactions accurately in such cases. This has been demonstrated previously for the special case of cytokines [123]. Moreover, it has been observed that functional residues such as those at the interface are more conserved than non-functional ones, both in sequence [26, 27, 55] and structure [124, 184]. Furthermore, it has been shown just recently that partial homology models, based only on interface alignments, are good candidates for templates used in docking studies [92]. Here we capitalize on these observations by performing threading on only the protein-protein interface after a suitable complex template is identified.

In this chapter, I introduce the program iWRAP (Interface Weighted RAPtor), which predicts whether two proteins interact by combining a novel linear programming approach for interface alignment with a boosting classifier [34] for interaction prediction. iWRAP simultaneously optimizes contacts in query sequences to templates of protein-protein interfaces, after constraining alignments to only those residues likely to be involved in the interaction. This approach is in contrast to existing threading approaches that align each sequence individually to an entire protein structure in the complex. We recently demonstrated the utility of interface threading on two cytokine receptor families by implementing LTHREADER [123], where we manually generated templates specific to this family and aligned each query sequence separately to each template. The driving hypothesis of iWRAP's approach is that more accurate prediction of protein-protein interfaces improves predictions of protein-protein interactions. We show in this chapter for general PPIs that (i) more accurate interface alignments lead to improved interface contact prediction, which in turn (ii) significantly improves PPI prediction. Thus, by optimizing the interface alignments after identifying a suitable template, iWRAP exploits functional conservation at the interface to predict PPIs.

We demonstrate the efficacy of these techniques on two datasets, SCOPPI, a database that classifies protein complexes in the PDB [173], and the yeast genome. First, we use SCOPPI as our gold standard database to confirm hypothesis (i): we show that interface threading, i.e. localized threading, leads to better interface contact prediction over full-complex threaders. For difficult alignment problems and a range of sequence identity values less than 40%, iWRAP outperforms standard threading and sequence-based methods , while for easier problems the methods are comparable (performance measured in terms of interface alignment accuracies and contact accuracies). Our results on the full yeast genome scan address hypothesis (ii): we demonstate that our method, which novelly uses boosting [34] [1] to classify iWRAP's interface threading scores for PPI prediction, outperforms methods based on whole-sequence alignments. In particular, we perform a full genome scan of yeast to predict interactions, and compare iWRAP's performance on experimental data to DBLRAP, which has been shown to have the best performance amongst available structure-based PPI prediction methods [143, 142].

As an application, through mapping of yeast-homologs of human cancer related genes and their putative interactions to the human genome, we identify interactions enriched relative to a recent yeast genetic interaction set [32]. We find that these interacting genes are involved in chromatin remodeling, ribonuclear complex assembly and nucleosome organization [59]; processes known to be critically involved in cancer. We focus on yeast cancer related genes and putative interactions since the function and interactions of yeast genes are much better understood than human genes [97]. Moreover, the malignant behavior of human cells is often caused by dis-regulation of cell cycle, growth and apoptosis processes that are conserved across eukaryotic organisms at the level of genes and their interactions [104].

[1] Boosting is an ensemble technique for classification problems. Instead of learning a single classifier, boosting involves learning many classifiers and combining the results of individual ones for the final prediction.

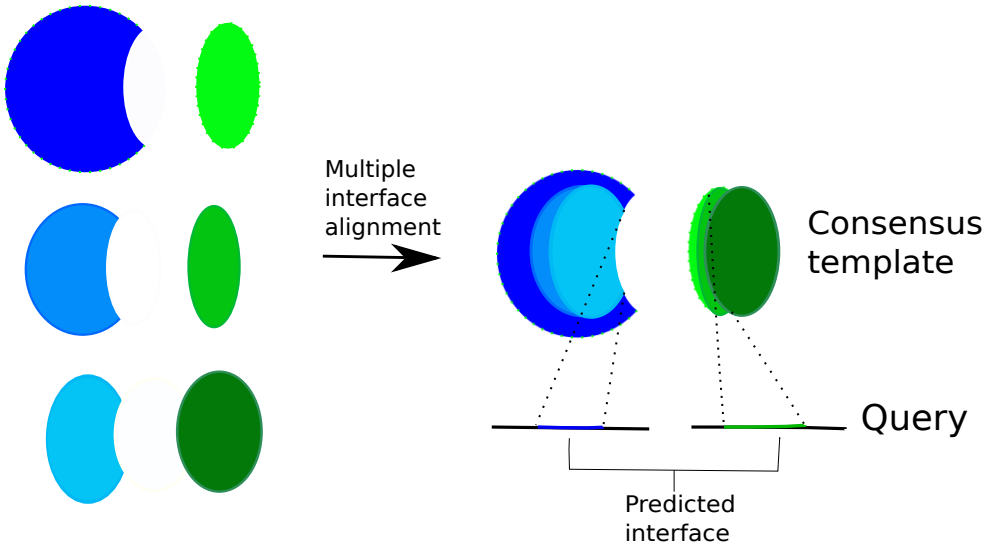
iWRAP's predictions are made publicly available at its website so that they can be used for further exploration or systems-level integrative approaches.

## 3.2 Results

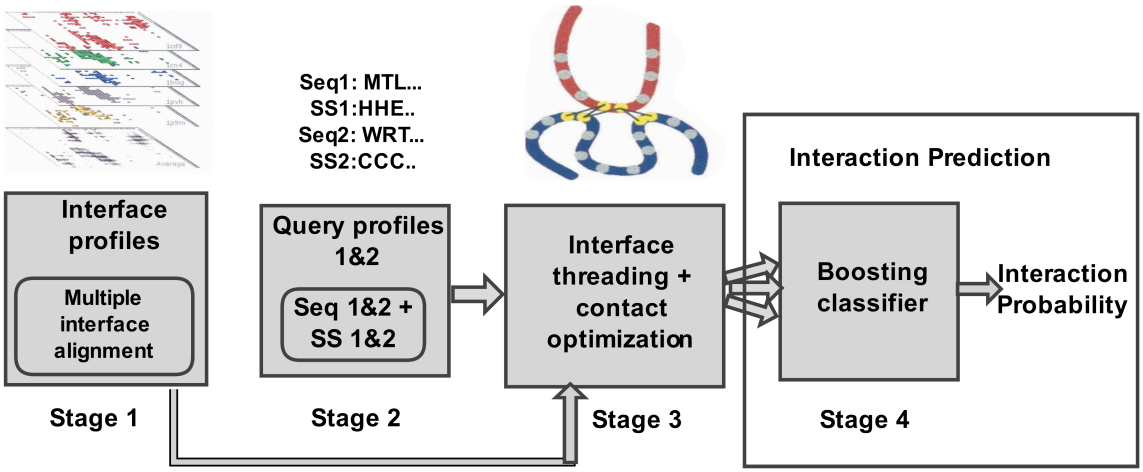
### 3.2.1 Overview of the threading algorithm

We develop iWRAP, an algorithm for threading query sequence pairs to only the interface of a suitable complex template (i.e the best template selected based on statistical significance of the alignment scores). Figure 3-1 is a schematic of iWRAP, displaying a flowchart of the various stages of the algorithm. In the first stage, template construction, from alignments of multiple protein-protein interfaces [124], we construct specific interface profiles (or consensus templates) based on amino acid propensities, secondary structure and solvent accessibilities for discrete environmental classes of the interface. The interface profiles are knowledge based propensities that capture the different biophysical forces (hydrophobic, electrostatics, structure) at the protein interfaces. Compatibility scores measured using these profiles indicate the biophysical suitability of the predicted interface. The hydrophobic effect is captured by the features of amino acids and solvent exposure. The electrostatics is captured by amino acid propensities. Structural constraints are encoded in secondary structure compatibilities.

In the second stage, alignment of a query sequence pair to a template, we utilize a profile-scoring scheme that captures amino acid sequence propensities and predicted secondary structure for the query sequences. We first identify a suitable template using a single domain threader- RAPTOR [177] (also see *PPI Prediction: yeast genome*). RAPTOR is used for whole genome scans of pairs of proteins to identify structures most compatible with each protein sequence. For each protein, we select ten top-scoring single domain structures with a threading z-score of at least 3 (see




**A.** iWRAP cartoon




**B.** Overview of iWRAP

Figure 3-1: A) Cartoon depicting how iWRAP's interface threading uses multiple templates to identify the putative interface region for the two query proteins. All the templates belong to the same SCOPPI family. B) Overview of the iWRAP's interface threading approach for PPI prediction [68].

Appendix A). We then rank the complex templates composed of these single domains based on the sum of their single-domain threading z-scores. When only one sequence of the query pair matches a domain in the complex, we do not discard it. This procedure selects for each query pair at most 10 possible complex templates for threading of the interface by iWRAP. For each of these selected complex templates, iWRAP uses a local alignment of the query sequence profile to the interface template profile; this directly reflects the quality of the interface alignment, without being influenced by alignments elsewhere in the structure. We select the best interface template using a z-score that evaluates iWRAP's interface score with respect to a distribution obtained by randomizing the interface contacts.

For the third stage, scoring the putative interaction, we begin by integrating stage 2's interface-specific alignment score into a general threading scoring scheme implemented similar to RAPTOR [177]. This produces an initial contact map, which we further refine through contact map optimization in the neighborhood of interacting residues. For the fourth stage, interaction prediction, we extract features of the predicted interface (e.g. interface energy, z-score, size) to input into a boosting classifier, which then computes a probability of interaction for the two query proteins. Note that this stage is employed only for our yeast genome scans, and not for our benchmarking tests on SCOPPI. See Materials and Methods for a more detailed description of each of these stages and training and test sets.

Our algorithm builds upon Pulim et al.'s method, LTHREADER [123], where the authors have shown that supervised construction of the interface templates, along with a localized scoring scheme based on sequence-specific profiles significantly improves alignment and prediction accuracies for the cytokine family. LTHREADER independently aligned each sequence to a profile representing one sequence of the interface template using a sliding-window approach. In contrast, iWRAP uses a linear programming approach (LP) to align pairs of sequences to a two-dimensional (2D)

profile of a protein-protein interface and utilizes pairwise quasi-chemical scores for evaluation and optimization. Additionally, LTHREADER focused the alignments on putative interaction cores determined by predicted secondary structure, while iWRAP does not make such an assumption; it uses the LP to decide the optimal interface region. iWRAP further optimizes an objective function based on the Hadamard product of 2D contact maps, thereby simultaneously adjusting interface residues of both interacting proteins. iWRAP rigorously deals with gaps in the alignment, whereas LTHREADER aligns the entire putative interaction core to the interface profile ignoring gaps altogether. Moreover, interface templates used by iWRAP are constructed by a fully-automated procedure that uses our recent multiple interface alignment algorithm CMAPi [124], while LTHREADER had to rely on time-consuming manually-constructed multiple interface alignments. In particular, LTHREADER chose parameters in its alignment algorithm to reflect the structural and physical constraints of the two cytokine families it was tested on. Extension of LTHREADER to other families would require the estimation of those parameters in a principled way. A detailed description of the nontrivial task of interface template construction from the CMAPi alignments is provided in Materials and Methods: *Template construction.* Finally, the combination of iWRAP's interface threading with a general single-domain threader (RAPTOR), the latter of which is used to identify most likely complexes for pairwise threading, allows PPI prediction on a genomic scale – a feature missing in LTHREADER.

### 3.2.2 Interface validation

We evaluate iWRAP on two challenges that one encounters using structural information to predict likely protein-protein interactions: sequence-interface alignment and interface contact prediction. For sequence-interface alignments, we first compare the performance of iWRAP with that of a full complex threader, DBLRAP [143], a profile-

| SCOPPI Family | Seq. ID (%) | LTHREADER (%) | MUSCLE (%) | DBLRAP (%) | iWRAP (%) |
|---|---|---|---|---|---|
| f.24.1.1_f.25.1.1 | 10 | 1 | 4 | 14 | **22** |
| b.47.1.2_g.8.1.1 | 18 | **34** | 24 | 0 | 32 |
| b.47.1.2_g.3.15.1 | 7 | 2 | 3 | 0 | **8** |
| a.56.1.1_d.133.1.1 | 5 | 12 | 3 | **30** | 27 |
| c.81.1.1_d.58.1.5 | 5 | 0 | 7 | 29 | **32** |
| a.74.1.1_d.144.1.7 | 11 | 16 | 10 | 19 | **26** |
| c.1.12.1_c.49.1.1 | 12 | 17 | **29** | 24 | 13 |
| c.55.1.1_d.109.1.1 | 21 | 2 | **19** | 13 | **19** |
| a.80.1.1_c.37.1.20 | 15 | 3 | 3 | 9 | **27** |
| d.133.1.1_d.87.2.1 | 11 | 0 | 0 | 15 | **24** |
| a.137.2.1_b.70.1.1 | 10 | 1 | 4 | 28 | **31** |
| d.171.1.1_h.1.8.1 | 28 | **28** | 13 | **28** | 19 |
| e.18.1.1_e.19.1.1 | 6 | 0 | 7 | 21 | **45** |
| c.2.1.4_c.23.12.1 | 15 | 1 | 20 | **25** | 21 |
| b.47.1.2_g.3.2.1 | 35 | 12 | 18 | 6 | **21** |
| d.122.1.2_d.14.1.3 | 12 | 1 | 5 | **15** | 10 |
| b.6.1.2_f.24.1.1 | 20 | 4 | 27 | 32 | **37** |
| **Average** | **14** | **8** | **11** | **18** | **24** |

Table 3.1: Comparison of iWRAP with other sequence and structure based techniques on cross validation tests in SCOPPI. The numbers indicate the alignment accuracies at the interface, with the true alignments taken as the ones given by CMAPi [124]

based alignment program MUSCLE [45] and our previous algorithm LTHREADER, in stringent cross-validation on SCOPPI. We then continue to compare the two superior alignment algorithms, iWRAP and DBLRAP, using several additional metrics that evaluate the absolute quality of the putative interface: Root Mean Square Deviation (RMSD) of the interface alignments, contact accuracy and interfacial energy (Definitions in Appendix A). See Materials and Methods for a detailed description of the training and test set construction. We emphasize that in cross-validation tests, we restrict ourselves to only difficult alignments (i.e. sequence identity $< 40\%$) because easier alignments are straightforward to address using conventional threading techniques or sequence alignment.

**Cross-validation within SCOPPI families.**

iWRAP performs better than or competitive to other sequence and structure-based techniques in terms of average alignment accuracies (Table 3.1). Average alignment accuracies are calculated by averaging the alignment accuracies computed by threading the test sequence pair to each template in the training set. iWRAP improves average alignment accuracies for roughly 80% of the families (in cross-validation tests) for which we can construct multiple interface alignments and sufficiently large training and test sets. For the remaining 20% of families, iWRAP gives equivalent or slightly lower accuracies than DBLRAP. Schematic describing the cross-validation tests on SCOPPI are shown in Figure 3-2. iWRAP performs much better than techniques based on sequence alone. We compared iWRAP with profile-based alignments computed using a state-of-the-art alignment program MUSCLE [45]. Profiles for the sequences were computed by running PSI-BLAST for 5 iterations with an E-value cutoff of 0.001 against the 'nr' protein database [6]. Profile-based alignments, rather than pairwise alignments, were used as they have been shown to be more accurate for remote homology detection [44]. iWRAP also performs much better than our earlier algorithm LTHREADER. To evaluate the additional value of iWRAP scoring function, we used our new interface profiles along with the threading approach employed by LTHREADER. Briefly, we first align the secondary structure tags of the query and template to roughly identify the interaction cores. Then we use predicted secondary structure and predicted solvent accessibilities in a scoring function similar to LTHREADER, confining the search space to within 5 residues of the secondary structure identified as the putative interaction core. In the three cases where iWRAP performs worse than any of the three previous methods, the overall sequence similarity is rather high giving these methods a slight advantage. Following on this observation, for whole genome scans, we combine DBLRAP with iWRAP.

Interfaces predicted by iWRAP are closer to true interfaces than those predicted

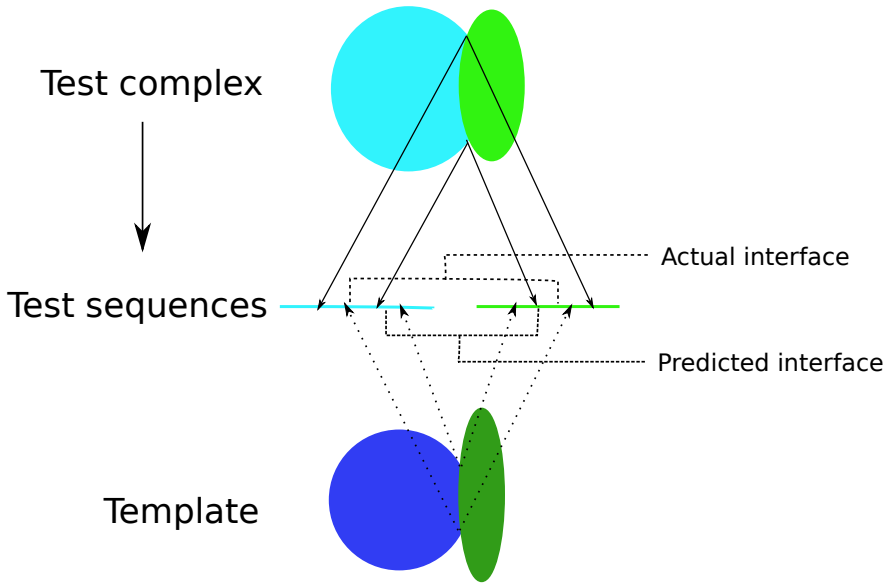



Figure 3-2: Schematic describing the cross-validation testing of iWRAP on the interface database SCOPPI. Sequences of a test complex belonging to the same family as the template, but less than 40% identical to it, are threaded onto the template using an alignment program. The predicted interface (from the threading alignment) is then compared with the actual interface (from the known structure) to compute accuracy. Dashed lines indicate the aligned interface computed using the alignment program, solid black lines indicate the actual interface mapped from the true structure.

by DBLRAP. Below we focus on comparing iWRAP and DBLRAP, since their contact accuracies are much better than that of MUSCLE and LTHREADER. As an example, Figure 3-3 illustrates the case of the interface formed in the PDB structure 1upc (Fig 3-3A) between chains A(12-195) and B(375-573). The template used for threading these two sequences is shown in Fig 3-3B, with the interface residues highlighted in green. DBLRAP completely misses the correct interface region as a result of poor alignment of chain B (Fig 3-3C), giving a contact accuracy of 0%. In contrast, iWRAP produces an initial interface closer to the true one, with a contact accuracy of 27% (Fig 3-3D). On further refinement of the contact map (see Materials and Methods: *Contact map optimization*), iWRAP's predicted interface (Fig 3-3E) is much closer to the true interface (Fig 3-3A), with 46% contact accuracy. The predicted structure of the true interface is shown in Fig 3-3F. It was constructed by mapping true interface residues (magenta, Fig 3-3A) to the template (Fig 3-3B) using alignments computed by iWRAP. iWRAP aligns the true interface residues to the interface of the template and is thus able to correctly identify the interacting residues. To emphasize the fact that iWRAP is an interface threading approach, rather than a full-complex approach, the rest of the structure is colored in light-gray. Additionally, the higher statistical significance of iWRAP's predicted interface energy (z-score=2.7), calculated by randomizing the interfacial contacts, as compared to DBLRAP's (z-score=-0.1), is further indicative of the improved interface prediction. The higher contact accuracies and associated z-scores enable iWRAP to improve PPI prediction over DBLRAP.

More generally, iWRAP outperforms other sequence-based and threading methods at correctly predicting interfacial contacts across all template-query pairs in the test set, except for a few very small interfaces (see Fig 3-4A). We find that iWRAP improves over DBLRAP in predicting interfacial contacts when the number of true contacts is greater than 25-30 (see Fig 3-4A, right of the solid vertical line). Even

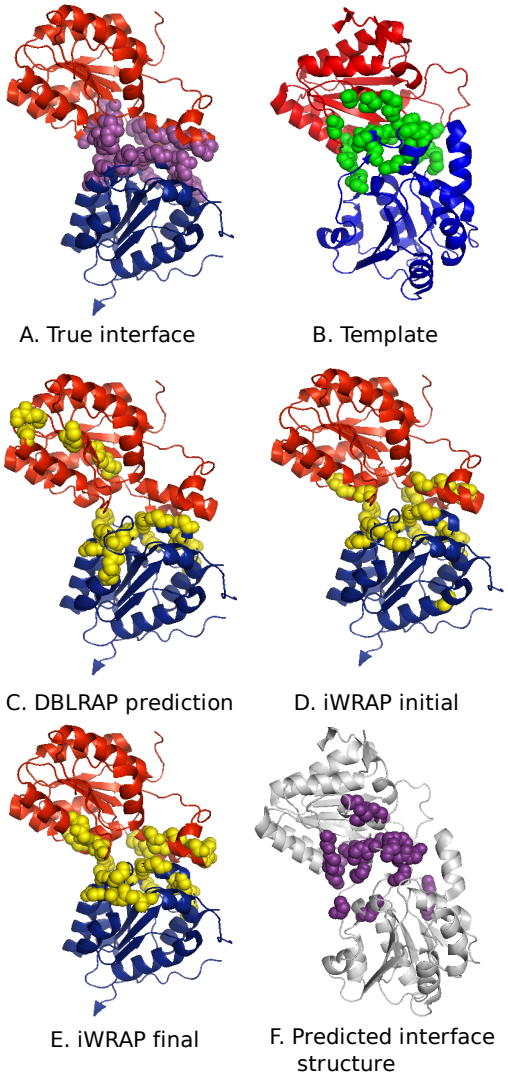



Figure 3-3: Example of improved contact predictions by iWRAP in within-family cross-validation. PDB 1upc chains A(12-195) and B(375-573) are threaded to the template 1qpbAB. A) The true interface computed from the PDB structure of 1upc has roughly 50 contacts. The interface residues are shown as purple spheres, chain B is shown in red and chain A in blue. B) The template (1qpbAB) used for threading the query sequences; the interface residues are shown in green. C) The interface residues (yellow spheres) predicted by DBLRAP. DBLRAP fails to align the interface region of one interacting partner due to low sequence homology between the query and template (contact accuracy = 0%). D) Initial interface (yellow spheres) predicted by iWRAP after threading (contact accuracy = 27%). iWRAP uses interface profiles constructed from a multiple alignment of the interfaces 1mczHG, 1jscAB, 1ozhDC and 1qpbAB; the profiles are then mapped onto the template 1qpbAB. E) Final interface (yellow spheres) predicted by iWRAP after contact map optimization. This step refines the contact map, resulting in contacts closer to the true interface. The final contact map is closer to the true contact map (contact accuracy = 46%). This is obtained by overlaying iWRAP predictions (yellow) on the actual structure of the interface (from A). F) Predicted interface structure obtained by mapping true interface residues from A onto the template structure in B using iWRAP alignments.

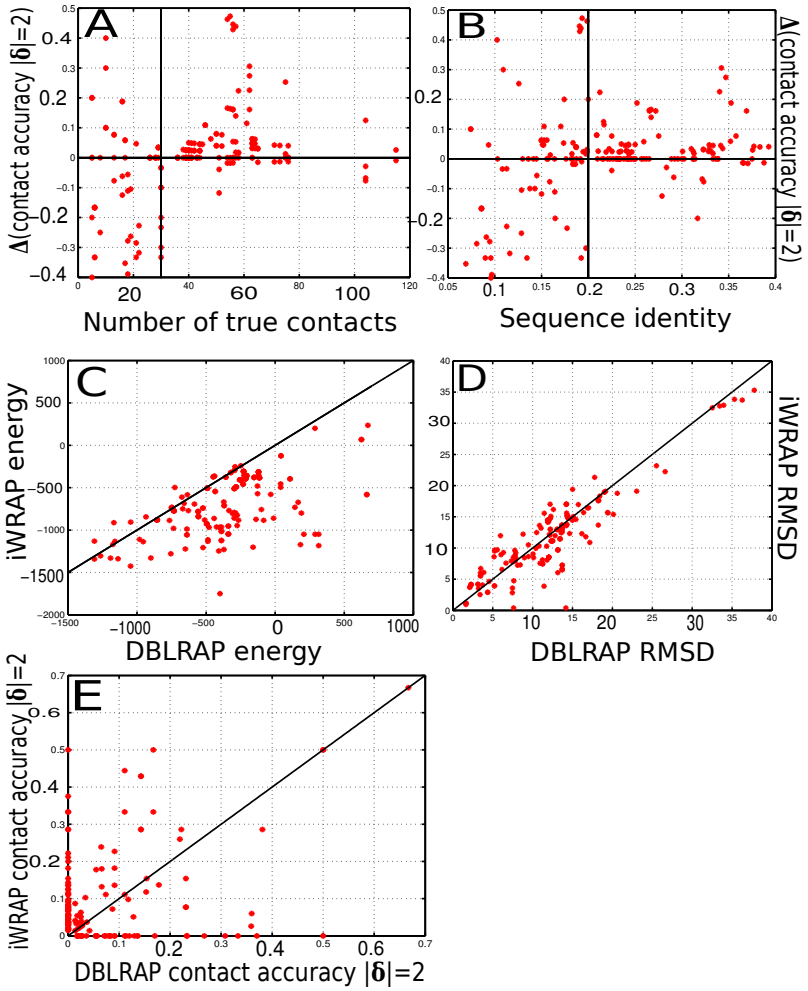



Figure 3-4: Interface alignment and contact validation. Panels A, B, C and D are cross-validation results on within SCOPPI family threading. $\Delta$(contact accuracy $|\delta|$=2) is the difference in contact accuracies ($|\delta|$=2) between iWRAP and DBLRAP. A) Contact accuracy improvement of iWRAP relative to DBLRAP as a function of number of true contacts at the interface. B) Contact accuracy improvement of iWRAP relative to DBLRAP as a function of sequence identity at the interface. C) iWRAP consistently achieves lower average interface energies as compared to DBLRAP. D) RMSD comparison between iWRAP and DBLRAP- better contact prediction by iWRAP does not affect RMSD of the predicted interface. E) Cross-validation results for interfaces sharing only one SCOP family (see *Cross-validation across SCOPPI families*). See Appendix A for calculation of contact accuracies and interface energy.

when DBLRAP fails to account for 10% of the contacts, iWRAP can predict 20-30% of the contacts.

We investigated the variation of contact accuracy with sequence similarity at the interface for the alignments in the cross-validation set. For sequence identities between 0.2 and 0.4, iWRAP significantly improves contact prediction (Fig 3-4B, right of the solid vertical line). However, when the sequence identity between the template and query becomes less than 0.15, there is no consistent improvement over DBLRAP (Fig 3-4B, left of the solid vertical line). We have also observed that other features of the interface, namely information content and iracc (see Materials and Methods: *Training and test sets*), do not significantly influence the contact predictions.

We sought to further investigate iWRAP's superior performance on medium to large contact maps (>25 contacts). We hypothesize this improvement is due to the localized character of our interface profiles. We evaluated the contact density for both methods on contact maps with greater than 25 contacts, where we presume iWRAP's profiles are aiding in its superior performance (Fig 3-4A). Following the contact-map mining techniques of Hu et al. [69], we characterized each contact by the pattern of contacts in a 5x5 residue neighborhood around it, where the average density is the number of contacts divided by 25. We observe that iWRAP contact predictions have a higher density (0.26) on average than DBLRAP predictions (0.22), on both the training and test sets. Furthermore, when the interface is small, there are many feasible alignments for the interface region; this makes it difficult for iWRAP to get accurate alignments without using restraints from the whole complex. Based on this analysis, we conclude that size and density are factors in the improved performance of iWRAP, and thus may be responsible for the decreased performance in the case of fewer than 20-25 contacts.

iWRAP consistently gives lower interface energies (normalized by the number of

predicted contacts) as compared to DBLRAP (Fig 3-4C). To predict protein interactions iWRAP and DBLRAP use the residue-level statistical potential developed by Lu et al. [102] to score putative interactions. The interaction score (energy) is obtained by summing over all the contacts in the putative interface.

We also evaluated alignments using the conventional metric of interface RMSD and confirmed that iWRAP alignments have similar or lower RMSD than DBLRAP's (Fig 3-4D). Thus iWRAP improvements in alignment and contact accuracy do not affect the RMSD of the predicted interface. Note that while optimizing the parameters, RMSD was not optimized for the threading alignments.

**Cross-validation across SCOPPI families.**

In addition to cross-validation tests within the same SCOPPI family we have tested the ability of iWRAP to accurately predict interfaces when threaded complexes are from SCOPPI family pairs sharing only one SCOP family (e.g. **b.47.1.2**_g.3.15.1 and **b.47.1.2**_g.68.1.1). For these across-family threading tests, we restricted ourselves to alignments having a high iracc score ($> 0.75$, see Materials and Methods: *Training and test sets*), thereby ensuring similar binding patterns. Successful threading of across-family pairs allows us to address PPI predictions when a template complex for the same SCOPPI family does not exist. However, in such cases, it is possible that the interaction can be predicted using a similar interface for another PPI. It is known that despite lack of overall structural similarity some proteins interact with different protein partner using a very similar interface; for example, interaction mimicry has been observed in host-pathogen interactions [152].

Most threading methods rely on a template database, which might not be completely representative and might not have an appropriate template for every query sequence. While traditional cross-validation strategies do not perform across-family tests, we do so in order to try to address the problem of the limited number of

templates available for genome-wide PPI predictions.

For across-family predictions iWRAP predicts the interacting residues more accurately than DBLRAP for 75% of SCOPPI family pairs in the cross-validation test. Despite the high iracc score (>0.75) for such alignments, the binding patterns might be relatively different, leading to a poorer overall prediction by DBLRAP. However, for cases when DBLRAP fails to predict even 10% of contacts, iWRAP can account for nearly 20-30% of the true contacts (see Fig 3-4E). This suggests that using iWRAP for PPI prediction with templates of complexes sharing one SCOP family can increase the coverage of predictions.

### 3.2.3 PPI Prediction: yeast genome

We have applied iWRAP for genome-scale analysis to predict the yeast interaction network. In cross-validation tests above, we used templates in the training set to thread query sequences in the test set. For the yeast genome scan we use a single sequence threader, RAPTOR, to identify suitable templates for each sequence in the query pair using z-score > 3.0. If we do not have an interface template for a SCOPPI family composed of the SCOP families corresponding to any combination of these templates, we use DBLRAP to thread the two sequences onto a conventional full-complex template (see Appendix A for details). Once the putative interface is determined, we use interface-specific scores to predict the interaction between the proteins (stage 4). See Materials and Methods for a detailed description of the classifier employed to predict an interaction.

In order to evaluate our predictions, we compute a receiver operating characteric (ROC) curve by varying the probability cutoff for predicting an interaction. When comparing ROC curves against other homology/structure-based PPI predictors, we find that iWRAP consistently outperforms HOMCOS, Multiprospector and DBLRAP. Multiprospector reports a sensitivity of 20% at a specificity of 80%, whereas

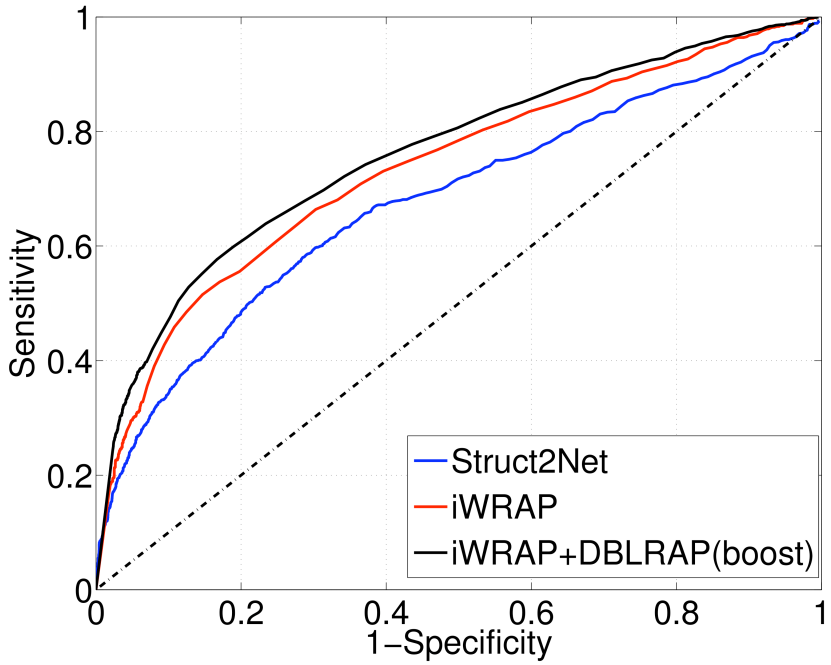



Figure 3-5: Results on the yeast genome. Sensitivity vs specificity for iWRAP, Struct2Net and iWRAP+DBLRAP (combined method). In the combined method, DBLRAP threading results are boosted and combined with iWRAP predictions. AUCs for the three methods are: 0.734 (iWRAP), 0.680 (Struct2Net) and 0.762 (combined). All the differences are statistically significant ($P < 10^{-10}$, *t-test*).Here sensitivity = (true positives)/ (true positives + false negatives) and specificity= (true negatives)/(true negatives + false positives).

iWRAP achieves a sensitivity of 56% at 80% specificity (see Fig 3-5). HOMCOS reports a recall of 80% with a precision of 10%. In contrast, iWRAP achieves a precision of 27% at the 80% recall level (see Fig A-2). Struct2Net [143, 142] uses the DBLRAP threading program for prediction of interactions from structural data. When comparing against Struct2Net (only yeast predictions), we find that iWRAP dominates Struct2Net at all accuracy levels (see Fig 3-5).

Interface threading requires multiple structural data for an interaction, which is not always available. By using interface threading in conjunction with DBLRAP, our method, i.e. iWRAP+DBLRAP(boost), achieves a coverage of 13% for the yeast interactome. This is close to a 50% increase in coverage over previous methods [142], without any compromise in sensitivity (Fig 3-5). Here, coverage is defined as the percentage of high-confidence interactions in Biogrid [151] for which a method can make a prediction. iWRAP makes predictions for 9752 high-confidence interactions in Biogrid (involving around 3400 proteins), whereas DBLRAP makes predictions for 5832 interactions (involving around 2700 proteins). 3920 are unique to iWRAP's interface threading predictions; this results in close to a 50% increase in coverage compared to DBLRAP. In addition, iWRAP predicts about 100,000 novel interactions in the yeast genome; the cutoff ($= 0.9$) for identifying a positive interaction is chosen based on the distribution of interaction probabilities. We note that around 60% of our predictions come from across family threading– that's not surprising given the limited template database; it is more likely to have a good match to one sequence of the query, than to both of them.

To further analyze iWRAP's performance, we looked at the 640 proteins involved in the high-confidence interactions from Biogrid uniquely predicted by iWRAP. One finding from a GO term enrichment analysis using Amigo [28] revealed that this set was enriched for proteins functioning as structural constituents of the ribosome (GO: 0003735, P-value $< 10e - 6$). Additionally, iWRAP makes predictions for pro-

teins within functional complexes involving nuclear proteins such as the 'U5 snRNP complex' and 'SAS complex'. Amongst the type of functional complexes that both iWRAP and DBLRAP predict, we find that iWRAP's predictions are significantly enriched for the following complexes (> 6 fold over DBLRAP): 'Rtt109p/Vps75p complex' (12 fold over DBLRAP), 'signal peptidase complex' (11 fold) and 'GPI-anchor transamidase complex' (9 fold). The full list of such complexes and complexes unique to iWRAP predictions is given in Appendix A: *Genomic Predictions*. The annotation of these complexes, including their memberships, were taken from a manually-curated dataset compiled by Pu et al. [122]. Finally, we investigated the templates selected for the unique predictions made by iWRAP. Table 3.2 gives a summary of the most frequent templates used for predicting these interactions. While DBLRAP selects one representative complex for each SCOPPI family, multiple templates can be selected by iWRAP from within a family. This contributes to iWRAP's improved prediction accuracy as features for only the most significant interface are considered for PPI prediction. Furthermore, as noted earlier in cross-validation tests, size of the interface template is correlated with iWRAP's accuracy: larger interfaces lead to more confident predictions. From Table 3.2, the average probability computed by iWRAP for interface templates of size less than 20 contacts (mean=0.20, std.dev=0.13) is half of the average probability computed for templates greater than 20 (mean=0.40, std.dev=0.20).

### 3.2.4 iWRAP predicts novel cancer-related interactions

We demonstrate that iWRAP can be used to identify important targets for experimental investigation through an application to yeast homologs of human cancer-related genes. We integrate enrichment and functional analysis to enumerate bona fide candidates for further investigation (Fig 3-6). Recently, a large scale double-mutant study has revealed a genetic interaction map for yeast [32]. However, the set of interesting

| SCOPPI Family | Template | Size of Interface | Number of interactions in test set | Average Probability |
|---|---|---|---|---|
| f.17.2.1_f.24.1.1 | 1m56H30-1m56G14 | 135 | 40 | 0.297 |
| f.17.2.1_f.24.1.1 | 1qleB1-1qleA17 | 132 | 23 | 0.398 |
| f.17.2.1_f.24.1.1 | 1v55B2-1v55A2 | 124 | 33 | 0.481 |
| f.17.2.1_f.24.1.1 | 1fftG27-1fftF52 | 96 | 18 | 0.183 |
| b.40.4.1_d.104.1.1 | 1asyA68-1asyB205 | 63 | 5 | 0.400 |
| b.40.4.1_d.104.1.1 | 1b8aB1001-1b8aA104 | 51 | 16 | 0.624 |
| b.40.4.1_d.104.1.1 | 1g51A1-1g51B1105 | 46 | 9 | 0.667 |
| b.40.4.1_d.104.1.1 | 1n9wB1-1n9wA111 | 43 | 14 | 0.428 |
| a.56.1.1_d.133.1.1 | 1jrpE85-1jrpF124 | 61 | 8 | 0.000 |
| c.55.1.1_d.109.1.1 | 1yagA147-1yagG1 | 45 | 8 | 0.732 |
| c.55.1.1_d.109.1.1 | 1h1vA147-1h1vG412 | 32 | 7 | 0.281 |
| b.40.2.2_d.19.1.1 | 1d5mC2-1d5mA4 | 41 | 12 | 0.180 |
| d.185.1.1_f.23.12.1 | 1bgyM234-1bgyQ1 | 22 | 24 | 0.333 |
| d.185.1.1_f.23.12.1 | 1bccA233-1bccE1 | 22 | 16 | 0.499 |
| a.39.1.5_c.37.1.9 | 1dfkZ3-1dfkA6 | 19 | 32 | 0.258 |
| a.39.1.5_c.37.1.9 | 1dflX4-1dflB5 | 14 | 23 | 0.277 |
| a.80.1.1_c.37.1.20 | 1sxjA548-1sxjB7 | 16 | 18 | 0.397 |
| a.80.1.1_c.37.1.20 | 1iqpC233-1iqpD2 | 15 | 10 | 0.100 |
| a.80.1.1_c.37.1.20 | 1jr3B243-1jr3E1 | 10 | 10 | 0.300 |
| d.185.1.1_f.23.12.1 | 1kb9A240-1kb9E31 | 7 | 6 | 0.000 |

Table 3.2: The most frequent templates used by iWRAP for threading sequences involved in high-confidence interactions in Biogrid unique to iWRAP. Column 2 gives the total number of pairs threaded using the template, column 3 gives the number of pairs in the test set and column 4 gives the average predicted probability of interactions in the test set. A template id '1v55B2-1v55A2' represents the interface formed by SCOP domains in chain B and chain A in the PDB complex '1v55'.

genes for any detailed study of a disease (e.g. cancer) is still large. In contrast to this approach, we use iWRAP predictions to identify the most important targets for further study. It has been shown that structure-based scores are one of the most significant predictors, as compared to co-localization, co-expression and GO term enrichment, for general PPI prediction [11, 143]. We employ these criteria to prioritize

and validate our targets (Fig 3-6A). For the set of yeast genes related to cancer identified in CYGD [62], we first filter the predicted interactions based on co-localization. iWRAP identifies 727 interactions for the disease genes (out of ~54000 possible interactions). After discarding predictions between proteins that are not co-localized; 301 putative interactions remain for further analysis. We then identify genes enriched for GO processes, with the genetic interaction set as the background. Note that this is a much more stringent criterion than using the whole genome as the background; the latter yields many more putative interacting genes. We used AmiGO [28] to filter genes based on a p-value cutoff of 0.01 (corrected for multiple hypothesis testing). The enrichment analysis narrows down the list of candidate genes to 28. Note that we are using both co-localization and enrichment as filters to select the most important candidate genes; we treat both of them as equally important. For genes that were significantly enriched (~4 fold), we used IsoBase [144] to identify their human funtional orthologs. To exploit the more comprehensive yeast genome annotation, we carried out the enrichment on the yeast predictions before mapping them onto the human genome. We found that these enriched genes are differentially expressed in cancer-vs-normal tissues [128]. Furthermore, using BLAST we were able to identify similar proteins (E-value $< 10$) in a database of cancer-related proteins [71]. We hypothesize that these novel interactions are directly involved in cancer-related pathways, and should be investigated further (Fig 3-6B).

Amongst the genes predicted by iWRAP as interacting with known cancer promoting genes, particularly interesting are genes coding for ribosomal proteins associated with either the small (RPS) or large (RPL) subunit (Fig 3-6B). Mutations in several of these proteins, including RPS17 and RPL5 identified by iWRAP, have been very recently implicated in congenital abnormalities and predisposition to cancer, known as Diamond Blackfan Anemia (DBA) [99]. The expression disregulation of RPS and RPL genes have also been observed in pancreatic cancer and stromal displasia [33]

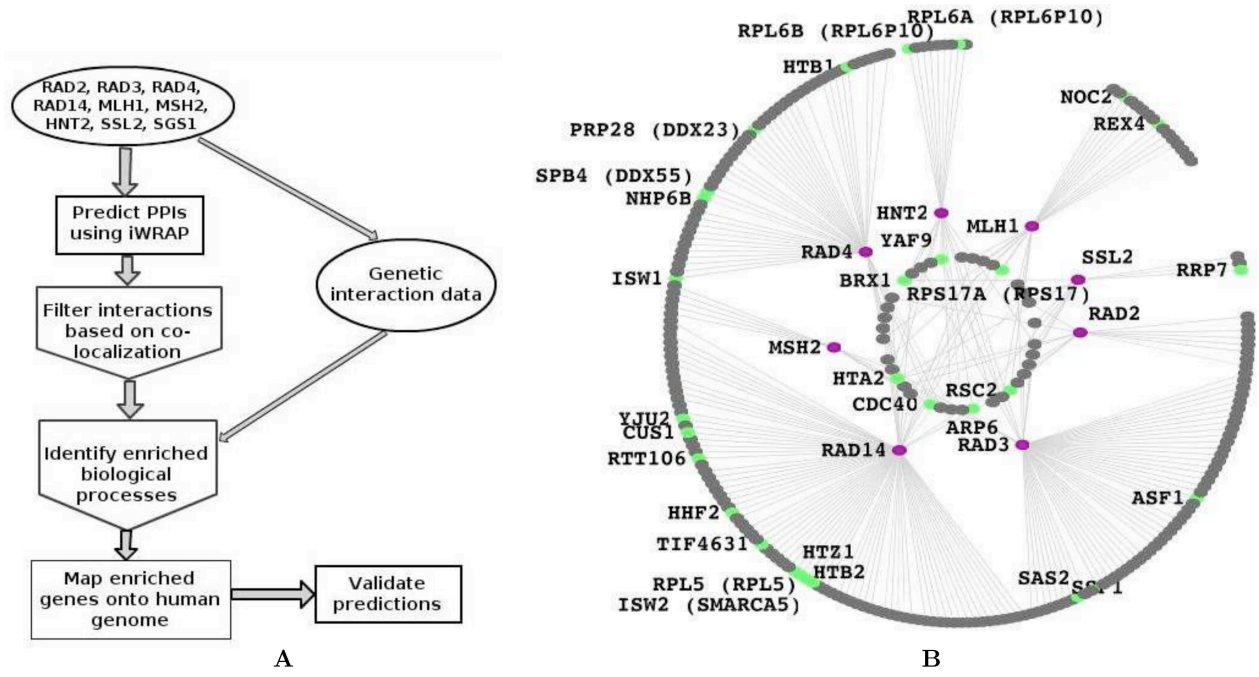



Figure 3-6: iWRAP predicts novel, bona fide interactions. A) Enrichment analysis was carried out to identify high-confidence interactions. Genes filtered by co-localization and significantly enriched compared to the genetic interaction set were validated using the Oncomine and HCPIN databases. Number of genes remaining after each stage are indicated in parantheses. B) The analysis in A reveals a set of high-confidence genes (green) predicted to be interacting with yeast homologs of cancer related genes (purple). Human orthologs of genes for which there is literature providing evidence of implication in cancer have been indicated in parentheses. Genes interacting with only one "cancer" (purple) gene are in the outermost circle, whereas those interacting with more are in the innermost circle. Genes which are not significantly enriched are colored in grey, however, these predicted interactions could also reveal novel biological insights. The figure was created using Cytoscape [136]

and in colorectal cancer [94]. In addition, there are two (human DEAD box) helicases DDX23 and DDX55 (Fig 3-6B) in the set of putative interactions. Even though there is limited research on various human helicases they are believed to be involved in embryogenesis and cell growth and have recently been shown to be involved in tumorigenesis [115]. Furthermore, iWRAP predicts an interaction between XPA (RAD14) and SMARCA5; the latter has been shown to be critical for regulating the genetic program required for normal differentiation [156].

## 3.3 Materials and Methods

### 3.3.1 Stage 1: Template construction

We utilize the SCOPPI classification of protein-protein interfaces to construct interface profiles. SCOPPI classifies interfaces based on sequence and structural similarity of the interface [173]. In addition, for each interacting SCOP family pair, SCOPPI provides a sequence alignment of other interfaces in the same SCOP family pair. Here we use this classification of interfaces to construct our own multiple interface alignments for each SCOP family pair using CMAPi [124]. CMAPi employs a contact-map representation to efficiently align multiple interfaces and thereby improves alignments, as compared with SCOPPI and other sequence/structure-based alignment programs, especially in cases where the sequence identity between aligned structures is low [124]. A contact map is a binary matrix representation of the residue-residue interactions between two proteins. If the distance between any two heavy atoms of the two residues is less than 4.5Å, the corresponding entry in the contact map is one, and zero otherwise.

We construct interface profiles from these interface alignments by computing a unique set of consensus environment classes, one for each interface alignment position (see Fig A-1). An environment class is a combination of a secondary structure (SS) class, an amino acid class and average solvent accessibility (across the alignment at that position). We use the classification as defined by Rice et al. (1997), which, briefly, consists of three SS classes, two solvent accessibility classes and seven amino acid classes. Rice et al. also provide a table, H3P2, which provides amino acid/SS preferences for these environmental classes. The profiles computed from a multiple interface alignment represent the environment information at the interface across the multiple structures in the alignment. Since the consensus contact map constructed by CMAPi includes all contacts across the aligned complexes, our interface profiles

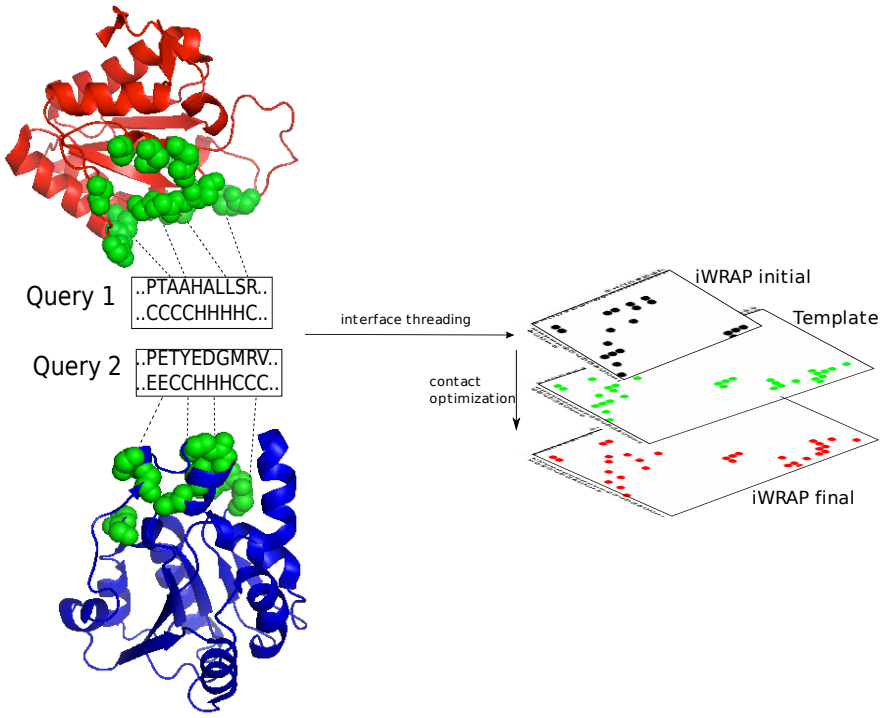



Figure 3-7: Schematic of interface threading and contact optimization. For the example shown in Figure 2, the query proteins are individually aligned to the template (left) using a local alignment to the interface (dashed lines). For scoring this alignment, we use the interface profiles computed from the multiple-interface alignments, predicted secondary structure for the query pair and the single-domain threading score of RAPTOR. Minimizing this alignment score produces an initial contact map, 'iWRAP initial', which is further refined using Hadamard product optimization and quasi-chemical pairwise residue potentials to produce 'iWRAP final' (right).

are robust to small variations in inter-residue distances.

### 3.3.2 Stage 2: Aligning query sequences to templates

The goal in this stage is to align query sequence profiles to interface template profiles, constructed in stage 1. We obtain query sequence profiles from PSIBLAST [6] and query secondary structure (SS) predictions from PSIPred [77]. Once we identify a suitable template, we score individual query-template alignments using Rice et al.'s H3P2 table (see above), which, in the context of single structure alignment, quantifies the preference of aligning a query sequence/SS profile to a template profile. However,

since our query SS's are predicted, we instead use H3P2 scores weighted by the PSIPred SS probability distribution at a query sequence position.

$$H3P2score(t, s(t)) = \sum_{ss=C,H,E} P(ss)H3P2(s(t), ss, t) \tag{3.1}$$

Here $t$ is the template position, $s(t)$ is the query sequence position aligned to template position $t$, $ss$ is C(coil), H(helix) or E(beta strand), $P(ss)$ is the probability of a secondary structure class at position $s(t)$ given by PSIPred and $H3P2(s(t), ss, t)$ is the H3P2 table score of aligning query $s(t)$ having $ss$ to the template position $t$. While Eq. 3.1 represents the score for one aligned position, the total alignment score is calculated by summing over all aligned positions. Note that we utilize only one state 'C' to model loops. We currently do not distinguish between coil and other structural loops such as beta turns or tight turns.

### 3.3.3 Stage 3: Interface scoring

The goal in this stage is to integrate the interface profile scoring scheme from stage 2 into a general threading approach to obtain a score for a putative interaction. Our solution employs a LP strategy motivated by that used by RAPTOR for single-domain threading. We begin by constructing our objective function. For each sequence in the query pair, in addition to the RAPTOR single-domain threading score, we include the interface profile score (see stages 1,2 above) of aligning the query sequence, $s$, with the interface template profile:

$$E_{iWRAP} = E_{RAP} - \alpha E_{CMAPi} - \omega_{gap} GAP_{RAP} \tag{3.2}$$

$$E_{CMAPi} = \sum_{t} H3P2score(t, s(t)) \tag{3.3}$$

$$E_{RAP} = E_m + E_s + E_g + E_p + E_{ss} \tag{3.4}$$

$E_{iWRAP}$ is the interface threading energy function (scoring function); $E_{CMAPi}$ is the interface profile score; $H3P2score$ is the alignment score from the H3P2 table (see stage 2, Eq 3.1); and $GAP_{RAP}$ is the total gap (opening+extension) score used by RAPTOR. $E_{RAP}$ is the threading score employed by RAPTOR. This includes environment fitness score $E_s$ based on solvent accessibility, secondary structure compatibility score $E_{ss}$, sequence profile scores calculated from PSI-BLAST $E_m$, an affine gap penalty $E_g$ and a pairwise within-domain interaction score $E_p$ [177]. To score an alignment to an interface template position represented by a gap state, we use the mean negative score in the H3P2 table (i.e. mean of the unfavorable alignment scores). To take into account possible gaps at the interface, we add a weighted negative penalty ($\omega_{gap}GAP_{RAP}$) to the score. Note that parameters $\alpha$ and $\omega_{gap}$ are optimized independently based on our training set, as described in *Training and test sets*. To obtain the alignment, the $E_{iWRAP}$ score is minimized independently for each of the two query sequences using the implementation of RAPTOR, which utilizes an open-source optimization library (COIN) [101] (Fig 3-7, left).

**Contact map optimization**

From the independent interface threading above, we produce an initial query contact map (Fig 3-7, right). We further refine this contact map by incorporating residue-residue interaction specificity and optimizing similarity of the binding patterns in query and template. We carry out optimization in the neighborhood of interacting residues using a residue-residue interaction score [102]. A 10x10 sub-matrix in the contact map around an interacting pair defines this local neighborhood. For each contact $(S_1, S_2)$ in the initial contact map, we maximize the Hadamard product between two matrices: one, a sub-matrix around the predicted contact in the query contact map ($Q_{cmap}$) and two, a sub-matrix around the corresponding template contact in the template contact map ($T_{cmap}$). If $(T_1, T_2)$ is the corresponding template contact,

then this optimization can be written as:

$$A = \arg \max_{s_1=S_1+d_1, s_2=S_2+d_2} \sum_{d_1,d_2 \,\epsilon[-5,5]} \delta(Q_{cmap}(s_1, s_2), T_{cmap}(T_1 + d_1, T_2 + d_2)) \quad (3.5)$$

where '$A$' represents the set of possible contacts that maximize the Hadamard product, $\delta$ is the kronecker-delta function and $d_1, d_2$ are the sub-matrix indices. This optimization maximizes (around each contact) the similarity of binding patterns in the template and query contact maps. For residues aligned to gaps, we allow the alignment to shift so that the nearest non-gapped position is used in the Hadamard product optimization. Since each Hadamard optimization is performed independently, one template contact could be mapped to multiple contacts in the query contact map. To avoid one to many mappings, for each template contact, we rank the possible predicted contacts using the quasi-chemical residue-residue interaction scoring potential of Lu et al. [102] ('$E_{pwqc}$') and choose the top ranking unique one:

$$optimizedContact = \arg\min_{c \,\epsilon A} \; E_{pwqc}(c) \quad (3.6)$$

The final contact map is the set of these *optimizedContact*s (Fig 3-7, right). Additionally, the significance of the predicted interaction score is measured by calculating a z-score with respect to a distribution generated by randomizing the interfacial contacts. The total score (energy) of the interface and the associated z-score are used in predicting interactions in stage 4.

### 3.3.4 Stage 4: PPI prediction

The goal in this stage is to predict whether the two query proteins interact based on the interface score computed in stage 3. Since only a few protein pairs interact *in vivo*, the main challenge here is to discriminate true interactions from false ones. To

achieve this goal, we extract a vector of scores '$X_{Interface}$' that quantifies the quality of the predicted interface [143] and feed this vector to a boosting classifier, which computes a probability 'p' of the interaction:

$$p = f(X_{Interface}) \tag{3.7}$$

$$\begin{aligned} X_{Interface} = \{&tA, tB, sA, sB, cmap, E, e, zA, zB, \\ &z\_e, tZ, E\_pi, cmap\_pi, piAB\} \end{aligned} \tag{3.8}$$

We extract the following features, i.e. '$X_{Interface}$', from the putative interface: template sequence lengths (*tA, tB*), query sequence lengths (*sA, sB*), predicted number of contacts (*cmap*), total interface energy computed from the pairwise potential ($E = \sum_{c \epsilon optimizedContacts} E_{pwqc}(c)$), normalized interface energy ($e$), z-scores for the threading alignments (*zA, zB*) and z-score for the interface energy ($z\_e$). In addition, we use the features sum of threading z-scores (*tZ*), square root of the product of sequence lengths (*piAB*), total interface energy normalized by *piAB* ($E\_pi$) and number of contacts normalized by *piAB* ($cmap\_pi$).

We train a boosting classifier on known high-confidence interactions from Biogrid to learn an accurate function '$f$'. Our method is based on AdaBoost, which involves improving the overall classification by appropriately weighting outputs of a series of rules of thumb, or base classifiers; we use classification trees as the base classifiers [34] (see Appendix A for details). Using this trained model a probability of interaction is computed, which indicates iWRAP's confidence in predicting an interaction between the query proteins: 1 indicates maximum confidence and 0 indicates no confidence. Note that this stage is used only for our genome scans, where we have no *a priori* knowledge of interaction between the query proteins.

### 3.3.5 Training and test sets

For each SCOPPI family (i.e. SCOP family pair), the set of complexes is divided into a training set and a test set; a leave-one-out cross-validation (LOOC) procedure is employed to optimize the parameters. A complex in the test set has an interface sequence identity less than 40% with each of the complexes in the training set. The complexes from the training set are used in constructing the multiple interface alignments with CMAPi, and subsequently the interface profiles. We use the training set to optimize the two parameters in the scoring function, $\alpha$ and $\omega_{gap}$ from Eq 3.2. The parameters are varied alternatively to maximize the alignment accuracy of the threading alignments, where CMAPi alignments are used as the gold-standard. At each iteration, $\alpha$ is varied in intervals of 5, and $\omega_{gap}$ is varied in intervals of 0.1. The parameter value which gives the maximum alignment accuracy is chosen at each iteration. After an initial broad sweep for $\alpha$, the parameters typically converge within 20 iterations.

In addition to LOOC testing within a SCOPPI family, we consider the performance of iWRAP on complexes having similar binding patterns (as given by an iracc of greater than 0.75) across families. Interacting residue accuracy (iracc) gives a measure of similarity in binding patterns between two interfaces: an iracc of one indicates very similar interfaces, and zero highly dissimilar interfaces [123]. For across-family cross-validation, we restrict ourselves to SCOPPI family pairs sharing one SCOP family. Notice that the parameter optimization has been carried out independently for each SCOPPI family, and hence alignments across SCOPPI family pairs are independent of the training process.

In order to train the classifier in stage 4 for our genomic scans, we constructed the set of training examples as in Struct2Net [143, 142]. Briefly, the set of positive examples was taken as the high-confidence interactions in Biogrid [151]. Any two proteins separated by at least three edges in the interaction network constructed

from Biogrid were considered as non-interacting, and included in the negative set. For our predictions on yeast, the training set consisted of 3500 positive and 16000 negative examples. Our test set had 720 positive and 3000 negative examples.

## 3.4 Discussion

We introduce the program iWRAP and show that integrating interface profiles into a localized scoring scheme aids in interfacial contact prediction. We introduce the use of across-family templates to mitigate the limited number of templates, and also capture convergently evolved interface motifs. We apply our approach to predict interacting proteins encoded by the entire yeast genome. Furthermore, by integrating our predictions in a combined functional and enrichment study of cancer related genes in yeast, we show that iWRAP can uncover novel, biologically relevant interactions.

While we have optimized the two new parameters ($\alpha$ and $\omega_{gap}$) in our threading scoring function that measure the biophysical compatibility at the interface (see Materials and Methods), it would be interesting to see if simultaneously optimizing the other parameters, already optimized separately in the fold recognition score of RAPTOR, improves accuracies even further. In particular, we expect the sequence profile and secondary structure scores to be the most important for very low sequence identities; as we have shown in Fig 3-4B, the interface profiles may not be sufficient to pinpoint the exact interaction core in such cases. As noted in *Cross-validation within SCOPPI families*, for sparse contacts and small interfaces in long sequences, the localized nature of iWRAP can miss the interaction core, thus identifying an incorrect interacting surface. In such cases, a pre-processing step with DBLRAP to roughly identify the interface region could be beneficial before using the localized threading algorithm.

In this paper, we have focused on SCOPPI families having more than three com-

plexes in a binding mode. In addition, we have not considered complexes formed by domains in the same SCOP family, which rules out homodimers (as handled by HOMCOS). Combining interface threading with DBLRAP effectively addresses limitations of small number of SCOPPI-derived interface templates. Furthermore, for families having only one solved complex, we plan to utilize interface profiles computed from PSI-BLAST as input to our localized algorithm. We believe that an expanded template database and a full optimization of the scoring function parameters will improve iWRAP's predictive abilities even further.

Our program iWRAP makes accurate PPI predictions that are independent of all the non-structure-based approaches and may thus be combined with any of them. iWRAP novelly uses physicochemical properties specific to protein interfaces to better identify interacting regions between the query proteins. iWRAP is designed to handle template-query pairs having low sequence similarity, making it complementary to other PPI databases like MODBase [119]. A key advantage of iWRAP is that, apart from the PDB data used for constructing templates, the prediction algorithm only requires protein sequence data as input. It can thus be applied to proteins for which no functional data is available.

# Chapter 4

# Coev2Net: a computational framework for boosting confidence in HTP PPI datasets

## 4.1 Background

Despite considerable improvements in HTP techniques, they are still prone to spurious errors and systematic biases, yielding a significant number of false-positives and false-negatives [148, 16, 165, 168, 8, 70]. This limits our ability to assess the true quality and coverage of the "interactome" [166, 182, 43].

Several attempts have been made to characterize the quality of the interactions obtained from HTP experiments [182, 157, 141, 43, 135, 30, 8, 70]. Experimental methods aim to limit false discovery by performing multiple iterations of the screen, which are time-consuming and expensive [135]. Secondary data, such as co-expression, co-localization, ontology correlation, topological features and orthology information are often used to further improve confidence in predicted interactions [143, 74]. In addition to non-trivial correlations between these features (i.e. co-expression need not

imply interaction), this data is not complete for all proteins. Furthermore, as more and more genomes are sequenced, only a fraction of proteins will have additional data to complement any experimental HTP study. Techniques developed from integrating interactions observed in common across multiple secondary experimental assays of an initial network are laborious, expensive and time-consuming. Moreover, as suggested by Venkatesan et al. and Cusick et al. [35], the low overlaps achieved across different datasets highlight the differences in sampling and biases in experimental techniques rather than pinpoint the true interactions. Moreover, in many experimental methods, the confidence of observations is evaluated for that specific technique – they are seldom generalizable. Thus cost-effective and high-confident strategies are clearly required to complete the human interactome.

As seen in Chapter 1 and 2, a number of algorithms have been developed to predict protein interactions by integrating complementary data such as sequence features and structural features [13, 15, 25, 39, 49, 138, 163, 164, 60, 96]. Also recently, computational approaches to PPI prediction using structural information have been gaining much attention due to the rapid growth of the Protein Data Bank (PDB) [5, 3, 7, 46, 56, 57, 68, 71, 86, 87, 92, 102, 103, 107, 143, 154, 153, 158, 159, 160, 161, 170, 172, 123].

In this chapter, I introduce a general framework to predict, assess and boost confidence in individual interactions inferred from a HTP experiment. Our contribution is three-fold – 1) we develop a novel computational algorithm to quantitatively predict interactions, given just the protein sequences; 2) we show how the algorithm can be used in a general framework to quantify confidence in observed interactions; and 3) we demonstrate the utility of our structure-based framework in providing biologically significant additional information about binding sites, which is not provided by any other HTP method (either computational or experimental). As compared to iWRAP and DBLRAP, Coev2Net improves accuracy and coverage further by making a pre-

diction even with limited structural data. Coev2Net samples correlated mutations at the interface of protein-protein interactions to enrich sequence and structure profiles for accurate prediction.

We first validate our method on a high-confidence network in the recently investigated human Mitogen Activated Protein Kinase (MAPK) interactome [10, 167]. We experimentally validate predicted high-confidence interactions for the MAPK interactome using a complementary assay and show that the concordance between prediction and experimental validation is as good as the overlaps achieved in previous protocols involving multiple secondary assays [22]. Finally, we show that the interfaces predicted by our algorithm are enriched for functionally important sites in the context of signaling networks; and utilize this information to hypothesize a novel regulatory mechanism involving cross talk between the insulin and stress-response pathways via interactions between proteins MAPK6, YWHAZ and FOXO3 proteins.

## 4.2 Results

### 4.2.1 The Coev2Net framework

We developed Coev2Net (Fig 4-1), a framework for assessing confidence in protein interactions. To quantify confidence in an interactome, we incorporate high-confidence data sources, namely low throughput interactions and structural information. The framework gives a confidence score for each interaction, along with a predicted model of the binding interface for the proteins (Fig 4-1).

Inputs to the framework are a high-confidence network (usually much smaller than the HTP screen) and the interactions identified from the HTP experiment for which one wishes to quantify confidence. For every pair of interaction in the HTP screen, Coev2Net provides a score to assess their likelihood of being co-evolved from interacting homologous sequences (see Methods). To do this, Coev2Net first predicts a likely

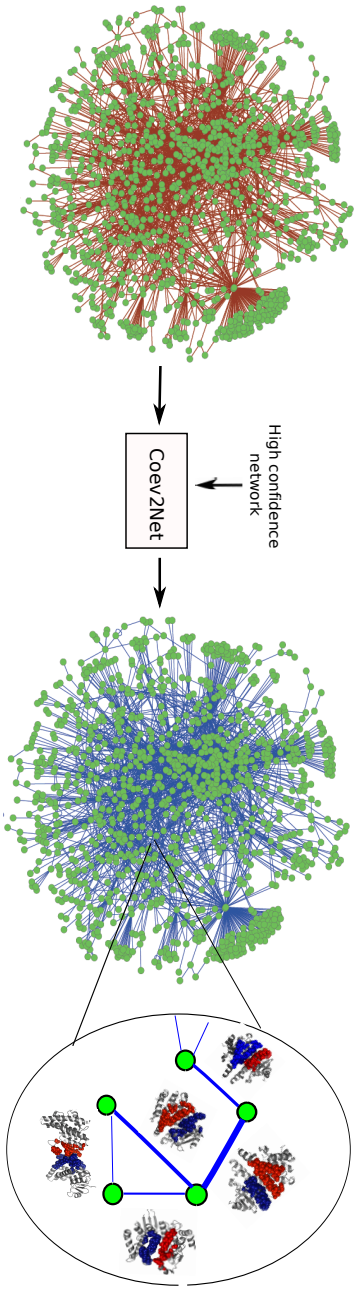



Figure 4-1: Framework for assessing confidence in a HTP PPI screen. Coev2Net, re-trained on a high-quality PPI network, is able to assign structure-based confidence scores for HTP PPI networks. Each node represents a protein and each edge the putative interaction between the two proteins. The thickness of an edge describes structure-based confidences of putative PPIs.

interface model for the two proteins, by threading [177] the sequences onto the best-fit template complex in our library. It then computes the likelihood of co-evolution of the two proteins (i.e. of the predicted interface) with respect to a probabilistic graphical model induced by the aligned interfaces of artificial orthologous sequences (Fig 4-2). By generating artificial sequences, we enrich the interfacial sequence/structure profile for those protein-pairs with sparse sequence profiles and thus improve protein interface scoring accuracy. These PGM scores are then input into a classifier trained on a small high-confidence network to compute a score between 0 and 1, representing the confidence of our method in that interaction (Fig 4-1). High-scoring interactions can then be investigated further using a secondary experimental assay or taken as true positives for subsequent analyses. Additionally, since Coev2Net is a structure-based algorithm, it also produces as output a putative interface for the interacting pair (Fig 4-2). This information can be analyzed to design site-directed experiments to further characterize the specificity of the interaction.

### 4.2.2 Benchmarking Coev2Net

*SCOPPI*: We first benchmark Coev2Net on SCOPPI [173], a protein complex database. The database is divided into interacting family pairs for which multiple complexes have been solved. Rigorous cross-validation tests on the database indicate that Coev2Net achieves high accuracies, thereby validating our approach of modeling interface co-evolution as a high-dimensional sampling problem (Fig B-2). For the cross-validation tests, we considered only those family pairs in SCOPPI that have at least three non-redundant (sequence id $< 50\%$) complexes. We randomly selected one as the test complex and used the other complexes within our Coev2Net protocol to simulate interacting homologs and construct the probabilistic graphical model (Fig 4-2). Furthermore, Coev2Net also performs well on SCOPPI family pairs not having more than two non-redundant complexes, indicating Coev2Net's ability to deal with

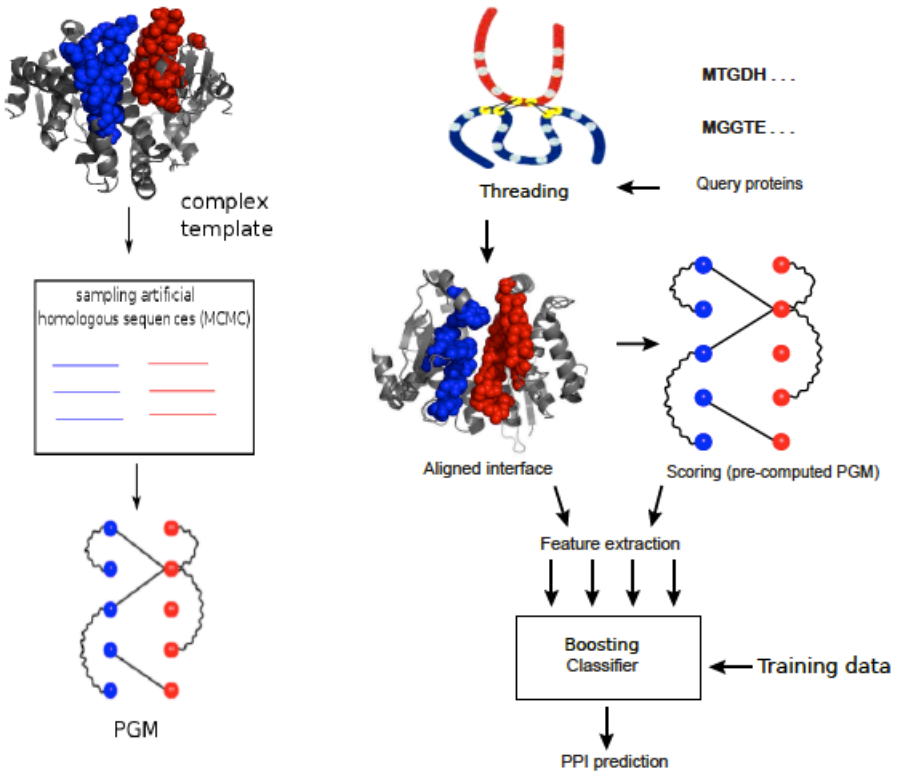



Figure 4-2: Flowchart of Coev2Net. Left: MCMC sampling to generate synthetic homologous sequences for each complex template. Right: 1) For given query protein pairs, the best template (from the structural library) is identified by user-defined protein threading; 2) structural and sequence features are extracted from the interfacial alignment and residue correlations scored w.r.t. the profile PGM; and 3) a classifier gives the probability of interaction for the query protein pair.

limitations of both structural and sequence training data (Fig B-2).

### 4.2.3 MAPK interactome validation

To test the framework's ability to predict interactions for which there is often no structural data available and to assign confidence values to interactions, we re-trained Coev2Net on a high-quality human MAPK PPI network [10] and tested it on another high-quality MAPK network [167](Fig 4-3A,B,C). Oddly, these two MAPK networks are almost disjoint with only 6 overlapping interactions out of 4904 total interactions. We found that the experimentally validated coverage of our method ( 55% with a probability cutoff of 0.6) is significantly higher than that reported by other prediction methods based on conservation, genomic data, GO annotation and literature extractions ( 14% to 28%) [135], although each method was evaluated on a different network.

Moreover, our predicted confidence scores are highly correlated with the experimental observation frequencies of Y2H screens on this network. To assess significance, we divided our predictions into high confidence and low confidence based on the probability cutoff of 0.6. To categorize interactions as true positive (TP) or true negative (TN) in the Y2H screens, we assumed the cutoffs employed in Schwartz et al. (for a False Discovery Rate FDR $< 5\%$, TP interactions should be observed at least twice when tested with $<5$ independent assays, and at least three times when tested with more assays) [135]. The predicted interactions correlate (P-value $< 0.01$) with those deemed likely true positives from an experimental standpoint. Encouragingly, the percentage of our framework's predicted TP interactions that are confirmed positive by the Vinayagam dataset is roughly 52% (294 TP, 571 predicted positive, a two-fold increase compared to previous methods on Y2H retesting of computational predictions [135]. Alternatively, training Coev2Net on the high confidence network in the Vinayagam dataset and testing it on the Bandyopadhyay core network yields similar

results. By predicting only a fraction of interactions with high confidence, Coev2Net enables us to focus on only the most likely interactions, enabling a more accurate understanding of the biology.

### 4.2.4 Experimental validation of predictions

The confidence scores given by our framework can be used to design additional experiments to enhance the quality of the initial interactome. We tested 19 randomly chosen high confidence interactions (predicted probability $> 0.6$) using a complementary assay (LUMIER) [12]. Of the 19 interactions we found that 14 interactions exhibited luciferase intensity greater than 1.5 times the control (Fig 4-3D). Multiple repeats of the assays allowed for a filtering of the interactions based on variance observed in the repeats. Interaction pairs for which the repeats were too variable to confidently confirm the outcome of the experiment (either positive or negative) were discarded, leaving 11 pairs out of the initial 19 tested. Notably, 10 out of the 11 were confirmed as true positive and one was confirmed as a false positive by the fold-change in luciferase intensity values. Overlaps achieved by our method compare favorably with previous approaches in which an initial positive reference set (PRS) was re-tested experimentally using a LUMIER assay (Table 4.1) [25].

| Yeast strains implementation | # validated (LUMIER) | Y2H PPIs | % overlap |
|---|---|---|---|
| Y strain 2m 1 reporter 1mM_3-AT | 19 | 33 | 57 |
| Y strain 2m 2 reporters 1mM_3-AT | 13 | 22 | 59 |
| Y strain CEN 1 reporter 1mM_3-AT | 17 | 23 | 74 |
| MaV CEN 2 reporters 20 mM_3-AT | 9 | 14 | 64 |
| Our prediction | 14 | 19 | 74 |
| Our prediction ($z > 1.5$) | 10 | 11 | 91 |

Table 4.1: Comparison of overlaps achieved by Braun et. al. and our method when some of the initial Y2H interaction pairs are re-tested using LUMIER assay.

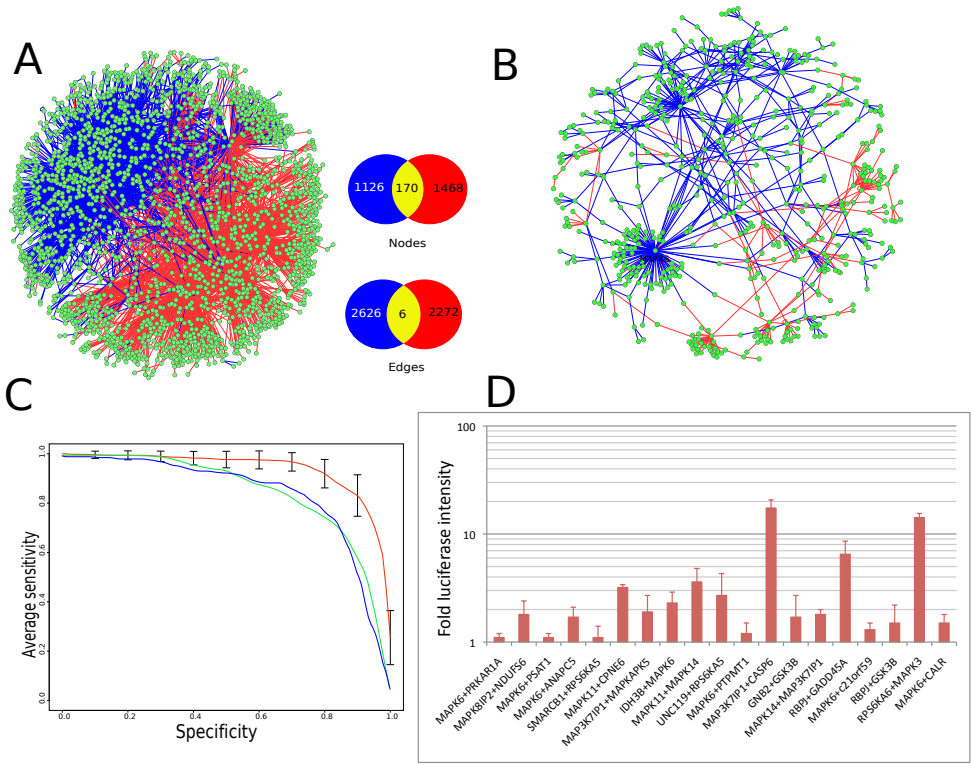
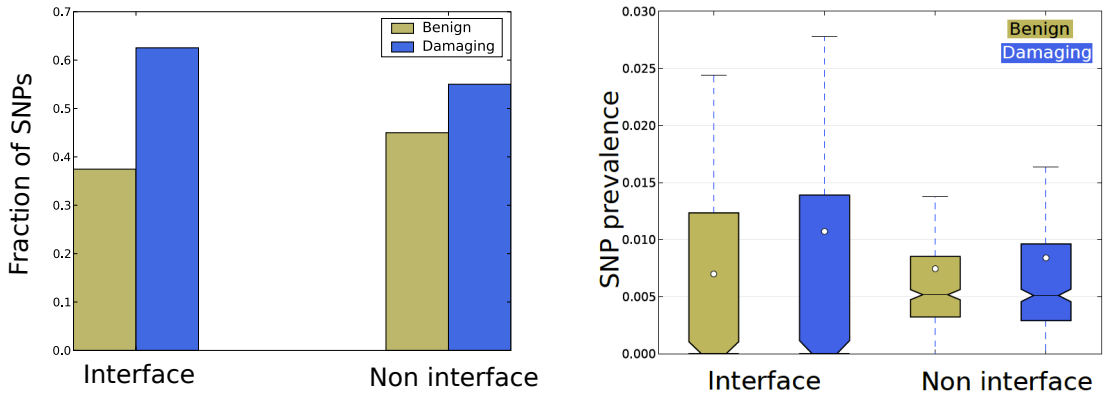



Figure 4-3: A) Overlap of the Vinayagam (blue) and Bandyopadhyay (red) datasets (left). The study by Bandyopadhyay et al. reveals 2269 interactions with 641 "core" interactions supported by multiple lines of evidence, whereas the Vinayagam dataset has 2626 interactions connecting 1126 proteins. Differences in the two experimental techniques are highlighted by the fact that only 170 nodes and 6 interactions overlap in the two sets. B) Coev2Net predicted high-confidence network is shown on the right. Edge colors correspond to the dataset they come from. MAPK6 has the highest degree, and its label is shown explicitly. C) Comparisons of performance on MAPK network for Coev2Net and previous Struct2Net (iWRAP+DBLRAP) [143, 142, 68] in terms of sensitivity and specificity. Coev2Net performs much better than Struct2Net on this dataset (core network of Bandyopadhyay et al.), and its performance is robust with respect to the randomness in MCMC sampling. The classifier (Fig 4-2) is trained and tested via 5-fold cross-validation on the core network. The MCMC procedure is repeated 5 times to assess robustness of the predictions. 'Baseline' method represents a logistic regression classifier with just the alignment features and no PPI (either Coev2Net or Struct2Net) features. D) Experimental validation of predicted high-confidence interactions using LUMIER assay. Typically a fold increase of 1.5 is considered as a true positive.

### 4.2.5 Abundance of missense SNPs at predicted interfaces

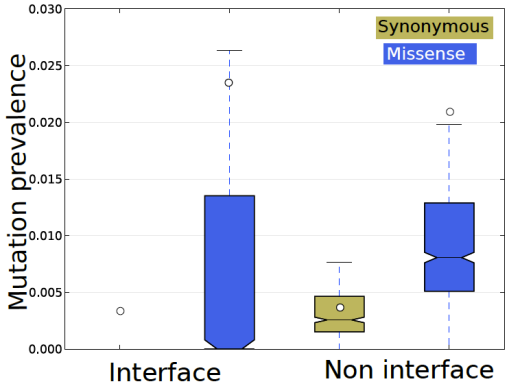
In addition to the confidence scores, Coev2Net also provides a putative interface for the interaction. These interfaces can yield novel mechanistic insights into the protein-protein interaction and provide hypotheses about disease-associated mutations that occur at the interface. Missense SNPs occurring at the interface can potentially disrupt the interaction between the proteins, leading to abnormal functioning of the cell. We analyzed the predicted interfaces for existence of PolyPhen2 annotated missense mutations in dbSNP (build 131) [139]. PolyPhen2 classifies a SNP as "benign", "probably damaging", "possibly damaging" or "unknown" based on various features including conservation score, monomeric structure score and physicochemical properties [127, 2]. It does not however account for SNPs occurring in potential interacting regions. Interestingly, SNPs annotated as damaging by PolyPhen2 are preferentially observed at the interface as compared to non-interfaces ($P = 0.0075$, Fisher exact test, Fig 4-4A). Furthermore, if we take into account the number of interface and non-interface sites, we find that the predicted interfaces are enriched for damaging SNPs as compared to the rest of the protein ($P < 7e\text{-}8$, Fisher exact test). The same analysis with SNPs classified as benign by PolyPhen2 does not show up as highly significant ($P = 0.06$). We further analyzed the distribution of the SNPs in terms of their density at the interface and non-interface. Here again, we find that damaging SNPs are preferentially located on the interface. We find that the average density of damaging SNPs at the predicted interfaces is significantly higher than their density at non-interface positions (Fig 4-4B; $P < 1e\text{-}10$, Mann-Whitney test); a bias also observed by Wang et al. recently [170]. For benign SNPs, the average density at the interface is lower than that at non-interfaces (Fig 4-4B; $P < 1e\text{-}10$, Mann-Whitney test). These analyses show that there is an evolutionary pressure to admit only benign SNPs at the interface, since any potentially damaging SNP will hinder the interaction.

To investigate the structural distribution of annotated mutations, we analyzed




**A.** SNP fractions

**B.** SNP prevalence




**C.** Cancer mutations

Figure 4-4: Predicted interfaces are enriched for SNPs in the Coev2Net predicted high-confidence MAPK network. A) Relative distribution of PolyPhen annotated mutations at the interface and non-interface. B) SNP (PolyPhen annotated) prevalence at the interface and non-interface. C) Somatic mutations characterized as "missense" preferentially fall on the interface (bottom). The white circles represent corresponding means. Error bars represent the 75%-25% data range.

somatic mutations characterized in cancer to see if there is any preference for their location on the protein. We analyzed annotated mutations in the coding region deposited in the Cosmic database for their predicted location [53]. We only considered mutations that are annotated as either synonymous, missense or nonsense. Interestingly, for these mutations we find that missense mutations are more prevalent on average at the PPI interface than synonymous mutations ($P < 10e\text{-}20$, Mann-Whitney test) (Fig 4-4C). This suggests that these mutations might be responsible for disruption of protein-protein interactions, and thus aberrant molecular signaling associated with cancer.

Finally, we looked at the predicted locations for some of the un-annotated mutations in kinases (from the MoKCa database [129]. As an example, we considered the BRAF protein as it contained the highest number of annotated mutations in the database. Coev2Net predicts an interaction between BRAF and PAK2, using the template structure 1G3N (chains E and F). Fig 4-5A shows the predicted interface for this interaction, with the annotated (magenta) and un-annotated (dark blue) mutations indicated. The presence of these mutations at the interface of the interacting proteins gives us an added insight into the investigation of such variations. Further study using this information can provide mechanistic details about how such mutations disrupt normal cellular signaling.

### 4.2.6 Novel potential cross-talk regulatory mechanism

Phosphorylation sites have been observed to be enriched at interfaces in solved structures [110]. This observation has mechanistic implications as the PPI can be used as an additional regulatory mechanism for phosphorylation, or the interaction could be a precursor to phosphorylation. An example for such a mechanism is found in the signaling protein YWHAZ [106]. Its phosphorylation is regulated by its dimerization, which buries the phospho-sites on YWHAZ [175]. Our predictions revealed

an interesting observation that suggests similar regulatory mechanisms in the MAPK interactome. Coev2Net predicts an interaction between MAPK6 and YWHAZ. Both are important signaling proteins, with much known about YWHAZ, including the experimental observation that MAPK8 regulates phosphorylation at S184 [179]. Relatively less is known about MAPK6's function and its substrates [81]. However, it is known that S189 is a phospho-site regulated by PAK1, PAK2 and PAK3 [40, 113, 38]. Interestingly, we found that the phosphorylation sites for both MAPK6 (S189) and YWHAZ (S184) lie within the predicted interface for the interaction (Fig 4-5B). This structural observation could imply that the interaction regulates downstream activities of MAPK6 and YWHAZ by controlling their phosphorylation. The most likely mechanism is that MAPK6 phosphorylates YWHAZ, thereby preventing its dimerization and regulating downstream activities of YWHAZ. Additionally, Coev2Net also predicts an interaction between MAPK6 and FOXO3. From a signaling context, these observations suggest a possible mechanism of cross talk between the MAPK and insulin pathways.

## 4.3 Methods

The Coev2Net algorithm can be roughly divided into three distinct stages, 1) prediction of the binding interface, 2) evaluation of the compatibility of the interface with an interface co-evolution based model (i.e. probabilistic graphical model) and 3) evaluation of the confidence score for the interaction.

*Prediction of the putative interface:* The two query sequences are threaded against a template library to search for the best template. We use a top-performing threader program "RAPTOR" [177, 117] to look for the best template match. Given a set of potential template matches, the best match is selected based on the z-score of the alignment. Quality metrics for the alignment, such as the mutation scores and

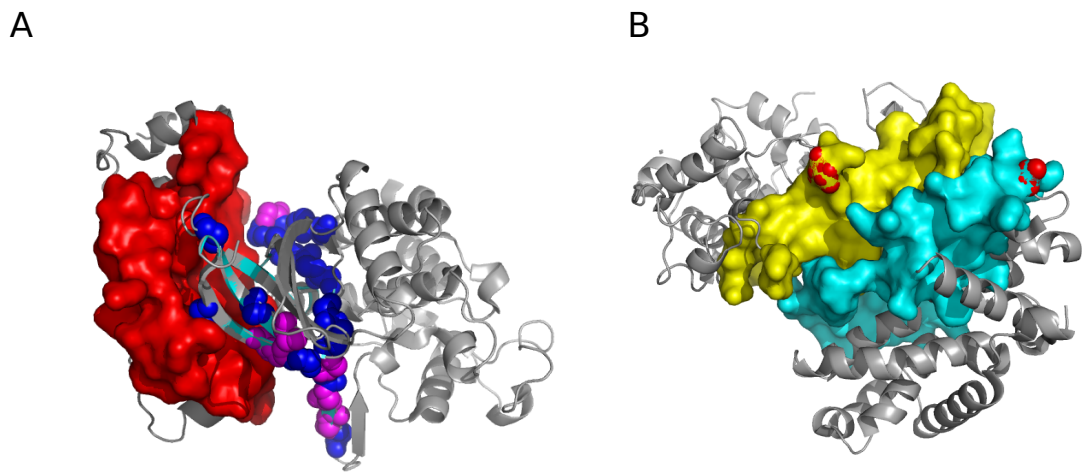



Figure 4-5: A) Predicted interface for the interaction between BRAF (light blue) and PAK2 (red surface). Cancer associated mutations that are annotated are shown in magenta. In dark blue we indicate mutations that are predicted to be associated with cancer but with no current annotations. Rest of the template structure is shown in gray. Mutations were taken from MoKCa database [129]. B) Predicted interface for the interaction between MAPK6 (yellow) and YWHAZ (cyan). Phosphorylation sites on the proteins are indicated in red (S189 for MAPK6 and S184 for YWHAZ). The template used for the prediction was 1F5Q (chains A and B).

secondary structure match scores are used as features in the classifier in the third stage of Coev2Net.

*Evaluating the interface:* Our intuition behind checking the "interacting propensity" of the predicted interfaces is that interacting proteins exhibit co-evolution at the interface. This co-evolution has been detected even in residues within 10-12 Angstrom at the interface [83, 172, 161, 126, 116, 114]. In Coev2Net, a probabilistic graphical model, pre-computed for each SCOPPI family (Pre-computed PGM), encodes the most significant pattern of interface correlations exhibited by the interacting members of the SCOPPI family. This model is computed by formulating interface co-evolution as a high-dimensional sampling problem. The predicted interface is evaluated by computing the log-likelihood of the interface residues with respect to this graphical model. A higher log-likelihood implies that the protein sequences show co-evolution at the interface, compatible with the model and are hence likely to interact.

*Computing confidence score:* Once we have the compatibility scores for the predicted interface, we use these as features to predict our confidence in the interaction. A logistic-regression classifier is trained on a high-confidence network, and is used to predict the confidence score for the interaction. Both alignment features (from stage 1) and interface features (from stage 2) are used as features in the classifier.

### 4.3.1 Simulating interface co-evolution

Coev2Net simulates the natural process of interface co-evolution which is thought to be responsible for maintaining physical and chemical compatibility at the interface between two interacting proteins. We formulate this simulation process as a sampling problem from a high-dimensional distribution.

#### Simulation algorithm

**Stage 1**: Seeding the co-evolution. To overcome sampling issues, we start from regions in the sequence space that we know are in high-probability interaction regions. Therefore, we seed the co-evolution with data from known complexes. For a given SCOPPI family, the set of training complexes are aligned using the alignment program CMAPi [124]. CMAPi employs a contact map representation to efficiently align multiple interfaces and thereby improve alignments as compared to other sequence and structure based techniques [124]. A contact map is a binary matrix representation of the residue-residue interactions between two proteins. If the distance between any two heavy atoms of the two residues is less than 4.5 Å(similar to the cutoffs used by others [103, 172]), the corresponding entry in the contact map is 1, and 0 otherwise. In the following steps, the aligned interface sequences are used for the initialization (seed) of co-evolution.

**Stage 2**: Simulating co-evolution. Similar to the natural process of evolution, our simulation has a mutation and a selection step for the evolved sequences.

**Mutation**. For each pair of aligned seed sequences (full proteins forming the complex), additional sequences are constructed via random mutations according to a probability distribution based on paired positions within interfaces of complexes (Fig B-1). For non-interface residues, the BLOSUM62 matrix is used. Starting from the aligned seed sequences, mutations are carried out on the aligned sequences, with each simulated sequence having the same gap structure as the original seed alignment. We randomly select 5% of interface residues to mutate, and 5% of the non-interface residues. These numbers were selected based on previous studies on simulating sequences for homology search [90, 91].

**Selection**. The new sequences are first aligned to the HMMs representing the corresponding families [147], and the alignment scores computed. They are then accepted or rejected in a stochastic manner, based on their joint "fitness" score. If $E^1$ and $E^2$ are the (negative) alignment scores for the two evolved sequences w.r.t the HMMs, then the following function $\alpha$ is computed and used to select new sequences:

$$\begin{aligned} \alpha &= (P^{new}\Pi_j p_j^{old})/(P^{old}\Pi_j p_j^{new}) \\ P &\propto exp(-E^1 - E^2) \\ p_j &= q_{uniprot},\ j \text{ is not an interface position} \\ &= q,\ \text{otherwise} \end{aligned} \tag{4.1}$$

where $q_{uniprot}$ is the amino-acid distribution in Uniprot; $q$ the amino-acid distribution at the interface from a selected non-redundant set of complexes (Fig B-1); and $\alpha$ the probability of the mutations at the interface being accepted. If $\alpha > 1$, the new sequences are accepted automatically. However, to incorporate diversity into the evolved sequences, we also accept sequences with a certain probability even if this ratio is low. A random number is drawn uniformly from [0,1], and the new

sequences are accepted if this number is less than $\alpha$. Intuitively, $\alpha$ represents how likely it is the sequences (interface) belong to the co-evolving families, as compared to a model that considers all positions independent. We show that simulated co-evolution, viewed through the lens of a high-dimensional sampling problem, leads to the same co-evolution and selection step (see proof in Appendix B). Along the course of the simulation, we monitor the sum of the entropies of all the sequence positions, and only retain sequences at an interval of 10 iterations after this value converges. These sequences are non-redundant representatives of their respective families, with the added feature that they are assumed to be interacting.

**Stage 3**: Correlation graph. A probabilistic graphical model (PGM) is then constructed for a particular interface alignment, based on correlations at the interface in sequences simulated by co-evolution. Nodes of the PGM are individual positions in the two proteins and edges indicate correlations. Once the MCMC has converged, we select 1000 interacting sequences per training complex as our interacting set. To model the correlations between residues of the interacting proteins, we use the Sanghavi-Tan-Willsky algorithm [132] to construct two trees– one for the simulated interacting proteins and one for background correlations. Briefly, for any two positions $i$ and $j$ in the simulated sequences, weights $w_{ij}^{+}$ and $w_{ij}^{-}$ are computed as:

$$w_{ij}^{+} = \sum_{x_i, x_j} (p(x_i, x_j) - q(x_i, x_j)) log \frac{p(x_i, x_j)}{p(x_i)p(x_j)} \tag{4.2}$$

$$w_{ij}^{-} = \sum_{x_i, x_j} (q(x_i, x_j) - p(x_i, x_j)) log \frac{q(x_i, x_j)}{q(x_i)q(x_j)} \tag{4.3}$$

where $x_r$ represents the amino acid at position $r$, $p$ is the empirical distribution computed from the simulated interacting sequences and $q$ is an empirical distribution computed by randomly pairing the simulated homologs (without regard to whether they were constructed together). $q$ thus represents the background correlations that would be expected due to limited sampling and other factors not really important

for interaction. Using these two sets of weights, two graphs (max-weight spanning trees) are learned over the set of nodes – one representing correlations important for interaction, the other background correlations. The maximum-weight spanning tree problem within STW is solved using NetworkX's implementation of Kruskal's algorithm [1]. Our choice of a tree graphical model is mainly due to the computational issues; trees are easy for both learning and inference. These PGMs are used to evaluate the interaction likelihood of the predicted interfaces.

## 4.4 Discussion

We have proposed a novel structure-based computational approach to identify protein-protein interactions on a genome-wide scale. Using structural features, we have demonstrated that our method can not only identify true-interactions better than previous approaches, but also provide key biological insights that are absent from HTP experiments. While it has been shown previously for some families that residues in and around the interface have correlated evolutionary histories, extracting such robust correlation signals for predictive purposes on a genome scale has remained difficult due to limited known interacting homologs. In the context of homology search for only monomers, enriching a multiple sequence alignment with artificial sequences has proven to be effective in the case of limited homologs [90, 91, 36]. Utilizing a statistical model for constructing evolutionarily correlated interacting homologs for a given interacting pair of proteins, we are able to simulate homologous sequences and predict PPIs from correlations at the interface of these homologs. The excellent performance of our method helps corroborate the hypothesis of residue-level correlations for a wide variety of protein-protein interactions and provides an efficient way of using these correlations for predictive purposes.

In contrast to iWRAP and DBLRAP that model the interface using a simple

contact-based scoring function, Coev2Net captures long-distance correlations that extend beyond the interacting regions. Improved performance of Coev2Net indicates the need to move beyond simple descriptions of the interface. Coev2Net also improves coverage by making it possible to make a prediction even when limited structural data is available. In particular, iWRAP requires multiple complexes to be available to build interface profiles, which is not often available. By sampling correlated mutations at the interface, Coev2Net enriches the sequence and structure profiles for such families, thereby making it possible to use them as templates for prediction.

As more and more HTP data for mapping the interactome are gathered, there would be a necessary demand for automatic protocols to evaluate the data quality and estimate confidence in individual interactions. In particular, transient interactions have been notoriously difficult to elucidate and validate. We have shown that confidence in protein-protein interactions investigated through high throughput techniques can be quantified and enhanced by our proposed complementary structure-based PPI prediction algorithm. Our PPI predictions on recent HTP human MAPK interactomes and further experimental validations have indicated the efficacy of our predicted confidence scores. Moreover, since our framework requires only the sequences of the two candidate proteins, it can be used as a complementary feature to other methods that rely on additional features [8, 70].

Limited studies have been undertaken to link structural features to genome-wide interactomes to gain a mechanistic understanding of underlying biological processes. Our threading-based approach enables us to extend coverage of structure-based studies further than that possible by homology models. As a result, the predicted structures are more reliable and provide a sound basis for mechanistic hypotheses. We provide an anecdotal example by analyzing the distribution of annotated missense SNPs in our predicted models. In agreement with a recent study [62], we show that such mutations are enriched at the interfaces. Furthermore, detailed analysis of phos-

phorylation sites enables us to propose a cross-talk mechanism involving an atypical kinase, MAPK6. Predictions made by our model for the potential interactors of MAPK6 provide the basis for further exploration of the role of this relatively less-studied kinase. These examples show how HTP techniques, in conjunction with our structure-based framework, can provide insights into transient interactions as well as static interactions.

# Chapter 5

# Conclusions

Protein-protein interactions are critical in a wide-range of biological processes ranging from maintenance of cellular integrity, metabolism, transcription/translation, and cell-cell communication. Thus elucidating PPIs is a fundamental step towards a deeper understanding of biology, and has the ability to make a significant impact on systems biology, genomics and therapeutics. The sheer number of interactions to test and validate makes it a very hard and expensive task. Although high-throughput PPI data is rapidly accumulating, building complete and confident datasets requires multiple replicates of expensive screens. This thesis develops new methods that significantly advance our efforts at structure-based approaches to better predict PPIs and boost confidence in emerging high-throughput data with the goal of comprehensive interactome mapping at lower cost.

Structure-based methods previously simply extended single-structure prediction methods for PPI prediction. In this thesis we further the state-of-art in PPI prediction by developing algorithms that utilize the biophysical and evolutionary features specific to protein interfaces to better predict interactions. This allows us to identify likely experimental errors in HTP datasets and helps develop novel testable hypothesis by providing a high-confidence network. In addition, improved accuracy of our interface

predictions provides mechanistic insights into how and why the interactions are taking place. This kind of information is not provided by any other HTP method, both computational and experimental.

In this thesis, I have described three structure-based methods to predict protein-protein interactions (Figure 5-1). The first method, DBLRAP, is a general, widely applicable tool that first identifies a suitable template for the two proteins and predicts their putative interface. It then formulates the PPI prediction problem as a classification problem and employs logistic regression to calculate an interaction probability. Our tests on yeast, fly and human genomes indicate that its predictive capabilities are better than sequence-only and other structure-based methods [143, 142]. However, it is known that the template identification and subsequently PPI prediction break down in the "twilight zone" of sequence identities. Furthermore, Struct2Net can make a prediction only if both the proteins thread well onto the template – this severely restricts the coverage of the predictions.

In order to improve coverage and quality of PPI prediction, we developed iWRAP [68]. Instead of threading two proteins onto a single template, iWRAP first builds an interface profile by aligning multiple complex templates. It then aligns the two sequences to this interface profile to predict the putative interface region. Cross-validation studies on an interface database [173] indicate that iWRAP is more accurate at identifying the true interface than other methods. iWRAP employs a boosting classifier trained on a high quality PPI dataset to predict interaction probability. While Struct2Net had limitations in coverage, we show that iWRAP can potentially handle templates in which one protein doesn't thread that well to the template. We show that iWRAP does much better than Struct2Net on predicting interactions in the yeast genome. In the process, we are able to improve accuracy and increase coverage by as much as 50% over Struct2Net [68]. We further demonstrate how iWRAP's predictions can be utilized to identify key genes involved in cancer. These genes are

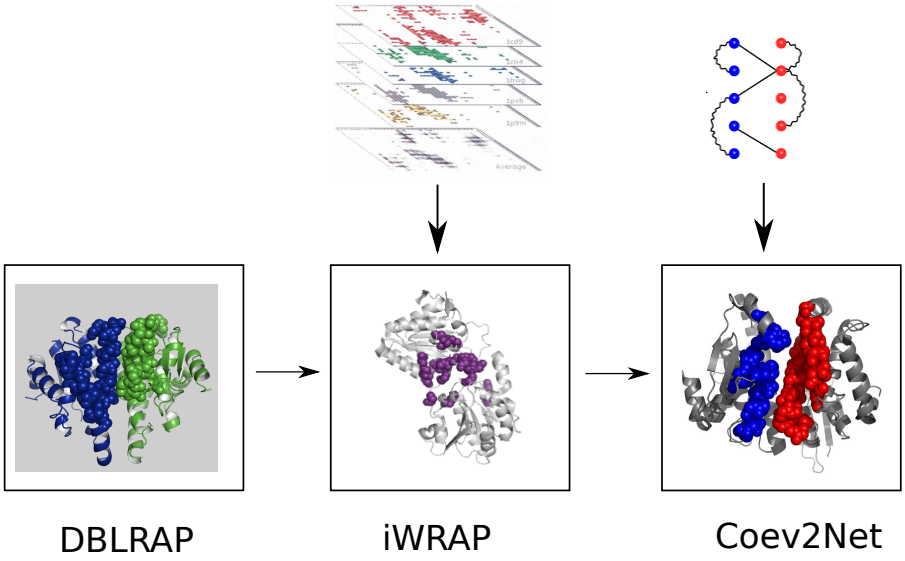



Figure 5-1: Methods introduced in this thesis. DBLRAP predicts the entire structure of the putative complex from the query sequences. iWRAP uses interface profiles that characterize biophysical properties of protein interfaces to predict just the interface residues. Coev2Net scores the predicted interface using a probabilistic graphical model that encodes long-distance correlations (i.e compatibilities) at the interface. The interface in this case can be obtained from any threading/alignment method.

candidates for further studies to determine if they can be used as novel therapeutic targets.

There exist a number of families for which multiple complexes haven't been solved yet. In such cases, it is not possible to use iWRAP. Furthermore, iWRAP does not do well when the interface is small, for example in the case of transient interactions. To overcome these limitations, we developed Coev2Net – an algorithm that utilizes long-distance correlations to predict PPIs. The intuition behind Coev2Net is that interacting families of proteins need to co-evolve to maintain the physicochemical compatibility at the interface. We formulate this as a high-dimensional sampling problem and provide a provably exact algorithm to extract artificial interacting homologous sequences. By enriching the sequence/structure profiles by simulations, we overcome the limited complex problem of iWRAP. Furthermore, we show that such a procedure allows one to predict PPIs as well as previous approaches on a gold-standard dataset.

Additionally, I demonstrated how Coev2Net can be integrated into a computational framework to assess confidence in binary interactions detected by large-scale high-throughput experiments. By analyzing two recent non-overlapping human mitogen activated protein kinase (MAPK) pathways, we show that Coev2Net can be used to address the false-positive and false negative issues in HTP datasets. Correlation between Coev2Net's predicted probabilities and frequency of observation in multiple repeats indicates that Coev2Net's score can be used to prune the list of interactions to test. We confirm this by experimentally validating some of the high-scoring interactions predicted by Coev2Net. The concordance between our prediction and experimental validation is as good as the overlaps achieved by previous protocols that use multiple secondary assays.

Finally, I show how Coev2Net's predicted interfaces can give us additional mechanistic insights that are not given by any other HTP technique. In agreement with a previous study, we find that missense SNPs annotated as "damaging" are enriched at the predicted interfaces [170]. Mutations found in tumor samples also follow a similar trend – they are preferentially found at the interfaces. This provides clues as to the role of those mutations in disrupting normal regulatory mechanisms. Furthermore, analysis of the predicted interfaces also aids in constructing hypothesis that, when verified, can lead to insights into novel regulatory processes.

Computational prediction of PPIs is one of the hardest and one of the most important problems in molecular biology. Diversity of protein interactions coupled with lack of high-quality, trustworthy data make it a challenging problem from a computational standpoint. From a biophysical point of view, PPIs are challenging since the main drivers of PPIs are very context dependent and different from protein folding (e.g electrostatics has only a minor role in protein folding). The rules governing protein association are still a matter of debate and there is no clear consensus, partly due to the diverse nature of the problem. The results presented in the thesis demonstrate

that there is significant added value in modeling protein interfaces separately from rest of the protein – either by a profile (iWRAP) or a probabilistic graph (Coev2Net). In addition, the predicted cancer-interactome should help identify targets for further experiments, which might lead to development of new drugs. Knowledge of the location of SNPs will help us better characterize their effect on the phenotype. This will pave the way for personalized medicine, where individual genotypes are treated differently based on their predicted phenotype.

To conclude, the PPI prediction problem is far from solved. Although the analysis of current networks has given us a wealth of information, getting a first draft of a high-quality "static" interactome is just the beginning. A large fraction of the interactions are context-dependent, i.e. occur only under a set of conditions/stimuli. Identifying such interactions and the contexts under which they occur is key to understanding cellular behavior. While current computational methods cannot handle this "dynamic" aspect of protein interactions, I believe ignoring this information could be harmful in future network analysis. After nearly a decade of work focusing on static interactomes, I believe it is high time we move on to elucidating the dynamic nature of these interactomes.

# Appendix A

# Appendix:iWRAP

## A.1 Evaluation of alignments

**Calculation of information content.** Besides sequence identity, information content is another popular metric used to quantify the difficulty of an alignment problem. The information content for an alignment is calculated by summing the information content of each column of the alignment. The information content of each column is calculated as given by the equation:

$$ic_j = \sum_i P_{ij} log(P_{ij}/Q_i)$$

In the above equation, $ic_j$ is the information content of column $j$, $P_{ij}$ the frequency of amino acid $i$ in column $j$ and $Q_i$ the background frequency of amino acid $i$. To get the frequency of each amino acid in a column, we count the number of occurrences of that amino acid and divide it by the length of the column. A pseudo-count of 0.01 is added to all counts to avoid zero count. The background distribution $Q$ is taken as the interface propensities of the amino acids [50]. This distribution is quite different from the frequencies of occurrence of individual amino acids in the entire SWISSPROT [9] database. However, for the purposes of this study, information content calculated based on this distribution captures the relative hardness of each alignment.

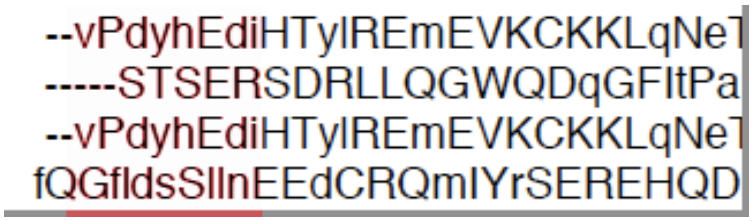
**Calculation of alignment accuracy.** For an alignment of a sequence S to a template T obtained using a threading approach, the number of correct alignments is calculated by counting the number of common pairs (t,s) between the threading alignment and the alignment generated by CMAPi for T and S. The accuracy is then obtained by normalizing this count by the length of the CMAPi alignment.

**Calculation of contact accuracy.** Three contact accuracies are calculated for each predicted contact map. The exact accuracy, i.e., the number of correctly predicted contacts divided by the total number of true contacts. The two other accuracies allow for a shift ($|\delta|$) in the predicted contacts. For example, if $(s_1, s_2)$ are positions of a true contact, we consider a predicted contact to be correct if it is within ($s_1 \pm \delta$, $s_2 \pm \delta$). We only report the contact accuracies with a shift of 2.

**Calculation of interface RMSD.** For an interface alignment of a sequence S to a template T obtained using a threading approach, the RMSD is calculated by considering only the $C_\alpha$ coordinates of the aligned residues. The Biopython module SuperImposer is used to calculate the minimum RMSD. Average RMSD per family pair is calculated by averaging the RMSDs for all possible template-sequence alignments within a family pair.

**Alignment Z score.** For a given optimal alignment, a background distribution of alignment scores is computed by fixing the alignment and randomizing the amino acids in the query at the aligned positions. Z scores are calculated by calculating the mean and standard deviation for 1000 such randomizations.

**Interface Z score.** For a given predicted interface, a background distribution of contact energies is calculated by randomizing the amino acids at the contacting positions in the interface. Z scores are calculated by calculating the mean and standard deviation for 1000 such randomizations.




**A.** Interface alignment

| Features/Position in alignment | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|
| Residues | VXG | PXF | DXL | YSD | HTS | SE | DLE |
| Sec. Struct. | C | C | C | C | H | H | H |
| Avg. Solv. Acc. | 58 | 78 | 59 | 22 | 25 | 99 | 69 |

**B.** Template construction

Figure A-1: Example of an interface template. A) An example of a multiple interface alignment from CMAPi (only one core is shown). The upper case letters represent the contacting residues in the interface, profiles constructed from residues highligted in red are shown in B. B) Interface template encoding the consensus residues, consensus secondary structure class and average solvent accessibility at the highlighted (in red) alignment positions in **A**. "X" represents the gap state in the alignment.

## A.2 Methods

### A.2.1 Templates

For each family pair in SCOPPI, the coordinates are obtained from the listed PDB IDs. In order to exclude interfaces formed due to crystallization, we select interfaces with more than five contacts. Furthermore, PDB models with resolutions lower than 2.5 Å are selected whenever possible. From an interface made up of two domains, three templates are constructed. One is the complex template (dimer), which consists of residue pairs (one on each domain) which have at least one of their heavy atoms at a distance less than 4.5 Å. Three templates are constructed from an interface in a PDB [14] file. A "dimer" template is the template describing the interface residues (see main text). Two additional templates are constructed by extracting the $C_\alpha$ and $C_\beta$ coordinates for individual domains from the PDB entry. In addition to spatial

coordinates, these two templates have information about solvent accessibilities and secondary structure, computed using the program DSSP [82]. These are in the form similar to the templates used by RAPTOR [177].

### A.2.2 Multiple Interface Alignment

Unlike profiles used in prediction of single chain protein structure, construction of profiles for PPI prediction is challenging because interactions between the two protein sequences complicates their treatment as independent alignments. In addition, profiles based on sequence alignments alone do not effectively capture the multiple binding modes exhibited within the same family. As demonstrated in Pulim et al. for the special case of cytokines [123], profiles based on a contact-map representation and alignment of interfaces are better suited for PPI prediction. Templates and profiles are constructed using these multiple interface alignments (see Template Construction in Main Text) for every family pair having atleast 3 "inter-domain" interfaces. These consist of domains on two different chains in the PDB file. Since we are interested in templates for PPI prediction, we consider only inter-domain interfaces. This has the added advantage of filtering out (dimer) interfaces formed due to crystallization. On the other hand, true homodimers will be excluded from our analysis.

### A.2.3 Genomic Predictions: *S.cerevisiae*

For genomic predictions, we used a two phase approach to identify templates for threading. In the first phase, each of the two query proteins is threaded (using RAPTOR) against the non-redundant database (<40% sequence identity) of proteins in SCOP1.75 [108]. This database contains around 10000 templates. We then select the top templates for each query protein by ranking them by z-scores and using a z-score cutoff of 3.0. At the end of this phase, we end up with 10-15 templates for each protein. In the second phase, we check to see if we have a dimer with

the SCOP domains represented by any one of these templates. In case we don't find such a template, we look for a dimer template which has one SCOP family common with one of the templates for the two query proteins. In case of multiple such dimer templates, we use the template with the highest sequence identity to the query proteins. This ensures that even for across-family threading, we utilize structurally similar templates. Our database has around 2000 total dimer templates (compared to around 2200 non-redundant dimers for Struct2Net).

Once we have the threading alignments for two yeast query proteins using iWRAP, we extract the following features from the results: template lengths (ltmpa,ltmb), sequence lengths (lseqa,lseqb), predicted number of contacts (cmap), total interface energy (total.energy), normalized interface energy (energy), z-scores for the threading alignments (alnza, alnzb) and z-score for the interface energy (z). In addition, we use the features sum of threading z-scores (total.z), square root of the product of sequence lengths (piab), total interface energy normalized by piab (energy_pi) and number of contacts normalized by piab (cmap_piab). The negative examples are generated as in Struct2Net [143].

The variable importance plot is shown in Fig A-2. As was observed by Singh et al. [143], the size of the sequences (piab), total interfacial energy (total.energy), normalized interfacial energy (energy_pi) are the most significant predictors in our boosting classifier. In addition, we find that sum of alignment z-scores (total.z) and the number of predicted contacts are important features which were not used in [143].

For the combined predictor, we used DBLRAP's threading alignments to extract features used in Struct2Net, and trained a classifier as above. The two predictions were combined by using a common cutoff to compute the combined ROC curve.

iWRAP makes predictions for proteins within the following functional complexes: 'cohesin loading factor complex', 'Bub2p/Bfa1p complex', 'eIF1/eIF1A/40S complex', 'Psr1p/Whi2p complex', 'Rot2p/Gtb1p complex', 'Reg1p/Glc7p complex', 'HAT-B

complex', 'U5 snRNP complex', 'ER V-ATPase assembly complex', 'Bud14p/Glc7p complex', 'Polzeta-Rev1p complex', 'nucleotide-excision repair factor 3 complex', 'Gip1p/Glc7p complex', 'protein farnesyltransferase complex', 'SAS Complex', 'Reg2p/Glc7p complex', 'Car1p/Arg3p complex', 'CURI complex', 'GAL3p/GAL80p complex', 'Fig4p/Vac14p complex', 'Nem1p/Spo7p complex', 'ribonuclease MRP complex', 'Cox14p/Cox1p/Mss51p complex', 'Gcn1p/Gcn20p complex', 'SF3b complex' and 'Bni4p/Glc7p complex'. Furthermore, iWRAP is able to make predictions for binary interactions within the following functional complexes: 'smc5p-Smc6p', 'Nem1p/Spo7p', 'Dig1p/Ste12p/Dig2p', 'GPI-anchor transamidase', 'i-AAA', 'Rot2p/Gtb1p', 'ribonuclease MRP', 'NatC', 'Npa2p-containing subcomplex', 'U5 snRNP' and 'signal peptidase'.

For functional complexes common to both iWRAP and DBLRAP, the following complexes are significantly enriched: 'Npa2p-containing subcomplex' (8 fold), 'alpha-1,6-mannosyltransferase complex(Anp1p/Mnn9p)' (8 fold), 'transcription factor TFIIF complex' (7 fold), 'Rad53p/Asf1p complex' (7 fold), 'MRX complex' (7 fold), 'Dig1p/Ste12p/Dig2p complex' (6 fold), 'DNA polymerase delta complex' (6 fold), 'Cdc28p/Clb1p complex' (6 fold), 'mitochondrial ribosomal large subunit' (6 fold), 'NuA4 histone acetyltransferase complex' (6 fold) and 'cytoplasmic ribosomal large subunit' (6 fold).

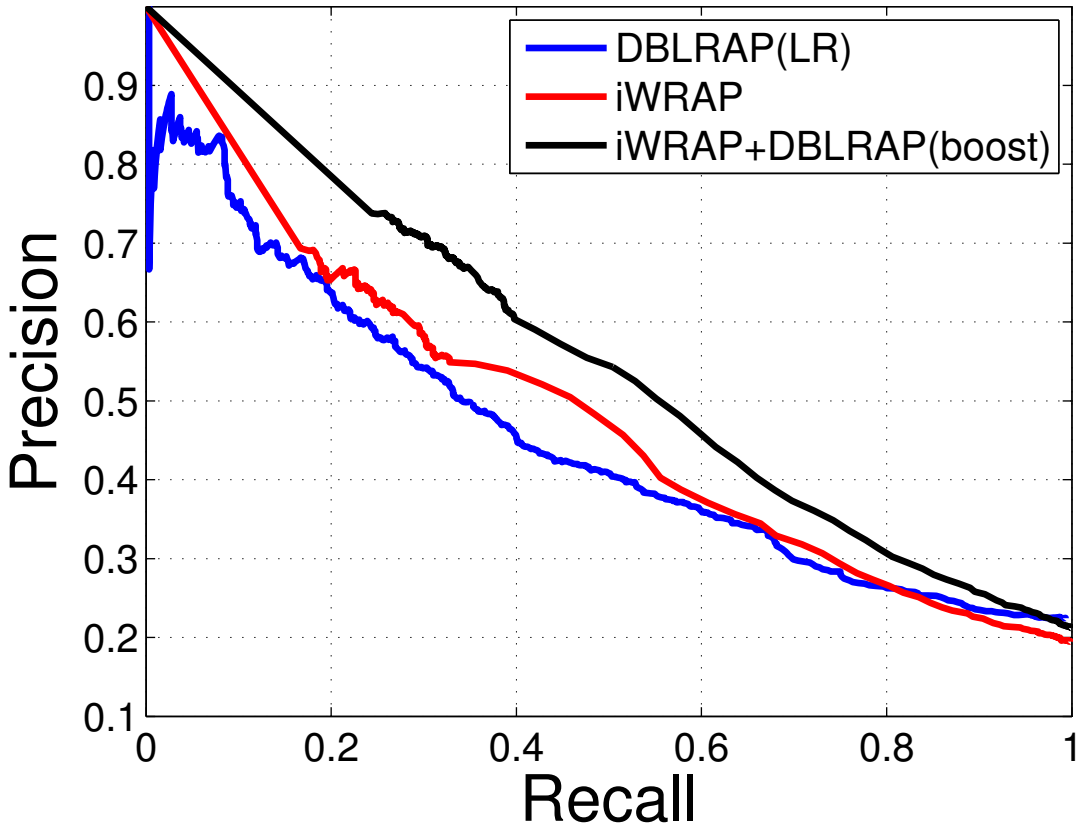



Figure A-2: Precision vs recall for the *S.cerevisiae* predictions. Here, precision=true positives/(true positives + false positives) and recall = true positives/(true positives + false negatives).

# Appendix B

# Appendix:Coev2Net

## B.1 Proof of equivalence of simulated co-evolution and high-dimensional sampling

Our procedure for simulated co-evolution is equivalent to a high-dimensional sampling problem. We can model evolution as nature drawing samples jointly from a complicated graphical model, which has the two HMMs, one for each family, and edges at the interface to couple the two HMMs together. In general, calculating the partition function and profiles (or marginals) is computationally intractable [149]; therefore we use a Markov Chain Monte Carlo (MCMC) technique to draw sample sequences from this distribution [100]. If $E^1$ and $E^2$ are the (negative) alignment scores for the two evolved sequences w.r.t the HMMs, then we assume the form of this distribution to be:

$$\begin{aligned} P^{eq} &\propto exp(-E^1 - E^2 - E^{int}) \\ E^{int} &= -\sum_{interface} log(Q(a,b)/q(a)q(b)) \end{aligned} \tag{B.1}$$

where the interface "energy" term $E^{int}$ is obtained by summing over all contacts $(a, b)$. $'Q'('q')$ is the pairwise (singleton) distribution shown in Fig B-1. Let $X_i^1, X_j^2, i = 1..n, j = 1..m$ be the amino acids at the interface ($< 10$ ) of the two interacting proteins (complex) that are in the contact map constructed by CMAPi. At each iteration of the MCMC, the goal is to construct $X_i^{1,new}, X_j^{2,new}$ from $X_i^{1,old}, X_j^{2,old}$ by mutating a fraction of the residues. For each contact $(i, j)$ at the interface, we first randomly select a protein from the pair and fix the corresponding amino acid in the contact. Let that protein be 1, say. The contacting amino acid in protein 2 (at position $j$) is then chosen from the following probability distribution (see Fig B-1):

$$\begin{aligned} X_j^{2,new} &\sim Q(.|X_i^{1,new}) \\ X_i^{1,new} &= X_i^{1,old} \end{aligned} \tag{B.2}$$

where the conditional probabilities are computed from the distributions in Fig B-1. For non-interface residues, the BLOSUM62 matrix is used (by computing the conditional probabilities) to mutate residues independently in the two proteins. The new sequences are then aligned to the HMMs representing their families and are accepted or rejected using a Metropolis-Hastings criterion based on their alignment scores and the interface energy $E^{int}$.

*Metropolis-Hastings criterion:*

Since we treat each contact independently while sampling, let us assume for the sake of simplicity that there is only one contact $(a, b)$. In the simulated sequences, this is evolved to $(a', b)$. Because we simulate co-evolution of the contact one residue at a time, the ratio of transition probabilities will be (old $\rightarrow$ new over new $\rightarrow$ old):

$$J^{int} = Q(a|b)/Q(a'|b) = Q(a, b)/Q(a', b) \tag{B.3}$$

where $Q$ is the pairwise distribution shown in Fig B-1. For the mutation of non-interface residues, since the two partners are mutated independently, the ratio of transition probabilities will just be the product across all non-interface positions:

$$\begin{aligned} J^{non-int} &= \Pi q_{uniprot}(x^{old}|x^{new})/\Pi q_{uniprot}(x^{new}|x^{old}) \\ &= \Pi q_{uniprot}(x^{old})/\Pi q_{uniprot}(x^{new}) \end{aligned} \tag{B.4}$$

where $q_{uniprot}$ is the Uniprot distribution. The Metropolis-Hastings criterion can then be written as:

$$\begin{aligned} \alpha &= P^{eq}(X^{new}) * J^{int} * J^{non-int}/P^{eq}(X^{old}) \\ P^{eq}(X^{new}) &= P^{new} * Q(a', b)/(q(a')q(b)) \\ P^{eq}(X^{old}) &= P^{old} * Q(a, b)/(q(a)q(b)) \\ P &\propto exp(-E^1 - E^2) \end{aligned} \tag{B.5}$$

Note that the pairwise probability terms, $Q(a, b)$ and $Q(a', b)$, in $P^{eq}(X^{new}) * J^{int}/P^{eq}(X^{old})$ cancel each other, leaving only the product of singleton probabilities. Therefore:

$$\begin{aligned} \alpha &= (P^{new}\Pi_j p_j^{old})/(P^{old}\Pi_j p_j^{new}) \\ P &\propto exp(-E^1 - E^2) \\ p_j &= q_{uniprot},\ j \text{ is not an interface position} \\ &= q,\ \text{otherwise} \end{aligned} \tag{B.6}$$

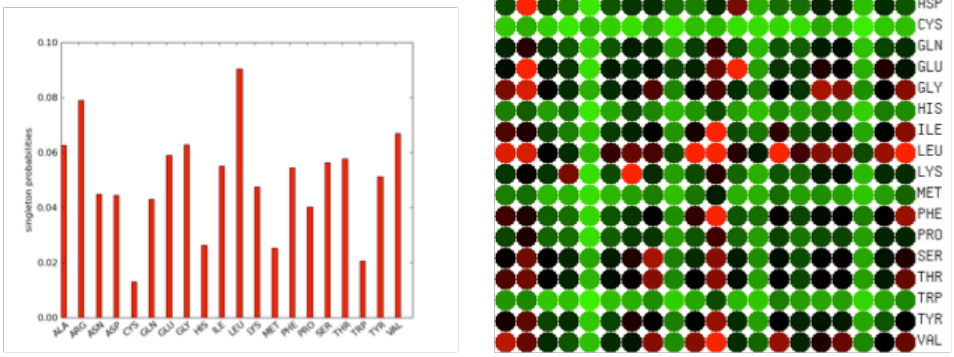
where recall that $q_{uniprot}$ is the amino-acid distribution in Uniprot; $q$ the amino-acid distribution at the interface from a selected non-redundant set of complexes (Fig B-1). This is exactly our "fitness" score used to select the co-evolved interfaces in the Selection step. QED

Note that this MCMC procedure allows us to efficiently compute any pairwise correlations, even those that are not contact based; a feature not possible without our sampling-based procedure.

## B.2 Datasets

All crystal structures were obtained from the Protein Data Bank (PDB). Singleton and pairwise amino-acid probabilities at the interface were calculated from a 50% non-redundant set of complexes downloaded from the 3DComplex database [98]. Here, two residues were assumed to be interacting if any heavy atom in one residue on one protein was at a distance of less than 5Å from any heavy atom on the other residue in the partner protein. The calculated singleton and pairwise probabilities calculated are shown in the Fig B-1. As one would expect, hydrophobic residues (A, V, L) are highly represented at the interface, whereas cysteine has the lowest propensity. Interestingly, Arg, Gly and Glu show up with a high propensity as well, indicating a preference for ionic and H-bond interactions at interfaces. This is in contrast to the general composition in globular proteins, where Arg is less frequent than Ala, Glu, and Gly is found at a much lower frequency .

All the MAPK PPI data was taken from Bandyopadhyay et al. (2010) and Vinayagam et al. (2011). The negative dataset used in evaluation of the classifier (PDB-negative) was downloaded from the negatome database [145]. In these datasets, only the sequences that could be aligned to templates belonging to families for which we could apply the simulated evolution protocol were considered. Sequences that had a z-score less than 5 for their alignment were discarded and such alignments were deemed not confident enough to give an accurate inference. In the Bandyopadhyay set, we could get predictions for 461 interactions; in the Vinu set, 860 interactions, and in the negatome (PDB-negative set), 330 non-interactors. The Bandyopadhyay




**A.** Singleton probabilities **B.** Pairwise probabilities

Figure B-1: Singleton (A) and Pairwise (B) probabilities at the interface calculated from a non-redundant set of complexes in [98]

set was further divided into a 173 "Core" set of interactions, defined by the authors, and the rest as "non-core".

## B.3 Results

### B.3.1 Coev2Net benchmarking

*Cross-validation on SCOPPI.* For each family in SCOPPI having three or more non-redundant complexes ($< 50\%$ sequence identity), we randomly select one as a Test Set and the remaining complexes as the Training Set. RAPTOR [177] is used to align the test sequences to the training templates, and the best alignment (based on RAPTOR's z score) selected for evaluation. Because of limited datasets ($\sim$ 45 families that meet our criterion of non-redundancy in SCOPPI and $\sim$ 300 negative pairs from the manually curated PDB-negative set (see Datasets)) [145], we use a 5-fold cross-validation to train and test the classifier.

*Limited complex families.* Additionally, for SCOPPI families that have only two non-redundant complexes, Coev2Net gives similar results (Fig B-2). To test on these

families, one complex (of the two) was chosen randomly, and the correlation graph computed as before, except for the multiple interface alignment stage. The classifier trained on multiple complex families was used to compute the probability of interaction of the test complex. As can be seen in Fig B-2, the algorithm is able to successfully use relevant correlations, even in the absence of multiple complexes for a given family, to help identify conserved structural features. Note that iWRAP cannot handle such families as it cannot build interface profiles due to the limited number of complexes.

### B.3.2 Abundance of SNPs

To compute association between PolyPhen annotations ('benign' and 'damaging') and our prediction of the SNP's location, we calculated the p-value using a 2x2 contingency table. Similarly to calculate association between SNPs and the location, we computed the p-value using a 2x2 contingency table with one grouping as total number of interface/non-interface residues and the other grouping as the occurrence/non-occurence of a SNP at that location. To verify abundance, we first normalized the occurrence of a SNP at a site by the number of such sites in the protein (a site is either an interface or a non-interface), and then performed a mann-whitney (paired) test to compute the p-value for the difference between the mean of the two densities (for the two types of sites).

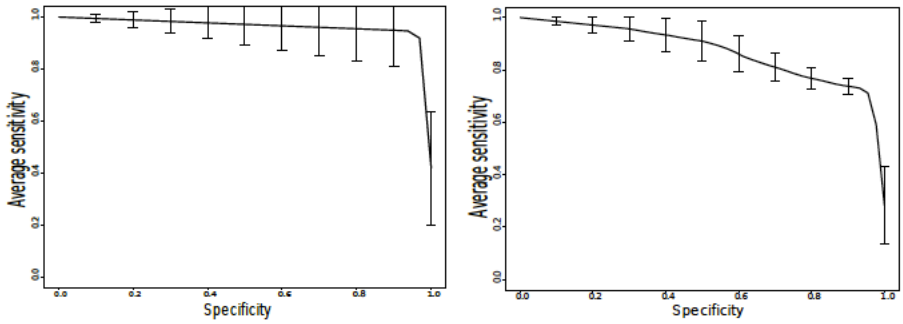



Figure B-2: Cross-validation results on SCOPPI. (left) Results on SCOPPI families having 3 or more complexes. (right) Results on SCOPPI families having only 2 complexes (1 training and 1 test)

# Bibliography


[1] 2009. `http://networkx.lanl.gov/`.

[2] I Adzhubei, S Schmidt, L Peshkin, V Ramensky, A Gerasimova, P Bork, A Kondrashov, and S Sunyaev. A method and server for predicting damaging missense mutations. *Nature methods*, 7:248–249, 2010.

[3] P Aloy and R Russell. Interrogating protein interactions networks through structural biology. *Proceedings of the National Academy of Sciences*, 99:5896–5901, 2002.

[4] P Aloy and R Russell. Interprets: protein interaction prediction through tertiary structure. *Bioinformatics*, 19:161–162, 2003.

[5] P Aloy and R Russell. Structural systems biology: modelling protein interactions. *Nature Reviews Molecular Cell Biology*, 7:188–197, 2006.

[6] S Altschul, T Madden, A Schaffer, J Zhang, Z Zhang, W Miller, and D Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, 25:3389–3402, 1997.

[7] A Aytuna, A Gursoy, and O Keskin. Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics*, 21:2850–2855, 2005.

[8] J Bader, A Chaudhuri, J Rothberg, and J Chant. Gaining confidence in high-throughput protein interaction networks. *Nature Biotech.*, 22:78–85, 2003.

[9] A Bairoch and R Apweiler. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, 28:45–48, 2000.

[10] S Bandyopadhyay, C Chiang, J Srivastava, M Gersten, S White, R Bell, C Kurschner, C Martin, M Smoot, S Sahasrabudhe, D Barber, S Chanda, and T Ideker. A human map kinase interactome. *Nature Methods*, 7:801–805, 2010.

[11] S Bandyopadhyay, R Kelley, N Krogan, and T Ideker. Functional maps of protein complexes from quantitative genetic interaction data. *PLOS Computational Biology*, 4:e1000065, 2008.

[12] M Barrios-Rodiles, K Brown, B Ozdamar, R Bose, Z Liu, R Donovan, F Shinjo, Y Liu, J Dembowy, I Taylor, V Luga, N Przulj, M Robinson, H Suzuki, Y Hayashizaki, I Jurisica, and J Wrana. High-throughput mapping of a dyamic signaling network in mammalian cells. *Science*, 307:1621–1625, 2005.

[13] A Ben-Hur and W Noble. Kernel methods for predicting protein-protein interactions. *Bioinformatics*, 21, Suppl 1:i38–46, 2005.

[14] H Berman, J Westbrook, Z Feng, G Gilliland, T Bhat, H Weissig, I Shindyalov, and P Bourne. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000.

[15] D Betel, K Breitkreuz, R Isserlin, D Dewar-Barch, M Tyers, and C Hogue. Structure-templated predictions of novel protein interactions from sequence information. *PLOS Computational Biology*, 3, 2007. e182.

[16] A Björkland, S Light, L Hedin, and A Elofsson. Quantitative assessment of the structural bias in protein-protein interaction assays. *Proteomics*, 8:4657–4667, 2008.

[17] J Bock and D Gough. Predicting proteinÐprotein interactions from primary structure. *Bioinformatics*, 17:455–460, 2001.

[18] P Bourne and H Weissig. *Structural Bioinformatics*. Wiley-Liss, Inc.,NJ, 2003.

[19] T Bouwmeester, A Bauch, H Ruffner, P Angrand, G Bergamini, K Croughton, C Cruciat, D Eberhard, J Gagneur, S Ghidelli, C Hopf, B Huhse, R Mangano, A Michon, M Schirle, J Schlegl, M Schwab, M Stein, A Bauer, G Casari, G Drewes, A Gavin, D Jackson, G Joberty, G Neubauer, J Rick, B Kuster, and G Superti-Furga. A physical and functional map of the human tnf-alpha/nf-kappa b signal transduction pathway. *Nature Cell Biol.*, 6:97–105, 2004.

[20] P Bradley, K Misura, and D Baker. Toward high-resolution de novo structure prediction for small proteins. *Science*, 309:1868–1871, 2005.

[21] C Branden and J Tooze. *Introduction to protein structure*. Garland Publishing, 1998.

[22] P Braun, M Tasan, M Dreze, M Barrios-Rodiles, I Lemmens, and H et. al. Yu. An experimentally derived confidence score for binary protein-protein interactions. *Nature methods*, 6:91–97, 2009.

[23] K Brown and I Jurisica. Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biology*, 8, 2007. R95.

[24] A Bryan Jr, J Starner-Kreinbrink, R Hosur, P Clark, and B Berger. Structure-based prediction reveals capping motifs that inhibit $\beta$-helix aggregation. *Proc. Natl. Acad. Sci.*, 108, 2011.

[25] L Burger and E Nimwegen. Accurate prediction of protein-protein interactions from sequence alignments using a bayesian method. *Molecular Systems Biology*, 4:165, 2008.

[26] D Caffrey, S Somaroo, J Hughes, J Mintseris, and E Huang. Are protein-protein interfaces more conserved in sequence than rest of the protein surface? *Protein Science*, 13:190–202, 2004.

[27] J Capra and M Singh. Predicting functionally important residues from sequence conservation. *Bioinformatics*, 23:1875–1882, 2007.

[28] S Carbon, A Ireland, C Mungall, S Shu, B Marshall, S Lewis, and Web Presence Working group. AmiGO HUB. Amigo: online access to ontology and annotation data. *Bioinformatics*, 25:288–289, 2009.

[29] A Ceol, A Chatr-aryamontri, E Santonico, R Sacco, L Castagnoli, and G Cesareni. Domino: a database of domain-peptide interactions. *Nucleic Acids Res. (Database*, 35:D557–560, 2007.

[30] H Choi, B Larsen, Z-Y Lin, A Breitkreutz, D Mellacheruvu, D Fermin, Z Qin, M Tyers, A-C Gingras, and A Nesvizhskii. SAINT: probabilistic scoring of affinity purification-mass spectrometry data. *Nature Methods*, 8:70–73, 2011.

[31] S. R. Collins, P. Kemmeren, X. C. Zhao, J. F. Greenblatt, F. Spencer, F. C. Holstege, J. S. Weissman, and N. J. Krogan. Toward a comprehensive atlas of the physical interactome of saccharomyces cerevisiae. *Mol Cell Proteomics*, 6(3):439–50, 2007.

[32] M Costanzo, A Baryshnikova, J Bellay, Y Kim, E Spear, and et al. The genetic landscape of a cell. *Science*, 327:425–431, 2010.

[33] T Crnogorac-Jurvevic, E Efthimiou, P Capelli, E Blaveri, A Baron, and et al. Gene expression profiles of pancreatic cancer and stromal desmoplasia. *Oncogene*, 20:7437–7446, 2001.

[34] M Culp, K Johnson, and G Michailidis. ada: A R package for stochastic boosting. *Journal of Statistical Software*, 17, 2006.

[35] M Cusick, H Yu, A Smolyar, K Venkatesan, and A et. al. Carvunis. Literature-curated protein interaction datasets. *Nature methods*, 6:39–46, 2009.

[36] N Daniels, R Hosur, B Berger, and L Cowen. Smurflite: combining simplified markov random fields with simulated evolution improves remote homology detection for beta-structural proteins into the twilight zone. *Bioinformatics*, 2012. doi: 10.1093/bioinformatics/bts110.

[37] F Davis and A Sali. Pibase: a comprehensive database of structurally defined protein interfaces. *Bioinformatics*, 21:1901–1907, 2005.

[38] P Deleris, M Trost, I Topsirovic, P Tanguay, K Borden, P Thibault, and S Meloche. Activation loop phosphorylation of erk3/erk4 by group i p21-activated kinases (paks) defines a novel pak-erk3/4-mapk-activated protein kinase 5 signaling pathway. *J. Biol. Chem.*, 286:6470–6478, 2011.

[39] M Deng, S Mehta, F Sun, and T Chen. Inferring domain-domain interactions from protein-protein interactions. *Genome Research*, 12:1540–1548, 2002.

[40] N Dephoure, C Zhou, J Villen, S Beausoleil, C Bakalarski, S Elledge, and S Gygi. A quantitative atlas of mitotic phosphorylation. *Proc. Natl. Acad. Sci. USA*, 105:10762–10767, 2008.

[41] C Dominguez, R Boelens, and A Bonvin. Haddock: a protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.*, 125:1731–1737, 2005.

[42] Jan Drenth. *Principles of Protein x-ray crystallography.* Springer-Verlag New York,Inc, 1999.

[43] M. Dreze, D. Monachello, C. Lurin, M. E. Cusick, D. E. Hill, M. Vidal, and P. Braun. High-quality binary interactome mapping. *Methods Enzymol*, 470:281–315, 2010.

[44] S Eddy. Hidden markov models. *Current Opinion in Structural Biology*, 6:361–365, 1996.

[45] R.C Edgar. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, 2004.

[46] A. M. Edwards, B. Kus, R. Jansen, D. Greenbaum, J. Greenblatt, and M. Gerstein. Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends Genet*, 18(10):529–36, 2002.

[47] M Eisenstein. Everything is illuminated. *Nature Methods*, 2:323, 2005.

[48] A. Elefsinioti, O. S. Sarac, A. Hegele, C. Plake, N. C. Hubner, I. Poser, M. Sarov, A. Hyman, M. Mann, M. Schroeder, U. Stelzl, and A. Beyer. Large-scale de novo prediction of physical protein-protein association. *Mol Cell Proteomics*, 10(11):M111 010629, 2011.

[49] JA Encinar, G Fernandez-Ballester, IE Sanchez, E Hurtado-Gomez, F Stricher, P Beltrao, and L Serrano. ADAN: a database for prediction of protein-protein interaction of modular domains mediated by linear motifs. *Bioinformatics*, 25:2418–2424, 2009.

[50] A Fernández, L Ridgway Scott, and H Scheraga. Amino-acid residues at protein-protein interfaces: Why is propensity so different from relative abundance? *J. Phys. Chem. B.*, 107:9929–9932, 2003.

[51] S. Fields and R. Sternglanz. The two-hybrid system: an assay for protein-protein interactions. *Trends Genet*, 10(8):286–92, 1994.

[52] R Finn, M Marshall, and A Bateman. ipfam: visualization of protein-protein interactions in pdb at domain and amino acid resolutions. *Bioinformatics*, 21:410–412, 2005.

[53] S Forbes, N Bindal, S Bamford, C Cole, C Kok, D Beare, M Jia, R Shepherd, K Leung, A Menzies, J Teague, P Campbell, M Stratton, and P Futreal. COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Research (Database Issue)*, 39:945–950, 2011.

[54] E. Formstecher, S. Aresta, V. Collura, A. Hamburger, A. Meil, A. Trehin, C. Reverdy, V. Betin, S. Maire, C. Brun, B. Jacq, M. Arpin, Y. Bellaiche, S. Bellusci, P. Benaroch, M. Bornens, R. Chanet, P. Chavrier, O. Delattre, V. Doye, R. Fehon, G. Faye, T. Galli, J. A. Girault, B. Goud, J. de Gunzburg, L. Johannes, M. P. Junier, V. Mirouse, A. Mukherjee, D. Papadopoulo, F. Perez, A. Plessis, C. Rosse, S. Saule, D. Stoppa-Lyonnet, A. Vincent, M. White, P. Legrain, J. Wojcik, J. Camonis, and L. Daviet. Protein interaction mapping: a drosophila case study. *Genome Res*, 15(3):376–84, 2005.

[55] H Fraser, A Hirsh, L Steinmetz, C Scharfe, and M Feldman. Evolutionary rate in the protein interaction network. *Science*, 296:750–752, 2002.

[56] N Fukuhara, N Go, and T Kawabata. Prediction of interacting proteins from homology-modeled complex structure using sequence and structure scores. *Biophysical Journal*, 3:13–26, 2007.

[57] N Fukuhara, N Go, and T Kawabata. HOMCOS: a server to predict interacting protein pairs and interacting sites by homology modeling of complex structures. *Nucleic Acids Research (Web Server Issue)*, 36:W185–189, 2008.

[58] A Gavin, M Bosche, R Krause, P Grandi, M Marzioch, A Bauer, J Schultz, J Rick, A Michon, C Cruciat, M Remor, C Hofert, M Schelder, M Brajenovic, H Ruffner, A Merino, K Klein, M Hudak, D Dickson, T Rudi, V Gnau, A Bauch, S Bastuck, B Huhse, C Leutwein, M Heurtier, R Copley, A Edelmann, E Querfurth, V Rybin, G Drewes, M Raida, T Bouwmeester, P Bork, B Seraphin, B Kuster, G Neubauer, and G Superti-Furga. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415:141–147, 2002.

[59] The GO Consortium. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

[60] S Gomez, W Noble, and A Rzhetsky. Learning to predict protein-protein interactions from protein sequences. *Bioinformatics*, 19:1875–1881, 2003.

[61] S Gong, G Yoon, I Jang, D Bolser, P Dafas, M Schroeder, H Choi, Y Cho, K Han, S Lee, H Choi, M Lappe, and et al. Psibase: a database of protein structural interactome map (psimap). *Bioinformatics*, 21:2541–2543, 2010.

[62] U Güldener, M Münsterkōtter, G Kastenmüller, N Strack, J van Helden, and et al. CYGD: the comprehensive yeast genome database. *Nucleic Acids Research*, 33:D364–D368, 2005.

[63] Y Guo, L Yu, Z Wen, and M Li. Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. *Nucleic Acids Res.*, 36:3025–30, 2008.

[64] K. G. Guruharsha, J. F. Rual, B. Zhai, J. Mintseris, P. Vaidya, N. Vaidya, C. Beekman, C. Wong, D. Y. Rhee, O. Cenaj, E. McKillip, S. Shah, M. Stapleton, K. H. Wan, C. Yu, B. Parsa, J. W. Carlson, X. Chen, B. Kapadia, K. VijayRaghavan, S. P. Gygi, S. E. Celniker, R. A. Obar, and S. Artavanis-Tsakonas. A protein complex network of drosophila melanogaster. *Cell*, 147(3):690–703, 2011.

[65] G Hart, A Ramani, and E Marcotte. How complete are current yeast and human protein interaction networks? *Genome Biol.*, 7:120, 2006.

[66] Y Ho, A Gruhler, A Heilbut, G Bader, L Moore, S Adams, A Millar, P Taylor, K Bennett, K Boutilier, L Yang, C Wolting, I Donaldson, S Schandorff, J Shewnarane, M Vo, J Taggart, M Goudreault, B Muskat, C Alfarano, D Dewar, Z Lin, K Michalickova, A Willems, H Sassi, P Nielsen, K Rasmussen, J Andersen, L Johansen, L Hansen, H Jespersen, A Podtelejnikov, E Nielsen, J Crawford, V Poulsen, B Sorensen, J Matthiesen, R Hendrickson, F Gleeson, T Pawson, M Moran, D Durocher, M Mann, C Hogue, D Figeys, and Tyers M. Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature*, 415:180–183, 2002.

[67] R Hosur, R Singh, and B Berger. Sparse estimation for structural variability. *Algorithms for molecular biology*, 6, 2011.

[68] R Hosur, J Xu, J Bienkowska, and B Berger. iwrap: an interface threading approach with application to cancer-related protein-protein interactions. *Journal of Molecular Biology*, 405:1295–1310, 2011.

[69] J Hu, X Shen, Y Shao, C Bystroff, and M Zaki. Mining protein contact maps. *BIOKDD02: Workshop on Data Mining in Bioinformatics*, 2002.

[70] H Huang and J Bader. Precision and recall estimates for two-hybrid screens. *Bioinformatics*, 25:372–378, 2009.

[71] Y Huang, D Hang, L Lu, L Tong, M Gerstein, and G Montelione. Targeting the human cancer pathway protein interaction network by strutural genomics. *Molecular and Cellular Proteomics*, 7:2048–2060, 2008.

[72] J. M. Izarzugaza, D. Juan, C. Pons, F. Pazos, and A. Valencia. Enhancing the prediction of protein pairings between interacting families using orthology information. *BMC Bioinformatics*, 9:35, 2008.

[73] J Janin, R Bahadur, and P Chakrabarti. Protein-protein interaction and quaternary structure. *Quarterly Reviews of Biophysics*, 41:133–180, 2008.

[74] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, S. Chung, A. Emili, M. Snyder, J. F. Greenblatt, and M. Gerstein. A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302(5644):449–53, 2003.

[75] E Jefferson, T Walsh, T Roberts, and G Barton. Snappi-db: a database and api of structures, interfaces and alignments for protein-protein interactions. *Nucleic Acids Research*, 35:D580–D589, 2007.

[76] L Jensen, M Kuhn, M Stark, S Chaffron, C Creevey, J Muller, T Doerks, P Julien, A Roth, M. Simonovic, and et al. String 8: a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res. (Database)*, 37:D412–416, 2009.

[77] DT Jones. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology.*, 292:195–202, 1999.

[78] S Jones and J Thornton. Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. USA*, 93:13–20, 1996.

[79] R Jothi, P Cherukuri, A Tasneem, and T Przytycka. Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-domain interactions mediating protein-protein interactions. *Journal of molecular biology*, 362:861–875, 2006.

[80] D. Juan, F. Pazos, and A. Valencia. High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proc Natl Acad Sci U S A*, 105(3):934–9, 2008.

[81] C Julien, P Coulombe, and S Meloche. Nuclear export of erk3 by a crm1-dependent mechanism regulates its inhibitory action on cell-cycle progression. *J. Biol. Chem.*, 278:42615–42624, 2003.

[82] W Kabsch and C Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637, 1983.

[83] M Kann, B Shoemaker, A Panchenko, and T Przytycka. Correlated evolution of interacting proteins: Looking behind the mirror tree. *Journal of molecular biology*, 385:91–98, 2009.

[84] G Kar, A Gursoy, and O Keskin. Human cancer protein-protein interaction network: A structural perspective. *PLOS Comp. Biol.*, 5, 2009. e1000601.

[85] S Kerrien, Y Alam-Faruque, B Aranda, I Bancarz, A Bridge, C Derow, E Dimmer, M Feuermann, A Friedrichsen, R Huntley, and et al. Intact – open source resource for molecular interaction data. *Nucleic Acids Res. (Database)*, 35:D561–565, 2007.

[86] P Kim, L Lu, Y Xia, and M Gerstein. Relating three-dimensional structures to protein networks provides evolutionary insights. *Science*, 314:1938–1941, 2006.

[87] W. Kittichotirat, M. Guerquin, R. E. Bumgarner, and R. Samudrala. Protinfo ppc: a web server for atomic level prediction of protein complexes. *Nucleic Acids Res*, 37(Web Server issue):W519–25, 2009.

[88] T. Kocher and G. Superti-Furga. Mass spectrometry-based functional proteomics: from molecular machines to protein networks. *Nat Methods*, 4(10):807–15, 2007.

[89] M Kotlyar. *Prediction of Protein-Protein Interactions and Essential Genes Through Data Integration*. PhD thesis, University of Toronto, 2011.

[90] A Kumar and L Cowen. Augmented training of hidden markov models to recognize remote homologs via simulated evolution. *Bioinformatics*, 25:1602–1608, 2009.

[91] A Kumar and L Cowen. Recognition of beta-structural motifs using hidden markov models trained with simulated evolution. *Bioinformatics*, 26:i287–i293, 2010.

[92] P Kundrotas and I Vakser. Accuracy of protein-protein binding sites in high-throughput template-based modeling. *PLOS Computational Biology*, 6:e1000727, 2010.

[93] J Kyte and R Doolittle. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, 157:105–132, 1982.

[94] M Lai and J Xu. Ribosomal proteins and colorectal cancer. *Current Genomics*, 8:43–49, 2007.

[95] R Lathrop, R Rogers Jr, J Bienkowska, B Bryant, J Buturovic, C Gaitatzes, R Nambudripad, J White, and T Smith. Analysis and algorithms for protein sequence-structure alignment. *Computational methods in molecular biology*, pages 227–283, 1998.

[96] J. G. Lees, J. K. Heriche, I. Morilla, J. A. Ranea, and C. A. Orengo. Systematic computational prediction of protein interaction networks. *Phys Biol*, 8(3):035008, 2011.

[97] B Lehner and A Fraser. A first-draft human protein-interaction map. *Genome Biology*, 5:63, 2004.

[98] E Levy, J Pereira-Leal, C Chothia, and S Teichmann. 3d complex: a structural classification of protein complexes. *PLoS Comput. Biol.*, 2, 2006. e155.

[99] J Lipton and S Ellis. Diamond blackfan anemia 2008-2009: broadening the scope of ribosome biogenesis disorders. *Current Opinion in Pediatrics*, 22:12–19, 2010.

[100] Jun S. Liu. *Monte Carlo strategies in scientific computing*. Springer series in statistics. Springer, New York, 2001.

[101] R Lougee-Heimer. The common optimization interface for operations research. *IBM Journal of Research and Development*, 47:57–66, 2003.

[102] H Lu, L Lu, and J Skolnick. Development of unified statistical potentials describing protein-protein interactions. *Biophysical Journal*, 84:1895–1901, 2003.

[103] L Lu, H Lu, and J Skolnick. MULTIPROSPECTOR: An algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins*, 49:350–364, 2002.

[104] F Madeo, E Herker, S Wissing, H Jungwirth, T Eisenberg, and K Fröhlich. Apoptosis in yeast. *Current Opinion in Microbiology*, 7:655–660, 2004.

[105] S Martin, D Roe, and J Faulon. Predicting protein-protein interactions using signature products. *Bioinformatics*, 21:218–26, 2005.

[106] D Morrison. The 14-3-3 proteins: integrators of diverse signaling cues that impact cell fate and cancer development. *Trends in Cell Biology*, 19:16–23, 2008.

[107] S. Mukherjee and Y. Zhang. Protein-protein complex structure predictions by multimeric threading and template recombination. *Structure*, 19(7):955–66, 2011.

[108] A.G Murzin, S.E Brenner, T Hubbard, and C Chothia. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247:536–540, 1995.

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

[110] H Nishi, K Hashimoto, and A Panchenko. Phosphorylation in protein-protein binding: effect on stability and function. *Structure*, 19:1807–1815, 2011.

[111] Y Niu, D Otasek, and I Jurisica. Evaluation of linguistic features useful in extraction of interactions from pubmed; application to annotating known, high-throughput and predicted interactions in i2d. *Bioinformatics*, 26:111–119, 2010.

[112] I Nooren and J Thornton. Diversity of protein-protein interactions. *EMBO J*, 22:3486–3492, 2003.

[113] F Opperman, F Gnad, J Olsen, R Hornberger, Z Greff, G Keri, M Mann, and H Daub. Large-scale proteomics analysis of the human kinome. *Mol. Cell. Proteomics*, 8:1751–1764, 2009.

[114] A. Panjkovich and P. Aloy. Predicting protein-protein interaction specificity through the integration of three-dimensional structural information and the evolutionary record of protein domains. *Mol Biosyst*, 6(4):741–9, 2010.

[115] A Parsyan, D Shahbazian, Y Martineau, E Petroulakis, T Alain, and et al. The helicase protein dhx29 promotes translation initiation, cell proliferation, and tumorigenesis. *Proceedings Of The National Academy Of Sciences*, 106:22217–22222, 2009.

[116] F. Pazos, D. Juan, J. M. Izarzugaza, E. Leon, and A. Valencia. Prediction of protein interaction based on similarity of phylogenetic trees. *Methods Mol Biol*, 484:523–35, 2008.

[117] J. Peng and J. Xu. Raptorx: exploiting structure information for protein alignment by statistical inference. *Proteins*, 79 Suppl 10:161–71, 2011.

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

[119] U Pieper, N Eswar, B Webb, D Eramian, L Kelly, D Barkan, H Carter, P Mankoo, R Karchin, M Marti-Renom, F Davis, and A Sali. MODBASE: A database of annotated comparative protein structure models and associated resources. *Nucleic Acids Research*, 37:D347–D354, 2009.

[120] S Pitre, C North, M Alamgir, M Jessulat, A Chan, X Luo, J Green, M Dumontier, F Dehne, and A Golshani. Global investigation of protein-protein interactions in yeast saccharomyces cerevisiae using re-occurring short polypeptide sequences. *Nucleic Acids Res.*, 36:4286–4294, 2008.

[121] C Prieto, Las De, and J Rivas. Structural domain-domain interactions: assessment and comparison with protein-protein interaction data to improve the interactome. *Nucleic Acids Research*, 34:W298–W302, 2006.

[122] S Pu, J Wong, B Turner, E Cho, and S Wodak. Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Research*, 37:825–831, 2009.

[123] L Pulim, J Bienkowska, and B Berger. LTHREADER: Prediction of extracellular ligand-receptor interactions in cytokines using localized threading. *Protein Science*, 17:279–292, 2008.

[124] V Pulim, J Bienkowska, and B Berger. Optimal contact map alignment of protein-protein interfaces. *Bioinformatics*, 24:2324–2328, 2008.

[125] Y Qi, Z Bar-Joseph, , and J Klein-Seetharaman. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins*, 63:490–500, 2006.

[126] A. K. Ramani and E. M. Marcotte. Exploiting the co-evolution of interacting proteins to discover interaction specificity. *J Mol Biol*, 327(1):273–84, 2003. Ramani, Arun K Marcotte, Edward M Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. England Journal of molecular biology J Mol Biol. 2003 Mar 14;327(1):273-84.

[127] V Ramensky, P Bork, and S Sunyaev. Human non-synonymous snps: server and survey. *Nucleic Acids Research*, 30:3894–3900, 2002.

[128] D Rhodes, J Yu, N Deshpande, R Varambally, D Ghosh, and et al. ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia*, 6:1–6, 2004.

[129] C Richardson, Q Gao, C Mitsopoulous, M Zvelebil, L Pearl, and F Pearl. MoKCa database – mutations of kinases in cancer. *Nucleic Acids Research (Database Issue)*, 37:824–831, 2009.

[130] J.F Rual, K Venkatesan, T Hao, T Hirozane-Kishikawa, A Dricot, and et al. Towards a proteome-scale map of human protein-protein interaction network. *Nature*, 437:1173–1178, 2005.

[131] J Rudolp. Inhibiting transient protein-protein interactions: lessons from the cdc25 protein tyrosine pho sphatases. *Nature Rev. Cancer*, 7:202–207, 2007.

[132] S Sanghvi, V Tan, and A Willsky. Learning graphical models for hypothesis testing. *Statistical Signal Processing Workshop (SSP)*, 2007.

[133] M. E. Sardiu and M. P. Washburn. Building protein-protein interaction networks with proteomics and informatics tools. *J Biol Chem*, 286(27):23645–51, 2011.

[134] O Schueler-Furman, C Wang, P Bradley, K Misura, and D Baker. Progress in modeling of protein structures and interactions. *Science*, 310:638–642, 2005.

[135] A Schwartz, J Yu, K Gardenour, R Finley Jr, and T Ideker. Cost-effective strategies for completing the interactome. *Nature methods*, 6:55–61, 2009.

[136] P Shannon, A Markiel, O Ozier, N Baliga, J Wang, and et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13:2498–2504, 2003.

[137] F Sheinerman, R Norel, and B Honig. Electrostatic aspects of protein-protein interactions. *Current opinion in structural biology*, 10:153–159, 2000.

[138] J Shen, J Zhang, X Luo, W Zhu, K Yu, K Chen, Y Li, and H Jiang. Predicting protein-protein interactions based only on sequences information. *Proceedings Of The National Academy Of Sciences*, 104:4337–4341, 2007.

[139] S Sherry, M Ward, M Kholodov, J Baker, L Phan, E Smigielski, and K Sirotkin. dbsnp: the ncbi database of genetic variation. *Nucleic Acids Research*, 29:308–311, 2001.

[140] B Shoemaker and A Panchenko. Deciphering protein-protein interactions. part i. experimental techniques and databases. *PLOS Computational Biology*, 3, 2007. e42.

[141] N. Simonis, J. F. Rual, A. R. Carvunis, M. Tasan, I. Lemmens, T. Hirozane-Kishikawa, T. Hao, J. M. Sahalie, K. Venkatesan, F. Gebreab, S. Cevik, N. Klitgord, C. Fan, P. Braun, N. Li, N. Ayivi-Guedehoussou, E. Dann, N. Bertin, D. Szeto, A. Dricot, M. A. Yildirim, C. Lin, A. S. de Smet, H. L. Kao, C. Simon, A. Smolyar, J. S. Ahn, M. Tewari, M. Boxem, S. Milstein, H. Yu, M. Dreze, J. Vandenhaute, K. C. Gunsalus, M. E. Cusick, D. E. Hill, J. Tavernier, F. P. Roth, and M. Vidal. Empirically controlled mapping of the caenorhabditis elegans protein-protein interactome network. *Nat Methods*, 6(1):47–54, 2009.

[142] R Singh, D Park, J Xu, R Hosur, and B Berger. Struct2net: a web service to predict protein-protein interactions using a structure-based approach. *Nucleic Acids Research (Web Server Issue)*, pages 1–8, 2010.

[143] R Singh, J Xu, and B Berger. Struct2net: Integrating structure into protein-protein interaction prediction. *Proceedings of the Pacific Symposium on Biocomputing*, 11:403–414, 2006. `http://struct2net.csail.mit.edu/`.

[144] R Singh, J Xu, and B Berger. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proceedings of the National Academy of Sciences*, 105:12763–12768, 2008. `http://isobase.csail.mit.edu/`.

[145] P. Smialowski, P. Pagel, P. Wong, B. Brauner, I. Dunger, G. Fobo, G. Frishman, C. Montrone, T. Rattei, D. Frishman, and A. Ruepp. The negatome database: a reference set of non-interacting protein pairs. *Nucleic Acids Research*, D38:D540–544, 2010.

[146] G Smith and M Sternberg. Prediction of protein-protein interactions by docking methods. *Curr. Opin. Struct. Biol.*, 12:28–35, 2002.

[147] J Sōding. Protein homology detection by hmm-hmm comparison. *Bioinformatics*, 21:951–960, 2005.

[148] D Sontag, R Singh, and B Berger. Probabilistic modeling of systematic errors in two-hybrid experiments. *Proceedings of the Pacific Symposium on Biocomputing*, 12:445–457, 2007.

[149] I Sorin. Statistical mechanics, three-dimensionality and np-completeness: I. universality of intractability of the partition functions of the ising model across non-planar lattices. *Proceeding of the* $32^{nd}$ *ACM symposium on the theory of computing (STOC00)*, pages 87–96, 2000.

[150] B Srinivasan, A Novak, J Flannick, S Batzoglou, and H McAdams. Integrated protein interaction networks for 11 microbes. *Lecture Notes in Computer Science*, 3909:1–14, 2006.

[151] C Stark, B Breitkreutz, T Reguly, L Boucher, A Brietkreutz, and M Tyers. BIOGRID: A general repository for interaction datasets. *Nucleic Acids Research*, 34:D535–539, 2006.

[152] C Stebins and J Galán. Structural mimicry in bacterial virulence. *Nature*, 412:701–705, 2001.

[153] A. Stein, R. Mosca, and P. Aloy. Three-dimensional modeling of protein interactions and complexes is going 'omics. *Curr Opin Struct Biol*, 21(2):200–8, 2011.

[154] A Stein, R Russell, and P Aloy. 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Research*, 33:D413–D417, 2005.

[155] U Stelzl, U Worm, M Lalowski, C Haenig, F Brembeck, H Goehler, M Stroedicke, M Zenkner, A Schoenherr, S Koeppen, and J et al. Timm. A human protein-protein interaction network: A resource for annotating the proteome. *Cell*, 122:957–968, 2005.

[156] T Stopka, D Zakova, O Fuchs, O Kubrova, J Blafkova, and et al. Chromatin remodeling gene smarca5 is dysregulated in primitive hematopoietic cells of acute leukemia. *Leukemia*, 14:1247–1252, 2000.

[157] S. Suthram, T. Shlomi, E. Ruppin, R. Sharan, and T. Ideker. A direct comparison of protein interaction confidence assignment schemes. *BMC Bioinformatics*, 7:360, 2006.

[158] N. Tuncbag, A. Gursoy, and O. Keskin. Prediction of protein-protein interactions: unifying evolution and structure at protein interfaces. *Phys Biol*, 8(3):035006, 2011.

[159] N. Tuncbag, A. Gursoy, R. Nussinov, and O. Keskin. Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using prism. *Nat Protoc*, 6(9):1341–54, 2011.

[160] M. Tyagi, K. Hashimoto, B. A. Shoemaker, S. Wuchty, and A. R. Panchenko. Large-scale mapping of human protein interactome using structural complexes. *EMBO Rep*, 2012.

[161] M. Tyagi, R. R. Thangudu, D. Zhang, S. H. Bryant, T. Madej, and A. R. Panchenko. Homology inference of protein-protein interactions via conserved binding sites. *PLoS One*, 7(1):e28896, 2012. Tyagi, Manoj Thangudu, Ratna R Zhang, Dachuan Bryant, Stephen H Madej, Thomas Panchenko, Anna R United States PloS one PLoS One. 2012;7(1):e28896. Epub 2012 Jan 31.

[162] P Uetz, L Giot, G Cagney, T Mansfield, R Judson, and et al. A comprehensive analysis of protein-protein interactions in *saccharomyces cerevisiae*. *Nature*, 403:623–627, 2000.

[163] A Valencia and F Pazos. Computational methods for the prediction of protein interactions. *Current Opinion in Structural Biology*, 12:368–373, 2002.

[164] A Valencia and F Pazos. *In silico* two-hybrid system for the selection of physically interacting protein pairs. *Proteins*, 47:219–227, 2002.

[165] A Vazquez, J Rual, and K Venkatesan. Quality control methodology for high-throughput protein-protein interaction screening. *Methods in Molecular Biology*, 781:279–294, 2011.

[166] K Venkatesan, J-F Rual, A Vazquez, and et al. An empirical framework for binary interactome mapping. *Nature Methods*, 6:83–90, 2008.

[167] A Vinayagam, U Stelzl, R Foulle, S Plassmann, M Zenkner, J Timm, H Assmus, M Andrade-Navarro, and E Wanker. A directed protein interaction network for investigating intracellular signal transduction. *Science Signaling*, 4, 2011.

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

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

[170] X Wang, X Wei, B Thijssen, J Das, S Lipkin, and H Yu. Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nature Biotechnology*, 2012.

[171] M Wass, G Fuentes, C Pons, F Pazos, and A Valencia. Towards the prediction of protein interaction partners using physical docking. *Molecular systems biology*, 7, 2011.

[172] M. N. Wass, A. David, and M. J. Sternberg. Challenges for the prediction of macromolecular interactions. *Curr Opin Struct Biol*, 21(3):382–90, 2011.

[173] C Winter, A Henschel, WK Kim, and M Schroeder. SCOPPI: a structural classification of protein-protein interfaces. *Nucleic Acids Research (Database issue)*, 34:310–314, 2006.

[174] S Wodak and R Mendez. Prediction of protein-protein interactions: the capri experiment, its evaluations and implications. *Curr. Opin. Struct. Biol.*, 14:242–249, 2004.

[175] J Woodcock, J Murphy, F Stomski, M Berndt, and A Lopez. The dimeric versus monomeric status of 14-3-3zeta is controlled by phosphoryltion of ser58 at the dimeric interface. *J. Biol. Chem.*, 278:36323–36327, 2003.

[176] C Wu, H Ma, K Brown, L Geisler, E Tzeng, C Jia, I Jurisica, and S Li. Systematic identification of sh3 domain-mediated human protein-protein interactions by peptide array target screening. *Proteomics*, 7:1775–1785, 2007.

[177] J Xu, M Li, D Kim, and Y Xu. RAPTOR: Optimal protein threading by linear programming. *J Bioinform Comput Biol*, 1:95–117, 2003.

[178] C Yan, F Wu, R Jernigan, D Dobbs, and V Honavar. Characterization of protein-protein interfaces. *The Protein Journal*, 27:59–70, 2008.

[179] K Yoshida, T Yamaguchi, T Natsume, D Kufe, and Y Miki. Jnk phosporylation of 14-3-3 proteins regulates nuclear targeting of c-abl in the apoptotic response to dna damage. *Nature Cell Biology*, 7:278–285, 2005.

[180] C Yu, L Chou, , and D Chang. Predicting protein-protein interactions in unbalanced data using the primary structure of proteins. *BMC Bioinformatics*, 11:167, 2010.

[181] H. Yu, P. Braun, M. A. Yildirim, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, T. Hao, J. F. Rual, A. Dricot, A. Vazquez, R. R. Murray, C. Simon, L. Tardivo, S. Tam, N. Svrzikapa, C. Fan, A. S. de Smet, A. Motyl, M. E. Hudson, J. Park, X. Xin, M. E. Cusick, T. Moore, C. Boone, M. Snyder, F. P. Roth, A. L. Barabasi, J. Tavernier, D. E. Hill, and M. Vidal. High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–10, 2008.

[182] J. Yu and Jr. Finley, R. L. Combining multiple positive training sets to generate confidence scores for protein-protein interactions. *Bioinformatics*, 25(1):105–11, 2009.

[183] M Zaki and C Bystroff. *Protein structure prediction*, volume 413. Springer, Humana Press, 2 edition, 2007.

[184] Q Zhang, D Petrey, R Norel, and B Honig. Protein interface conservation across structure space. *Proceedings of the National Academy of Sciences*, 107:10896–10901, 2010.

[185] H Zhu, M Bilgin, R Bangham, D Hall, A Casamayor, P Bertone, N Lan, R Jansen, S Bidlingmaier, T Houfek, T Mitchell, P Miller, R Dean, M Gerstein, and M Snyder. Global analysis of protein activities using proteome chips. *Science*, 293:2101–2105, 2001.

[186] H Zhu, J Klemic, S Chang, P Bertone, A Casamayor, K Klemic, D Smith, M Gerstein, M Reed, and M Snyder. Analysis of yeast protein kinases using protein chips. *Nature Genetics*, 26:283–289, 2000.